PROVIDER AND PEER SELECTION IN THE EVOLVING INTERNET ECOSYSTEM

A Thesis
Presented to
The Academic Faculty

by

Amogh Dhamdhere

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
College of Computing

Georgia Institute of Technology
May 2009

PROVIDER AND PEER SELECTION IN THE EVOLVING INTERNET ECOSYSTEM

Approved by:

Dr. Constantine Dovrolis, Advisor
College of Computing
Georgia Institute of Technology

Dr. Mostafa Ammar
College of Computing
Georgia Institute of Technology

Dr. Nick Feamster
College of Computing
Georgia Institute of Technology

Dr. Ellen Zegura
College of Computing
Georgia Institute of Technology

Dr. Walter Willinger
AT&T Labs - Research

Date Approved: 02 Apr 2009

To my family, for their support and encouragement.

ACKNOWLEDGEMENTS

This thesis would not have been possible without the presence of many people who have influenced my life and work over the years. I take this opportunity to thank some of them.

First and foremost, I would like to thank my advisor Constantine Dovrolis for his guidance and support over the years. His patience, attention to detail, and depth of thought continue to amaze me. Most of all, I never got the feeling of working *for* my advisor; I worked *with* him. Everything I have learned about research has been from him.

Thanks also to the NTG faculty – Mostafa Ammar, Ellen Zegura, Jim Xu and Nick Feamster. They are wonderful people to interact with on a personal level, critical and insightful when it comes to research, and great teachers. All the credit goes to them for creating the perfect environment in the NTG. Thanks also to Walter Willinger for serving on my proposal and thesis committees, and for the tremendously useful comments on proposal and thesis drafts that I kept bugging him with. I would also like to thank my mentors at summer internships – Christophe Diot and Renata Teixeira at Thomson and Nick Duffield, Shubho Sen, Lee Breslau, Alex Gerber, Carsten Lund and Cheng Ee at AT&T Labs – for giving me the opportunity to work on challenging research problems not directly related to my thesis.

I couldn't imagine spending the years at Georgia Tech without the company of some of the nicest, smartest and most hardworking coworkers I've known; The "seniors" Sridhar, Pradnya, Shashi, Ruomei, Qi, Minaxi and Christos were around when I joined the group and did their best to make me feel a part of the NTG. I learned a lot just standing around and listening to them discuss their research. Then there are the folks with whom I spent the bulk of my PhD years – Ravi, Manish, Abhishek, Yarong, Srini and Mukarram. These guys made coming back to the lab every day something to look forward to. I will never forget all those coffee breaks, cricket sessions, lunch gatherings and SRGs at GCATT. Finally, the bunch of new arrivals – Partha, Anirudh, Murtaza, Ahmed and others – who never

stopped reminding me of my age (in the Ph.D. program), and staked their claims to the most coveted desk in the lab. They also never refused to sit through various practice talks and dry runs, kept the lab awake at ungodly hours of the night, and in general ensured a lively work environment in the lab. I am fortunate to be able to count many of these current and former labmates among my close friends. Special thanks to Pradnya, Ruomei, Ravi and Manish for the great times!

I have been lucky to have met many interesting people outside of the lab during my years at Georgia Tech. These are too many to name them all, but special thanks to George, VKG, Glen, Harpreet, Agni, Avanish, Mahesh and Sushant. These guys were great outlets when computer science, networking and research all became a bit too much..

Finally, words cannot adequately express my gratitude towards my family, in particular my parents, brother Ashay and sister-in-law Gyanam. If not for their their constant encouragement, help and support over the years, I might never have had the chance to write this. I dedicate this thesis to them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

The Internet consists of thousands of autonomous networks connected together to provide end-to-end reachability. Networks of different sizes, and with different functions and business objectives, interact and co-exist in the evolving "Internet Ecosystem." The Internet ecosystem is highly dynamic, experiencing growth (birth of new networks), rewiring (changes in the connectivity of existing networks), as well as deaths (of existing networks). The dynamics of the Internet ecosystem are determined both by external "environmental" factors (such as the state of the global economy or the popularity of new Internet applications) and the complex incentives and objectives of each network. These dynamics have major implications on how the future Internet will look like. How does the Internet evolve? What is the Internet heading towards, in terms of topological, performance, and economic organization? How do given optimization strategies affect the profitability of different networks? How do these strategies affect the Internet in terms of topology, economics, and performance?

In this thesis, we take some steps towards answering the above questions using a combination of measurement and modeling approaches. We first study the evolution of the Autonomous System (AS) topology over the last decade. In particular, we classify ASes and inter-AS links according to their business function, and study separately their evolution over the last 10 years. Next, we focus on enterprise customers and content providers at the edge of the Internet, and propose algorithms for a stub network to choose its upstream providers to maximize its utility (either monetary cost, reliability or performance). Third, we develop a model for interdomain network formation, incorporating the effects of economics, geography, and the provider/peer selections strategies of different types of networks. We use this model to examine the "outcome" of these strategies, in terms of the topology, economics and performance of the resulting internetwork. We also investigate the effect of external factors, such as the nature of the interdomain traffic matrix, customer

preferences in provider selection, and pricing/cost structures. Finally, we focus on a recent trend due to the increasing amount of traffic flowing from content providers (who generate content), to access providers (who serve end users). This has led to a tussle between content providers and access providers, who have threatened to prioritize certain types of traffic, or charge content providers directly – strategies that are viewed as violations of "network neutrality". In our work, we evaluate various pricing and connection strategies that access providers can use to remain profitable without violating network neutrality.

CHAPTER I

INTRODUCTION

The Internet, commonly described as a "network of networks", consists of thousands of autonomous networks connected together to provide global reachability. Each network is independently operated and managed, and has its own (possibly different) incentives and requirements in connecting to the Internet. Networks with different sizes, functions, and business objectives interact and co-exist in the "Internet ecosystem". Networks are *selfish*, meaning that they are concerned only with maximizing their own utility from connecting to the Internet. Further, the Internet is distributed, where no single entity has global knowledge about the actions and objectives of other networks. As such, we can think of the Internet as a distributed, multi-agent system with strictly local interactions between agents.

An important characteristic of the Internet is that it constantly evolves. A plausible cause for the evolution and dynamics in the Internet is that networks change their connectivity to optimize a utility function, and also respond to external effects such as economic conditions and regulation. The Internet, when viewed as a graph at the interdomain level, is thus a *dynamic graph* that shows birth and death of networks and rewiring of the connectivity of existing networks. A key feature is that this dynamic graph evolves through *local* optimizations, as networks change their set of providers and peers. The dynamics of the Internet ecosystem are also influenced by external "environmental" factors (such as the state of the global economy or the popularity of new Internet applications). Much previous work on interdomain topologies has studied the static properties of the Internet graph, such as the degree distribution or clustering coefficient, without studying how this graph evolves over time. Several important questions thus remain unanswered: How does the Internet evolve? Which types of networks account for most of the growth of the Internet? Are most of the dynamics (links created or destroyed) due to the growth of the Internet or changes

in the connectivity of existing networks? What is the Internet headed towards, in terms of topological and economic organization?

When we view the Internet as a graph, it is important to recognize that all nodes and links are not the same. The networks that constitute the Internet have very different objectives and incentives. For instance, the objective of a transit provider may be to maximize its profit, and it may approach this goal through competitive pricing policies and selective peering. The objective of a content provider, on the other hand, may be to have highly reliable Internet access and to minimize transit expenses, and it may pursue these goals through aggressive multihoming and an open peering policy. Further, interdomain links also have certain semantics associated with them. In particular, networks engage in transit (or customer-provider) relations, and also peering relations. These relations transfer not only traffic but also economic value between networks. Most previous work on interdomain topology modeling has viewed the Internet graph as "flat", where all nodes and links are alike. Further, these modeling efforts were "top-down", meaning that they try to explain certain structural properties of the Internet graph, e.g., the power law degree distribution [10, 15, 90, 107, 113]. This body of work does not try to model the Internet as the outcome of the optimization strategies used by individual networks, and hence cannot provide any insight into which strategies different types of networks should use to maximize their utility. For example, one would like to know which provider and peer selection strategy is most likely to maximize the profitability for different types of networks. Also of interest is the global effect of the strategies used by these networks, e.g., the effect of these strategies on user-perceived cost or performance.

In this thesis, we first measure the evolution of the Internet ecosystem over the last decade. We then develop a first-principles model for interdomain network formation, based on the interactions between different types of networks. We use this model to evaluate the effects of the provider and peer selection strategies used by different types of networks. Our approach differs in several ways from previous research. First, we are interested in the dynamic properties of the topology, rather than static characteristics such as degree distributions or clustering coefficients. Second, we follow a bottom-up approach, modeling

the behavior of different network types as they try to optimize their utility functions, and then observing the emerging global properties. Finally, we recognize the fact that the Internet ecosystem is diverse in the types of networks and interdomain links, and we take into account the different business functions of networks and the semantics associated with interdomain links.

Understanding the evolution of the Internet ecosystem is important for several reasons. First, we believe that there is a need to develop *bottom-up* models for Internet topology evolution and dynamics that capture the complex interactions between different types of networks. As such, it is necessary to study the differences between the types of networks that form this ecosystem, in terms of business function and incentives. Creating such models will give us the ability to better understand global phenomenon in the Internet, and also to study the global effects of local actions. Second, understanding the evolution of the Internet is critical for studying the performance of protocols and applications over time. For instance, to answer questions like "How will BGP perform 10 years from now?", we need to know the properties that the Internet's interdomain topology is likely to show in the future. Studying the evolution of the Internet can help to predict what the Internet may look like in the future. Third, there is much recent interest in generating synthetic interdomain topologies for use in simulations and analysis, e.g, evaluating the scalability of a new routing protocol. A study of the evolution of the Internet can provide valuable inputs to such topology generators, such as how various types of networks connect to each other, and their topological and behavioral properties over time. Further, in light of the recent interest in re-designing the Internet with "clean-slate" approaches, it is crucial to understand how the existing Internet has evolved. Doing so could help design new architectures and mechanisms with the goal of "evolvability", meaning that they have an intrinsic capability to evolve towards states that are desirable in terms of economics, reliability or performance. Finally, from a practical perspective, Internet Service Providers (ISPs) would benefit from a better understanding of Internet evolution. Doing so would help them choose their provider and peer selection strategies that are likely to maximize their utility in terms of monetary profit, costs or performance.

A summary of the main components of this thesis follows:

- First, we study the evolution of the Internet ecosystem over the last decade, using snapshots of the Autonomous System (AS) topology over the last 10 years. We are interested in the *dynamic* properties of the AS graph, rather than static measures. Further, we account for the heterogeneity in the types of ASes and inter-AS links in the Internet, and highlight the need to study these separately. We classify ASes according to their business function (Enterprise Customers, Transit Providers and Content/Access Providers), and study the behavior of these AS types separately. We classify inter-AS links as customer-provider (where the customer pays the provider for Internet connectivity) and settlement-free (where peers agree to exchange traffic for free). We highlight several important trends in the global Internet graph over the last decade, such as densification, constant path lengths, and growth that occurs mostly at the edges. We also identify trends in the behavior of different AS types, in terms of their activity (how often they change their connectivity), multihoming preferences, and the geographical region in which they are present.

- Next, we focus on stub networks (Enterprise customers and Content Providers) at the edges of the Internet. Enterprise Customers (EC) are mostly concerned with minimizing their monetary costs, while Content Providers (CP) try to optimize the performance of their egress traffic. The choice of upstream providers can significantly impact the costs that these networks incur and the end-to-end performance they achieve (to/from their major sources/destinations of traffic). Further, once a stub network selects a set of upstream providers, it needs to determine how to route its egress traffic using that set of providers. In this part of the thesis, we propose algorithms for a stub network to optimize its set of upstream providers. The optimization objective is to minimize the monetary cost incurred while achieving good performance (short AS-level paths and high path diversity) to the major destinations of egress traffic. We show that our proposed algorithms can choose the set of upstream providers that are close to optimal in terms of the resulting costs, AS-level

path lengths and path diversity. In the second part of this work, we propose an algorithm for egress path selection that finds a congestion-free allocation of egress flows to upstream providers (if it exists) with minimum cost for the source network.

- Next, we focus on transit providers in the Internet ecosystem, which are mainly concerned with maximizing their revenue. They may achieve this objective by competitive pricing schemes, intelligent provider selection and selective peering with other transit providers or content providers. In this part of the thesis, we develop a model for interdomain network formation that captures the interdependence between topology, traffic flow and revenue in the Internet. We also account for the interdependence between provider and peer selection by a network. We model the effect of external factors such as economics, geography, and the nature of the interdomain traffic matrix. We then use agent-based simulations to computationally find the equilibrium internetwork, as it is too complex to do using analytical or game-theoretic approaches. We then study the global effects of various provider and peer selection strategies used by transit providers. Anecdotal evidence suggests that there are commonly accepted rules of thumb that ASes use to engage in peering relationships. For example, large transit providers engage in "restrictive" peering, whereby they do not peer unless it is necessary to maintain global reachability. Smaller transit providers typically peer if the traffic they exchange with their peers is roughly balanced (commonly referred to as the "traffic-ratio" requirement). We use our model to determine the conditions under which these strategies are profitable for small and large transit providers. We also study the effects of these strategies on the resulting network in terms of topology (which networks tend to attract customers or peers?), economics (which providers are profitable?), and performance (average interdomain path lengths).

- Finally, we take a technical look at the recent debate over "network neutrality", which concerns the tussle between content providers and access providers. The increasing penetration of broadband access, faster last-mile links, and the rise of Internet video and peer-to-peer file sharing mean that residential and SOHO (Small Office, Home

Office) users download increasingly more traffic. This traffic is delivered to users by Access Providers (APs). APs earn their revenues mostly from their users, and they incur costs to operate their network and to purchase upstream connectivity from transit providers. A much discussed trend in recent times is that APs are not profitable, as the increasing volume of transit traffic leads to escalating costs, while the intense competition in the access market and the commoditization of Internet access leads to falling prices, typically in the form of a flat monthly fee [43, 50, 84]. On the other hand, content providers (CPs) are often seen as being profitable, which has led to considerable tension between APs and CPs. In this work, we use a simple model to study the possible reasons for the non-profitability of access providers. We evaluate the effectiveness of different pricing and connection strategies that the AP can use to remain profitable. Our results indicate that AP strategies that rely on differential pricing mechanisms or non-neutral behavior (directly charging the largest CPs for better performance) are unlikely to succeed in the face of competition in the access market.

1.0.1 Thesis organization

The rest of this thesis is structured as follows. In chapter 2, we study the evolution of the Internet ecosystem over the last decade, highlighting important trends for the entire Internet and also for individual classes of ASes. In chapter 3, we focus on stub networks and content providers at the edges of the Internet, and present algorithms for these networks to optimize their upstream connectivity. In chapter 4, we propose a model for interdomain network formation, capturing the effects of topology, traffic, and the peer selection strategies of transit providers at the core of the Internet. We validate the ability of this model to reproduce some of the features observed in the real Internet, and study the effect of various provider and peer selection strategies on the equilibrium internetwork. In chapter 5, we approach the recent debate on "network neutrality" from a technical standpoint, focusing on strategies for access providers to remain profitable. We conclude by outlining the contributions of this thesis and proposing directions for future work in chapter 6.

CHAPTER II

MEASURING THE EVOLUTION OF THE INTERNET ECOSYSTEM

2.1 Introduction

The Internet, as a network of Autonomous Systems (ASes), resembles in several ways a natural ecosystem. ASes of different sizes, functions, and business objectives form a number of *AS species* that interact to jointly form what we know as the global Internet. ASes engage in competitive transit (or customer-provider) relations, and also in symbiotic peering relations[1]. These relations, which are represented as inter-AS logical links, transfer not only traffic but also economic value between ASes. The Internet AS ecosystem is highly dynamic, experiencing *growth* (birth of new ASes), *rewiring* (changes in the connectivity of existing ASes), as well as *deaths* (of existing ASes). The dynamics of the AS ecosystem are determined both by external "environmental" factors (such as the state of the global economy or the popularity of new Internet applications) and by complex incentives and objectives of each AS. Specifically, ASes attempt to optimize their utility or financial gains by dynamically changing, directly or indirectly, the ASes they interact with. For instance, the objective of a transit provider may be to maximize its profit, and it may approach this goal through competitive pricing and selective peering. The objective of a content provider, on the other hand, may be to have highly reliable Internet access and minimal transit expenses, and it may pursue these goals through aggressive multihoming and an open peering policy.

Our study is motivated by the desire to better understand this complex ecosystem, the behavior of entities that constitute it (ASes), and the nature of interactions between those entities (AS links). How has the Internet ecosystem been growing? Is growth more important than rewiring in terms of the formation of new links? Is the population of

[1]We refer to "settlement free interconnection" as a "peering relation" and "paid transit" as a "customer-provider" relation.

transit providers increasing (implying diversification of the transit market) or decreasing (consolidation of the transit market)? Given that the Internet grows in size, does the average AS-path length also increase? Which ASes engage in aggressive multihoming? What is the preferred type of transit provider for AS customers? Which ASes tend to constantly adjust their set of providers? Are there regional differences in how the Internet evolves? These are some of the questions we ask in this part of the thesis.

Understanding the evolution of the Internet ecosystem is important for several reasons. First, we believe that there is a need to develop *bottom-up* models of Internet topology evolution that capture the complex interactions between the constituent entities. As such, it is necessary to study the differences between the types of ASes that form this ecosystem in terms of business function and incentives. Second, understanding the evolution of the Internet is critical for studying the performance of protocols and applications over time. For instance, to answer the question "How will BGP perform 10 years from now?" we first need to answer the question "How will the Internet look 10 years from now?". Third, there is much recent interest in generating synthetic AS graphs for simulation and analysis. A study of the evolution of the Internet can provide valuable inputs to such topology generators, such as the types of ASes in the Internet and their topological and behavioral properties over time. Finally, in light of the recent interest in re-designing the Internet with "clean-slate" approaches, it is crucial to understand how the existing Internet has evolved. Doing so could help us identify new architectures and mechanisms that have an intrinsic capability to evolve towards desirable economic, reliability and performance conditions.

There is an extensive literature on AS-level topology measurement and modeling. A large portion of that literature, however, takes a graph-theoretic perspective, viewing all ASes as nodes in a graph and all inter-AS relations as edges, without considering the type of relation (customer-provider versus peering) or the role of the participating ASes (customer versus provider). Viewing all ASes as the same type of node ignores the major differences in the function and objectives of different ASes. Further, even though most of the previous work on AS-level topology modeling mentions the terms 'evolution" or "dynamics", the

main focus has been on measurements and modeling of growth, ignoring rewiring. The latter is very important, however, as it represents the attempt of individual ASes to optimize their connectivity. Finally, most of the earlier work on AS-level topologies has focused on macroscopic properties and metrics, such as the degree distribution, the clustering coefficient or the graph diameter, without considering the local policy and semantics of inter-AS relations. The latter are very important as they control the flow of traffic and value in the AS ecosystem.

In this part of the thesis, we attempt to measure and understand the evolution of the Internet ecosystem during the last decade (1998-2007). We propose a method to classify ASes into a number of types depending on their function and business type, using observable topological properties of those ASes. The AS types we consider are large transit providers, small transit providers, content/access/hosting providers, and enterprise networks. We are able to classify ASes into these AS types with an accuracy of 80-85%. We focus on *primary* inter-AS links, meaning links that are used under "normal operating conditions", to distinguish with backup links that appear under failure conditions or routing convergence. We also consider the semantics of inter-AS links, in terms of customer-provider (CP) versus peering (PP) relations, and distinguish between the customer, provider and peering role of an AS in each relation. Unfortunately, we find that the available historical datasets from RouteViews and RIPE are *not sufficient to infer the population and evolution of peering links.* So we restrict the focus of this study to the evolution of the population of AS types and of customer-provider links.

The rest of this chapter is structured as follows. In Section 2.2, we describe the data collection and filtering methodology. In Section 2.3, we focus on the evolution of the global Internet. In Section 2.4, we present a classification scheme of ASes into four AS types based on their expected business function. Then, we examine the evolution of each AS type at a global scale as well as regionally. In Sections 2.5 and 2.6, we investigate the evolution of customer-provider relations in the Internet, from the perspective of the customer and provider, respectively. In Section 2.7, we present some results on the evolution of the Internet peering ecosystem. These results should be viewed as "conjectures" because of the

limitations in detecting the complete set of peering links. We conclude with a summary of our main findings in Section 2.9.

2.2 Datasets and methodology

A study of the evolution of the Internet ecosystem needs frequent snapshots of the AS-level Internet topology, annotated with policy information for each link. Given that such historical information is not available, we have to rely on measurement and inference, collecting data from multiple sources and considering the limitations of each dataset. This section describes the datasets we use and the subsequent filtering and validation processes.

We collected BGP AS-paths from BGP table dumps obtained from the two major publicly available repositories at RouteViews [96] and RIPE [94]. The RouteViews collection process started in November 1997, providing an invaluable resource in the past ten years. The first RIPE collector became active in October 1999. We rely only on these two repositories because no other source of topological/routing data (routing registries, traceroutes, looking glass servers, etc.) provides historical information. Note that the use of AS-paths has been shown to be inadequate to expose the *complete Internet topology* [29, 32, 59]. In particular, even though most ASes are detected, a significant fraction of peering and backup links at the edges of the Internet are missed [26, 59, 111]. In fact, it has been estimated that there are at least 40% more peering links in the Internet than those obtained from AS-paths [26, 32]. We are well informed of these limitations, which are further exposed later in this section. There are, however, three important points to consider. First, *we do not aim to detect backup links*; instead, we are only interested in primary Internet links, used most of the time (as opposed to backup links that are only used upon failures or overload conditions). We describe later how to avoid backup links in the data filtering process. Second, *the main focus of this evolutionary study is customer-provider links*. As we show later in this section, the available monitors from RouteViews and RIPE are not enough to detect all peering links or the births and deaths of those links. Third, even though missing links can be detrimental for complex inference applications (such as AS path prediction or BGP root-cause analysis), it has been shown recently that they are *less critical in topology*

inference [112].

Filtering of backup and transient links: Next, we describe how to only detect primary links, avoiding backup links and false AS-paths that often appear during BGP convergence. First, note that short-term failures and routing transient events can "confuse" an evolutionary study, misinterpreting link disappearances and appearances due to transient failures as link deaths and births respectively. For instance, suppose that the primary link l_p between AS-x and AS-y fails at time t_1, causing the activation of a backup link l_b between AS-x and AS-z. l_p is repaired at t_2 and the connectivity returns to its original state. Since we focus on primary links, our goal is to ignore the transient event during (t_1, t_2) and to *not* detect l_b. On the other hand, a change of routing policy that exchanges the role of links l_p and l_b (so that l_b becomes the primary link) should be detected as the death of l_p and the simultaneous birth of l_b.

To achieve the previous objective we follow the "majority filtering" approach described next. Note that a *snapshot*, in the following discussion, does not refer to a time instant but to a period of 21 days. During a certain snapshot, we collect at N different times the unique AS-paths that are exported from all active RouteViews and RIPE monitors. The period between these successive *samples* is T_s, with $N\,T_s$=21 days. Then, *we keep only those AS-paths that appear in the majority of the samples and ignore the rest.* This process is designed to filter out links that appear due to routing transient events, as well as due to "hard" failures of interdomain links (e.g. due to router crashes or fiber cuts). Routing transients typically persist for less than a few hours, while it is reasonable to expect that hard failures are repaired within 10 days. In each of these cases, the majority filtering rule successfully filters out the transient links.[2] Note that if a certain link X-Y is used as primary in one AS path but as backup in another path, it will be included in our snapshot.

To select an appropriate value of N, we do the following. We collect all visible AS-paths for each day of January 1998. Next, we divide the month into N blocks of the same duration, and collect the set of visible AS-paths from a randomly selected instant in each of

[2] A similar process was used by Dimitropoulos et al. [40], but considering an AS-path only if it appears in *all* N samples.

the N blocks. Then, we perform majority filtering, considering only AS-paths that appear in the majority of the N samples. Finally, we measure the number of visible AS links. We vary N from 1 to 10, and repeat the previous process multiple times for each value of N. As N increases, the average number of visible links decreases (from about 5850 to 5725 during that month) because fewer backup links become visible. Additionally, the variability in the number of visible links decreases. We observe that N=5 results in about the same average as higher values of N, and reasonably low variance (standard deviation of 12 links). In the rest of this study, N=5 samples.

The trade-off behind the selection of the snapshot duration (21 days in our study) is explained next. If the snapshot duration is too long (say more than a month), then we may miss several birth-death (or death-birth) transitions of the same link. On the other hand, if the snapshot duration is too small (say a few days), then the majority filtering mechanism may not be able to filter out backup links that appear during long-lasting failures such as fiber cuts. Finally, a new snapshot is collected every three months, providing us with 40 snapshots (10 years) from January 1998 to October 2007.

Variations in the number of active monitors: Another issue we need to consider is that the number of BGP monitors in both RouteViews and RIPE has been increasing significantly over the last ten years, from about 10 in 1997 to almost 400 at the end of 2007. The increase in the number of monitors has been less than 20% in 35 out of the 39 pairs of successive snapshots. As the number of monitors increases, some previously existing links may become visible for the first time at a certain snapshot. How do we distinguish those first appearances of existing links from genuine link births? Similarly, sometimes monitors are removed. How do we distinguish between the disappearance of existing links from genuine link deaths? Also, can we bound the estimation error in the number of link births and deaths between each pair of successive snapshots?

To answer the last question we perform the following analysis. Let the set of monitors at snapshots T_1 and T_2 be $\mathcal{M}_1$ and $\mathcal{M}_2$ respectively. Let $\mathcal{L}_1$ and $\mathcal{L}_2$ be the set of links observed at T_1 and T_2, respectively. $\mathcal{L}_2 - \mathcal{L}_1$ is the set of new link appearances between T_1 and T_2. This is our estimate for the set of new link births. This set includes the links

that were genuinely born between T_1 and T_2, but it may also include an error term that is the set of links that were present at T_1 but became visible at T_2 due to the monitor set increase. To derive an upper bound for the latter, we do the following.

First, determine the set of links $\mathcal{L}'_2$ that would be observed at T_2 *using the set of monitors that were common between* $\mathcal{M}_1$ *and* $\mathcal{M}_2$, i.e., $\mathcal{M}_1 \cap \mathcal{M}_2$. The set $\mathcal{L}'_2 - \mathcal{L}_1$ (where $\mathcal{L}'_2 - \mathcal{L}_1 \subseteq \mathcal{L}_2 - \mathcal{L}_1$) includes links that were definitely born between T_1 and T_2, and hence it gives a *lower bound* on the number of actual link births. On the other hand, the number of links in the set $(\mathcal{L}_2 - \mathcal{L}_1) - (\mathcal{L}'_2 - \mathcal{L}_1)$ is an upper bound for the error between the estimated and actual number of link births. So, the *worst case relative error* (WCRE) in the number of link births between T_1 and T_2 is:

$$\text{WCRE} = \frac{|(\mathcal{L}_2 - \mathcal{L}_1)| - |(\mathcal{L}'_2 - \mathcal{L}_1)|}{|(\mathcal{L}'_2 - \mathcal{L}_1)|} \tag{1}$$

We measured the WCRE for every pair of snapshots. In 30 out of the 39 snapshots pairs, the WCRE is less than 10%. For all but one pair, the WCRE is less than 20%. In the remainder of this paper, we omit the pair of snapshots for which the WCRE was larger than 20% (Jan-Apr 2000). We also measured the WCRE separately for customer-provider (CP) links and peering (PP) links. Unfortunately, the WCRE is very high for peering links and in 9 out of 39 snapshots it is greater than 100%. On the other hand, the WCRE for CP links is quite low, and for all except one pair of snapshots (Jan-Apr 2000), it is less than 10%.[3]

The previous analysis considers the effect of an increased set of monitors on the measurement of link births. A similar problem occurs while measuring link deaths, as some monitors are occasionally disconnected temporarily or permanently from the RouteViews and RIPE collectors. We performed a similar analysis to determine the effect of monitor deaths on the estimated number of link deaths. We find that the WCRE in the estimated number of link deaths is less than 10% for 37 out of the 39 snapshot pairs.

The previous WCRE analysis showed that, even though we can estimate well (within 10%) the link births/deaths of CP links, we do not get a reasonable accuracy for the

[3]Note that the WCRE is calculated for every pair of snapshots, and so it does not accumulate over time.

link births/deaths of PP links. This is a negative but significant result, which should be considered by future studies that rely on RouteViews and RIPE topological data. It also implies that the conclusions of several previous topological studies should be re-examined.

Sensitivity of population counts to number of monitors: We next examine the visibility of CP and PP links, as well as of ASes, when we vary the number of used monitors. Consider first the population of ASes. Let n_{AS} be the set of visible ASes if we use all available monitors at a given snapshot. We then randomly select a fraction f of the available monitors, and determine the population of ASes that is visible using that subset of monitors. We repeat this experiment 100 times, and determine for each run the number of ASes visible with a fraction f of the available monitors $n_{AS}(f)$. Figure 1 shows the median, 10th and 90th percentile values of the ratio $n_{AS}(f)/n_{AS}$ for the snapshot Jan 2007, together with the corresponding ratios for the populations of CP links and PP links. We repeated this analysis for all snapshots, and the results are quantitatively similar across time, without any noticeable trends.

Notice that the number of visible ASes is strongly insensitive to the number of available monitors. Even with 10% of the monitors we practically see the same set of ASes that is visible with all monitors. The fraction of CP links is also insensitive to the number of available monitors, as long as we use more than 60-70% of the available monitors in the given snapshot. So, we expect that a 10-20% increase in the number of available monitors across successive snapshots will not cause a significant variation in the number of visible CP links. The situation is very different with PP links however. The fraction of visible PP links increases roughly linearly with the fraction of used monitors. This means that if we had more monitors we would probably see significantly more PP links. So, *the estimated population size of PP links should be viewed as lower bound on the actual population size.* Similar observations were recently reported by Oliveira et al. [87].

The previous observations have two major implications. First, on the positive side, *it appears that the RouteViews and RIPE historical datasets contain enough monitors to detect the ASes and CP links in a robust manner.* Even though we cannot be certain that we see *all* ASes or CP links, we at least have evidence that these populations would not differ by

a large number if we had more monitors. Second, on the negative side, *it is clear that the RouteViews and RIPE datasets are not sufficient to detect the population or the birth/death rates of PP links.* Consequently, in the rest of the paper we focus on the evolution of CP links. When we present some results for PP links, the reader should recall that those figures are lower bounds on the actual number of PP links.

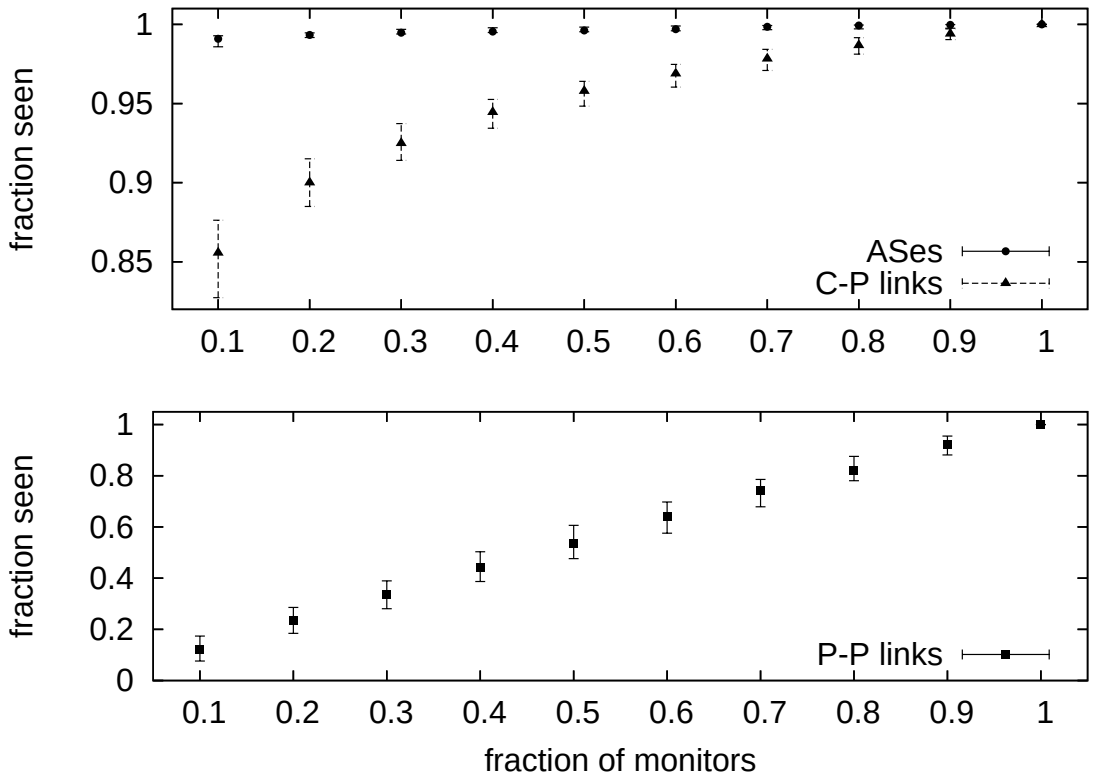

Figure 1: Visibility of ASes, CP and PP links as a function of the number of monitors used in a snapshot.

Policy inference: After collecting and filtering the data as described earlier, the final data processing step is to use the AS-paths in each snapshot (those that passed the majority filtering process) to infer the underlying AS topology and the relationships between adjacent ASes. For this purpose we use the well-tested algorithm described by Gao in [51]. Despite the significant follow-up work on AS relationship inference [40, 101], we prefer Gao's algorithm because of its ability to infer relationships using only observed AS paths, without any additional information such as data from routing registries or active probes. Comparison

studies for the accuracy of related algorithms in [101] and [51] have shown that Gao's algorithm is more accurate in identifying peering relationships. Further, a recent study [112] showed that the AS relationship inferences made by this algorithm are quite stable with respect to variations in the observed AS-paths. Gao's algorithm results in four types of AS relationships: Customer-Provider (CP), Peering (PP), Sibling, and Unknown. We ignore the last two categories, as they account for less than 2% of the visible links in any snapshot.

Finally, the AS topology and relationship matrix provide an annotated graph for each snapshot. The differences between successive snapshots show the evolutionary events of link and node births and deaths, which form the core of the analysis in the following sections. Note that if a certain link has changed role at some snapshot (say from CP to PP), we view that event as the death of a CP link and the simultaneous birth of a PP link between the corresponding ASes. The reader may be wondering about the frequency of link type changes, from CP to PP or the opposite. Even though we cannot answer this question in a definite manner (due to the visibility problem with PP links), we measured that 9% of the PP links in a snapshot become CP links in the next snapshot (This number is the average over all pairs of snapshots). The fraction of CP links that become PP links appears to be much less (1%) but that is probably due to the poor visibility of PP links. Also, these changes are not cumulative, as we run the relationship inference algorithm separately for each snapshot.

2.3 Growth and rewiring trends

We first examine the evolution of some major characteristics of the global Internet.

Growth of ASes and inter-AS links: Figure 2 shows the number of ASes and inter-AS links in each snapshot. Due to the previously discussed issues with measuring PP links, we only count the number of CP links in each snapshot. A first observation is that, despite the economic recession of 2001-03 and the well documented turmoil in the telecom market, *the Internet AS-level topology has been increasing in size over the last ten years.* Second, it appears that the Internet has gone through two distinct growth phases so far: *an initial phase, up to mid-2001, in which the Internet grew exponentially in terms of the number of*

ASes and links (of the form $y = a*e^{bx}$). Then, the growth process switched to linear for both the number of ASes and links (of the form $y = ax + b$). We find that the number of ASes from 1998 to mid-2001 can be modeled as $y = 3150*e^{0.094x}$, where x is the snapshot number ($x = 0, 1, \dots$). In the last six years, the number of ASes can be modeled as $y = 2537+604x$. Regarding the number of CP links, the corresponding functions are $y = 5462 * e^{0.102x}$ and $y = 1499x - 35$. Each of the previous regression formulae gives a correlation coefficient that is at least 99%. To eliminate the possibility that this trend shift is an artifact of the measurement infrastructure (e.g. the changing set of monitors), we measured the number of visible ASes and CP links with a set of monitors that remained the same in the last ten years. The results, even though revealing a lower number of links, still show a trajectory change from exponential to linear in mid-2001. Huston [61] observed a similar trend shift in the number of ASes (but not CP links) around mid-2001.

To determine the boundary at which the trajectory shifted from exponential to linear, we perform the following test. We assume that the number of CP links and ASes can be modeled as $y = a\,e^{b\,x}$ when $x \leq z$ and $y = a\,x + b$ when $x > z$. We then compute the value z_{min} that minimizes the total sum-of-squares error (SSE) for the above regression formula. z_{min} is our estimate for the snapshot where the growth trajectory changed from exponential to linear. It appears that the exponential phase lasted for the first 15 snapshots for ASes and 16 snapshots for CP links, *ending in mid/late 2001.* Figure 2 also shows the exponential and linear regression curves for the number of ASes and CP links.

Evolution of CP link count (and lower bound estimates of PP link count): Next, we distinguish between CP and PP links, and examine the growth trends separately for these two link types. We emphasize again that the number of PP links we report here should be viewed as a *lower bound* on the actual number of peering links. Figure 3 shows the number of CP and PP links, as well as their fractions, over time. Both link types have been increasing in absolute numbers. As shown earlier, the number of CP links shows an initial exponential growth followed by a linear growth after 2001. Modeling the growth of PP links is difficult with the given measurements. It appears, however, that that growth process has followed a different trajectory than that of CP links.

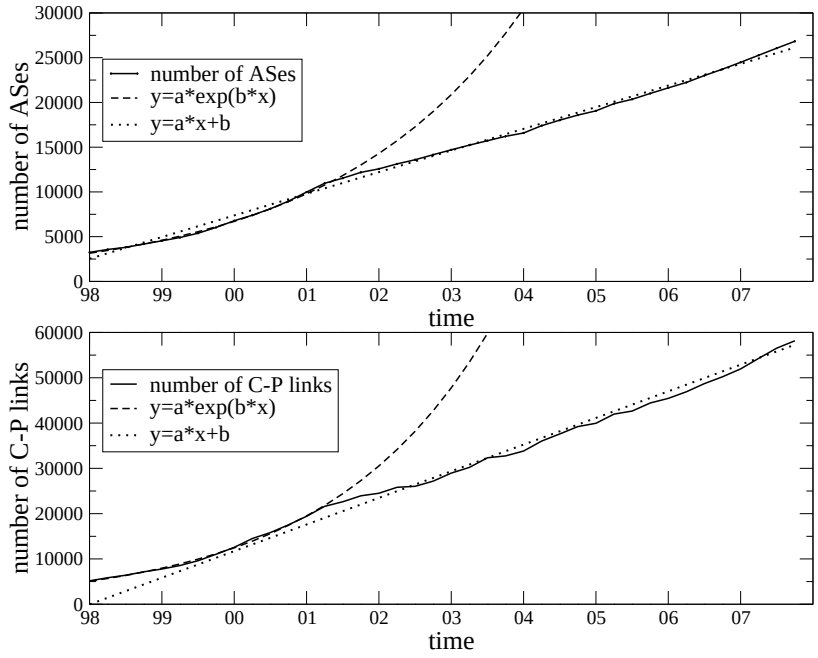

Figure 2: Evolution of the number of ASes and CP links. The regression curves are also shown.

The bottom panel in Figure 3 shows the fraction of CP and PP links. *The fraction of PP links has been increasing steadily after 2001*, even though the growth rate of CP links is larger than that of PP links. The reason is that the relative increase rate of PP links is larger than that of CP links. Given that we probably underestimate the number of PP links, *the fraction of PP links at the end of 2007 is at least 20%.*

Evolution of AS-path length and multihoming trends: Next, we investigate the evolution of the average AS-path length (after removing AS-path prepending). We do so by calculating the average length of AS-paths observed at Routeviews and RIPE collectors in each snapshot. The upper panel in Figure 4 shows that *the average path length measured in this manner has remained practically constant (at 4.2 AS hops) over the last 10 years.* Note that the AS-paths measured here are those that are seen by the Routeviews and RIPE vantage points. The path advertisements seen by these monitors are mostly those exported over customer-provider links. Peering links low in the hierarchy can make paths shorter, as they provide shortcuts in end-to-end paths. Those AS-paths, however, would not be seen from the set of vantage points at Routeviews and RIPE. Consequently, what we measure

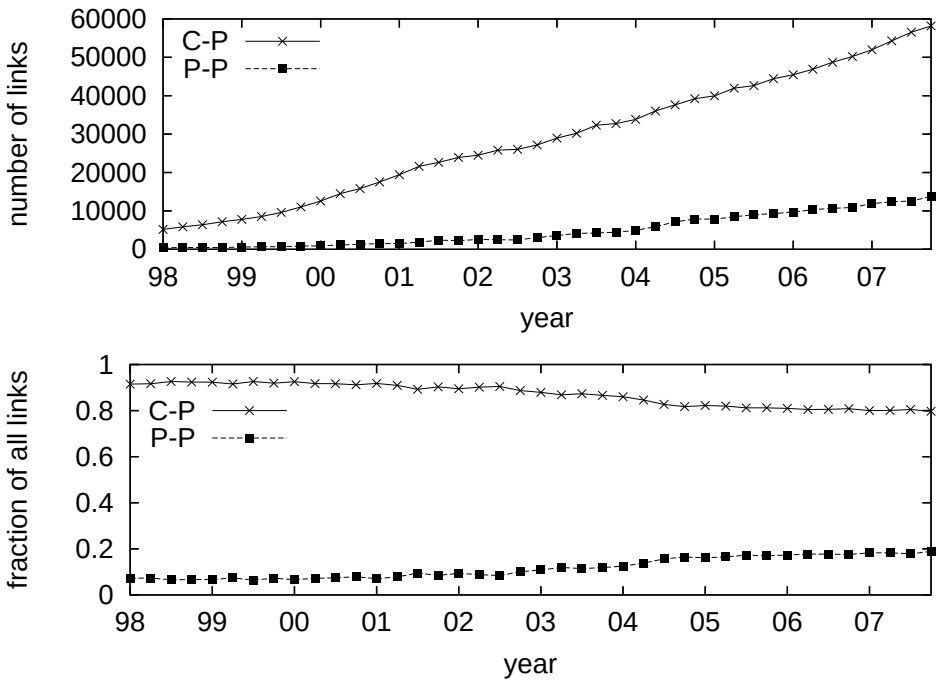

Figure 3: Evolution of CP and PP links in absolute numbers and as a fraction of the total number of links.

may be an overestimation of the AS-path length; the average path length could actually be *decreasing* over time. This is interesting, given the significant growth of the underlying network. Earlier modeling work, such as the preferential attachment growth model of Albert and Barabasi [10], predicted an average path length that grows slowly with the size of the network (O(ln ln n)), when a newly attached node has at least two edges. Such a growth model would result in an increase in the average path length from 4.2 to 4.7 over the last 10 years, contrary to the constant average path length of 4.2 that we observed.

There are two plausible effects that could lead to constant or decreasing AS-path lengths. The first is the increasing presence of "shortcut" peering links, especially between providers at lower tiers in the hierarchy. Due to the aforementioned visibility problem, however, AS-paths that we measure at Routeviews and RIPE collectors would not show the effect of the increasing number of peering links. Studying the effect of peering links on average path lengths would need more accurate topology data with a good visibility of peering links. The second effect that could lead to constant path lengths is a *densification process* that

increases the average degree (considering only CP links) of ASes. As most CP links are visible from Routeviews and RIPE monitors, it is possible to measure the densification of the graph of CP links. Indeed, the upper panel of Figure 4 shows that the average AS degree, counting only CP links, has increased consistently over time, from 3.2 links to 4.3 links per AS. The median degree (not shown) is dominated by small networks that have just 1 or 2 providers, and hence it does not show an increasing trend. This densification process has also been studied by Leskovec et al. [72], who observed that the effective diameter [4] of the AS graph *slowly decreases* with time.

A plausible explanation for the densification of the Internet is the increasing popularity of *multihoming* for economic, reliability and performance reasons. The bottom panel of Figure 4 shows the average *multihoming degree*,[5] defined as the number of providers of a given AS, for two broad classes of ASes: *stubs* (i.e., ASes that never had customers during their observed lifetime), and *non-stubs* (the rest of ASes). We find that *the average multihoming degree has been increasing in both classes.* Non-stubs, however, have been increasing their average multihoming degree much faster than stubs (from 1.5 to about 3.5), in particular after 2003. This may be because non-stubs, which are typically content/access/hosting/transit providers, attempt to optimize their connectivity, and at the same time improve their reliability, by multihoming to several upstream transit providers. For many stubs, on the other hand, one or two (primary) transit providers is often enough.

Growth versus rewiring: Next, we seek to understand the relative significance of *growth versus rewiring.* Growth refers to the addition of new ASes in the network (together with their corresponding links), while *rewiring refers to changes in the connectivity of existing ASes.* Specifically, we focus on the number of CP link births due to AS births (growth) versus CP link births due to rewiring. We also look at the number of CP link deaths due to AS deaths versus CP link deaths due to rewiring. The top panel of figure 5 shows, for each pair of snapshots, the number of CP link births due to AS births and due to rewiring. Initially, the CP link births due to AS births and rewiring were comparable in

[4]The effective diameter of a graph is the minimum value of d such that at least 90% of the connected node-pairs are at distance at most d. A smoothed version of this metric is used in [72].

[5]Multiple physical links between two ASes are counted as a single inter-AS link.

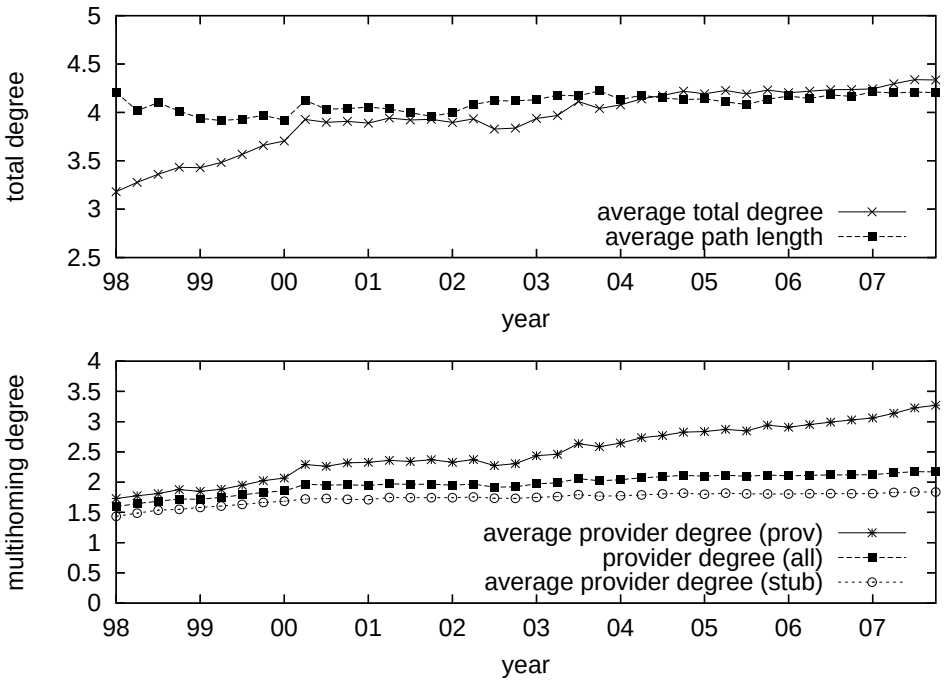

Figure 4: Evolution of average AS degree, AS-path length, and multihoming degree.

number. Since 2001, however, we find that the number of CP link births due to internal rewiring has increased much faster than that due to AS birth. Currently, *around 75% of link births are associated with existing ASes (rewiring).* A similar analysis, shown in the bottom panel, shows that the number of CP link deaths due to rewiring is significantly higher than that due to AS deaths. About 80% of the link deaths are due to rewiring and this fraction is increasing. These observations are important for two reasons. First, most of the literature on AS topology modeling has focused on growth, ignoring rewiring. Second, rewiring represents the effort of individual ASes to optimize their performance, reliability, profitability or other objectives. An intriguing possibility is that rewiring implies that the Internet, as a multi-agent and self-organized system, attempts to optimize a certain, still unknown, global objective in a distributed manner. This possibility has also been discussed by Chang et al. [24].

Given the increasing significance of rewiring, we next focus on the births and deaths of links between existing nodes in two successive snapshots. Let G_1 and G_2 be the graphs representing the primary AS topology in two consecutive snapshots. We construct G_1' from

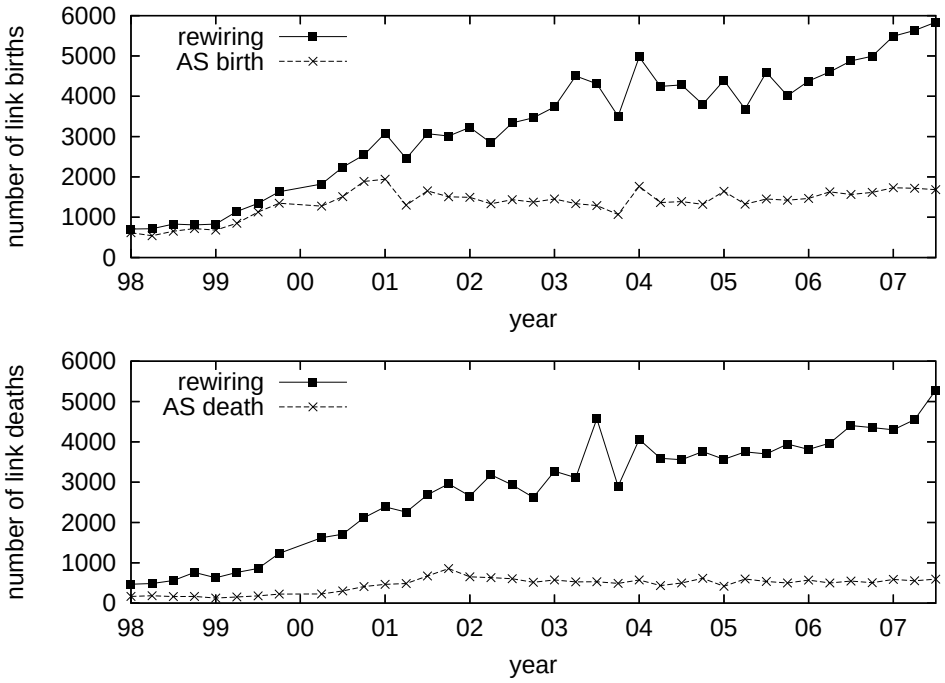

Figure 5: Evolution of the number of CP link births (and deaths) due to node births (and deaths) versus rewiring.

G_1 by removing all nodes that are not present in G_2; similarly construct G_2' from G_2. Note that G_1' and G_2' have the same set of nodes. Let E_1' and E_2' be the set of links in G_1' and G_2' respectively. We use the following graph-level metric, referred to as **Jaccard Distance**, to quantify the rewiring between G_1' and G_2'.

$$s(E_1', E_2') = \frac{|(E_1' - E_2') \cup (E_2' - E_1')|}{|E_1' \cup E_2'|} \tag{2}$$

Note that $s(E_1', E_2')$ captures both link births and deaths between the two snapshots. The Jaccard distance thus quantifies the difference between the sets of links in two consecutive snapshots. For example, a Jaccard distance of 0.5 indicates that 50% of the links seen in the two snapshots were either born before the second snapshot or died after the first.

We calculate the Jaccard distance separately, first, on the CP graph where the customer is a stub, and second, on the CP graph where the customer is a non-stub. Figure 6 shows these metrics for each pair of snapshots over the last 10 years. We find that the Jaccard distance is much smaller for the CP graph where the customer is a stub, as compared to the CP graph where the customer is a non-stub. This indicates that *non-stubs have*

consistently been more aggressive than stubs in changing their upstream connectivity. We further investigate this effect after proposing a finer classification of AS types in the Internet in the next section.

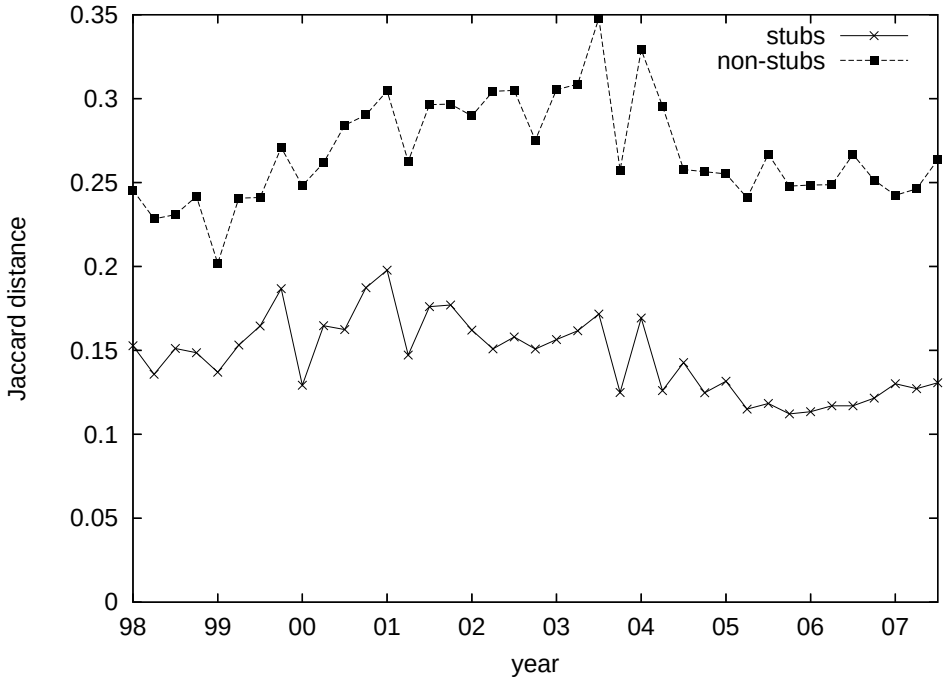

Figure 6: The Jaccard distance for CP links where the customer is stub versus non-stub.

2.4 Evolution of AS types

When we think of the Internet as a graph, it is important to recognize that not all nodes are the same. ASes connect to the Internet with different requirements and business interests, and hence optimize their connectivity in different ways [43]. The topology changes that we observe represent the outcome of a complex multi-constraint optimization process that individual ASes conduct.

AS classification scheme: We propose a simple classification scheme for ASes according to their business type. The initial classification consists of the following five AS types.

Enterprise Customers (EC) represent various organizations, universities and companies at the network edge that are mostly users, rather than providers of Internet access, transit or content. Typically, ECs do not have AS customers.

Small Transit Providers (STP) are often regional ISPs that provide Internet access and transit services. STPs aim to maximize their customer base in their geographical area and to reduce their upstream transit costs through *selective peering* with other regional ISPs. STPs often peer selectively rather than openly to avoid peering with ASes already in their customer tree, or ASes that are likely to become customers at a future time. We count national and academic/research transit networks also as STPs.

Large Transit Providers (LTP) are international ISPs with a large footprint, both in terms of number of AS customers and geographical presence. LTPs aim to maximize their customer base, peering with other ASes only when it is necessary to maintain reachability (*restrictive peering*).

Access/Hosting Providers (AHP) are ISPs that offer Internet access (e.g., DSL, cable modem, dial-up, leased lines) and/or server hosting. Their access customers can be residential users or enterprises that do not have AS numbers, while their server hosting customers are content/service providers that also do not have AS numbers[6]. AHPs often engage in selective peering to minimize the transit costs paid to their upstream providers.

Content Providers (CP) are not in the business of offering Internet transit or access. Instead, their revenues result from providing content that users pay for. CPs aim to minimize transit costs, and so often have *open peering* policies.

Similar classifications have been proposed in previous work. Chang et al. [28] classified ASes (for the purposes of determining interdomain traffic matrices) into "web hosting", "residential access" and "business access". Dimitropoulos et al. [41] classified ASes into large and small ISPs, customer networks, universities, Internet exchange points and network information centers. We chose the previous five AS types based on the terminology used in discussions on the NANOG mailing list and in W. Norton's white papers [84].

Note that the difference between LTPs and STPs is quantitative, as both AS types have the same business function. LTPs are basically the major ISPs that are often referred to, rather informally, as "'tier-1" transit providers. The "tier-1" label is often associated with

[6]A limitation of AS topologies derived from BGP tables is that they include only the organizations that have AS numbers.

10-20 ASes. We choose to be more inclusive, defining as LTPs *the top-30 ASes* in terms of the average number of customers during the time period in which an AS was seen in the last decade. That average is larger than 140 AS customers for the LTPs in our datasets.

This leaves us with around 27,000 ASes (in the latest snapshot) that cannot be classified manually. Instead, we first pick *a training set of 50 ASes* for each of the remaining four AS types (EC, STP, AHP and CP) that are definitely of the corresponding type (based on information obtained from their webpages). For ECs, we pick well-known universities and corporations. For STPs, we choose transit providers that are mostly regional in terms of their coverage and customer size. For CPs and AHPs, we pick well-known content providers, hosting sites, and large broadband/dial-up residential/business access ISPs. Next, we observe the topological properties of the ASes in each training set, in terms of the *average* number of customers C, providers P, and peers R for that AS in the last decade. We found significant overlap in the number of providers among the four AS types, and so we do not rely on that metric. On the other hand, the number of customers and peers (C, R) allows us to distinguish between ECs, STPs and CPs. Unfortunately, we are unable to distinguish CPs from AHPs. These two AS types, even though have different business roles, largely overlap in terms of both C and R. So, in the rest of the paper we merge these two AS types in what will be referred to as **Content/Access/Hosting Providers (CAHPs)**. Figure 7 shows the average number of customers and peers for ASes in the four training sets. Most ECs have zero customers and peers, and they are not shown in this graph.

The next step is to determine a set of boundaries in the two-dimensional (C, R) space that separate the training sets of the four AS types with the minimum number of misclassifications. We apply the well known machine learning technique of *decision trees* on the training samples to obtain the following C and R coordinate boundaries for each AS type:

$$\text{EC: } C < 2.1,\ R <= 1$$

$$\text{STP: } 2.1 \leq C < 140,\ R < 3.5 \text{ and } 33.1 \leq C < 140,\ R \geq 3.5$$

$$\text{LTP: } C \geq 140$$

$$\text{CAHP: } C < 2.1,\ R > 1 \text{ and } 2.1 \leq C < 33.1,\ R \geq 3.5$$

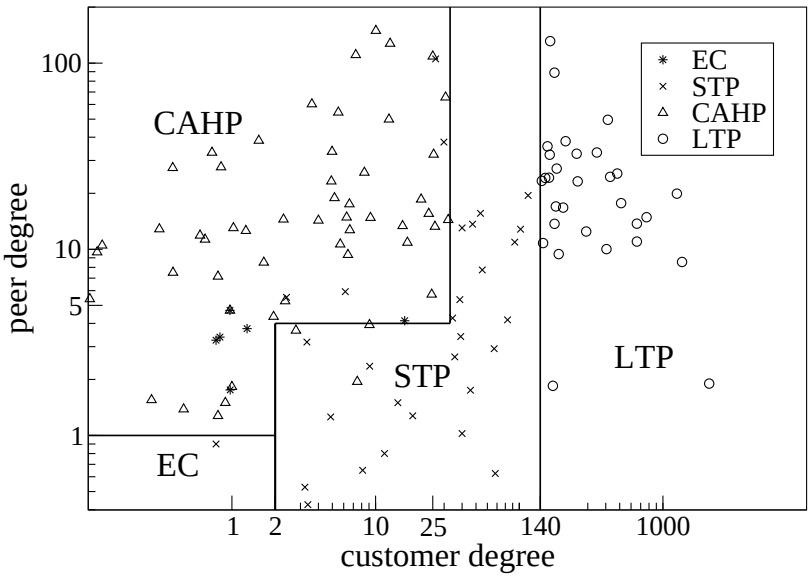

Figure 7: Coordinate boundaries for the four AS types we consider.

Based on the previous boundaries, we next use the average C and R values of each AS (measured over the snapshots in which that AS was present in the ten-year dataset) to classify it into one of the four AS types. Note that the AS types we consider are quite distinct from each other in terms of their function and business goals. It is thus reasonable to expect that ASes do not change from one AS type to another during their lifetime. To examine this hypothesis, we performed the following test. We rerun the decision tree algorithm to classify each AS using a two-year dataset from 2006 and 2007. We then compared this more recent classification with that based on the ten-year dataset. We found that only 3% of the ASes that appear in both datasets were classified differently. In most of these cases, it appears that the classification change was due to a large shift in the customer and peer degrees of that AS. For example, AS-1 has a large average customer degree over the ten-year dataset and is classified as an LTP. However, in the two-year dataset it has a customer degree of 0, and is classified as an EC. AS1 was originally owned by Genuity Inc., a large global ISP. In 2004, Genuity sold AS-1 to Level3 Communications, also a global ISP. Level3 does not use that AS number for its transit services, and this is why that AS has no customers in the last couple of years.

To evaluate the accuracy of the previous classification scheme, we perform the following.

We select a random sample of 150 ASes (50 ECs, 50 STPs and 50 CAHPs), and mix these samples to remove any information about the classification of these ASes (to avoid any subjective bias in the validation process). Then, we use information from WHOIS servers and the webpages of those ASes to infer their main business function. If the actual business function does not match the classification produced by our algorithm, we count that AS as a misclassification. We find that the classification accuracy for ECs is 78%. The errors in this category are due to some residential access providers that are classified as ECs because they have no AS customers and no peers. The accuracy for STPs is 86%. The errors here are due to ASes that mainly offer content hosting services. These providers have few AS customers and a small number (or none) of peers and hence they get classified as STPs. The classification accuracy for CAHPs is 86%. The errors in this case are mostly due to some academic/research backbones that get classified as CAHPs due to their large number of peers. Dimitropoulos et al. [41] reported a similar accuracy figure (78%) for their AS classification scheme.

Population trends for each AS type: Figure 8 shows the population of each AS type over the last ten years. These curves show two distinct phases, similar to the global growth trends observed in Section 2.3, with a change of slope around 2001. The STP population shows a small growth rate (increase by factor of 1.23 over the last six years). The LTP population remains almost 30 by definition. The EC population shows a strong growth trend (increase by factor of 2.33 in the last six years), contributing most of the growth in the number of ASes. The CAHP population, even though much smaller in absolute numbers than ECs, has also been growing significantly (increase by factor of 1.6 in the last six years). ECs and CAHPs represent the periphery of the network, where the users and content reside. If we judge by the population of this AS type, *the Internet edge grows at a significant and stable pace.* On the other hand, LTPs and STPs represent the core of the Internet. Even though the STP population was growing significantly before 2001, their growth rate in the last few years has decreased. This may be an indication that the number of transit providers is stabilizing.

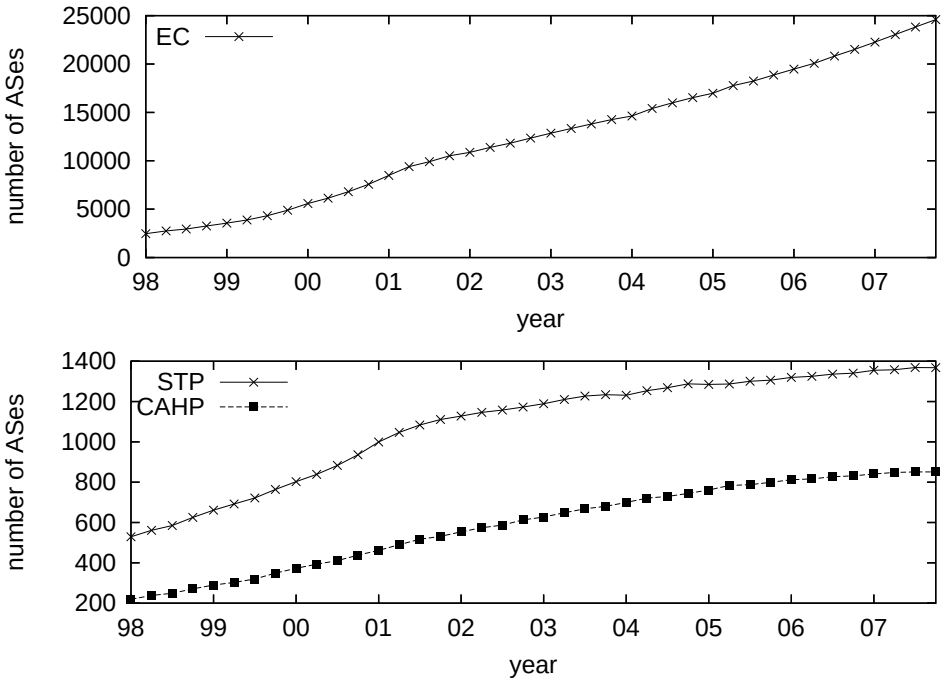

Figure 8: Evolution of the population of AS types.

Geographical trends for each AS type: To classify ASes into broad geographical regions, we use the "registry" field of the corresponding WHOIS entries. Figure 9 shows the fraction of ASes of each AS type that were registered to ARIN (North America) and RIPE (mostly Europe). The other registries (APNIC, LACNIC and AFRINIC) account for the remaining small fraction, and are not shown here. Interestingly, we see that the population of ECs in the two continents (NA and Europe) converges. It is likely that in the next few years there will be more ECs registered in Europe than in North America. This has already happened in the case of STPs, and the number of STPs is now slightly higher in Europe. LTPs, though, are mostly still based in North America. On the other hand, the fraction of CAHPs in Europe has always been higher than in North America, probably because of the many regional access providers (several per country) in Europe. These trends imply that *the Internet market, in terms of the number of access/hosting, transit and content providers will soon be larger in Europe than in North America*, if this is not happening already.

Rewiring activity for each AS type: The differences in the business function and

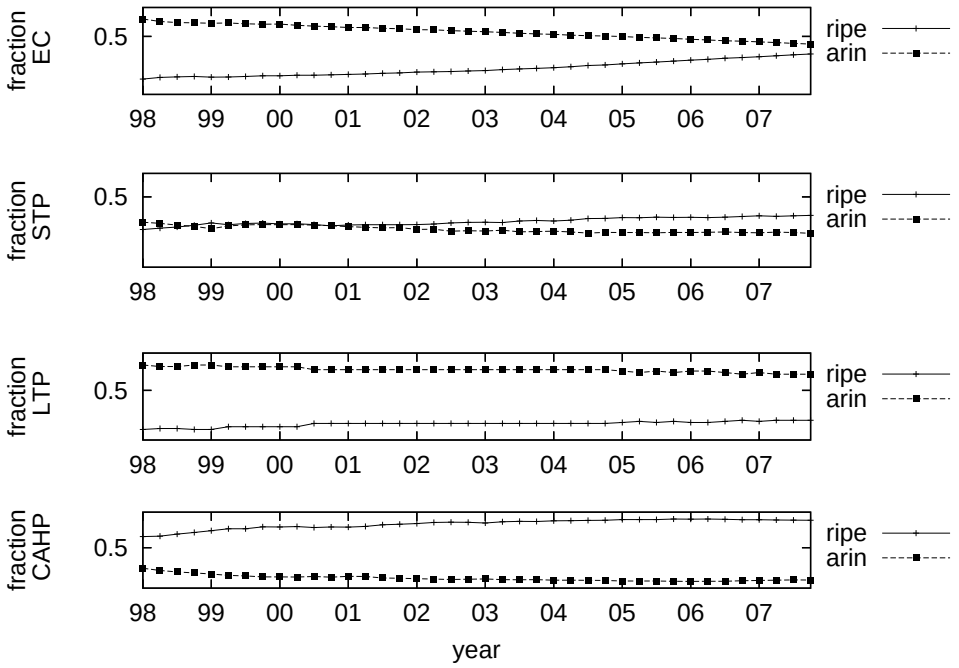

Figure 9: Regional distribution of AS types over time.

incentives of the four AS types could also appear in their rewiring activity. To measure this quantity between a pair of snapshots, we calculate the Jaccard distance for the set of CP links of each AS. We then compute the average Jaccard distance for all ASes of the same AS type. The top panel of Figure 10 shows these averages over time. We see that, clearly, *ECs show the lowest rewiring activity* throughout the last ten years. *STPs and LTPs have similar rewiring activity, while CAHPs exhibit the highest rewiring especially since 2001.* CAHPs rewire their CP links frequently, as they attempt to minimize their transit costs and provide good performance/reliability to their customers.

A related metric is the fraction of nodes in each AS type that are *inert*, meaning that they do not undergo any change in their set of CP links between two successive snapshots. The bottom panel of Figure 10 shows the fraction of inert nodes for different AS types over time. We find that the fraction of inert ECs increased slightly with time, from 74% in 2001 to 80% currently. This implies that ECs at the network edge are becoming increasingly stable with respect to the connectivity to their providers. The fraction of inert STPs has stayed almost constant since 2001 (between 25% and 30%). We examined the set of STPs that are

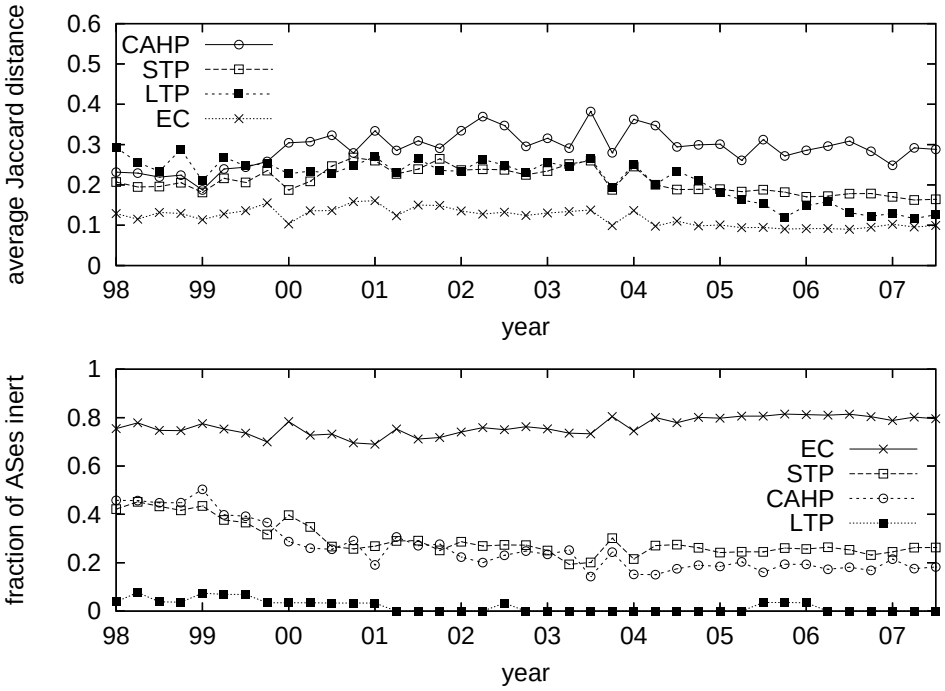

Figure 10: Rewiring activity and fraction of inert ASes for each AS type.

inert in every pair of snapshots since 2001, and found that several of the inert STPs are national monopoly providers or research and educational backbone networks. Such STPs have a fairly stable customer base, and do not have the incentive to constantly optimize their connectivity. As expected, the fraction of inert LTPs is very low and it approaches zero, because large transit providers have a constant churn in their customers. The most interesting trend is that the fraction of inert CAHPs has decreased significantly, from 46% to 18%. This again suggests that such access/hosting/content providers have an incentive to constantly optimize their connectivity.

2.5 Evolution of CP relations: customer-side properties

Number of providers per AS type: Figure 11 shows the average number of providers per customer (or the average multihoming degree) for each AS type. The median number of providers (not shown) shows similar trends. *The multihoming degree for ECs has increased very slowly over the last decade (from 1.5 to 1.9), and is almost constant since 2001. On the other hand, the multihoming degree for STPs has increased significantly (from 1.9 to*

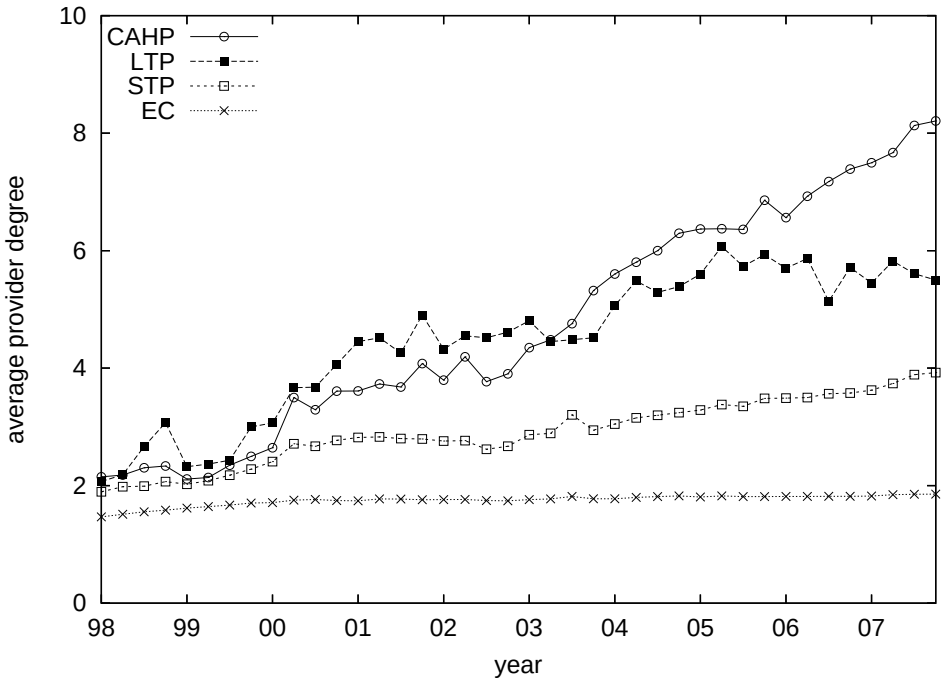

Figure 11: Evolution of average number of providers for each AS type.

3.9), LTPs[7] *(from 2 to 5.5), and CAHPs (from 2.1 to 8.2).* The dramatic increase in the multihoming degree of content/hosting/access providers and transit providers is probably the main reason behind the densification of the Internet, discussed earlier.

We further study the *distribution* of the number of providers of different AS types. We find that the distribution of the number of providers for ECs has not changed significantly in the last 10 years. On the other hand, the largest change is for CAHPs. Figure 12 shows the distribution of the number of providers for CAHPs in 5 snapshots over the last 10 years. We see that the distribution has been shifting consistently towards the right, indicating an increase in the number of providers for CAHPs. Further, we find that the median number of providers for CAHPs has been quite close to the average, and 50% of CAHPs in the latest snapshot (2007.10) have more than 7 providers. This means that the average number of providers for CAHPs seen in Figure 11 is not biased by a small number of CAHPs that have many providers.

[7] *Tier-1 ASes are commonly attributed as not having any providers. Recall, however, that we define LTPs as the top-30 providers in terms of average number of AS customers. This set includes ASes that have providers.*

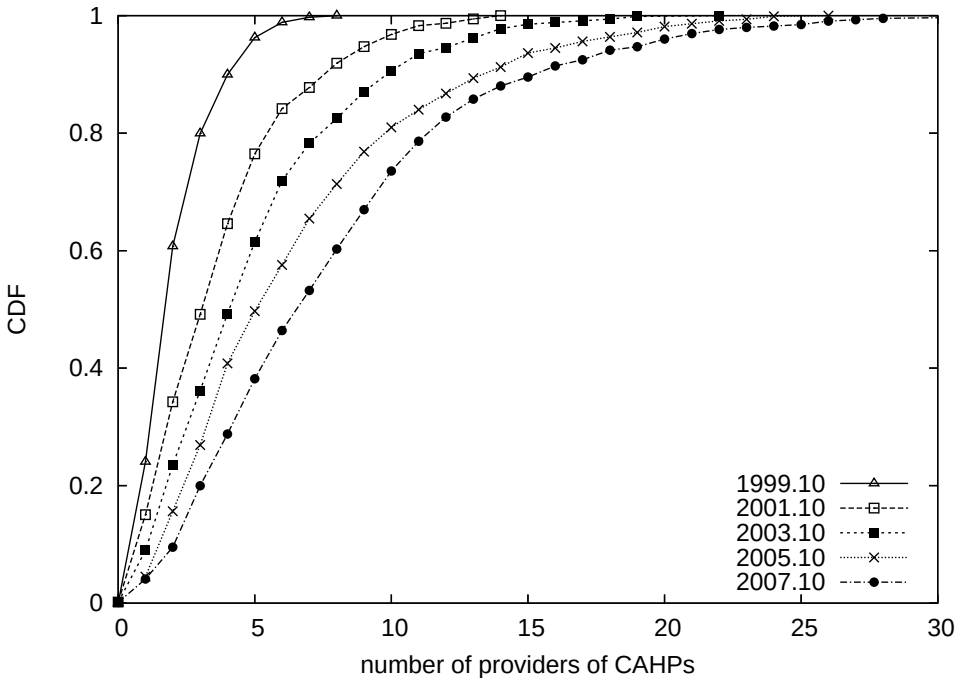

Figure 12: Evolution of the distribution of the number of providers of CAHPs.

STPs versus LTPs: We are also interested in differences in *the type of provider* that each AS type connects to when acting as the customer in a customer-provider relation. Figure 13 shows the number of links in each transit category over time. Interestingly, we find that both EC-LTP links (meaning, the customer is an EC and the provider is an LTP) and EC-STP links show an exponential increase up to 2001, which matches the trend of the total number of CP links. Thereafter the growth rate slowed down, following a linear increase. We find that until 2004 the number of EC-STP links was almost the same as the number of EC-LTP links. After 2004, the growth rate of EC-STP links has been higher than that of EC-LTP links (240 links/month vs 106 links/month), meaning that ECs increasingly prefer to connect to smaller, regional providers. There are several possible reasons why ECs may prefer STPs over LTPs. One possibility is that STPs are cheaper than LTPs. Another possibility is that ECs connect to STPs due to regional factors such as national monopolies and regulations, or region-specific marketing by STPs.

The middle panel of Figure 13 shows the evolution of provider links for CAHP customers, while the bottom panel shows the number of provider links for STP customers. The numbers

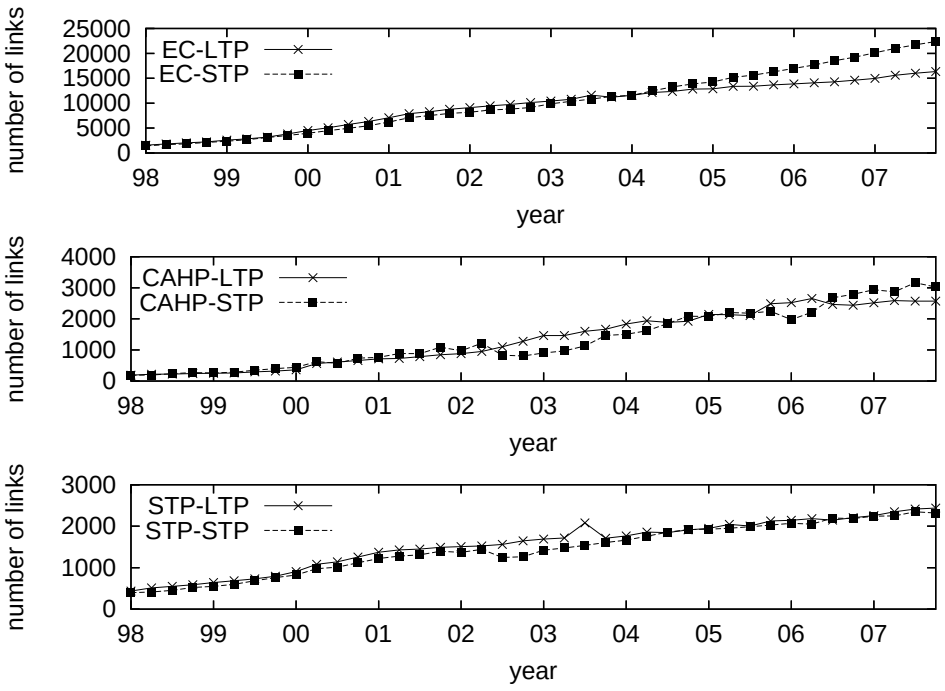

Figure 13: Evolution of CP links between different pairs of AS types.

of CAHP-LTP and CAHP-STP (also STP-STP and STP-LTP) links have been increasing at roughly the same rate. Unlike ECs, CAHP and STP customers do not prefer one type of provider over the other.

Rewiring activity of AS customers: Next, we investigate the rewiring activity of AS customers according to the broad geographical region in which they belong. Specifically, we first find the set of active customers (customers that made some change to their set of providers) between pairs of successive snapshots. Then, we calculate the fraction of those active customers that belong to each geographical region. Figure 14 shows these trends. The fractions for Asia-Pacific (APNIC), Latin America (LACNIC) and Africa (AFRINIC) are practically constant and significantly lower than for Europe (RIPE) and North America (ARIN). Interestingly, we find that after 2004-2005, there are more active customers based in Europe than in North America. In Section 2.4, we showed that Europe is catching up with North America in terms of the population of ECs, and the population of STPs is already larger in Europe. We conjecture that this has created a more competitive market in Europe than in North America, with European customer ASes being more active in adjusting their

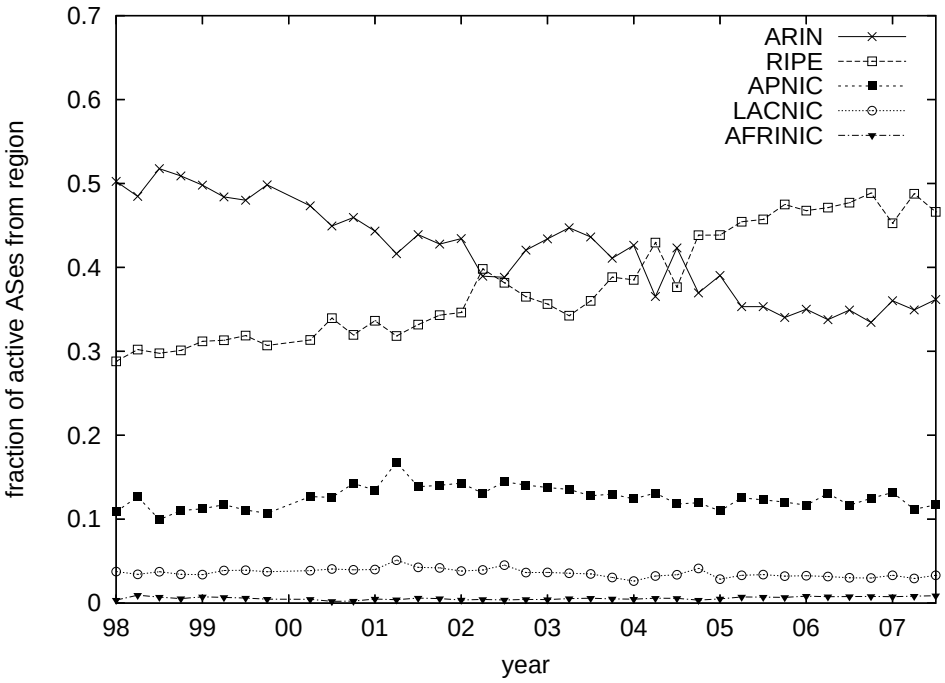

Figure 14: Fraction of active customer ASes in each geographical region.

upstream connectivity.

2.6 Evolution of CP relations: provider-side properties

Preferential attachment and preferential detachment: First, we measure the total number of CP links that were born and died between two consecutive snapshots. We define the *attractiveness* A_p of a provider p as *the fraction of CP links born in the second snapshot that connected to provider* p. Similarly, the *repulsiveness* R_p of a provider p is *the fraction of CP links that died in the second snapshot and that belonged to provider* p. These two metrics, attractiveness and repulsiveness, associate a business property (the ability to attract and retain customers) with a topological property (number of customer links of a provider AS).

Figure 15 shows the scatter plots of attractiveness and repulsiveness versus the number of customers, for a recent pair of snapshots in 2007. The left plot shows that *the likelihood with which a provider gains a CP link shows positive correlation with the customer degree of that provider*, as one would probably expect from the "rich get richer" principle. However, there are several outliers, and the correlation coefficient is only 64%. The low correlation indicates

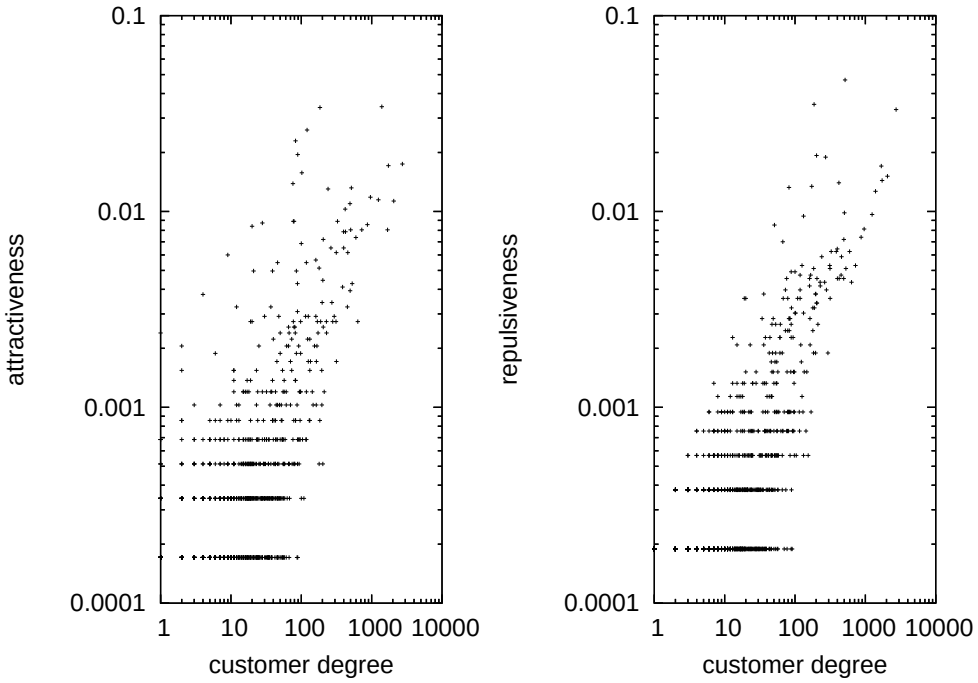

Figure 15: Attractiveness and repulsiveness versus customer degree.

that a simple model in which the attractiveness of a node is proportional to its customer degree would not be very accurate. The graph at the right is also interesting because it shows *an equally strong positive correlation between the repulsiveness of a provider and its customer degree.* Thus, when we consider the rewiring of CP links, we observe not only a "preferential attachment" behavior, but also, an equally strong *preferential detachment* behavior. Preferential detachment has been largely ignored in the earlier literature, with the exception of a brief mention in [99].

Attractors and repellers: Figure 15 also shows that *there are a few providers that have very large attractiveness and repulsiveness.* We are interested in the properties of these *attractors* and *repellers* of AS customers, and use the following approach to identify them. For each pair of snapshots, we calculate A_p and R_p for each provider p. We find that in all snapshot pairs, around 50-100 providers account for more than 60% of the total number of CP link births in the Internet. Henceforth, we identify the *attractors* of a snapshot pair as the set of providers with the highest attractiveness that account for at least 60% of the total CP link births. Similarly, we identify the *repellers*, based on the set of maximum

repulsiveness providers that account for at least 60% of the total CP link deaths.[8]

Next, we examine the number of attractors and repellers between each pair of snapshots over time. Figure 16 shows the evolution of the total number of attractors and repellers, distributed among AS types. A decreasing trend in the number of attractors would imply that the customer gains are shared by a decreasing set of providers, indicating a shift towards an oligopoly or even monopoly. What we see, however, is that *the number of attractors and repellers shows an increasing trend.* This is significant because it implies that the gains and losses of customers are increasingly shared by a larger set of providers. In other words, the Internet is not heading towards an oligopoly or consolidation of providers; instead, the market of competing providers is increasing in size. We find that since 2001, the number of LTPs in the set of attractors and repellers has stayed almost constant. This is because around 25 out of the 30 LTPs appear in these sets in any given snapshot pair. *The increase in the number of attractors and repellers is mainly due to an increasing number of STPs in these sets.*

Figure 17 shows the number of attractors and repellers in different geographical regions. Initially, it was the case that most attractors and repellers were registered in North America. Since 2003-04, however, providers from Europe have outnumbered those from North America in the attractor and repeller sets.

In addition to the number of attractors and repellers in each geographical region, we examine the total attractiveness and repulsiveness in different regions. The total attractiveness (repulsiveness) of a set of providers is the fraction of CP link births (deaths) that are contributed by providers in that set. The top (bottom) panel of Figure 18 shows the total attractiveness (repulsiveness) of the attractors (repellers) in each geographical region. From 1998 until 2003-04, the attractors in North America had a greater total attractiveness than those in Europe (coinciding with the period in which the number of attractors in North America was larger than that in Europe). It is interesting, however, that after 2003-04 the attractors in Europe and North America have similar total attractiveness. This means that even though the number of attractors is larger in Europe, they account for a similar fraction

[8]Choosing different values for this threshold yields qualitatively similar results.

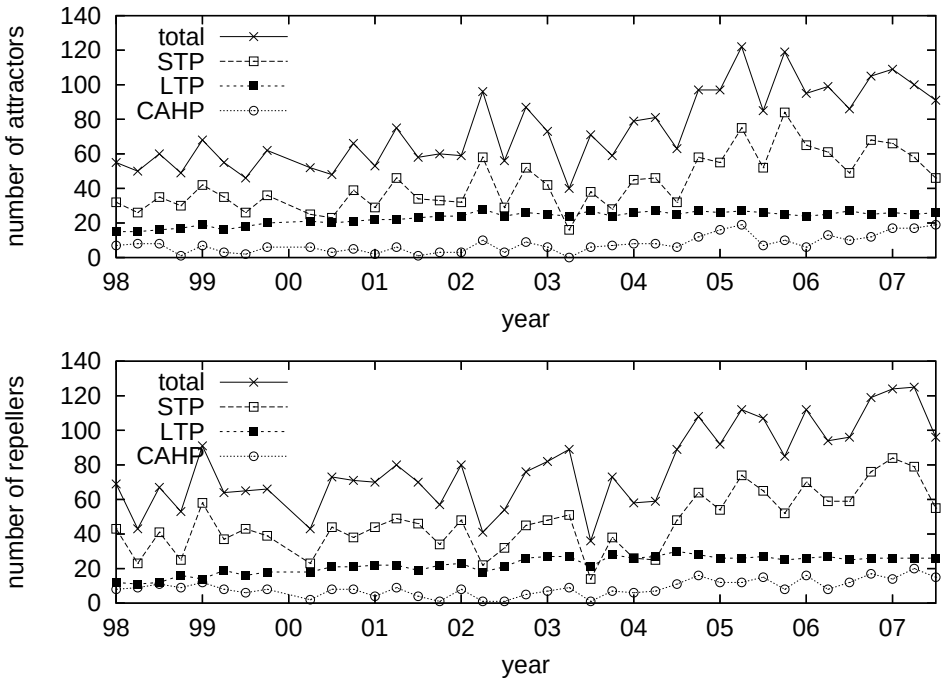

Figure 16: Evolution of the number of attractors and repellers (total and among AS types).

of the total CP link births than the attractors in North America. Similar trends are seen for the total repulsiveness in Europe and North America.

Correlation of attractiveness and repulsiveness for the same AS: We have seen that providers can act as attractors or repellers of AS customers. Here, we examine whether a correlation exists between these two properties of the same provider. If so, how do these correlations vary at different time lags? To answer these questions, we calculate the crosscorrelation of the attractiveness $A_p(t)$ and repulsiveness $R_p(t)$ timeseries of the same provider at different lags. Instead of examining all providers, we restrict this analysis only to those providers that were classified as either attractors or repellers (according to the 60% rule described earlier) at some point in their lifetime. We refer to this set of providers as $\mathcal{AR}$, where $|\mathcal{AR}|$=638. For each provider in $\mathcal{AR}$, we compute the crosscorrelation at different lags, and also the confidence bounds at 99% significance level. The confidence bounds are used to determine whether there is a significant correlation between the attractiveness and repulsiveness time series at a particular lag. We find 317 providers for which a significant

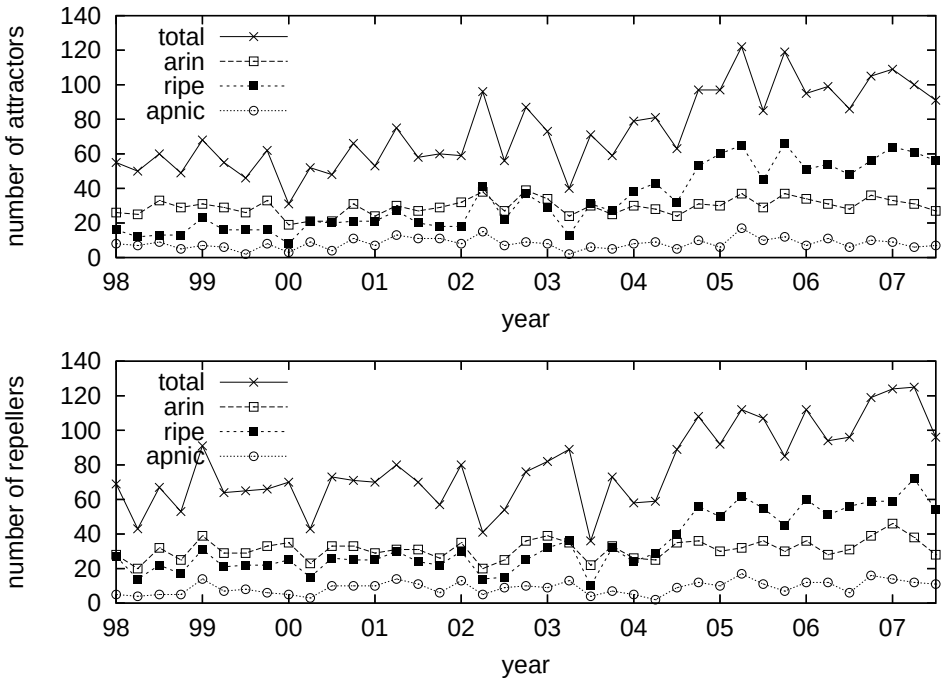

Figure 17: Evolution of number of attractors and repellers in each geographical region.

correlation exists at some lag. For each of those providers, we then determine the lag that shows the maximum absolute correlation.

Figure 19 shows, for the previous 317 providers, the lag at which the maximum (in absolute value) correlation occurred. Interestingly, we find that *in almost all cases the correlation is positive*. Further, we find that in 85% of the cases, the maximum correlation occurs at positive lags. In particular, most of the mass is at lags 1, 2 and 3 snapshots (44.7%, 13.5% and 9.1% of the providers, respectively). Note that a positive lag l means that we correlate the attractiveness at time t with the repulsiveness at time $t + l$, and each lag corresponds to 3 months. So, for a large number of providers, *strong attractiveness precedes strong repulsiveness by a period of 3-9 months*. There are several possible reasons for this effect. We conjecture that some providers attract many new customers due to advertising and promotions. These providers are not always able to keep their new customers, leading to significant repulsiveness a few months later. This may be due to customers that change providers frequently (such as CAHPs), or due to follow-up advertising/promotions from competitors.

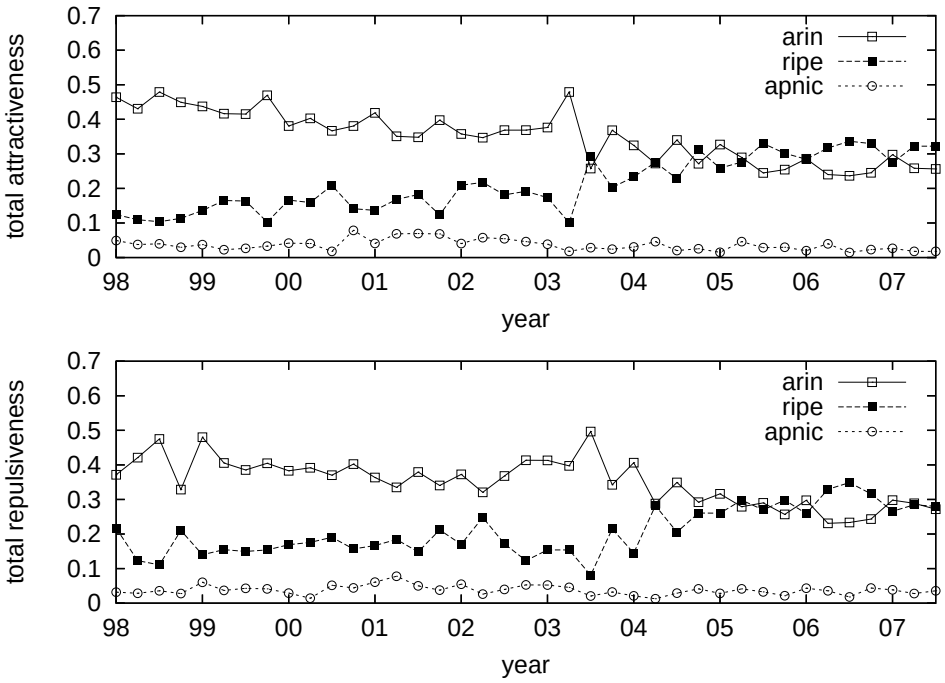

Figure 18: Evolution of total attractiveness of attractors and repellers in each geographical region.

2.7 Conjectures on the evolution of peering

Given that a large fraction of peering links may not be visible through RouteViews and RIPE routing tables, we do not study in detail the evolution of peering relations in this paper. In this section, we only present some tentative results, which should be viewed as unproven "conjectures" about the evolution of peering. The following observations need to be re-examined in a future study, when the research community obtains sufficient visibility of the peering links in the Internet.

Figure 20 shows the median peering degree for each of the four AS types. We prefer to use the median degree in this case because the average peering degree is heavily influenced by a single LTP provider (AS13237) that appears to have over 200 peers. ECs and STPs have median peering degrees of zero. It is interesting that *the median peering degree of CAHPs has increased significantly since 2003, from 2 to 10.* It is not surprising that LTPs establish many peering links; those links are needed for global reachability when it is not possible to directly reach some destinations through customers. CAHPs, on the other hand,

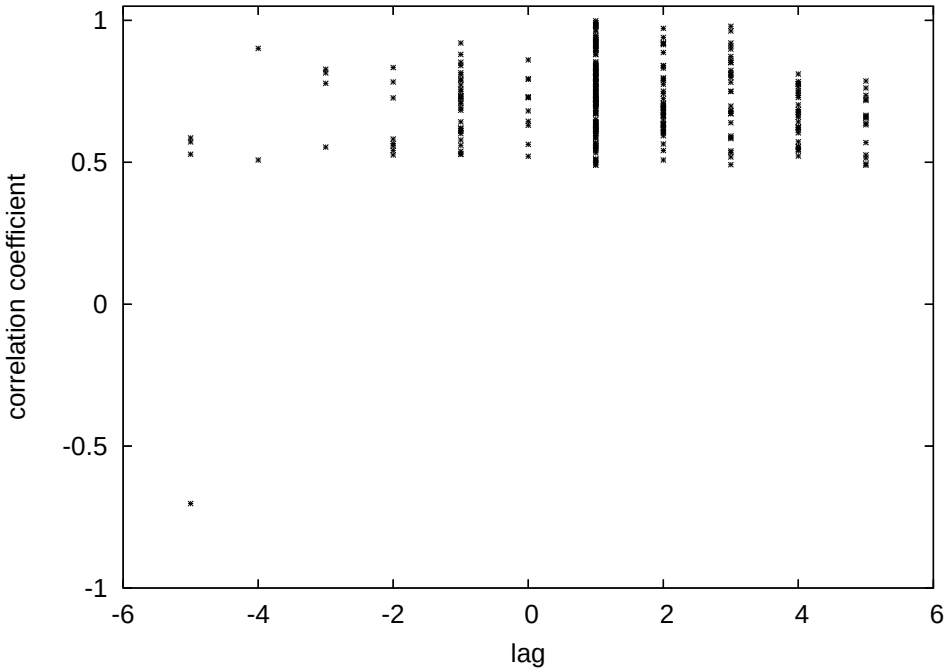

Figure 19: Lag of maximum absolute correlation for each AS provider in $\mathcal{AR}$.

have the incentive to create many peering links to reduce their transit costs paid to upstream providers. Their revenues result from the content they offer or from access/hosting fees.

Figure 21 shows the number of peering links in each category over time. We see several interesting trends. First, the number of peering links that involve CAHPs (CAHP-CAHP, EC-CAHP, STP-CAHP) increased significantly between 2001-2005, and it shows a persistent growth rate thereafter. The exception is for the links of type LTP-CAHP, which are almost constant in number since 2003. The largest number, as well as the highest growth rate, is for links of the type CAHP-CAHP and CAHP-STP. This could be because content/hosting/access providers have the incentive to get as close as possible to the destinations/sources of their traffic. These destinations/sources of traffic are other CAHPs or they are networks that are reachable through STPs. Another interesting observation is that the number of STP-LTP peering links has remained almost constant over the last 5-6 years. We conjecture that this is due to the "restrictive" peering policy of most large transit networks. The previous observations confirm the anecdotal evidence, mentioned in various white papers (see [84] and related references), that content/access providers are

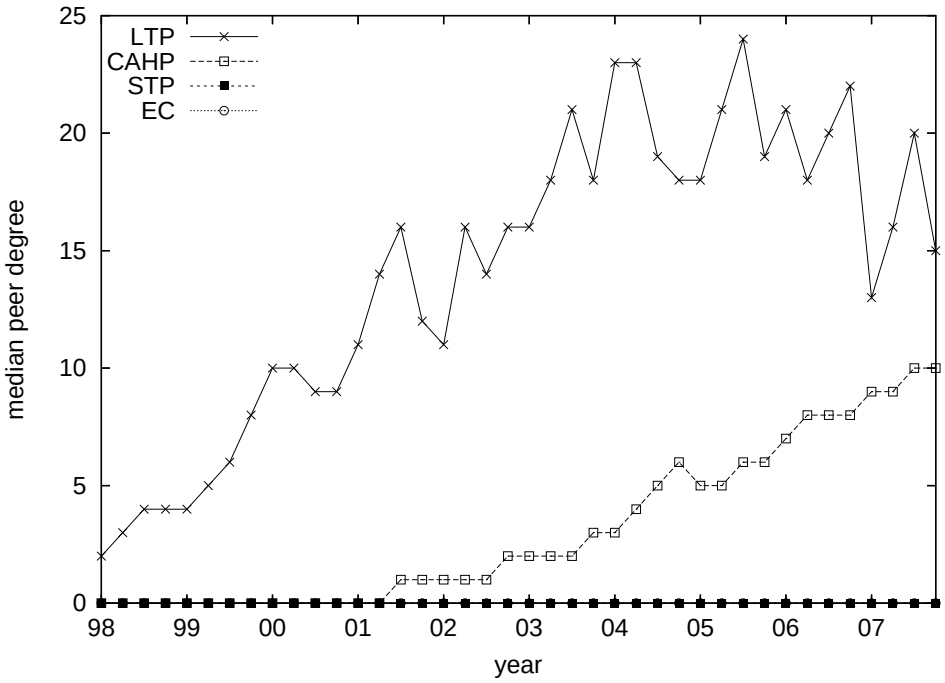

Figure 20: Median number of peers for each AS type over time.

rising in the peering ecosystem as the dominant players. The underlying reason is that such ASes mostly have an *open peering policy*, while transit providers have *selective or restrictive* policies, peering by necessity rather than by choice.

2.8 Related work

A major research effort during the last decade aimed to characterize the AS-level topology. One of the most well cited papers, by Faloutsos *et al.* [47], argued that the Internet AS-level topology is "scale-free". This observation was questioned by Chen *et al.* [29], who showed that the degree distribution in the Internet, though heavy-tailed, does not obey a strict power-law distribution. Tangmunarunkit *et al.* [102] attempted to explain the heavy-tailed degree distribution, and conjectured that this could simply be due to the heavy-tailed AS size distribution. Most previous measurement studies focused on static topological properties of the Internet, such as degree distribution or clustering, and did not examine the evolution of the topology over time.

A recent study measured the average degree and effective diameter of the Internet AS

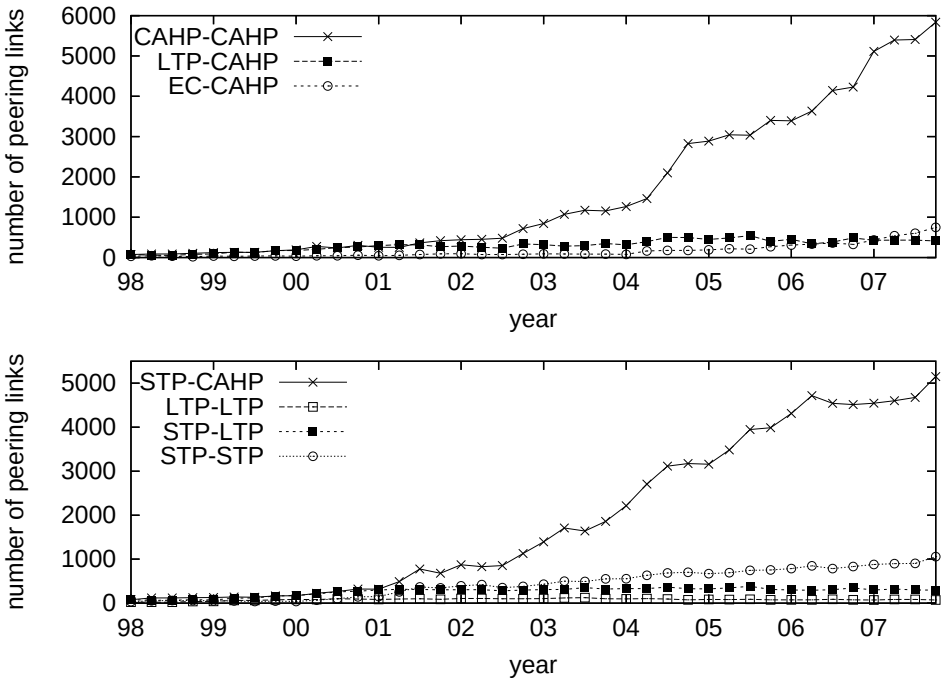

Figure 21: Number of PP links of the most common types.

graph and concluded that the AS graph is *densifying* [72]. Two other measurement studies [76, 99] studied the evolution of the Internet topology in the period 1997-2001 with respect to several microscopic and macroscopic properties. Siganos *et al.* [99] observed the exponential growth of the Internet during that time period, and showed that a rich-get-richer form of preferential attachment leads to exponential growth in the number of edges. Magoni *et al.* [76] examined the evolution of some global Internet characteristics and found exponential growth in the number of ASes and links during that time period.

The observation that the degree distribution follows a power-law led to several topology generation models that could produce such distributions. These models focused on "growing" a topology that could match the Internet topology with respect to certain measurable graph metrics. The most well known work in this area is the preferential attachment model of Barabasi *et al.* [15]. An extension of that model [10] incorporated the random rewiring of a fixed number of existing links. Several variants of preferential attachment models were later proposed [21, 110, 113]. Park *et al.* [90] compared different growth models for Internet

topology with respect to several metrics such as the average diameter and clustering coefficient. The models in this research thread have been mostly descriptive. More recent work has attempted to incorporate the effect of economic factors in the evolution of the Internet topology, most notably [97, 107].

The previous descriptive models received considerable criticism (for instance, see [67, 70]) because they mostly focus on the degree distribution and clustering, ignoring important features of the Internet topology such as hierarchy or the presence of links of different types (transit versus peering). Further, the previous models do not explain how the Internet topology is evolving. This led to new models that view the Internet topology as the outcome of optimization-driven activity of individual ASes. These concepts were first introduced by Carlson and Doyle [23], and later applied in the context of the Internet in [46] and elsewhere. Chang *et al.* [24] used domain-specific information about the Internet to model AS interconnection practices. A recent work by Chang *et al.* [25] models the behavior of an AS in two distinct economic roles (customer and peer), and examines the topological effects of actions of individual ASes when acting in different roles. A recent editorial [57] stresses the need to further understand the dynamics of the AS topology. Norton [84] discusses, mainly using anecdotal evidence, how economic and competitive interests influence peering and transit connectivity in the Internet. Economides [43] discusses the economics of the Internet backbone (without looking at topology dynamics).

Several measurement studies have highlighted the incompleteness of the topologies inferred from publicly available routing data [26, 32, 59, 77, 111]. Given that the inferred topologies are incomplete, much attention has been devoted to methods that capture as much of the Internet topology as possible, most notably the work by Zhang *et al.* [111] and He *et al.* [59]. Zhang *et al.* [112] investigated the effect of the selection of route monitors on different applications such as topology inference and AS path prediction. Their main observation was that for applications such as topology and relationship inference, the publicly available BGP data are reasonably accurate, and data from an increasing set of monitors is only marginally useful. Oliviera *et al.* [88] tackle the problems of topology liveness and completeness, i.e., how to distinguish between topology changes that are genuine link births

and deaths, versus those that are caused due to link appearance and disappearance during routing transients.

2.9 Conclusions

We measured the evolution of the AS-level topology over the last 10 years in terms of growth as well as rewiring, four distinct economic/business classes of ASes, and customer-provider links. Our findings highlight some important trends, trend shifts, and sketch what the Internet may be heading towards. The main findings are summarized next.

The Internet has gone through two growth phases in terms of ASes and transit links: an initial exponential phase up to mid/late-2001, followed by a linear phase. Even as the network grows, the average path length remains practically constant, meaning that the network *densifies*. We find that currently, around 75% of link births are associated with existing ASes rather than new ASes (rewiring versus growth); similarly, about 80% of the link deaths are due to rewiring.

We proposed a simple AS classification scheme according to economic considerations and business types. In terms of the population of these AS types, we find that the ASes at the network edge (ECs) contribute most of the overall growth. The average multihoming degree has remained roughly constant for ECs, has increased significantly for STPs, LTPs and CAHPs. The aforementioned densification process is thus driven by transit providers and access/hosting/content providers. In terms of rewiring activity, CAHPs are the most active, while ECs are the least active.

We introduced two provider metrics, attractiveness and repulsiveness, to measure the ability of a provider to attract and retain customers. We see positive correlations between the attractiveness and repulsiveness of a provider and its customer degree. Also, for many providers, strong attractiveness precedes strong repulsiveness by a period of 3-9 months. There are a few providers that have very large attractiveness and repulsiveness (attractors and repellers). The total number of attractors and repellers between successive snapshots shows an increasing trend.

In terms of regional growth, we find that the Internet market, in terms of the number

of access/hosting/content and transit providers will soon be larger in Europe than in North America. This is also reflected by the fact that since 2004-2005, a larger fraction of active customers are based in Europe than in North America, and providers from Europe increasingly feature in the set of attractors and repellers. This is important, because much of the regulatory and policy debate about the Internet has been focused on North America. Our measurements hint at an increasing European influence on the Internet ecosystem.

We have explained the previous measurement results with conjectures about the causes of the observed densification, the high activity of CAHPs, and the incentives that lead certain AS types to connect to other AS types. Unfortunately it is hard to validate these conjectures, mainly due to the lack of pricing and other economic data about various AS types. Obtaining such data and further explaining the previous observations in an economic or optimization basis is a valuable direction for future work.

Our results are important for the following reasons. First, these insights are a step towards creating better informed models of topology evolution. We show that such a model must account for the rewiring of links between existing ASes, as this process accounts for more link births (deaths) than node births (deaths). Further, such a model must take into account the different incentives and business functions of the constituent ASes, as they lead to significant differences in the evolutionary behavior of those ASes. Also, the model should account for the semantic differences between different link types. Our measurements reveal that the Internet graph grows mainly at the edges, while the core grows slowly. The average path length, however, is almost constant, and this densification is driven by aggressive multihoming of providers at the core. These observations are important inputs for a study of, for example, the projected future performance of an Internet protocol or application. We find in our measurements that content/access providers are becoming major players in the peering ecosystem, and peer directly with the sources/destinations of their content. Taken to the extreme, this could lead to a situation where transit providers become redundant. This means that transit providers may need to rethink their strategies to continue to be profitable.

CHAPTER III

THE VIEW FROM THE EDGE: ISP SELECTION FOR MULTIHOMED NETWORKS

3.1 Introduction

In chapter 2, we measured the evolution of the graph consisting of customer-provider links in the Internet over the last 10 years. Our major results from that study point to the increasing prevalence of multihoming, and the fact that certain networks tend to be very active in changing the set of providers with which they are multihomed. Multihoming refers to the connection of a stub/edge network to more than one Internet Service Providers (ISP) [16]. In the most common deployment scenario, one ISP is used as the primary provider while another is used as backup when the primary fails. Switching from the primary to the backup can be performed automatically by the border router of the multihomed network when it detects loss of connectivity with the primary ISP. As seen in Chapter 2, the use of multihoming has seen a dramatic increase in the last few years, and we estimated that more than 70% of the stub networks in the are multihomed to at least two ISPs. In particular, we observed that content providers have been the most aggressive in increasing their multihoming degrees in recent times. The widespread proliferation of multihoming is due to several reasons. First, as more and more enterprises rely heavily on the Internet for their transactions, reliability and availability become of primary importance. Second, multihoming can be used to drive down the costs of Internet access. This is the case when the multihomed network can use a lower-cost ISP for their bulk traffic and a higher-cost but better ISP for more performance-sensitive traffic.

Multihoming capabilities have expanded tremendously with the use of "Intelligent Route Control" (IRC). Multihoming-IRC systems allow a stub network to automatically switch parts of their ingress or egress traffic from one provider to another, driven by cost and/or performance considerations. A number of vendors currently provide such systems [31, 45,

48, 63, 81, 91, 92, 93, 95, 100]. IRC products typically assume that the set of upstream ISPs has been already chosen and it is fixed. A stub network, however, has several choices of upstream ISPs. The choice of the exact set of upstream ISPs is critical, as it can directly influence the monetary cost incurred by the network and the performance that the network can achieve to the rest of the Internet. In this part of the thesis, we put ourselves in the shoes of the network operator of a stub network (which is assumed to generate a significant amount of content), and devise algorithms to select the best possible set of upstream providers for this network.

The problem of which ISPs to select has been largely ignored so far, or it has been viewed as a non-technical decision (a notable exception is the recent work by Wang et al. [106]). Intuitively, a good set of ISPs should provide low cost, good performance, and significant path diversity (for robustness to network failures), at least for the major traffic destinations. In the first part of this work, we propose a methodology to select a set of upstream ISPs taking into account monetary cost, inter-domain level performance, and path diversity considerations. We show, based on traffic and topology measurement data, that the proposed algorithm can improve robustness to single inter-AS link failures by selecting the best possible combination of ISPs. The algorithm also performs well in the presence of double and triple inter-AS link failures.

Once the set of ISPs has been selected, the egress traffic can be routed so that we minimize the cost incurred by the source network, subject to the important constraint that the chosen outgoing paths do not experience persistent congestion. This is different from the current IRC practice, which is to change dynamically the traffic allocation in a reactive manner, to avoid transient periods of congestion. In the second part of this work, we propose an algorithm for egress path selection that determines a congestion-free solution (if it exists) with minimum cost for the source network. This is a challenging problem, because the source network does not have a priori knowledge about the topology and capacity of the upstream paths that reach each of its major destinations. We propose a stochastic search algorithm based on simulated annealing to find a feasible egress path for each major destination. Using simulations, we show that this algorithm performs well in meeting the

objectives, when a feasible path selection exists.

This rest of the chapter is structured as follows. Section 3.2 states our objectives and describes the network and traffic models. In Section 3.3, we describe the ISP selection problem and the proposed methodology, which is then evaluated in Section 3.4. In Section 3.5, we describe the egress path selection problem and our stochastic search solution, which is then evaluated in Section 3.6. We discuss the related literature in Section 3.7 and conclude in Section 3.8.

3.2 Problem Description and Objectives

The aim of this work is to provision the multihoming configuration of a source network S. We assume that S is a content provider, i.e., mostly a source rather than destination of traffic. Hence, we are concerned with the *egress* traffic from S. We also assume that before exploring multihoming options, the administrator of S has a good idea about the outgoing traffic profile at S. In particular, the operator knows the set of M "major destinations" that account for a large fraction of the outgoing traffic from S. These major destinations can be large networks. So, what we refer to as "flows" destined to these major destinations are large aggregates, rather than individual end-to-end flows. The operator of S should also have an estimate for the average rate of the traffic that is sent to each of these major destinations. It is possible to assemble this traffic profile through passive measurements at the outgoing links of S. Figure 22 shows our basic underlying model: a source network connected to the Internet through K ISPs, with each ISP providing a network path (potentially different) to each of the M destinations. Provisioning the multihoming configuration of S involves the following two tasks:

1. Choose the ISPs that S will subscribe to. Typically, several ISPs would have a point-of-presence at the location of S. These ISPs may differ based on their performance, the level of reliability that they guarantee, and of course their pricing policies. S will choose K of these ISPs as upstream providers. K depends on the level of reliability that S desires, and also on monetary or other practical constraints (e.g., number of available ports at the border router). We assume that K is given. We term this part

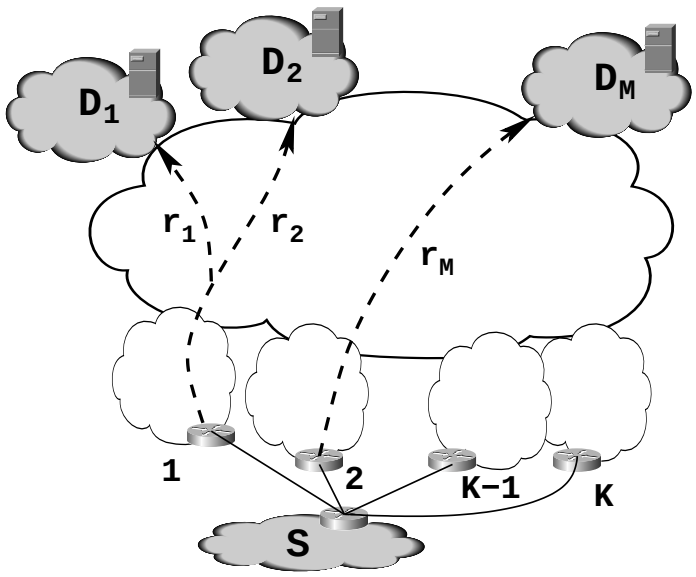

Figure 22: A multihomed network with K upstream ISPs, and M major destinations

of the problem as *ISP Selection*.

2. Once the set of K ISPs has been chosen, the operator of S can determine the ISP that should be used to reach each of the M major destinations. The objective is to minimize the total cost paid to the K ISPs, subject to the constraint that none of the chosen paths to the major destinations is congested. We term this part of the problem as *Egress Path Selection*.

Note that the ISP selection phase is "semi-static", meaning that it would be typically performed over very long timescales, say months. The set of upstream ISPs would be modified only if there are major changes in the destinations of the outgoing traffic or in the underlying ISP market.

The egress path selection problem aims to find an allocation of destinations to ISPs that minimizes cost and avoids long-term congestion. However, there may still be periods of short-term congestion in which some paths are overloaded and see poor performance. These short-term congestion events can be dealt with in a reactive way using dynamic path switching, as done by most IRC products. However, such dynamic path switching can result in a sub-optimal configuration. Hence, the egress path selection algorithm should be run

periodically so that S returns to a more optimized configuration. We envision that the path selection phase would be repeated every few hours, while dynamic path switching could take place in the time scales of seconds to deal with short-term congestion.

Our provisioning objectives are determined by the following factors:

1. **Cost:** S aims to minimize the total cost incurred for routing its egress traffic through the chosen set of upstream ISPs. Each ISP has a pricing function that is used to determine the cost it charges to S. A common practice is that the cost charged by an ISP depends on the total amount of traffic that it receives from a customer (volume-based pricing). In this work, we avoid any specific pricing function, mostly because different ISPs use significantly different pricing models. In our simulator, each pricing function is an increasing and concave function of the total traffic volume routed through that ISP.

2. **Performance:** S aims to avoid routing its traffic through congested paths. This is the objective of the path selection phase. We focus on long-term congestion, resulting from gross misallocation of traffic. For example, routing the traffic to a major destination, say 10Mbps, through an ISP that reaches that destination with a 5Mbps link.

3. **Robustness:** S should select ISPs that provide significant *path diversity* to its major destinations. Selecting ISPs with path diversity provides resiliency to upstream network failures. With a selection of ISPs that provides significant path diversity, a working alternate path is likely to exist in case the primary path to a destination fails or becomes congested. On the other hand, if the K paths to a certain destination are largely overlapping, then it is likely that none of the upstream paths can avoid the point of failure or congestion.

3.3 Phase I - ISP Selection

3.3.1 Problem statement

In this section, we focus on the problem of selecting the egress ISPs for a source network S. Let $\mathcal{I}$ be the set of possible ISPs to which S can subscribe, out of which K will be

selected. $\mathcal{D}$ represents the set of M major destinations of S. $\mathcal{R}=\{r_i\}$ is the average traffic rate destined to each destination in $\mathcal{D}$. We assume that the K links have the same capacity A. The operator of S can determine an appropriate value of A based on the average rate of the egress traffic, the number K, and the desired utilization level under ideal load balancing. For example, if the egress rate is 400Mbps and $K = 4$, the average load of each link would be 100Mbps. Since link capacities are typically available in a few discrete steps (e.g., Fast Ethernet, OC-3, OC-12), S would probably subscribe to four OC-3 links in this example. The assumption of equal capacities simplifies the ISP and egress path selection problems, and it may be necessary for practical reasons (e.g., homogeneous border router interfaces). Our approach can be easily modified for the case that S connects to each ISP with a different (but known) capacity.

Both monetary cost and performance should be taken into account when selecting ISPs. An important issue is whether S can evaluate the performance of an ISP *before* it actually subscribes to that ISP. Typically, an ISP would not allow a network S to perform extensive probing and performance measurements before S becomes its customer. However, most ISPs maintain public Looking Glass Servers (LGS). LGSs are routers inside an ISP that report AS-level paths to given destination networks. Many ISPs today deploy LGSs at different points-of-presence, mostly for the diagnosis of inter-domain routing problems. Here, we assume that each ISP in $\mathcal{I}$ has a LGS from which S can determine the AS-level paths to destinations in $\mathcal{D}$.

The ISP selection problem takes into account the following three factors:

Monetary Cost: The cost of routing the traffic of S through the selected set of ISPs should be minimized. The actual cost cannot be determined until S subscribes to a set of ISPs and allocates its traffic among them. This is because the cost charged by an ISP depends on the total amount of traffic that is routed through it. However, as we show next, there is a way for S to estimate the cost that would be incurred with each selection of ISPs.

AS-level path length: The AS-level paths to the networks in $\mathcal{D}$ through the chosen ISPs should be as short as possible. Long paths tend to translate into larger delays, and are more vulnerable to interdomain routing failures and pathologies.

Path diversity: The K chosen ISPs should be such that S has the maximum possible path diversity to the destinations in $\mathcal{D}$. Large path diversity improves the robustness to upstream network failures and congestion events. We define an AS-level path diversity metric later in this section.

Since we have to pick K ISPs out of $|\mathcal{I}|$ choices, there are $\binom{|\mathcal{I}|}{K}$ possible selections. With each selection we associate a cost metric for each of the three aforementioned factors. So, the ISP selection process can be viewed as an optimization problem with the objective to minimize the total (generalized) cost. $|\mathcal{I}|$ would typically not be higher than 10-20 ISPs. Hence, it is tractable to enumerate all possible combinations of K out of $|\mathcal{I}|$, and then choose the selection that minimizes the total cost. For example, if $|\mathcal{I}| = 15$ and $K = 4$, there are 1365 possible combinations. Recall that the ISP selection process is performed over very long timescales, and so exhaustive search of all 1365 combinations should be feasible.

For each combination $\mathcal{C}$ of K ISPs, let $c_m(\mathcal{C})$ be the monetary cost, $c_p(\mathcal{C})$ be the cost associated with the AS-level path length, and $c_d(\mathcal{C})$ be the cost associated with path diversity. These three cost terms will be defined in the following paragraphs. To get the total cost $c_t(\mathcal{C})$ for the selection $\mathcal{C}$, we form a weighted sum of the previous three costs as follows

$$c_t(\mathcal{C}) = \alpha_m c_m(\mathcal{C}) + \alpha_p c_p(\mathcal{C}) + \alpha_d c_d(\mathcal{C}) \tag{3}$$

where α_m, α_p and α_d are the corresponding normalization factors. The administrator of S can choose the values of these factors depending on the relative importance of each factor.

Let $\mathcal{C}^{\mathcal{T}}$ be the set of all possible combinations of K ISPs from the set $\mathcal{I}$, with $|\mathcal{C}^{\mathcal{T}}| = \binom{|\mathcal{I}|}{K}$. For each selection $\mathcal{C} \in \mathcal{C}^{\mathcal{T}}$, we calculate $c_t(\mathcal{C})$, and the optimal choice of ISPs is the selection $\mathcal{C}^*$ with the minimum total cost, i.e.,

$$\mathcal{C}^* = \text{argmin}_{\mathcal{C} \in \mathcal{C}^{\mathcal{T}}} c_t(\mathcal{C}) \tag{4}$$

Note that it is not necessary to use this particular additive cost function given in Equation (3). Since we use a "brute force" approach to find the best set of ISPs, any cost function that expresses the total cost in terms of the three different cost components can be used instead.

3.3.2 Monetary cost

Each ISP j has a pricing function $f_j(T_j)$, where T_j is the total traffic routed through it. The total monetary cost of a selection $\mathcal{C}$ is $c_m(\mathcal{C})$=$\sum_{j\in\mathcal{C}} f_j(T_j)$. Note that T_j, and hence the total cost, depends on how the traffic of S is allocated to the ISPs in $\mathcal{C}$. This allocation however is not known before S subscribes to $\mathcal{C}$. To deal with this problem, we estimate $c_m(\mathcal{C})$ as the *minimum* cost that would be incurred to route the traffic of S through the set $\mathcal{C}$ of ISPs. To compute this cost we need to solve the following lower-level optimization.

For a given ISP selection $\mathcal{C}$, let $j = G(i)$ represent the ISP that carries the traffic to destination i; we refer to the function $G(\cdot)$ as a "mapping". There are K^M possible ways to map the M flows in $\mathcal{D}$ to $\mathcal{C}$. Let $\mathcal{G}$ be the set of all such mappings. Some of these mappings may be infeasible, because the amount of traffic routed through one or more ISPs exceeds the corresponding access capacity. So, the minimum monetary cost for the selection $\mathcal{C}$ is

$$c_m(\mathcal{C}) = \min_{G\in\mathcal{G}} \sum_{j\in\mathcal{C}} f_j(T_j) \tag{5}$$

where the minimization is performed over all possible mappings in $\mathcal{G}$, subject to the constraints

$$T_j < A, \quad j = 1 \dots K \tag{6}$$

This is a variation of the bin-packing problem with M items of size r_i $(i = 1 \dots M)$ and K bins, each of capacity A. The bin packing problem is NP-hard and so we need to use a heuristic solution, especially if the number of destinations is large. The heuristic that we use is similar to the *First Fit Decreasing* (FFD) algorithm. The basic idea is to start with the largest destination, in terms of rate, and route it through the lowest-cost ISP in which it satisfies the capacity constraint. Algorithm 1 shows the pseudo-code for our heuristic. The total running time of the algorithm is $O(MlogM) + O(MKlogK)$. Simulations showing the performance of the FFD algorithm, in terms of being able to find a solution when one exists, and in terms of finding the optimal solution, are presented in Section 3.6.

Finally, the monetary cost of selection $\mathcal{C}$ is given by

$$c_m(\mathcal{C}) = \sum_{j\in\mathcal{C}} f_j(T_j) \tag{7}$$

Algorithm 1 FFD-like heuristic

Require: Rates $\mathcal{R} = \{r_i\}$, $i = 1 \dots M$
Require: Access capacity A of each ISP
Require: Pricing functions $\mathcal{F} = \{f_j\}$, $j = 1 \dots K$
Initialize $G = null$ {Constructed mapping}
Initialize $T_j = 0$, $j = 1 \dots K$ {Total rate through each ISP}
Initialize $\bar{A}_j = A$, $j = 1 \dots K$ {Residual access capacity of each ISP}
Sort destinations in decreasing order of r_i
for each destination i in sorted sequence **do**
    $c_j = f_j(T_j + r_i)$, $j = 1 \dots K$ {Cost if destination i was mapped to ISP j}
    Sort c_j in increasing order
    for each ISP j in sorted sequence **do**
        if $\bar{A}_j > r_i$ **then**
            $T_j = T_j + r_i$
            $G(i) = j$ {Map destination i to ISP j}
            $\bar{A}_j = \bar{A}_j - r_i$
            break {Route next destination}
        end if
    end for
end for
if there is a destination that could not be routed **then**
    return null
else
    return G {final mapping}
end if

where the traffic through ISP j is the sum of the rates of the destinations that are routed through ISP j, i.e.,

$$T_j = \sum_{i:G^*(i)=j} r_i \tag{8}$$

and G^* is the mapping reported by Algorithm-1

$$G^* = \text{FFD}(\mathcal{R}, A, \mathcal{F}) \tag{9}$$

It is possible that Algorithm-1 will fail to find a feasible mapping. The simulations in Section 3.6 show that that happens, almost always, when there is no feasible mapping. Also, the same simulation results show that the cost of the mapping reported by Algorithm-1 is within 5% of the minimum cost.

3.3.3 AS-level path length cost

The calculation of the AS-level path length cost for a selection $\mathcal{C}$ proceeds along similar lines as the monetary cost. Let $p_j(i)$ denote the AS-level path length to reach a destination i through ISP j. Thus, we can think of $p_j(i)$ as a cost for a given destination-ISP pair. As we did for monetary cost, the total path length cost of an ISP selection $\mathcal{C}$ can be estimated as the *minimum* that can be obtained with $\mathcal{C}$. The minimization is performed over all possible mappings in $\mathcal{G}$, given the ISPs in $\mathcal{C}$, i.e.,

$$c_p(\mathcal{C}) = \min_{G \in \mathcal{G}} \sum_{i=1\ldots M, j=G(i)} p_j(i) \tag{10}$$

subject to the constraints

$$T_j < A, \quad j = 1 \ldots K \tag{11}$$

This is identical to the monetary cost problem, except that the cost function in this case is given by path lengths. We use the same FFD algorithm to compute the optimal mapping G^*

$$G^* = \text{FFD}(\mathcal{R}, A, \mathcal{P}) \tag{12}$$

where $\mathcal{P}$ is the set of path length costs from the ISPs in $\mathcal{C}$ to the destinations in $\mathcal{D}$. If such a mapping exists, the minimum path length cost is given by

$$c_p(\mathcal{C}) = \sum_{i=1\ldots M} p_j(i) \quad \text{where } j = G^*(i) \tag{13}$$

3.3.4 Path diversity cost

A selection $\mathcal{C}$ of ISPs provides K paths to each destination i. We focus on AS-level paths, because these paths can be directly observed through LGSs. If an inter-AS link is shared by *all* K paths to i, then a failure of that link will make i unreachable. For single-link failures, a destination i will become unreachable *only* by the failure of a link shared by all K paths. We call such links as *K-shared*. Obviously, a selection of ISPs that has fewer K-shared links will provide better resiliency to inter-AS link failures.

For a destination i and a selection of ISPs $\mathcal{C}$, we define the path diversity metric $\kappa(i, \mathcal{C})$ as the number of K-shared links to destination i though the ISPs in $\mathcal{C}$. We use $\kappa(i, \mathcal{C})$ as the cost term for path diversity,

$$c_d(i, \mathcal{C}) = \kappa(i, \mathcal{C}) \tag{14}$$

The path diversity cost for a selection of ISPs $\mathcal{C}$ is then given by the sum of $c_d(i, \mathcal{C})$ over all destinations, weighted by the rate of each destination,

$$c_d(\mathcal{C}) = \sum_{i=1...M} r_i * \kappa(i, \mathcal{C}) \tag{15}$$

We choose this weighted average, so that there is a higher cost associated with the potential failure of large destinations.

3.4 Phase I - Path Diversity

The evaluation of Algorithm-1, used in the minimization of the monetary and AS-level path length costs, appears in Section 3.6. In this section, we focus on the path diversity cost instead.

3.4.1 Destination networks and rate distribution

To evaluate the achievable path diversity from a real network, we first need to characterize its main destinations of traffic and the corresponding rate distribution. We analyzed a large packet trace from the Internet egress link of our university's network. This campus network is a significant source of traffic because it hosts some popular Web and FTP servers. For each destination IP address, we find the corresponding destination network by searching

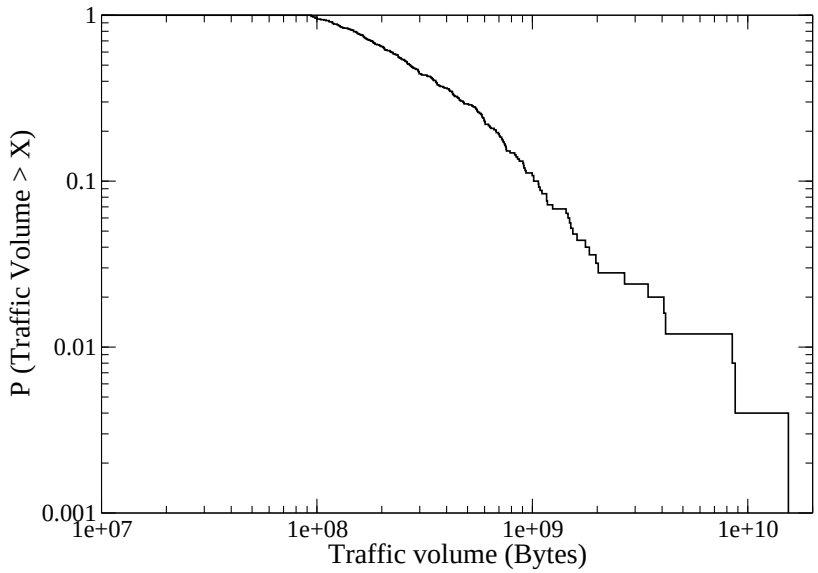

Figure 23: Complementary CDF of egress traffic to the 250 largest destination networks.

for the longest matching prefix in the BGP routing table of the border router. We then measure the total traffic to each destination network and rank those destinations based on their aggregate rate. The top 500-1000 destination networks account for about 80-90% of the total egress traffic, while the top 250 destinations account for about 65% of the traffic. In the following, we work with those 250 largest destinations (M=250). Figure 23 shows the complementary CDF of the traffic volume to those destinations. The approximately linear distribution on the log-log plot indicates a Pareto distribution. We estimated the shape parameter as $\alpha = 1.08$ using the *aest* tool [34].

3.4.2 AS-level paths

To determine the cost with respect to path length (c_p) and path diversity (c_d), we need to know the AS-level paths to the major destinations in $\mathcal{D}$ through each of the ISPs in the set $\mathcal{I}$. For this purpose, we make use of the LGSs that many ISPs provide. In this paper, we considered nine ISPs with points-of-presence in Atlanta (Qwest, Level 3, SAVVIS, Broadwing, Williams Communications, Teleglobe, Cogent, Global Crossing, and 1A Networks). Each of them provides an LGS. We queried these nine LGSs for each of the 250 destination prefixes. The collected AS-level paths provide us the required information for calculating

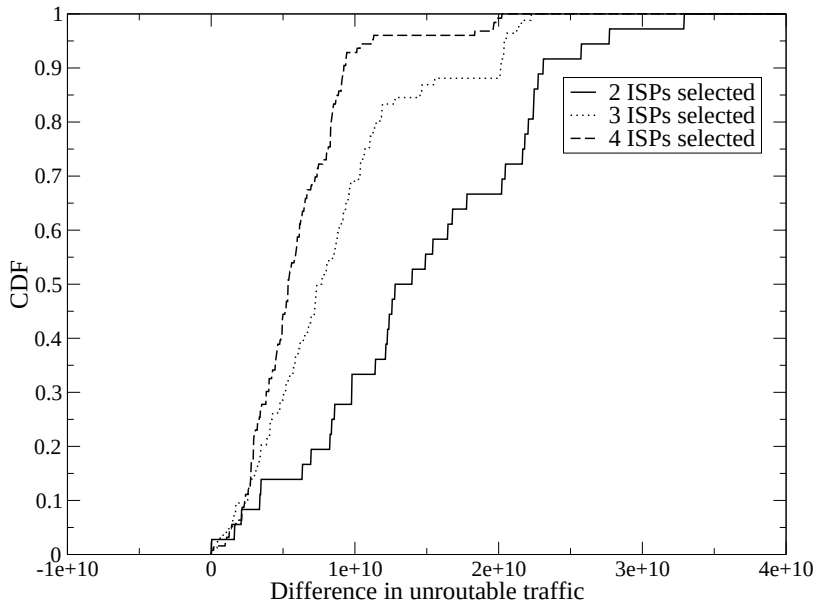

Figure 24: CDF of Δu for single-link failures.

c_p and c_d.

3.4.3 Evaluation of path diversity

We wrote a simple simulator that takes as input the AS-level topology from each of the nine potential ISPs to each of the 250 destination networks. This topology is basically nine different trees with the same 250 leaves, rooted in each of the ISPs. The internal nodes represent traversed ASes and the edges represent inter-AS links. The nine trees share some internal nodes and edges. We next calculate the path diversity metric $\kappa(i, \mathcal{C})$ for each destination i and selection of ISPs $\mathcal{C}$.

Given a specific number K, we compute the selection $\mathcal{C}^*$ of K ISPs with the minimum path diversity cost, as described in Section 3.3. To evaluate the robustness of that selection, the simulator considers each link in the topology and counts the amount of traffic that would not be routable if that link will fail. Recall that with single-link failures, a destination is unreachable only if a K-shared link fails.

Let $u(\mathcal{C})$ be the total amount of traffic that would not be routable with a selection of ISPs $\mathcal{C}$, considering all possible single-link failures in the topology. The difference between

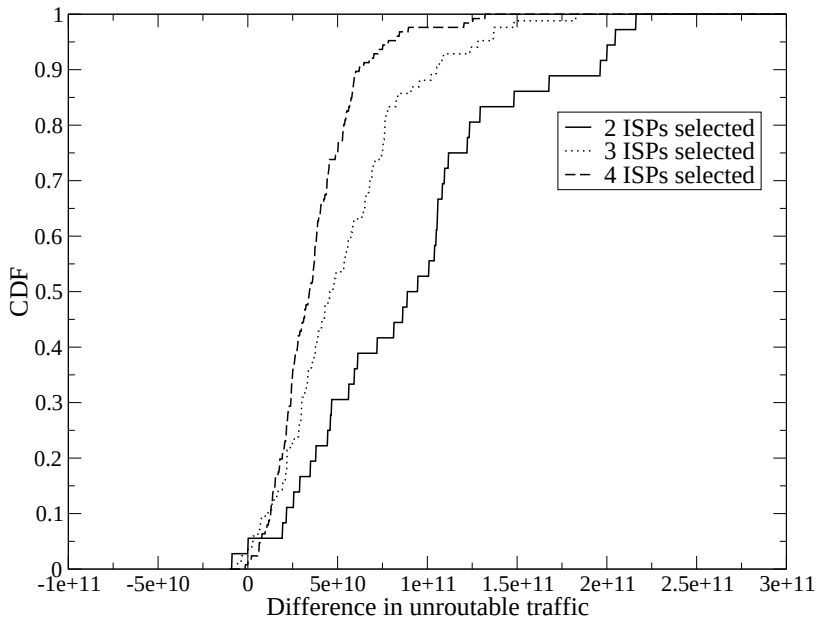

Figure 25: CDF of Δu for double-link failures.

the unroutable traffic with an arbitrary selection $\mathcal{C}$ and our selection $\mathcal{C}^*$ is

$$\Delta u(\mathcal{C}) = u(\mathcal{C}) - u(\mathcal{C}^*) \tag{16}$$

Figure 24 shows the CDF of $\Delta u(\mathcal{C})$ for K=2, 3 and 4 ISPs. Note that all differences $\Delta u(\mathcal{C})$ are positive or zero, confirming that our selection $\mathcal{C}^*$ is optimal in terms of providing robustness to single-link failures. This is not surprising, as $\mathcal{C}^*$ minimizes the number of K-shared links, which represent the Achilles' heel for single-link failures. Furthermore, $\Delta u(\mathcal{C})$ can often be very large, meaning that a selection of ISPs that ignores path diversity can lead to poor availability.

We repeated the previous experiments for double and triple link failures. As the number of possible failures in that case is very large, the simulator randomly picks 1000 cases of double or triple link failures, and measures the traffic that would not be routable through the K chosen ISPs in an arbitrary $\mathcal{C}$. Figures 25 and 26 show the corresponding CDFs of $\Delta u(\mathcal{C})$. In this case, the selection $\mathcal{C}^*$ is not always optimal. Especially when we only have two ISPs (K=2), there are some selections of ISPs that are better than $\mathcal{C}^*$ in terms of unroutable traffic. Nevertheless, $\mathcal{C}^*$ is still among the best 5% of ISP selections in terms of robustness to double and triple link failures, and it is almost optimal when we have three

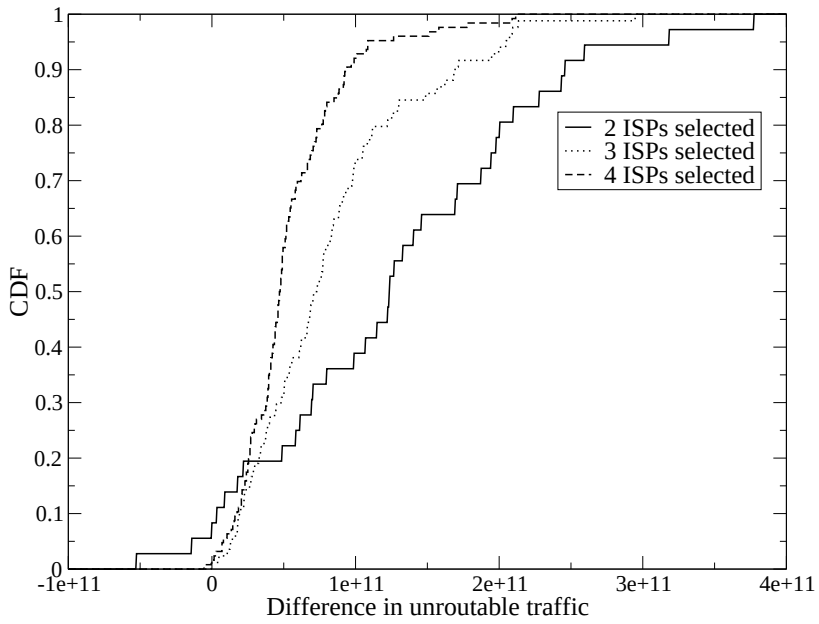

Figure 26: CDF of Δu for triple-link failures.

or four ISPs.

3.5 Phase II - Egress Path Selection

3.5.1 Problem statement

Once Phase-I is completed, S is connected to the Internet through the set $\mathcal{C}^*$ of the K best ISPs. In Phase-II, we aim to determine an optimal path allocation for each of the M destinations in $\mathcal{D}$. The notation in Phase-II remains the same as in Phase-I. In particular, $G(\cdot)$ is the "mapping function" such that destination i is mapped to ISP $j = G(i)$ ($j \in \mathcal{C}^*$), and $\mathcal{G}$ is the set of all possible mappings. The main objective in Phase-II is to determine a mapping in $\mathcal{G}$ that minimizes the total cost, given by $\sum_{j \in \mathcal{C}^*} c_j(T_j)$, subject to the constraint that none of the paths $P_{i,j}$ to the destinations in $\mathcal{D}$ is congested. A path $P_{i,j}$ to destination i through ISP j is congested if it has a positive loss rate $l(P_{i,j}) >0$. Note that congestion can occur either at the access links of S or in the upstream networks. Hence, Phase-II can be stated as the following problem:

Determine the mapping $G \in \mathcal{G}$ that minimizes the cost

$$\min \sum_{j \in \mathcal{C}^*} c_j(T_j)$$

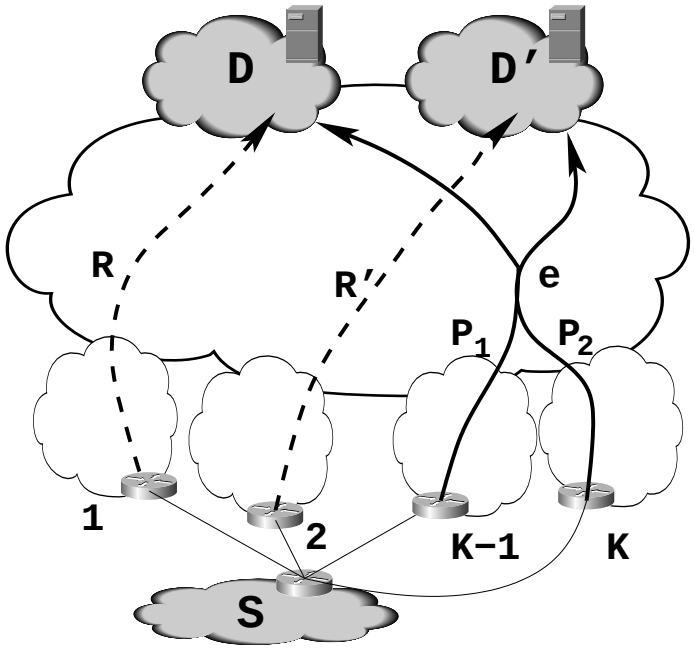

Figure 27: Link e is not the bottleneck of paths P_1 and P_2, but it can become the joint bottleneck of the two paths when they are used simultaneously.

subject to the constraint:

$$l(P_{i,j}) = 0 \quad \text{for all } i \in \mathcal{D}.$$

The previous problem may appear at first as a classical network flow problem. This is not the case however. Network flow and optimal routing problems assume that the topology of the network is given, and that the capacity of each link is known. In our context, this is not the case. Even though S knows the capacity of its own access links to the K ISPs, it does not know the topology or the capacity of the upstream network paths $P_{i,j}$. This is a key issue in Phase-II and it is the main reason we consider stochastic search techniques in the following.

Note that even though "traceroute" measurements can be used to infer the topology of upstream paths, such measurements do not provide us information about the capacity of those paths. Without such information it is not possible to determine whether a given mapping will result in congestion-free paths.

Also, even though there are techniques to estimate the available bandwidth in a path

through end-to-end measurements, those techniques cannot determine the available bandwidth in a bottleneck link that is simultaneously used by two or more paths from S. To illustrate this point, suppose that two paths from S share a link e with available bandwidth $A(e)$, as shown in Figure 27. The available bandwidth in the two paths is $A(P_1)$ and $A(P_2)$, and let us assume that $A(P_1) + A(P_2) > A(e)$ while $A(P_i) < A(e)$, $i = 1, 2$. This means than e is *not* the bottleneck of the two paths when they are considered in isolation, but it is their shared bottleneck when they are jointly used. Existing available bandwidth estimation tools can measure $A(P_i)$, but they cannot measure $A(e)$. Obviously, we need an estimate of $A(e)$ to infer the maximum traffic load that the two paths can jointly carry.

In conclusion, if we cannot know a priori whether a given mapping will be congestion-free or not, we need to consider *iterative routing approaches*. By iterative routing we mean that S routes its egress traffic based on a certain mapping for some time while measuring the loss rate in the corresponding paths. If any of these paths is congested the traffic is rerouted based on a different mapping. The loss rate $l(P_{i,j})$ in path $P_{i,j}$ can be estimated with active probing or passive measurements at the border router of S.

An iterative routing approach has the drawback that it causes rerouting. This can be a problem for TCP-based or streaming applications. Consequently, when the minimum-cost mapping is not congestion-free and routing iterations are necessary, we allow a certain cost increase while trying to keep the amount of rerouted and dropped traffic as low as possible. Note that if the minimum-cost mapping is congestion-free, then the path allocation problem is solved without routing iterations.

3.5.2 The algorithm

We propose a two-step algorithm. In the first step, we assume that the bottlenecks of all paths $P_{i,j}$ are the K access links of S. So, if the minimum-cost mapping is such that the traffic T_j routed through ISP j is less than the access link capacity A, that allocation will also be congestion-free. Under this assumption, the optimal path allocation problem is reduced to a variation of the bin-packing problem, for which we have already presented an efficient heuristic (Algorithm-1).

In the second step of the algorithm, we route the traffic based on the minimum-cost mapping and examine whether any of the egress paths is congested. If that is the case, our earlier assumption about the location of the bottlenecks is false, and the congestion occurs somewhere in the upstream networks. A simulated annealing algorithm is then invoked to search for a congestion-free mapping in the vicinity of the minimum-cost solution, while trying to reduce the amount of rerouted and dropped traffic.

The key idea behind this two-step approach is that in many cases the bottlenecks are the access links. The reason is that most core networks and private peering points between major ISPs are currently overprovisioned. Consequently, we expect that this assumption will usually result in a good, if not optimal, mapping. If some paths are congested elsewhere in the network, then the stochastic search component of the algorithm performs a local modification of the initial mapping, rerouting only the congested egress flows.

3.5.3 Initial mapping

The initial mapping is computed with Algorithm-1. In Phase I, we used that algorithm to examine whether a set of K ISPs provides a feasible mapping, and to determine the minimum cost of that mapping. Here, we use the set C^* that resulted from Phase I to route the egress traffic towards the destinations in $\mathcal{D}$. Note that we only consider the M largest destinations of outgoing traffic. The rest of the destinations should be also accounted for. We assume that that part of S's traffic is evenly distributed among the K ISPs, and so the access capacity A in Algorithm-1 refers to the residual capacity that is available for the largest M destinations.

3.5.4 Stochastic search and simulated annealing

Simulated annealing was first proposed by Kirkpatrick [69] as a general methodology within the area of stochastic search and optimization. The underlying idea is based on the physical process of annealing (cooling) in the chemical industry. Simulated annealing has been applied with slight variations to many combinatorial optimization problems (for instance, see [62, 80, 12]). In its most general form, the algorithm starts with an initial solution and an initial temperature. At each iteration, it first evaluates the cost of the current solution.

If the cost is found to be unacceptable, the algorithm generates a new candidate solution, typically modifying certain aspects of the current solution. If the cost of the new solution is lower ("downhill move"), the solution is always accepted. Otherwise, the solution is accepted with a probability $e^{\frac{-\Delta c}{T}}$ ("uphill move"), where Δc is the cost increase due to the new solution and T is the current temperature. This is called the Metropolis criterion [78]. Accepting a move with increasing cost helps the algorithm to avoid local minima. The temperature T decreases across successive iterations, diminishing the possibility of uphill moves and forcing the algorithm to eventually terminate. The algorithm exits when a solution with an acceptable cost is found, or when the temperature has reached a certain "freezing point". The pseudocode of the basic simulated annealing algorithm is shown in Algorithm-2. Some parts of the algorithm that are more specific to our problem are described in the following paragraphs.

Algorithm 2 Simulated annealing pseudocode for Phase II

```
1:  Calculate initial temperature T
2:  Get initial mapping G from Algorithm I
3:  Route traffic as in mapping G
4:  c_curr = cost(G)
5:  repeat
6:      if c_curr = 0 then
7:          return G {congestion-free solution}
8:      else
9:          Generate new solution G_new {as described in text}
10:         Route traffic as in mapping G_new
11:         c_new = cost(G_new)
12:         if c_new − c_curr ≤ 0 then
13:             G = G_new
14:             c_curr = c_new {new mapping is better}
15:         else
16:             With probability e^{−(c_new − c_curr)/T},
17:             G = G_new and c_curr = c_new {Metropolis criterion}
18:         end if
19:         T = ρT {cooling rate}
20:     end if
21: until T ≈ 0
```

Cost function: Recall that our objective is to find a congestion-free mapping in the vicinity of the minimum-cost mapping provided by the bin-packing step of the algorithm. Consequently, we consider a cost function that measures the overall congestion experienced

by S's egress traffic. Specifically, suppose that $l(P_{i,j})$ is the loss probability measured at path $P_{i,j}$ after the last routing iteration. The cost (overall congestion) $c_c(G)$ of a mapping G is the total rate of dropped traffic, across all destinations in $\mathcal{D}$,

$$c_c(G) = \sum_{i=1...M} r_i l(P_{i,j}) \tag{17}$$

where r_i is the average rate to destination i.

Initial temperature: In [68], Kirkpatrick suggests that the initial temperature should be chosen so that the probability of accepting an uphill move from the initial solution G_0 is about 0.8. We also assume that, initially, the worst move that should be accepted is one that at most doubles the initial cost. Hence, the initial temperature T_0 is set to $T_0 = \frac{-c_c(G_0)}{ln(0.8)}$.

Generating a new solution: A critical part of the algorithm is to determine a new mapping G_{new}, with lower cost than the current mapping G_{curr}. Since our cost function is congestion-related, we consider ways to reroute one or more congested destinations. We first simulated various schemes that reroute multiple congested flows at the same time. Those schemes performed consistently worse than schemes that move a single flow at a time. Hence, we examine mappings in which G_{new} and G_{curr} differ in the path of a single destination. Second, each time we reroute a destination, we allocate it to the ISP that will result in the *minimum cost increase*, among the set of ISPs with sufficient residual access capacity. The third issue is to determine the particular destination that should be rerouted. We evaluated the following three heuristics:

1. **Max-cost:** Reroute the congested destination with the highest cost in the current mapping.

2. **Max-loss:** Reroute the congested destination i with the highest loss rate $r_i l(P_{i,j})$ in the current mapping.

3. **Min-rate:** Reroute the congested destination i with the lowest rate r_i.

To summarize the results of this simulation study, we found that the Max-loss performs best in terms of minimizing the amount of dropped traffic (as we would expect), but it also performs best in terms of the number of routing iterations. The Min-rate algorithm is better

in terms of minimizing the amount of rerouted traffic, but the Max-loss algorithm does not do significantly worse. Consequently, in the following, we use the Max-loss algorithm.

Annealing Schedule: The annealing schedule determines the rate at which the temperature is decreased. The related literature proposes mostly geometric cooling for large combinatorial optimization problems [12]. Our simulations showed that an annealing schedule with very slow cooling rate, such as $\rho = 0.99$, is more suitable for our problem.

Termination conditions: Depending on the capacity and topology of the underlying network, a congestion-free mapping may not exist for a given traffic load. Allowing the simulated annealing algorithm to keep searching for a feasible solution until the temperature drops to zero may cause significant rerouting. Consequently, we set the following additional termination conditions.

1. The monetary cost of any considered mapping should not be too large. Specifically, if the monetary cost of a mapping becomes larger than a factor $cost\_thresh$=2 of the initial cost, the search terminates.

2. If the congestion cost has not decreased significantly over a number of iterations, it is likely that there is no feasible solution. Specifically, if the congestion cost has not decreased by at least a factor $cong\_thresh$=1.1 in any of the last $iter\_inc\_cong$=10 iterations, the search terminates.

3.6 Phase II - Evaluation

The objectives of this section are twofold. First, to evaluate Algorithm-1, needed in both Phase-I and Phase-II. Second, to evaluate Algorithm-2 of Phase-II, as well as some simpler algorithms for solving the same problem. The evaluation of the two algorithms is performed with flow-level simulations that use measured datasets for the outgoing traffic distribution and the underlying IP-layer topology.

3.6.1 Measured traffic and topology datasets

To simulate the model of Figure 1 more realistically, we rely on the following three measured datasets. First, as described in Section V-A, we collected traces of the outgoing traffic from

our university network to determine the 250 largest destinations (accounting for about 65% of the total traffic) and the distribution of traffic among them. We found that that distribution can be modeled as Pareto with shape parameter 1.08.

The second dataset is related to the network topology of the "upstream cloud" from S to the major destination networks. To simulate the upstream ISPs of S, we used three Planetlab nodes that are geographically located in the state of New York: Columbia University, New York University, and Cornell University. These three hosts provide us with different IP-layer paths from the same (roughly) geographical area to different destinations of traffic.

Third, to simulate the IP-layer paths from the upstream ISPs to the major destination networks we used two approaches. First, we collected traceroute data from the previous three Planetlab nodes to the 250 largest destination networks in our university packet trace. Unfortunately, traceroute cannot report the entire route to several destination networks. However, if the traceroute outcomes from all three Planetlab nodes merge at a common intermediate node after a certain point, we consider that node (router) as the traffic destination. In the second approach, we run traceroute from the three Planetlab nodes to 100 randomly chosen destination IP addresses from Caida's Skitter datasets [22]. Of course, the drawback of this approach is that randomly picked destinations may not be large traffic sinks in reality.

3.6.2 Simulator parameters

The simulator aims to route M flows, modeling the traffic to each of the destination networks, through a given network topology. Each flow can originate from one out of K=3 access links. The flows are modeled as constant fluids, i.e., they are completely specified by a rate; we do not consider the short-term effects of traffic variability. The key simulation parameters are the following:

Flow rates and destination ranking: As previously noted, the distribution of flow rates follows the Pareto distribution. The average rate, across all M flows, controls the load in the network. Different simulation random seeds result in different flow rates. We

always assign the largest rate to the same destination, the second largest rate to another destination, and so on. In other words, we assume that even though the traffic to each destination varies across simulations, the flows retain their ranking. We believe that this property resembles the characteristics of real egress traffic better than assuming that, for example, the 10th largest destination today can be the 100th largest destination tomorrow.

Location of bottlenecks: Each path $P_{i,j}$ from S to destination i through ISP j has a bottleneck link, which is the link with the minimum capacity. A parameter *bneck_loc*, between 0 and 1, controls the location of that bottleneck. This parameter determines the link of $P_{i,j}$ that has the largest probability of being the bottleneck. Neighboring links also have a probability of being the bottleneck, which decreases geometrically with their distance from the most likely bottleneck. *bneck_loc*=0 means that the access links of S are the most likely bottlenecks. *bneck_loc*=1 means that the access links of the destination networks are the most likely bottlenecks. A value of *bneck_loc* around 0.5 will bring the bottlenecks close to the core of the upstream network.

Shared bottlenecks: As illustrated in Figure 6, a network link that is shared by two or more paths can become their joint bottleneck when those paths are used simultaneously. The presence of shared bottlenecks can cause strong coupling between paths to different destinations. For example, switching the traffic of destination i from ISP j to ISP j' can cause congestion in other paths and destinations, not routed through j'. A parameter *bneck_shar*, between 0 and 1, controls the probability that a link which is shared by two or more paths is their joint bottleneck. By joint bottleneck we mean that that link becomes congested only when *all* those paths are active, i.e., used to reach the corresponding destinations. If only some of those paths are active, the shared link will not be congested. *bneck_shar*=0 means that a shared link is never a joint bottleneck, while *bneck_shar*=1 means that a shared link is always a joint bottleneck.

3.6.3 Evaluation of Algorithm-1

In this section, we focus on the evaluation of Algorithm-1. Recall that that algorithm attempts to identify the minimum-cost allocation of M destinations to K access links,

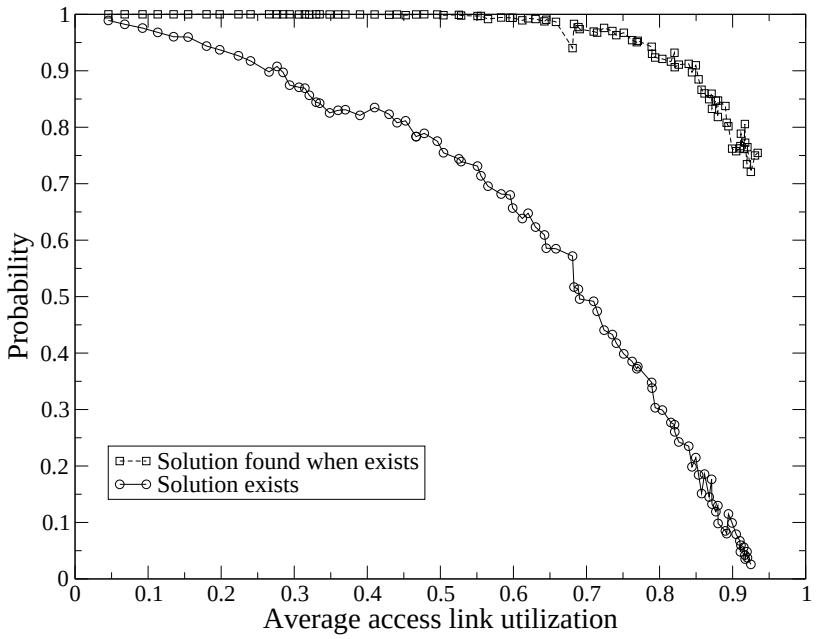

Figure 28: Probability that solution exists, and probability that solution is found by Algorithm-1 when it exists.

subject to the same capacity constraint for each link. We are interested in two major questions. First, can Algorithm-1 find a solution to the previous problem, when a solution exists? And second, what is cost of the solution reported by Algorithm-1 relative to the cost of the optimal solution? Both questions require exhaustive search in order to know whether a solution exists and, when that is the case, the cost of the optimal solution. For this reason, in this part of the evaluation study we limit the topology to 10 randomly picked destinations from the 250 destinations included in our topology.

The *bneck_loc* factor is set to 0, meaning that the destinations are bottlenecked at the K access links. For each value of the average flow rate, Algorithm-1 and the exhaustive search routine are run 5,000 times. We then estimate the following metrics for each value of the average flow rate. First, the probability that a solution exists. Second, the probability that Algorithm-1 will find a solution, when a solution exists. Third, the cost ratio of the solution reported by Algorithm-1 and the optimal solution, when a solution exists.

Figure 28 shows the first two metrics, as a function of the average utilization of the access links (each value of the average flow rate corresponds to a different utilization). Note

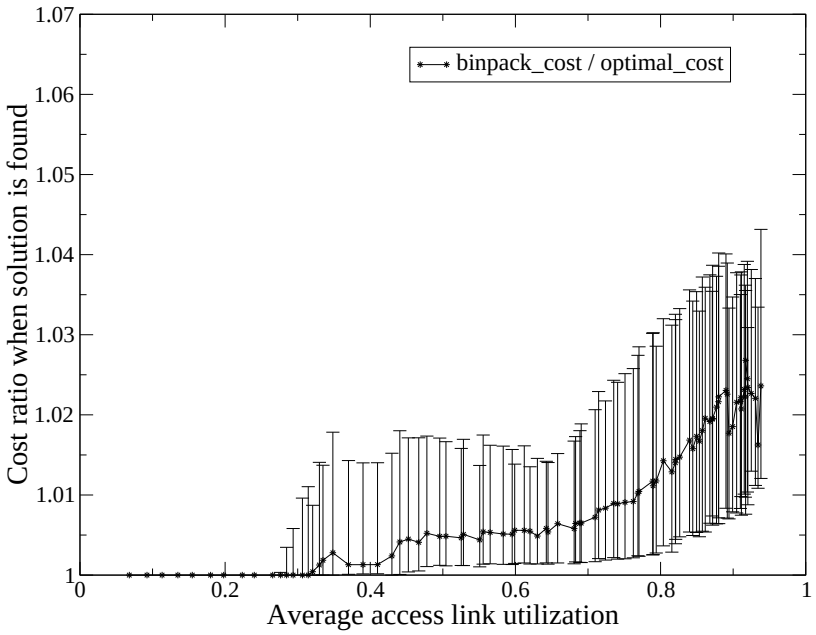

Figure 29: Cost ratio between Algorithm-1 solution and optimal solution.

that as the load increases to more than 20%-30%, the probability to find a feasible allocation drops significantly, to less than 80%-90%. This "early saturation" effect is a result of the heavy-tailed nature of the Pareto distribution: a few flows have very large rate relative to the individual link capacities. The good news is that Algorithm-1 can identify a solution with very high probability (practically 100%) when a solution exists, as long as the average load is below 60%-70%.

Another positive result is that Algorithm-1 results in almost the minimum-cost solution, when a solution exists. Figure 29 shows the median and the inter-quartile range for the cost ratio between the Algorithm-1 solution and the optimal solution. We see that the median cost ratio is very close to 1, and the 75^{th} percentile value is less than 1.05.

3.6.4 Evaluation of Algorithm-2

In this section, we focus on the evaluation of Algorithm-2. Recall that the objective of that algorithm is the minimum-cost assignment of each egress flow to an upstream ISP, subject to the constraint that none of the egress paths is congested. Algorithm-2 is based on stochastic search and it may need several routing iterations before it finds a solution. We

refer to the time period during which the algorithm searches for a solution as the "transient phase". We are interested in the following questions. First, what is the probability that Algorithm-2 will find a solution, as we increase the offered load from S? Second, how long is the transient phase in terms of the required number of iterations? Third, what is the total amount of dropped traffic due to congestion during the transient phase? And fourth, what is the total amount of rerouted traffic due to routing iterations during the transient phase? In the following evaluation study, we use the entire topology (250 destinations). An exhaustive search in a topology of this scale would be computationally prohibitive. Hence, we do not present results for the probability that a solution exists or for the cost of the optimal solution.

Instead of presenting results only for Algorithm-2, we also evaluate the following simpler algorithms. The objective of these comparisons is to get some insight in the relative performance of Algorithm-2 and to understand the significance of simulated annealing compared to methods that follow the "greedy search" paradigm.

1. **Access link bottlenecked (access-link):** In the simplest case, Phase-II can assume that all egress flows are bottlenecked at the K access links. In that case, Algorithm-1 can be used to approximate the minimum-cost congestion-free allocation. This algorithm does not require routing iterations.

2. **Greedy, moving a single flow at each iteration (greedy-single):** This is similar to Algorithm-2, but without the simulated annealing component. In each iteration, the flow with the highest loss rate (Max-loss) is moved to the ISP that will cause the minimum cost increase, among the ISPs with sufficient access link capacity. The algorithm never accepts uphill moves.

3. **Greedy, moving multiple flows at each iteration (greedy-mult):** In each iteration, the congested flows are first ordered in decreasing order of their loss rate. Then, each flow in that sequence is moved to the minimum-cost ISP that has sufficient access link capacity. Note that this is the only algorithm that moves more than one flow in the same iteration.

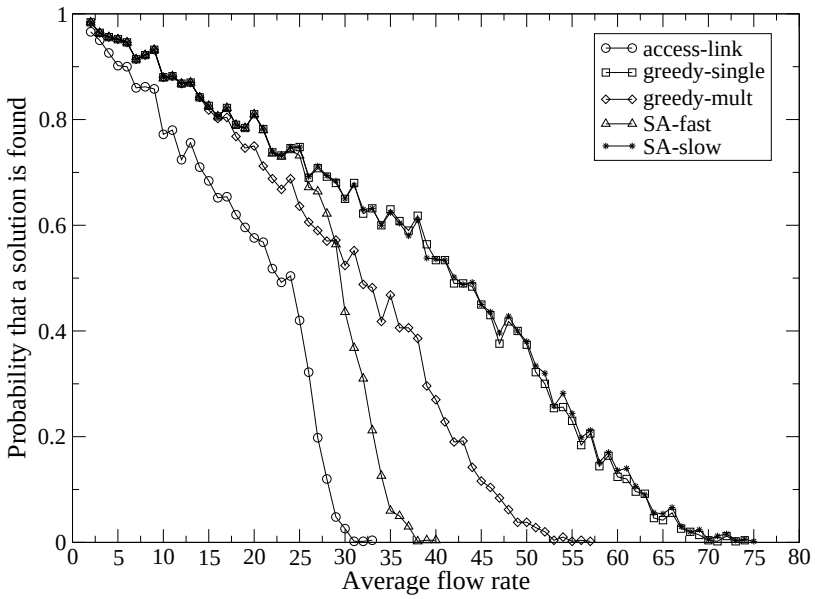

Figure 30: Probability that a solution is found.

4. **Simulated annealing variations (SA-fast and SA-slow):** We examine two variations of Algorithm-2, one with very fast cooling (ρ=0.5), called SA-fast, and another with very slow cooling (ρ=0.99), called SA-slow.

In the first set of simulations, we set the path bottlenecks at the access links of the three upstream ISPs. In this case, if a solution exists, we expect that Algorithm-1 will find it and Algorithm-2 will terminate without any routing iterations. If a solution does not exist, on the other hand, then the iterative routing approach of Algorithm-2, or of any other iterative algorithm, would obviously not help. The simulation results confirmed this intuition.

In the second set of simulations, we set the path bottlenecks inside the network ($bneck\_loc$=0.5). The shared links are not joint bottlenecks ($bneck\_shar$=0). Figures 30 to 33 show the simulation results for this configuration. In terms of the probability to find a solution, Figure 30 shows that *SA-slow* and *greedy-single* perform better than the other algorithms, and they actually give very similar results. On the other hand, *SA-fast* has significantly lower success probability than *SA-slow* after the load has become significant. The reason is that, when cooling happens very fast, a simulated annealing algorithm terminates too early, before it has the chance to find a solution. Also note that *greedy-mult* does not perform as well

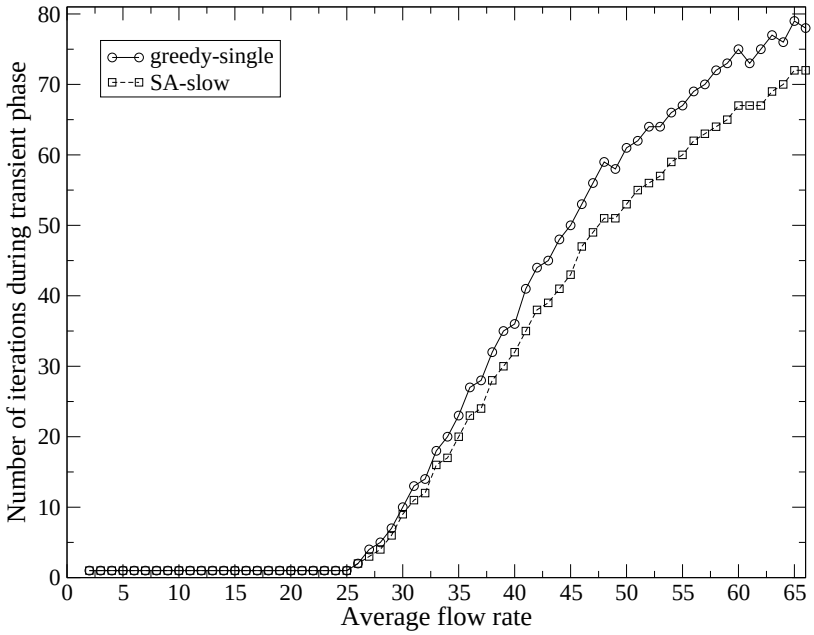

Figure 31: Number of iterations during transient phase.

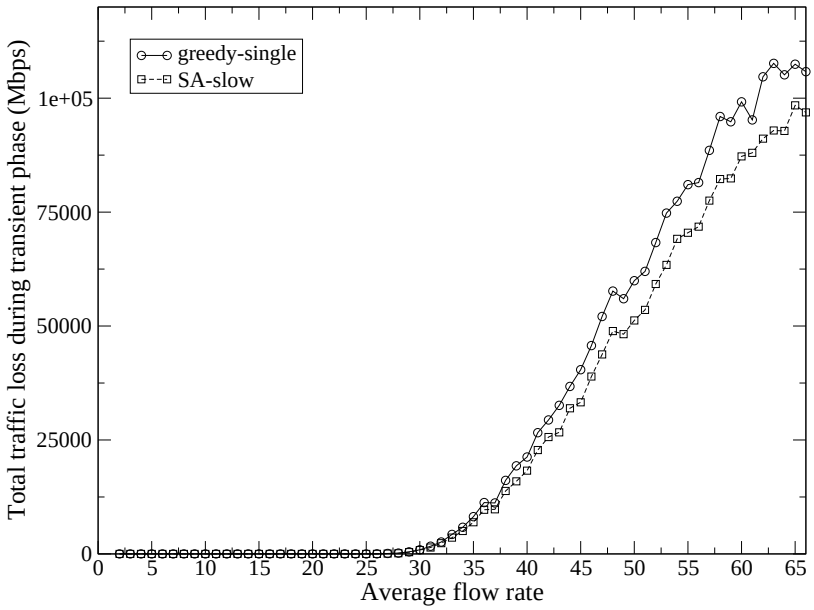

Figure 32: Total traffic loss during transient phase.

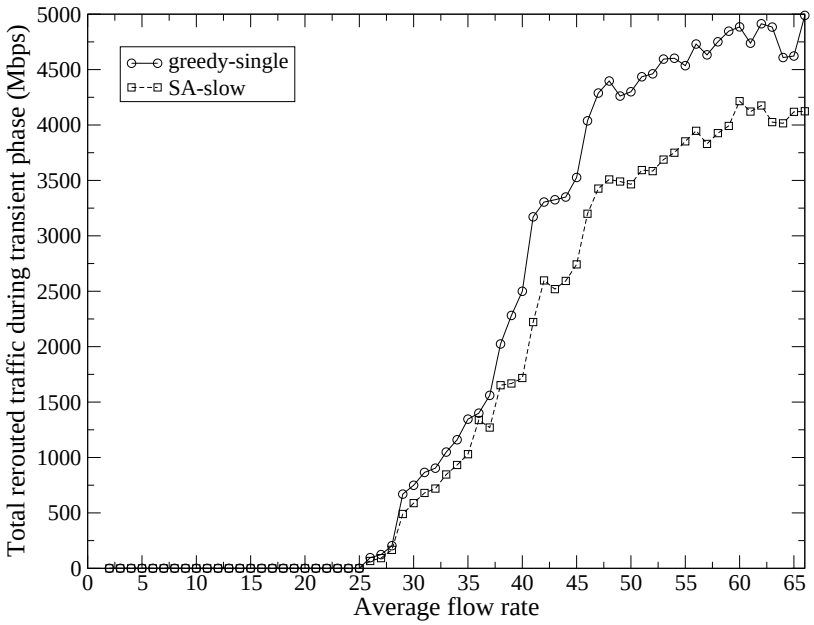

Figure 33: Total rerouted traffic during transient phase.

as the *greedy-single* algorithm. It is also interesting that even though *access-link*, *SA-fast*, and *greedy-mult* show a rapid decrease of the success probability after a certain load, the algorithms *SA-slow* and *greedy-single* observe a much more gradual degradation. Based on these results, in the following we focus on *SA-slow* and *greedy-single*.

Figure 31 shows the median number of iterations, across 500 simulation runs, until the algorithm terminates. Note that *SA-slow* performs better than *greedy-single*, especially in heavy load conditions, i.e., when the probability of success with these two algorithms is less than about 50%. We also calculated the confidence intervals for the number of routing iterations (not shown here). The two algorithms have wide and significantly overlapping confidence intervals. This means, first, that there is significant variability across different simulations (flow rates). The reason for this is that different rates across simulation runs could lead to different initial mappings produced by Algorithm-1. This, in turn could lead to a different set of congested flows and loss rates, significantly affecting the dynamics of the iterative algorithms. Consequently, even though *SA-slow* needs fewer iterations on the average, there is a significant fraction of simulations in which *greedy-single* performs better.

The amount of rerouting that an algorithm introduces is also important. We calculate

the cumulative rate of rerouted traffic during the transient phase by counting after each iteration the total rate of the flows that were rerouted. Figure 33 shows a similar trend with the number of routing iterations: *SA-slow* performs better than *greedy-single*, especially in heavy load conditions. The confidence intervals, however, again show significant variability and overlap. We see similar trends for the total rate of dropped traffic during the transient phase, shown in Figure 32.

In summary, the simulated annealing algorithm (with slow cooling) performs better, at least on the average, than a greedy algorithm which never accepts uphill moves. The difference between the two algorithms is more significant in terms of the number of required routing iterations and the amount of rerouted and dropped traffic in heavy load conditions, when the probability of finding a solution is less than 50%.

We also performed simulations in which the shared links are also joint bottlenecks ($bneck\_shar$=1). The results follow similar trends as the previous graphs, implying that the performance of the *SA-slow* and *greedy-single* algorithms does not depend significantly on the presence of joint bottlenecks.

3.7 Related Work

We discuss two bodies of related work. The first focuses on the problem of ISP selection for a multihomed network. The second deals with the allocation of egress traffic to the selected set of ISPs to improve performance and/or cost.

The work of Orda and Rom [89] was one of the first to consider the optimization of multihomed networks. That early work took a topological approach, with nodes representing potential attachment points for subscribers, and constraints on the number of subscribers that a node allows. The objective was to find the optimal set of nodes that each subscriber should connect to with the aim of minimizing the distance between the subscriber and every other node in the network. After several years, Wang *et al.* [106] studied this problem from the perspective of an ISP subscription problem. That work proposed algorithms for selecting a set of upstream ISPs with a (monetary) cost minimization objective. A main difference between the methodology of Wang *et al.* and our work is that we also consider

performance and path diversity objectives. In particular, we are interested in choosing ISPs that provide significantly diverse paths to the major destinations of traffic. Also, the results of [106] are rather specific to a class of pricing functions that are based on the "percentile charging" model.

Existing IRC products choose egress paths dynamically, avoiding congestion in a reactive manner. Even though most commercial multihoming-IRC systems do not expose much technical information about their internal operation, one of them (the ISMD device of Rether Networks) is described with significant detail in a research publication [56]. Another good description and evaluation of an operational multihoming-IRC system is given in [5]. These papers, however, do not consider the monetary cost of allocating traffic to different upstream ISPs. The recent work by Goldenberg *et al.* [55] approaches the traffic allocation problem with the objective of optimizing performance given a maximum cost constraint. That work considers latency as the performance metric, and proposes algorithms to dynamically reroute egress traffic upon transient congestion periods. The results of [55] are also based on the percentile charging model, while we use a more general pricing model.

An experimental study, based on measurements from the Akamai content distribution network, showed that multihoming can lead to significant benefits in terms of both availability and performance for both ingress and egress traffic [6]. The authors also showed that having up to four upstream providers is enough to gain the full benefit of multihoming. Another experimental work that evaluated the benefits of multihoming is described in [103].

3.8 Conclusions

Multihoming is becoming inceasingly popular for edge networks that generate large amounts of content (70% of stub networks are now multihomed, as measured in Chapter 2), as these networks try to optimize their costs and performance. In particular, we observed that content providers have been multihoming quite aggressively, and are also the most active in changing their upstream providers. The exact choice of upstream providers for an edge network can significantly impact the performance that the network can obtain to the rest

of the Internet, or the costs that this network must incur. In this part of the thesis, we proposed a systematic methodology for edge networks to choose the optimal set of upstream providers using information that can be collected *offline*, without actually connecting to those providers.

Our proposed methodology consists of two phases. In Phase-I, the objective is to choose the optimal set of upstream ISPs given a coarse traffic profile that captures the average rate and the AS path to each major destination network. The optimality of the resulting set of ISPs is determined by a weighted average of their monetary cost, AS path length and path diversity that any set of ISPs provides. In Phase-II, the objective is to assign the traffic towards each destination to an upstream ISP so that the total monetary cost is minimized, without experiencing long-term congestion. The main difficulty in Phase-II is that the available bandwidth of the upstream network paths is generally unknown. Hence, we need to use stochastic search techniques and iterative routing. We envision that Phase-I can be repeated in long time scales, from weeks to months, while Phase-II can be repeated whenever there is a major change in the egress traffic distribution.

CHAPTER IV

A MODEL FOR INTERDOMAIN NETWORK FORMATION, ECONOMICS AND ROUTING

4.1 Introduction

The Internet at the interdomain level is a system of interacting autonomous networks (ANs)[1] that connect to each other to provide end-to-end connectivity and access to various forms of content. In Chapter 2, we measured the dynamics of the "Internet ecosystem", which consists of networks with different business objectives that must interact and co-exist with each other. A major observation from Chapter 2 was that the Internet is highly dynamic, as ANs continually change the set of providers and peers that they connect to. A plausible reason for the dynamics in the Internet is that ANs attempt to optimize, in a distributed manner, utility functions such as monetary profit, cost or performance. The utility that ANs are able to obtain depends both on "environmental" factors (transit prices, peering costs or the popularity of new Internet applications) and on their choice of providers and peers[2]. In practice, however, the process of provider and peer selection is often treated as "black art", even by network operators of large ISPs. These ISPs select their providers and peers using rules of thumb such as "peer by traffic ratios" or "peer restrictively". The conditions under which such strategies are actually profitable for different types of networks has not been well studied. Further, ANs have no way to reason about the effects of their provider and peer selection strategies on the global Internet.

The motivation behind this work was to create a framework that can be used to study the effects of provider/peer selection strategies used by different types of ANs. We do this by developing a model for interdomain network formation that determines what the

[1] ANs are similar to Autonomous Systems in BGP in the sense that they are independently operated, except that they also include networks that do not have AS numbers.

[2] A "provider" is a network that provides transit, or access to the rest of the Internet, to its customers. Two networks are "peers" if they engage in settlement-free interconnection, whereby they provide access to each other's customers for free.

internetwork converges to as ANs try to optimize their individual utility functions. Our goal is not to recommend which exact networks an AN should choose as its providers or peers (Commercial offerings such as Renesys Market Intelligence [3] provide such a service to their customers; their algorithms and data sources, however, are proprietary). We also do not try to model the evolution of the Internet ecosystem in terms of which interdomain links appear or disappear. To do either in sufficient detail would require a precise knowledge of the strategy of every other AN, the interdomain traffic matrix, and pricing/cost parameters. Instead, the goal is to study the internetwork after it has converged, and to evaluate the effect of strategies such as "AN i peers with any network with which it exchanges roughly equal traffic". In this case, we do not aim to measure specifics such as which exact links are present, or the exact set of providers and peers for each AN. Instead, we try to gain more general insights into the effects of AN strategies on the utility of various network types and resulting global properties. We are interested in both *local effects* (how these strategies affect the ANs that use them), and *global effects* (how they affect the overall Internet).

Studying the effects of provider and peer selection by different types of networks is interesting for several reasons. First, individual networks would like to know which strategy maximizes their utility (either monetary profits, costs or performance). Second, we would like to know the effects of these strategies on the global Internet, in terms of topological structure, profitability of various network types, and the risk of emerging monopolies or oligopolies. Third, it is important for ANs to understand how their provider/peer selection strategies perform under different conditions, such as diverse traffic characteristics and application popularity, different pricing structures, or new technology (*e.g.*, inexpensive transmission capacity).

The main contribution of this part of the thesis is a model, ITER, that provides the framework for answering questions of the aforementioned type. ITER is based on first-principles, and models the provider and peer selection process for different classes of ANs – Enterprise Customers (EC), Small and Large Transit Providers (STP and LTP), and Content Providers (CP). ITER takes as input the interdomain traffic matrix, routing policies, geographical constraints, and the economics of transit, peering and local costs. ITER

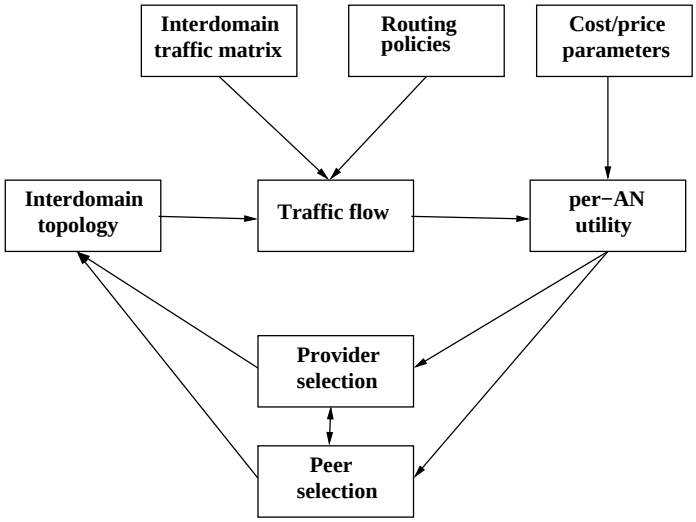

Figure 34: The interdependence between topology, traffic flow and per-AN utility in the Internet ecosystem

models the interdependence between traffic flow, topology and the provider/peer selection strategies of ANs, as shown in Figure 34. The interdomain traffic matrix, topology and routing policies determine the flow of traffic in the Internet. The traffic flow and economic factors together determine the utility of each AN (profit for transit providers and monetary cost/performance for ECs). ANs optimize their utility by changing their set of providers and peers, effectively changing the topology, which in turn can affect the utility of other ANs. The question we try to answer is, "Does this process to converge, and if so, where?", *i.e.*, we are interested in "solving" the model to find a state where no AN has the incentive to make further changes to its connectivity (if such a state exists). As ITER is intractable to solve analytically, we devise a method to solve it computationally, using agent-based simulations. We also study the existence and uniqueness of the resulting equilibrium. We emphasize again that *ITER is not an evolutionary model*; it does not model the long-term evolution of the Internet ecosystem with the birth and death of ANs, changes in the popularity of various applicaitons, and fluctuating economic conditions. Though it is important to study the evolution of the Internet, we argue that for the purpose of evaluating provider/peer selection strategies of ANs, the equilibrium of the static ITER model can give valuable insights.

In this part of the thesis, we focus on a first practical application of the ITER model, that of studying the properties of the equilibrium internetwork, given different provider/peer selection strategies used by ANs. In particular, we focus on two provider selection strategies (choose cheapest providers, or choose providers that are not competitors), and three peer selection strategies (peer only when necessary to maintain reachability, peer by traffic ratios, and peer when the potential benefit of peering is larger than the estimated cost) for small and large transit providers. We measure properties of the resulting network in terms of topology (*e.g.*, path lengths and diameter), traffic flow, profitability of different types of providers, and the number of providers that are profitable. We also analyze the effect of factors such as the interdomain traffic matrix, geography, and customer preference on the resulting internetwork. Specifically, we investigate what happens when the interdomain traffic matrix consists of mostly peer-to-peer (P2P) traffic, or if ANs at the edge of the Internet choose their providers based on path lengths, or if content providers replicate their content in all geographical regions. We envision several other applications of ITER, discussed in Section 4.10, which we plan to pursue in future work. We summarize the main findings from this part of the thesis:

- We find that if networks at the edge are price-conscious, then LTPs can benefit by peering with CPs, and can significantly harm the profitability of STPs; this comes at the cost of longer end-to-end paths (Section 4.5).

- We find that the STP strategy of peering using "balanced traffic ratios" is profitable only if they also use price-based provider selection. In this case, STPs should peer avoid peering with content providers. The choice of the best peering strategy for STPs is heavily influenced by their provider selection strategy (Section 4.5).

- We find that two conditions that are quite plausible in the future Internet – an interdomain traffic matrix with mostly P2P traffic , and content providers that replicate their content in all regions – result in increased profitability for STPs (Sections 4.6 and 4.8).

- We find that performance-aware provider selection by edge networks results in a situation where end-to-end paths are short and LTPs are profitable (Section 4.7).

The rest of this chapter is organized as follows. Section 4.2 presents the details of the ITER model. Section 4.3 describes our approach for solving ITER using agent-based simulations. In Section 4.4 we validate the model against some well-known static and dynamic properties of the Internet. In Section 4.5, we present results for the default model, which we view as the current state of the Internet. In Section 4.6, we evaluate a deviation of the default model with a predominantly P2P traffic matrix. We evaluate a deviation where edge networks choose their providers based on performance in Section 4.7, and a deviation where Content Providers replicate their content in all geographical regions in Section 4.8. We survey the related work in Section 4.9, and conclude in Section 4.10 with a discussion of future applications of ITER.

4.2 Model description

In this section, we summarize the key features of ITER. Table 1 defines the various terms used in the remainder of this chapter.

4.2.1 Network types

Enterprise Customers (EC): ECs are stub networks that normally act as either mostly sources of traffic (*e.g.*, web hosting companies), or mostly sinks of traffic (*e.g.*, campus, corporate or residential access networks). In ITER, ECs do peer and they do not have customers; their only action is provider selection. We model a fraction of ECs as traffic sinks (*sink-ECs*), while the remaining as traffic sources (*source-ECs*).

Content providers (CP): CPs are also stub networks that differ from ECs in two ways. First, they are sources of traffic (*e.g.*, Yahoo!, Google). Second, they can engage in peering relations, following an "open peering" policy (peer with any network that agrees to peer with them).

Small Transit Providers (STP) and Large Transit Providers (LTP): Transit providers are networks whose main business function is to provide Internet connectivity to

Table 1: Definitions of acronyms used

acronym	definition
AN	Autonomous Network
EC	Enterprise Customer
STP	Small Transit Provider
LTP	Large Transit Provider
CP	Content Provider
CS	Client-Server
P2P	Peer-to-Peer
PR	Price-based provider selection
PF	Performance-based provider selection
SEL	Price-based Selective provider selection
NC	Peering by necessity
TR	Peering by traffic ratios
CB	Peering by cost-benefit analysis
DF	Default Model
P2P	Deviation: P2P traffic matrix
EP	Deviation: edge networks use performance based provider selection
GEO	Deviation: content providers present in each geographical region

their customers. In ITER, transit providers do not act as sources or sinks of traffic; they only carry traffic on behalf of other networks. Transit providers aim to maximize their profit and so they select their providers and peers with this economic objective. Their peering policies are often described as "restrictive" or "selective", in practice. STPs are transit providers with limited geographical presence (*e.g.*, Rogers Telecom or China Telecom), while LTPs are transit providers with practically global presence (*e.g.*, AT&T or Level3).

In the default ITER model, we simulate 180 ECs, 10 CPs, 16 STPs, and 4 LTPs. 20% of the ECs act as source-ECs, while the rest are sink-ECs. This 210-node internetwork is of course small compared to the real Internet (the number of Autonomous Systems is about 30,000 today) to keep the simulation time tractable; we will return to this scalability issue in Section 4.3.

4.2.2 Traffic model

The traffic model concerns the generation of an inter-AN traffic matrix. This matrix determines the amount of traffic sent from each AN to every other AN. In ITER, we consider two

types of traffic: Client-Server (CS) traffic flows from traffic sources, which are either CPs or source-ECs, to sink-ECs (*e.g.*, YouTube or RapidShare). Peer-to-Peer (P2P) traffic flows between sink-ECs (*e.g.*, BitTorrent). Without showing the actual mathematical expressions, the key points of the traffic model are the following. The total traffic volume (both CS and P2P) destined to each traffic sink is heavy-tailed (Pareto distributed with shape parameter=1.1), *i.e.*, few sink-ECs are much larger traffic consumers than most other sink-ECs. Traffic sources are ranked based on a *popularity index*. CPs have higher popularity index than source-ECs. The fraction of traffic from a given source to any sink-EC follows a Zipf distribution (with exponent 0.8), determined by the previous popularity ranking. The Zipf distribution implies that few traffic sources, mostly CPs, are much heavier traffic producers than most other sources. For simplicity, we assume that the popularity of a source is the same for all sink-ECs, ignoring any regional content preferences. A similar popularity index for each sink-EC determines the distribution of P2P traffic between sink-ECs. In the default ITER model, 80% of the overall traffic is CS while the rest is P2P.

4.2.3 Geographical constraints

In ITER, each AN is geographically present in a certain set of locations (*e.g.*, exchange points or "GigaPoPs"). Two ANs cannot establish a customer-provider or peering relation unless they are present in a common location. In the default ITER model, 210 ANs are distributed in 5 locations. ECs and CPs are present in one location, STPs in 2, and LTPs in all 5 locations. This distribution of different network types among regions is designed to capture real-world constraints faced by these networks, while also taking into account the scale of our simulations (restricted to 210 ANs). LTPs represent the large tier-1 providers, and it is realistic to assume that these networks have a presence in most (or all) regions of the world. Stub networks that are universities or corporations typically have a single location and the same is true of CPs (though a recent trend is that CPs expand their geographical scope to be present in many regions – considered later as a deviation from the default model). STPs are transit providers with a mostly regional scope, and given that we consider a total of 5 regions, it is reasonable to place them in two regions.

4.2.4 Routing and traffic flow

ITER captures the salient features of interdomain routing. Specifically, traffic follows the "no-valley" policy, (traffic from a provider cannot be sent to another provider, and traffic from a peer cannot be sent to another peer), as well as the "prefer-customer" policy (prefer a route that goes through a customer; if not available, prefer a route that goes through a peer; otherwise route through a provider). Whenever multiple preferred neighbors offer a route, choose the shortest path; break ties deterministically based on the neighbor's AN number. It should be noted that calculating policy-compliant shortest paths between all pairs of nodes is computationally expensive ($O(N^3)$, where N is the number of ANs in the internetwork). We have simplified the routing computation, without violating the previous policies, with an algorithm inspired by the method proposed by Gao and Wang [52]. We simplify the routing computation by assuming that stub nodes do not form peering links. We can then calculate the shortest policy compliant paths among providers. This can be done efficiently in time $O(N_p E_p)$, where N_p is the number of providers and E_p is the number of edges among providers. Following this step, each provider p learns the best path towards each stub s, via the provider p' of s for which p has the shortest path. This can be done in time $O(N_p N_s d_s)$, where d_{p_s} is the multihoming degree of stubs. Finally, each stub s determines the best path towards stub s'. To do this, s chooses the provider p from among its set of providers that gives the shortest path towards s'. The final step can be done in time $O(N_s^2 d_s)$.

Given the inter-AN traffic matrix, the interdomain topology and the routing model, we can then calculate the traffic flow in the internetwork. The traffic flow determines the aggregate amount of traffic that flows over each link and AN. These per-link traffic loads are then used by the economic model, described next.

4.2.5 Economic model

The economic component of ITER focuses on the profit of transit providers. STPs and LTPs adjust their provider and peering selections so that they maximize their profit. The profit of a transit provider is calculated as the total revenue from its customers, minus

the transit fees to its providers (if any), the peering costs (if any), and the local costs to maintain and operate its network. Let $\mathcal{C}_i$ be the set of customers, $\mathcal{P}_i$ the set of providers and $\mathcal{R}_i$ the set of peers of a transit provider i. Its profit f_i is:

$$f_i = \sum_{c \in \mathcal{C}_i} T_i(v_{ic}) - \sum_{p \in \mathcal{P}_i} T_p(v_{ip}) - \sum_{r \in \mathcal{R}_i} R_i(v_{ir}) - L_i(v_i)$$

$T_i(v)$ represents the pricing function used by provider i for a transit volume v (*i.e.*, volume-based pricing). $T_i(v_{ic})$ gives the transit payment made by customer c to provider i when the aggregate traffic exchanged by the two networks is v_{ic}. $T_p(v_{pi})$ is the transit payment made by i to its provider p for the traffic volume v_{pi}. $R_i(v_{ir})$ is the cost of maintaining a peering link between i and its peer r when the corresponding traffic volume is v_{ir}. This fee is not paid by one peer to the other; rather, it represents costs to setup (amortized over time) or maintain that peering link. $L_i(v_i)$ determines local costs incurred by AN i (such as operations, staff, equipment) when the aggregate traffic handled by i is v_i.

In practice, transit prices show *economies of scale* meaning that the per-bit cost of Internet transit decreases as the volume of traffic increases. In ITER, we use concave increasing functions for transit, peering and local cost functions. Specifically, the pricing function of a transit provider p for traffic volume v is given by

$$T_p(v) = m_{t,p} * v^{e_t} \tag{18}$$

The exponent e_t controls the extent of the economies of scale associated with the various costs; a lower value of the exponent results in larger economies of scale. All transit providers have the same exponent e_t but they differ in the multipliers $m_{t,p}$. This is consistent with pricing data we collected from Norton [85] and Chang [25]. Similarly, peering costs are calculated as:

$$R_i(v_{ir}) = m_{r,i} * v_{ir}^{e_r} \tag{19}$$

while local costs also include a traffic-independent term l_i:

$$L_i(v_i) = l_i + m_{l,i} * v_i^{e_l} \tag{20}$$

All transit providers are assigned the same exponents for their peering and local cost functions, but they differ in the multipliers $m_{r,i}$, $m_{l,i}$, and in the l_i term.

To the extent possible, we parameterized the economic model using real-world data. Chang et al. [25] report that the exponent for the transit pricing functions e_t is around 0.75, while the peering exponent e_r is around 0.25. The transit price multipliers $m_{t,i}$ of STPs vary between [30,140], while those of LTPs vary between [80,150], *i.e.*, LTPs tend to be more expensive than STPs, but not always. These values are based on data reported by Norton [85] in 2006. The peering cost multipliers $m_{r,i}$ vary in [300,400]. The local cost exponent e_l is set to 0.5, while the local cost multipliers are set differently for STPs and LTPs: [100,200] for STPs and [300,400] for LTPs. The traffic-independent costs for LTPs are greater than those for STPs; this reflects that LTPs have larger networks, and hence larger operational costs. The local cost parameters are assigned so that the traffic-dependent and traffic-independent costs account for roughly equal fractions. The transit, peering and local cost parameters are assigned so that, for the same traffic volume, peering costs are the lowest, followed by traffic-dependent local costs, while transit costs are the highest.

4.2.6 Provider selection methods

The interdomain topology is formed when each AN selects its provider(s), and potentially its peers. In ITER, we consider three provider selection methods and three peer selection methods. Even though these methods are, to some degree, abstractions of a wide diversity of service agreements in the Internet, we believe that they capture the most common practices.

Regarding provider selection, an AN i first determines the set of candidate providers. These are transit providers (STPs or LTPs) that have at least one region in common with i and that are *not* in the customer tree of i. Then, i uses one of the following three methods to select the final provider (or set of providers, in case of multihoming):

Price-based (PR): The goal of i is to choose the cheapest provider(s). The metric used for comparing providers is the transit price multiplier $m_{t,j}$ associated with provider j.

Selective price-based (SEL): A transit provider i would not want to select as provider a network that may become its peer or customer in the future. In ITER, this implies that an STP would not want to select another STP as provider, and so it would choose only among

LTPs. Similarly, an LTP would not select an STP as provider, even if it is cheaper than LTP candidate providers. Among the remaining candidates, i would again select provider(s) based on price. SEL is applicable only to STPs and LTPs.

Performance-based (PF): A network may select providers based on the performance they offer. In ITER, we consider a performance metric that is related to the weighted path length from i to all sources and destinations of its traffic. This method is applicable only to ECs and CPs, not to transit providers (the latter would certainly not ignore pricing). For each destination j of i, let A_{ij} be the total traffic sent and received by i to/from j. Let l_{kj} be the path length from provider k to destination j. The performance metric associated with provider k is given by $L_i(k) = \sum_j A_{ij} l_{kj} / \sum_j A_{ij}$.

4.2.7 Multihoming

Multihoming, which refers to the practice of choosing multiple transit providers, is increasingly used, particularly by transit providers [39]. In ITER, AN i is assigned a *Maximum Multihoming Degree* (MMD), *i.e.*, a maximum number of providers, depending on its type. This upper bound is typically determined by the desired redundancy level. In practice, it may not be possible to always find MMD candidate providers. AN i ranks its set of candidate providers, based on one of the previous three selection methods, and selects up to MMD providers. In the default ITER model, we set the MMD to 1 for ECs, 3 for CPs, 2 for STPs and 3 for LTPs.

4.2.8 Peer selection methods

For any AN, the objective for peering is to save transit costs, by reducing the traffic volume that needs to be routed through providers. Further, peering is required in some cases to maintain reachability with the rest of the Internet. We consider three peer selection methods, modeling the most common approaches found in practice.

Peering by necessity (NC): With NC, networks i and j peer only if that is necessary to maintain global reachability; otherwise i will not be able to reach some of j's customers and vise versa. Neither AN can "force" the other to become its customer. Also, in some cases i and j would choose each other as provider based on their provider selection method.

When that is the case, they decide to peer instead.

Peering by traffic ratios (TR): A common approach for peering is to rely on "traffic ratios". Here, two ANs i and j agree to peer if they exchange "roughly equal" volumes of traffic. In practice, this is implemented by measuring the ratio of the traffic that flows from i to j and from j to i. If this ratio is close to one (within a factor of 2 in default ITER), the two ANs agree to peer.

Peering by cost-benefit analysis (CB): Here, AN i assesses both the costs associated with a given peering link and the potential benefits that can be achieved by that link. The costs associated with peering are due to the fixed and traffic-dependent costs of establishing a peering link. The benefits are due to reduced transit fees. AN i chooses to peer with j if the estimated benefits are greater than the estimated costs. In practice, i would need to estimate the "peerable traffic volume" with network j to use CB.

4.2.9 Initialization

We construct the initial internetwork so that it matches certain known properties of the Internet's interdomain topology. First, LTPs are assumed to be present in each geographical region and are fully-meshed with peering links. This is similar to the well-known clique of Tier-1 Internet providers. These are the only peering links in the initial topology. Regarding the initial customer-provider links, a recent study [39] measured the provider preference of different network types in the Internet and found that 60% of the providers of ECs are STPs while 40% are LTPs. On the other hand, approximately half of the providers of STPs and CPs are STPs. So, we connect STPs to other STPs and LTPs so that the number of links between STPs and LTPs is the same with that between STPs and STPs. To connect ECs and CPs, we follow a procedure that is similar to preferential attachment. We add ECs and CPs sequentially, choosing a provider (STP or LTP) with a probability that is proportional to the existing customer degree of that provider.

We define a *scenario* as a specification of the provider and peer selection strategies used by STPs and LTPs. In a scenario, we assume that *all providers belonging to the same class*

follow the same strategy. For example, the notation

$$\{DF, (SEL, TR), (SEL, NC)\}$$

represents a scenario with the default ITER model (DF), STPs use SEL provider selection and TR-peering, and LTPs use SEL-provider selection and NC-peering.

4.3 Solving the model

Our goal is to to "solve" the model, determining the internetwork that results as each AN changes its set of providers and peers to optimize a certain utility function. ANs play *sequentially*, and each AN i can observe how the actions of previous ANs affect i's traffic flow and economics.

4.3.1 AN actions

We present the steps used by an AN in each move.

1. **Provider selection:** First, an AN i identifies the set of preferred providers, according to its provider selection criteria. Let this set be P_i.

2. **Try to peer with providers:** If AN i does not engage in peering, skip to step 3. Else, i tries to convert each of its provider links to peering links. For this purpose, we evaluate the provider selection criteria of j, and find the set P_j. If $j \in P_i$ and $i \in P_j$, then i and j become peers "due to necessity". This condition captures the situation where i and j cannot agree on who should be the provider of whom. In this case, they need to peer to maintain global reachability for their customers. AN i then removes transit links to providers that are also in the customer tree of j. The intuition for this is as follows: When i and j form a peering link, some providers from P_i may be in the customer tree of j. i will never use such providers to reach nodes in the customer tree of j, since the direct path through the peering link is preferred. Figure 35 represents a case where i can safely remove providers k and l after forming a peering link with j. [3]

[3] A corner case can occur when i needs providers to reach ANs that are not in the customer tree of j, but all of i's providers are also in the customer tree of j. Rather than selecting arbitrarily which provider to keep, we impose the condition that *i keeps both k and l.*

3. **Check for potential peering candidates:** AN i maintains a list of possible peering candidates, R_i. As ECs do not peer in our model, the set of peering candidates of i is restricted to LTPs, STPs, and CPs that have a geographical region in common with i. For each possible peering candidate k, i performs the following actions: If k is already a peer of i, then i *unilaterally* verifies whether the peering requirements with i are satisfied. AN i also verifies if it needs to peer with k due to necessity. If these peering criteria are not satisfied, then i *de-peers* k and exits the peering loop. If i and k are not peers, then i examines whether it is possible to establish a new peering link with k. This is a bilateral decision, and hence the peering criteria of both i and k must be satisfied for a peering link to be created. If the peering link is formed, then i again executes the procedure for removing providers that are in the customer tree of k (see step 2). If the peering link is formed, i exits the peering loop. Note that in one move, i may add or remove only one peering link.

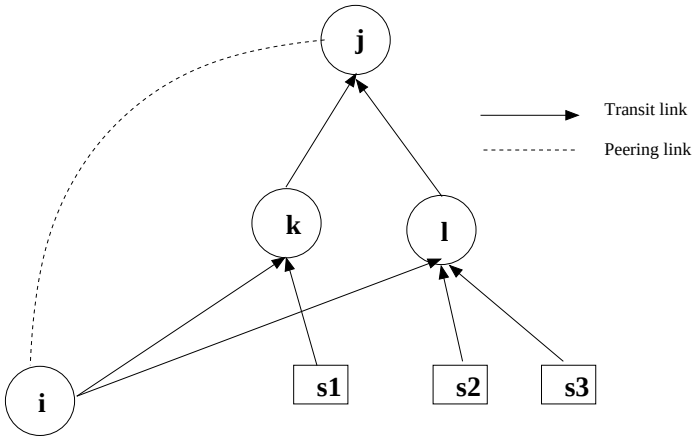

Figure 35: AN i can remove providers k and l after forming a peering link with provider j.

Note that all the actions performed by an AN in each move are *completely deterministic*. This is in contrast to previous evolutionary models of Internet topology (such as those based on preferential attachment [15]). Those models generate a random graph that has certain structural properties such as a desired degree distribution. The ITER model is not intended to be a topology generator. Instead, ITER models the optimizations performed by ANs, in

terms of selecting providers and peers. These optimizations are essentially deterministic in nature, as each AN attempts to unilaterally maximize its utility function.

4.3.2 Computing equilibrium

Our goal is to "solve" the ITER model, computing an equilibrium, given the initialization and the strategy of each AN. An equilibrium, if it exists, is a situation where no AN has the incentive to unilaterally change its set of providers or peers. We solve ITER computationally, as it is too complex to solve analytically. Solving ITER involves iteratively allowing an AN to play (according to its pre-defined strategy in each move), until we reach a stage where no AN has the incentive to change its connectivity. This state is analogous to the concept of Nash Equilibria (or pairwise stable equilibria when bilateral peering contracts are involved) in game theoretic models. We assume that nodes play in a particular sequence, with a randomly chosen starting node. We use the following procedure to compute the equilibrium for ITER.

1. Pick the next AN i in the playing sequence.
2. Complete the move of AN i, as described in section 4.3.1.
3. If the move of AN i causes the topology to change, recompute the routing tables, traffic flow and fitness function of each AN.
4. Check termination criteria. If each AN has had a chance to play and has not changed its connectivity, then stop.

An important issue is the time complexity involved in finding an equilibrium using agent-based approach described above. Figure 36 shows the simulation time [4] for the scenario {DF, (SEL,TR), (SEL,NC)} as we increase the number of ANs, keeping the relative proportions of different AN types fixed. We find that the running time of the model scales super-linearly with the number of ANs. The main reasons for this are the complexity of computing the interdomain traffic flow, and the number of iterations to reach equilibrium. As a result, it

[4]These simulations were run on a machine with with a 3GHz Intel Xeon processor and 2GB of memory.

is computationally infeasible to run the model at a scale larger than a few hundred ANs, particularly as we need to run multiple simulations to investigate a wide parameter space and different variations of the default model.

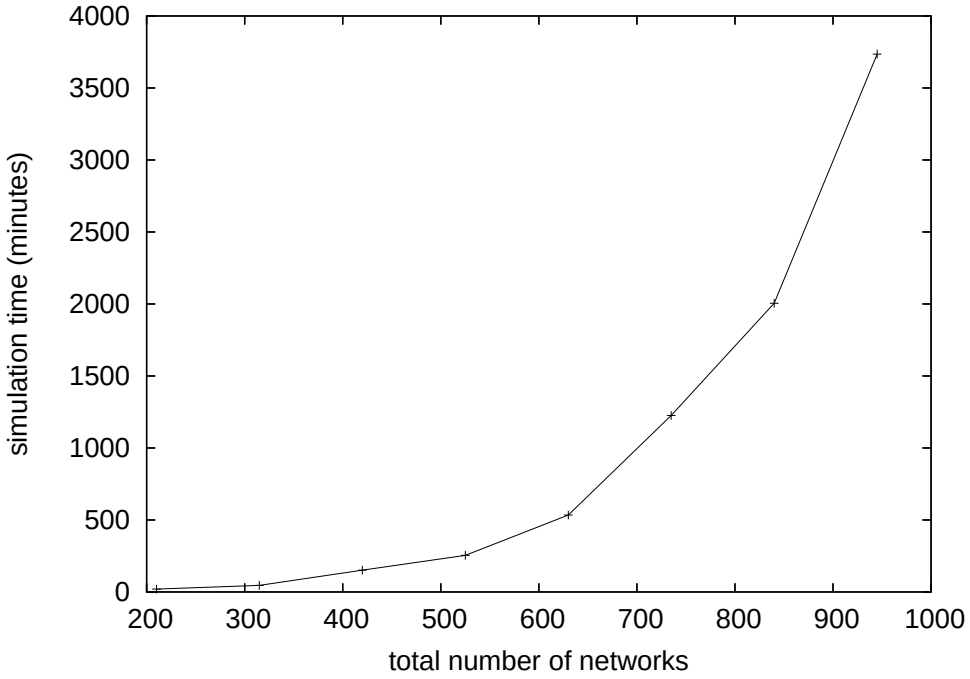

Figure 36: Simulation time to find an equilibrium vs. the number of ANs.

4.3.3 Existence of equilibrium

An important question is whether the agent-based simulation described in Section 4.3 is always able to find an equilibrium for ITER. We find empirically that in more than 95% of the simulation instances, we are able to solve ITER to find an an equilibrium. We find that 80% of the cases where we cannot find an equilibrium occur when STPs use CB-peering. In cases where we cannot find an equilibrium, the oscillation is caused by a small number of ANs, and *this oscillation is an expected outcome of the interaction between provider and peer selection, and traffic flow, and performance in the internetwork.* Next, we present some cases where ITER does not have an equilibrium, focusing on the fundamental reasons behind the oscillations.

In figure 37(a), AN 25 (a content provider) is connected to its preferred providers 1, 5

and 10, and the peering link with 12 does not exist. AN 25 uses its provider link to 10 to reach ANs in the customer tree of 10. When 12 uses CB-peering, it finds that peering with 25 leads to a higher fitness. This is because 25 now uses the (free) peering link with 12 to reach ANs in the customer tree of 10, due to which 12 earns revenues from 10. After the peering link between 25 and 12 is formed, 25 no longer needs 10 as a provider, and removes the link to provider 10. When 25 removes the provider link to 10, 12 no longer sees a benefit in peering with 25, and de-peers 25. As the peering link betwen 12 and 25 is removed, 25 is again able to choose its preferred providers, which includes AN 10. The above sequence then repeats. The fundamental factor that causes this oscillation is the interaction between provider and peer selection. An AN that creates a peering link with a provider does not need to retain providers that are in the customer tree of peers.

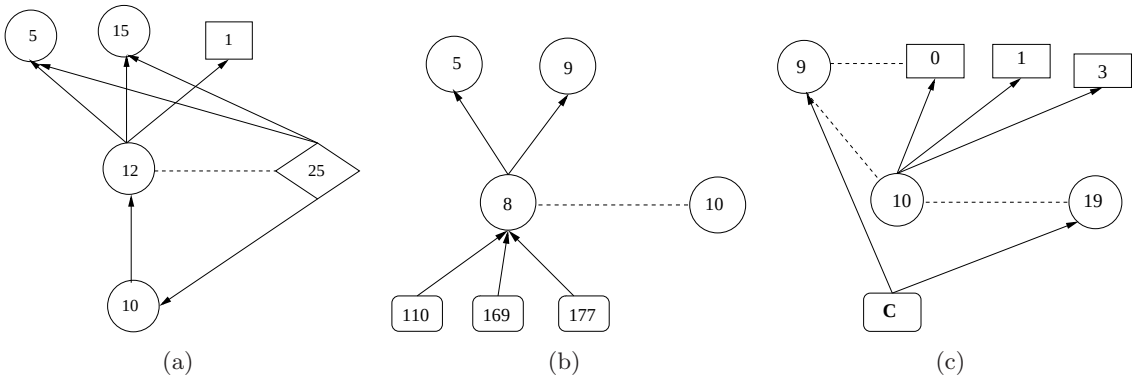

Figure 37: Examples of cases that lead to oscillations

In figure 37(b), AN 8 and 10 both use TR-peering. Content stubs 110, 169 and 177 use PF provider selection, and are initially not connected to 8. In this situation, the traffic ratio between 8 and 10 is balanced, and 8 is able to peer with 10. Due to this peering link, 8 obtains shortcut paths to nodes in the customer tree of 10, and becomes more attractive for content stubs 110,169 and 177 due to shorter weighted path lengths. These content stubs connect to 8 as customers. This affects the traffic flow between 8 and 10, whereby 8 sends more traffic 10 on the peering link. When 10 evaluates the peering link, it finds that the traffic ratios are no longer balanced. This causes 10 to de-peer 8. Consequently, 8 loses the advantage (attractiveness for performance-oriented customers) from peering with

10, and the content stubs 110, 169 and 177 no longer prefer to connect to 8 as provider. After the content stubs depart, the traffic ratio between 8 and 10 is again balanced, and 8 can peer with 10. The above sequence then repeats. The fundamental reason for this oscillation is that the creation of a peering link between two providers can improve (or harm) the weighted path lengths that either provider can offer to customers. The peering criterion (either traffic ratio or cost-benefit analysis) between the two peers could now fail as customers are attracted (or repelled) from this provider.

In a third example, the topology is as shown in figure 37(c). STPs 9 and 10 both use CB-peering, and initially, the peering links 9-10 and 10-19 are not present. Traffic from customers of 10 to the common customers of 9 and 19 (such as C) initially follows the path 10-0-9-C. Using CB-peering, STP 10 adds 19 as a peer, as both see a benefit. Now traffic from 10 to C flows over the peering link between 10 and 19 (path 10-19-C). This causes traffic to shift away from 9, leading to a loss of revenue. Using CB-analysis, STP 9 finds that creating a peering link with 10 would serve to bring traffic back to 9, leading to better fitness. Consequently, 9 and 10 form a peering link using CB-peering. Once the peering link between 9 and 10 is formed, 10 does not see a benefit in keeping the peering link to 19. After the link 10-19 is removed, 9 finds that it would achieve better fitness without the peering link with 10. Hence, 9 de-peers 10. The above sequence then repeats. The underlying reason for this oscillation is that the creation of a peering link alters the traffic flow, affecting the profitability of other networks and leading to the creation/removal of other peering links.

4.3.4 Uniqueness of equilibrium

An important issue is the uniqueness of the equilibrium that results from solving ITER using the method described in Section 4.3. We find that for a given initial topology and set of AN strategies, *the equilibria can depend on the order in which ANs make their moves*. In some cases ANs make the "right move at the right time", such as forming a particular peering link or choosing a certain provider, causing different equilibria. The presence of multiple equilibria is analogous to game theoretic models where the Nash equilibrium is

not unique. To account for this uncertainty, we run multiple simulations for a given initial topology and set of strategies by changing the order of play for ANs. We then study the expected value of the properties of the resulting equilibrium network. For example, the expected fitness for AN i is the fitness of AN i at equilibrium, averaged over a number of permutations with different orders of play.

4.4 Model validation

A major problem with any model that aims to capture, not only the interdomain topology, but also the economics and the traffic flow in the Internet, is how to validate it. ISPs are secretive about their economic and traffic data, while the ground truth for the Internet topology remains elusive (especially for peering links) [26]. In this section, we present a "best-effort" approach to validate ITER, comparing its predictions with known quantitative and qualitative characteristics of the Internet. These characteristics span topological properties, as well as some basic facts about Internet economics and distribution of traffic load. Clearly, however, the following results cannot be viewed as a definitive validation, given that other models may also be able to reproduce the same properties.

The following results are based on the following scenarios, {DF, (SEL,CB), (SEL,NC)}, {DF, (PR,TR), (SEL,NC)} and {DF, (PR,CB), (SEL,NC)}, which we view as the most common provider/peer selection scenarios in practice. The differences between these three scenarios, in terms of the following observations, are minimal.

Degree distribution: Figure 38 shows the complementary CDF (C-CDF) of the degree distribution for the scenario {DF, (SEL,CB), (SEL,NC)} with 945 networks. Even though it is not possible to be rigorous about the presence of a power-law in such a small scale, it is clear that the degree distribution is heavy-tailed. Of course this should not be surprising. In the default ITER, we set the multihoming degree of ECs and CPs to 1-3 providers, while STPs and LTPs can attract many customers at their regions, and so few of them will necessarily end up with large degrees. We also see the presence of networks with intermediate degrees, indicating that a single "attractor" network does not end up with all other networks as its customers or peers.

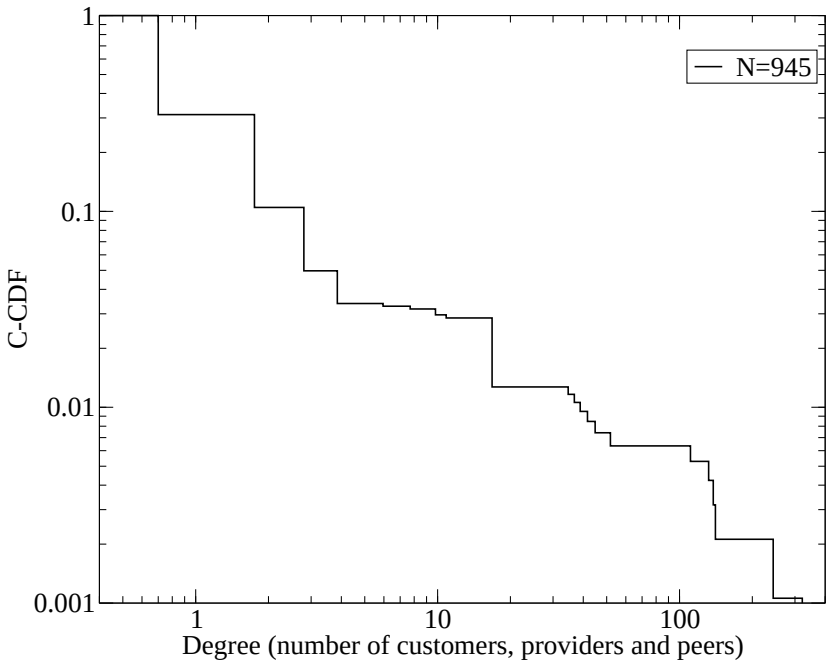

Figure 38: Degree distribution for an internetwork with 945 ANs {DF, SEL,CB), (SEL,NC)}.

Average path length: Another property of the Internet is that the average path length, in terms of AS links, has remained almost constant (at about 4 AS hops) during the last decade [72, 39]. We have reproduced the same behavior in ITER. Figure 39 shows the average path length in the network for the scenario {DF, (SEL,CB), (SEL,NC)} as the number of ANs is increased from 210 to 945. We find that the average path length between any two ANs remains close to 4 hops (with a variation range between 3 to 5 hops, which also does not vary with the size of the internetwork).

Economic structure of transit market: We also examine the profitability of transit providers in the resulting ITER internetwork. We find that a significant fraction of STPs and LTPs fail to attract enough customers, and so they end up with negative "profits". In an evolutionary version of ITER, these ANs would be removed as bankrupt or "dead", similar to what often happens in the real Internet. On the other hand, there are several profitable STPs and LTPs, meaning that the ITER internetwork does not converge to a monopoly or oligopoly. This is in agreement with a recent measurement study [39] which

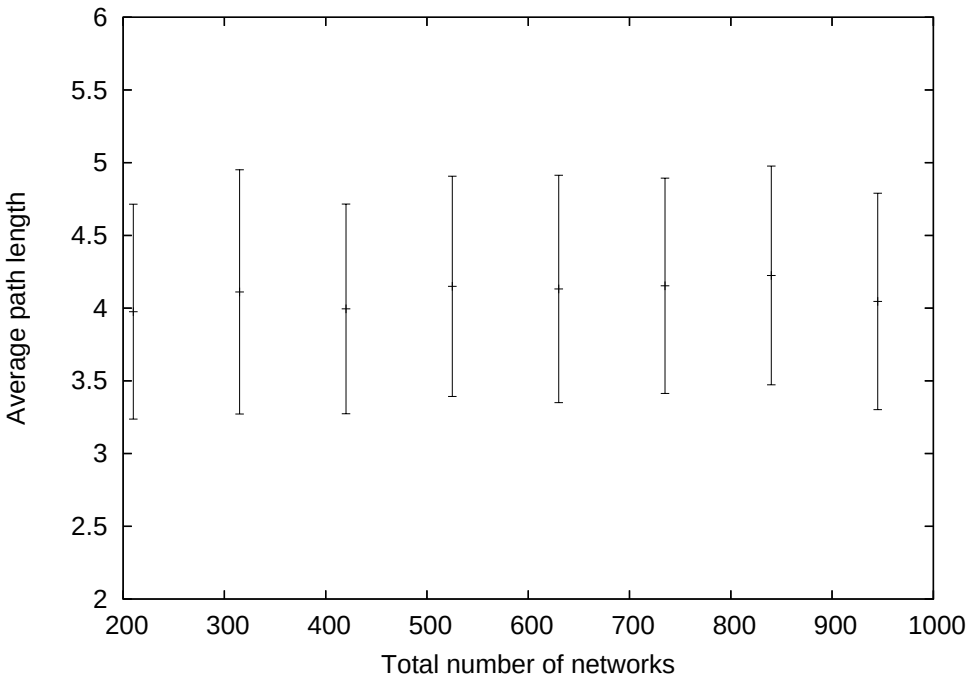

Figure 39: Average path length as the number of ANs is increased for scenario {DF, (SEL,CB), (SEL,NC)}.

showed that the number of transit providers that are active (meaning that they attract customers) is significant, indicating that the Internet transit market is not heading towards a monopoly or oligopoly.

Distribution of link load: We also measure the traffic volume carried by each link in the ITER internetwork. Figure 40 shows the C-CDF of the link loads on each interdomain link for the scenario {DF, (PR,TR), (SEL,NC)}. Most links carry small traffic loads; these are links mostly at ECs and CPs at the edge of the Internet. On the other hand, there are few links that carry very large traffic volumes; these are customer-provider and peering links between transit providers. Akella et al [8] observed a qualitatively similar phenomenon in the Internet. They reported that links between transit providers high in the hierarchy are typically of higher capacity than those between providers lower in the hierarchy.

4.5 The default model

In the rest of this paper, our goal is to understand the impact of different provider/peer selection methods on the topology, traffic flow, economics and performance of the resulting

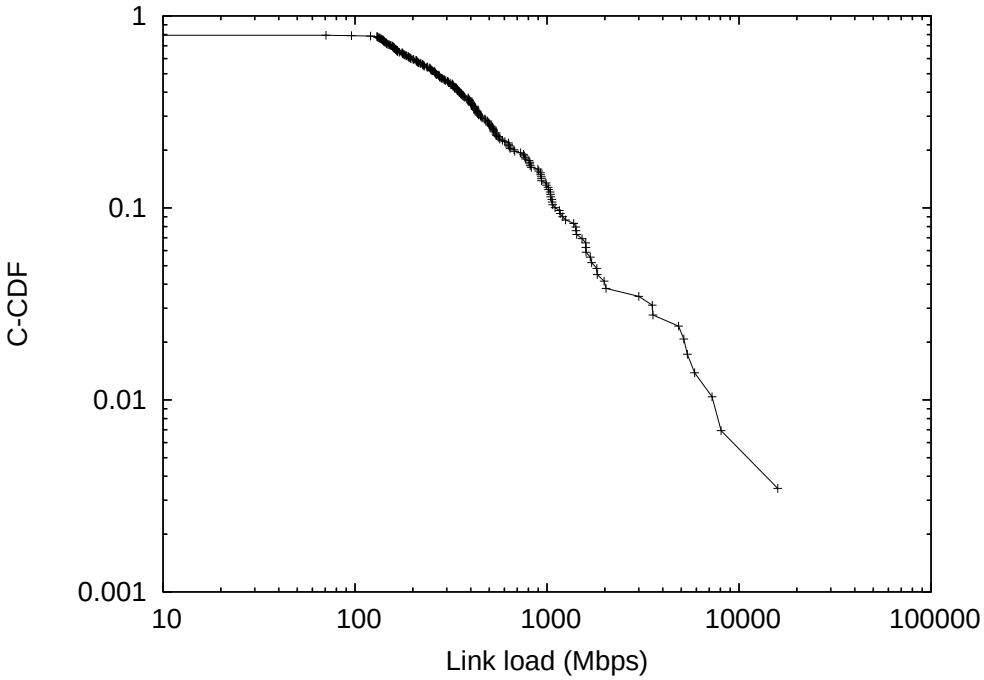

Figure 40: C-CDF of traffic volume on each link for scenario {DF, (PR,TR), (SEL,NC)}.

internetwork. In this section, we focus on the Default ITER model. In the following three sections, we consider a number of deviations from the Default model, in terms of the traffic matrix, the edge network provider preferences and the geographical presence of CPs.

In the default ITER model, ECs use PR provider selection and they do not peer with other ANs. CPs also use PR provider selection, but they peer using the CB method. For STPs, as well as for LTPs, we consider two provider selection methods, PR and SEL, and three peer selection methods: NC, TR and CB. All ANs of the same type choose the same provider and peer selection method. This agrees with what we see in the Peering Database [1], for instance, where networks of the same business function and size tend to use the same type of peering policy.

An *ITER scenario* refers to the selection of a specific pair of provider and peer selection methods for STPs and of another such pair for LTPs. Since we have 6 provider/peer combinations for STPs and 6 identical combinations for LTPs, the total number of scenarios we need to consider is 36. Table 2 shows the output metrics for each of these 36 scenarios in the default ITER model. For each scenario, we run 20 ITER simulations. In each

simulation, we use a different random permutation of the sequence in which ANs move during the ITER transient phase.

We measure several metrics that characterize the equilibrium network: The average path length between each pair of ANs (unweighted as well as weighted by the traffic that flows between those ANs), the aggregate fitness of STPs and LTPs, the number of fit STPs and LTPs, the fraction of peering links and the fraction of total traffic that flows over peering links. The results in Table 2 are averaged over that subset of the 20 runs in which ITER converged to a stable internetwork. The standard error for each metric is also shown. We compare various scenarios only when the corresponding confidence intervals are non-overlapping.

4.5.1 Path Lengths

We report the weighted path length (column "wPL" in Table 2) and the unweighted path length (column "uPL" in Table 2) for each scenario of the Default model. Note that the average path length in the resulting internetwork is close to 4 hops for all scenarios except when LTPs use CB; paths tend to be longer when LTPs use CB-peering. In particular, the scenario {{DF, (PR,NC), (SEL,CB)} results in average path length of 4.2, compared to 3.9 in other scenarios. When LTPs peer with CPs the traffic from CPs goes through peering links to LTPs, and from there to ECs potentially through one or more STPs. Figure 41 illustrates this case for scenario {DF, (PR,NC), (SEL,CB)}. We see paths of the following nature: LTPs peer with several CPs using the CB method. The path from these CPs to destination ECs (which are customers of say STP A) is of the form CP-LTP-STP_B-STP_A-EC. If LTPs do not use CB, they will not form peering links with CPs (TR peering would not work because CPs always generate much more traffic than they consume). The CPs would then probably choose STPs as providers, as they tend to be less expensive than LTPs. This leads to paths of the form CP-STP_A-EC or CP-STP_A-STP_B-EC that are shorter than the path observed when the LTP peers with CPs. Also, these longer paths are from the major sources of traffic (CPs) to their destinations (ECs). So, the weighted path length (4.2) is longer than the unweighted path length (4.0).

Table 2: Output metrics for the default model (DF), averaged over 20 simulation runs.

STP str	LTP str	wPL	uPL	dia	prof a-STP ($k)	prof a-LTP ($k)	prof f-STP ($k)	prof f-LTP ($k)	num f-STP	num f-LTP	num UA	Traf UA	%PP	Traf PP
PR,NC	PR,NC	3.9	3.9	6.1	331	409	446	527	4.4	1.6	1.6	0.1	2.3	0.1
PR,NC	PR,TR	3.9	3.9	6.1	331	409	446	527	4.4	1.6	1.6	0.1	2.3	0.1
PR,NC	PR,CB	4.2	4.0	6.6	40	439	180	521	4.0	2.0	2.1	0.2	6.3	0.4
PR,NC	SEL,NC	3.9	3.9	6.1	355	368	465	495	4.8	1.5	1.4	0.0	2.9	0.1
PR,NC	SEL,TR	3.9	3.9	6.1	355	368	465	495	4.8	1.5	1.4	0.0	2.9	0.1
PR,NC	SEL,CB	4.2	4.0	6.6	39	441	179	523	3.9	2.0	2.2	0.2	6.3	0.4
PR,TR	PR,NC	3.9	3.9	6.2	335	393	451	504	4.5	1.7	1.6	0.0	2.5	0.1
PR,TR	PR,TR	3.9	3.9	5.9	317	426	433	544	4.2	1.6	1.6	0.0	2.2	0.2
PR,TR	PR,CB	4.1	3.9	6.3	55	458	200	545	3.0	2.1	1.9	0.2	6.0	0.5
PR,TR	SEL,NC	3.9	3.9	6.2	347	369	459	491	4.9	1.5	1.4	0.0	3.0	0.1
PR,TR	SEL,TR	3.9	3.9	6.0	301	431	416	546	4.3	1.7	1.4	0.0	3.1	0.2
PR,TR	SEL,CB	4.1	3.9	6.3	24	480	173	554	3.1	2.2	2.0	0.2	6.7	0.4
PR,CB	PR,NC	3.9	3.9	6.0	333	392	445	502	4.5	1.7	1.5	0.0	3.3	0.2
PR,CB	PR,TR	3.9	3.9	5.8	229	498	344	602	4.1	1.9	1.0	0.0	3.4	0.2
PR,CB	PR,CB	3.9	3.9	6.0	63	472	209	538	2.5	2.5	1.4	0.2	7.4	0.5
PR,CB	SEL,NC	3.9	3.9	6.0	243	471	352	576	4.6	1.9	1.3	0.0	4.0	0.2
PR,CB	SEL,TR	3.9	3.9	5.9	226	501	340	605	4.2	1.9	1.0	0.0	3.9	0.2
PR,CB	SEL,CB	3.9	3.9	5.9	33	492	183	551	2.4	2.5	1.7	0.1	8.4	0.5
SEL,NC	PR,NC	3.9	3.9	5.0	-48	851	92	951	3.0	2.0	1.0	0.0	2.1	0.0
SEL,NC	PR,TR	3.9	3.9	5.0	-48	851	92	951	3.0	2.0	1.0	0.0	2.1	0.0
SEL,NC	PR,CB	4.0	3.9	5.0	-185	787	2	873	1.0	2.0	3.0	0.5	5.9	0.3
SEL,NC	SEL,NC	3.9	3.9	5.0	-48	851	92	951	3.0	2.0	1.0	0.0	2.1	0.0
SEL,NC	SEL,TR	3.9	3.9	5.0	-48	851	92	951	3.0	2.0	1.0	0.0	2.1	0.0
SEL,NC	SEL,CB	4.0	3.9	5.0	-185	787	2	873	1.0	2.0	3.0	0.5	5.9	0.3
SEL,TR	PR,NC	3.9	3.9	5.0	-48	851	92	951	3.0	2.0	1.0	0.0	2.1	0.0
SEL,TR	PR,TR	3.9	3.9	5.0	-6	799	134	899	2.9	2.0	1.1	0.0	2.0	0.1
SEL,TR	PR,CB	3.9	3.9	5.0	-104	701	83	772	1.2	2.3	2.5	0.4	5.7	0.4
SEL,TR	SEL,NC	3.9	3.9	5.0	-48	851	92	951	3.0	2.0	1.0	0.0	2.1	0.0
SEL,TR	SEL,TR	3.9	3.9	5.0	-10	806	129	906	3.0	2.0	0.9	0.0	2.3	0.1
SEL,TR	SEL,CB	3.9	3.9	5.0	-110	709	76	782	1.0	2.3	2.7	0.5	6.1	0.4
SEL,CB	PR,NC	3.8	3.9	5.0	65	734	182	834	4.0	2.0	1.0	0.0	3.6	0.1
SEL,CB	PR,TR	3.8	3.9	5.0	113	680	233	776	3.7	2.0	1.3	0.0	3.1	0.2
SEL,CB	PR,CB	3.9	3.9	5.0	-40	597	122	653	2.0	2.7	1.4	0.2	6.2	0.5
SEL,CB	SEL,NC	3.8	3.9	5.0	65	734	182	834	4.0	2.0	1.0	0.0	3.6	0.1
SEL,CB	SEL,TR	3.8	3.9	5.0	115	680	231	780	3.8	2.0	1.3	0.0	3.7	0.2
SEL,CB	SEL,CB	3.9	3.9	5.0	-46	622	115	687	2.0	2.4	1.6	0.2	6.7	0.5
standard error		0.02	0.01	0.07	24	29	23	28	0.13	0.07	0.14	0.01	0.13	0.02

4.5.2 Peering links

As expected, we see that there is a positive correlation between the percentage of peering links ("%PP") and the fraction of the total traffic flow that traverses at least one peering link ("Traf-PP"). Both of these metrics are maximized when STPs and LTPs both use CB-peering. In those scenarios, 6-8% of all links are PP links, and 50% of the total end-to-end traffic flows over those links. In those scenarios, both STPs and LTPs are able to peer with CPs. The large traffic volume from CPs to ECs now flows through those peering links.

4.5.3 "Unprofitable-but-Active" (UA) providers

We evaluate a metric that measures the long-term economic stability of the resulting internetwork. Some transit providers attract customers due to either lower prices or better performance, but are not profitable because their local and transit costs are higher than their revenues. Such an economic situation would not be sustainable in the long-term, as these providers would either go bankrupt or they would have to increase their prices. We measure the number of providers that are Unprofitable-but-Active ("num UA" in Table 2). We also measure the maximum traffic volume (as a fraction of the total traffic flow) carried by UA transit providers ("Traf UA"). First, note that the two metrics are positively correlated: a larger number of UA providers results in a larger traffic volume handled by UA providers. Second, we have more UA providers when LTPs peer with CPs (see, for example, scenario {DF, (PR,NC), (SEL,CB)}). In that scenario, traffic from CPs to ECs flows through a hierarchy of STPs. STPs at the top of the hierarchy can become UA providers, as they pay large transit fees to the generally more expensive LTPs. The largest number of UA providers results when STPs use SEL provider selection and peer using either NC or TR (up to 3 UA providers, and 50% of the total traffic carried by those providers). Then, STPs cannot peer with CPs and they also choose only LTPs as providers. All traffic from CPs to ECs flows through customer-provider links between STPs and LTPs, and this creates even more UA STPs.

4.5.4 Provider profitability when STPs use PR:

LTPs can harm STP profitability by peering with CPs:

When most edge networks choose providers based on price, cheaper STPs are able to attract a large fraction of the edge networks. Due to the overlapping prices of STPs and LTPs, however, LTPs can also attract some edge networks as customers. When STPs use PR provider selection, a hierarchy of STPs is formed. When LTPs use PR as well, they may be forced to connect to STP providers, and the peering clique of LTPs may no longer be sustainable. In such situations, both STPs and LTPs see approximately equal aggregate fitness. LTPs can, however, significantly harm the aggregate profits of STPs by using CB-peering, which allows them to peer with CPs. In this case some LTPs are able to form a large number of peering links with CPs. Consequently, these CPs reach most of their destinations through peering links with LTPs, followed by a hierarchy of STPs. In the default model, CPs source a large fraction of the traffic that goes to ECs. Consequently, the LTPs can significantly reduce the fitness of STPs by engaging in CB-peering. The conventional wisdom for LTPs is to only peer with other LTPs. This result shows that *CB-peering by LTPs can lead to a situation where LTPs significantly reduce STP profits, and are able to increase aggregate LTP fitness.*

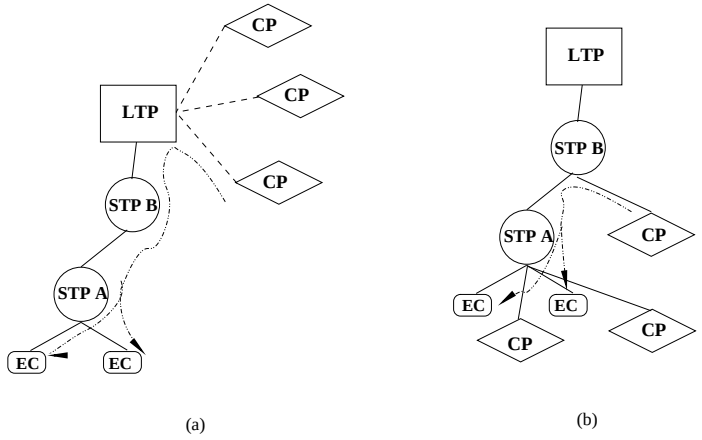

Figure 41: Peering between LTPs and CPs increases LTP profitability, but also increases weighted path lengths. The arrows indicate the paths followed by large traffic flows.

STPs should use TR-peering:

We find that the best peering strategy for STPs depends on the peering method used by LTPs. First consider the case where LTPs do not peer with CPs. In this scenario, we find that STPs achieve higher aggregate fitness by using TR-peering (though the total number of fit STPs is smaller). For example, the aggregate STP fitness in scenario {DF, (PR,TR), (PR,TR)} is \$317k, while it is \$229k in scenario {DF, (PR,CB), (PR,TR)}. This indicates that the conventional wisdom of TR-peering results in higher STP fitness, *when STPs use price-based provider selection*. The reason for this is as follows. If STPs use CB-peering, then some CPs become their peers. On the other hand, if STPs use TR-peering, they cannot peer with CPs, as CPs always generate more traffic than they consume. These CPs would eventually become customers of STPs, as most edge networks choose providers based on price. This increases the fitness of STPs. On the other hand, when STPs use CB-peering, they can peer with CPs. In this case the traffic flow is of the form CP-STP-EC; less traffic flows on the customer-provider links in the hierarchy of STPs, leading to lower aggregate fitness for STPs.

Next, consider the case where LTPs use CB-peering. In this case, STPs are more profitable by using CB-peering than with TR-peering. If CPs peer with LTPs, then they do not need to choose STPs as providers. Given that these CPs will not become their customers, STPs can improve their profitability by peering with them. This can happen only if STPs use CB-peering.

4.5.5 Provider profitability when STPs use SEL

STP fitness is determined by LTP prices:

In this scenario, STPs do not choose other STPs as providers, because they consider them as potential peers or competitors. STPs still attract the price-conscious ECs and CPs. All STPs connect directly to LTPs due to SEL provider selection. This results in higher fitness for LTPs than the scenarios where STPs use PR provider selection. In case STPs peer only by necessity, the aggregate STP fitness can be negative, and there are no fit STPs. As these STPs carry traffic to/from their customers, we see a larger number of UA providers, and a larger fraction of traffic flowing through such providers. The aggregate fitness of STPs

depends on the relative prices of STPs and LTPs. In our simulation setting, LTP prices are slightly higher than those of STPs, leading to a situation where STPs pay more in transit prices than they can recover from their customers. If LTP and STP prices are comparable, the aggregate STP profit can still be positive. The key point is that *if STPs use SEL provider selection, LTPs are in a position to use their market power to charge higher prices, and potentially make STPs unprofitable.*

STPs should use CB-peering:

When STPs use SEL provider selection, they achieve higher profits using CB-peering than TR-peering (*e.g.*, the aggregate STP profit is \$65k in scenario {DF, (SEL,CB), (PR,NC)}, while it is \$-48k in {DF, (SEL,TR), (PR,NC)}). This is in contrast to the case where STPs use PR provider selection, where they are better off using TR-peering. The reason for this is as follows. When STPs use SEL provider selection, they only connect to LTPs. Due to the higher prices of LTPs, it is beneficial for STPs to send as little traffic as possible to their upstream providers. If an STP S peers with a CP C, S only carries traffic destined from C to ECs in the customer tree of S. S does not send any of this traffic to its providers, and this traffic is profit-generating. This allows the STP to remain profitable even if LTPs charge high transit prices. A further benefit of CB-peering is that it allows "content-heavy" and "access-heavy" STPs to peer. Content-heavy STPs have many CPs as customers, while access-heavy STPs have many ECs as customers. These two types of STPs can peer only with CB-peering (traffic ratios will always be unbalanced), and results in increased fitness for STPs. This makes the case that *content and access heavy STPs should peer with each other to be profitable.*

4.6 Deviation-1: P2P Traffic matrix

In the default model, the interdomain traffic matrix consists mostly of CS traffic (80%). In this section, we consider a deviation where the traffic matrix consists mostly (80%) of P2P traffic. Edge networks still choose their providers using PR, as in the default model. We call this deviation "P2P". The tables with the detailed results for P2P and subsequent deviations are in the appendix.

Peer-to-peer traffic helps STPs:

In the default model, most edge networks choose providers based on price. In Section 4.5, we observed that LTPs can significantly diminish the aggregate profit of STPs by using CB-peering. When the traffic matrix consists mostly of P2P traffic, the traffic volume from CPs to ECs is relatively smaller. As a result, the benefit for LTPs from peering with CPs is lower. The aggregate fitness of STPs is \$187k with scenario {P2P, (PR,NC), (PR,CB)}, while it is \$40k for the scenario {DF, (PR,NC), (PR,CB)}. *A traffic matrix that consists of mostly P2P traffic thus benefits STPs.*

Smaller increase in weighted path lengths when LTPs peer with CPs:

In the default model, we observed that when LTPs peer with CPs, weighted path lengths are longer than unweighted path lengths. For example, the weighted path length is 4.2 in scenario {DF, (PR,NC), (PR,CB)}, while the unweighted path length is 4.0. For scenario {P2P, (PR,NC), (PR,CB)} the weighted path length is 4.1, while the unweighted path length is 4.0, *i.e.*, we observe a similar phenomenon with the P2P traffic matrix, though the difference between the weighted and unweighted path lengths is smaller. When the traffic matrix is predominantly P2P, the volume of traffic flowing from CPs to ECs (over the long paths caused when LTPs peer with CPs) is smaller than in the default model.

TR-peering is more profitable for STPs:

In the default model, when STPs use SEL provider selection, we found that either NC or TR-peering led to negative aggregate STP fitness (*e.g.*, aggregate STP profit is \$-104k in {DF, (SEL,TR), (PR,CB)}). This is because the traffic matrix has mostly CS traffic, and only a small fraction of the traffic flows between ECs (which are customers of STPs). Consequently, peering by STPs does not give significant benefit. With the P2P traffic matrix, however, a larger fraction of the end-to-end traffic flows between ECs. STPs can save significant transit fees if they use TR-peering (aggregate STP profit is \$58k in {P2P, (SEL,TR), (PR,CB)}). The likelihood of STPs being able to peer using TR-peering depends also on the peering strategy of LTPs; in particular, whether LTPs peer with CPs. We illustrate this with a specific example in Figure 42. In subfigure (a), the LTP peers with CPs. The traffic between STP A and STP B is now balanced, allowing them to peer using

TR-peering. Subfigure (b) shows the case where LTPs do not peer with CPs. These CPs become customers of STPs, which are cheaper than LTPs. This can lead to the emergence of "content-heavy" STPs (STPs with content customers) and "access-heavy" STPs (STPs with access customers). Content and access heavy STPs cannot peer with each other using TR-peering, as the traffic is always imbalanced (more traffic from CPs to ECs). Thus, *if the traffic matrix consists of mostly P2P traffic, then STPs can save significant transit costs with TR-peering.* Further, *peering between LTPs and CPs favors STPs, as it results in more balanced traffic between STPs, giving them more opportunities to peer.*

Traffic flow over UA providers:

In the P2P model, the traffic flow through UA providers is reduced, particularly when STPs use PR provider selection. In the default model, if LTPs peer with CPs, a number of STPs become unprofitable. As stated earlier, P2P traffic helps STPs, and the ability of LTPs to decrease aggregate STP fitness is reduced. This leads to a smaller number of STPs that are "unprofitable but active". In particular, for scenario {DF, (PR,NC), (PR,CB)}, 20% of the end-to-end traffic flows over UA providers, while for scenario {P2P, (PR,NC), (PR,CB)}, this value is around 4%.

An exception to the above result is when STPs use (SEL,NC). In this case, STPs connect directly to LTPs, and do not peer with other STPs. With P2P traffic, a large amount of traffic flows from ECs to other ECs. When STPs use SEL provider selection, this traffic traverses the customer-provider links from STPs to LTPs. This results in a larger number of UA providers and a larger fraction of traffic handled by those UA providers; 30% of the total traffic flows over UA providers in {P2P, (SEL,NC), (PR,NC)}, while that number is close to 0 for {DF, (SEL,NC), (PR,NC)}.

4.7 Deviation-2: PF provider selection by edge networks

In the default model, 80% of ECs and CPs use PR provider selection. In this section, we consider a deviation where 80% of edge networks choose their providers using the PF method described in Section 4.2.6. We call this deviation "EP". The interdomain traffic matrix still consists of mostly (80%) CS traffic, as in the default model.

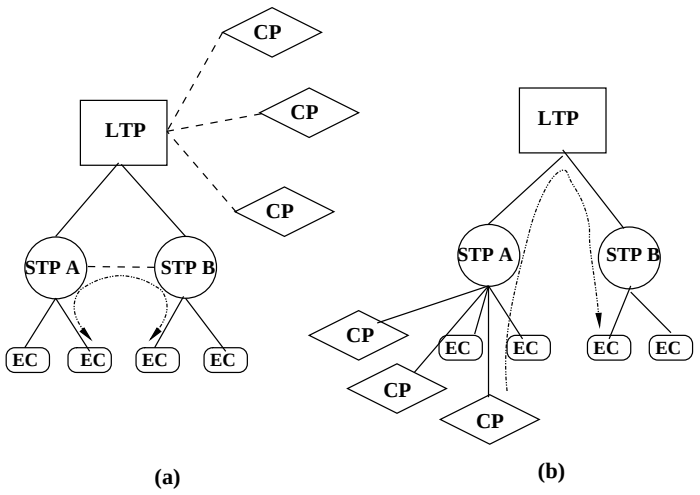

Figure 42: Peering between STPs more likely with P2P traffic and especially when LTPs peer with CPs. The arrows indicate the paths followed by large traffic flows.

PF provider selection favors LTPs:

ECs and CPs that use PF provider selection are attracted to LTPs, and eventually, most ECs and CPs connect directly to LTPs. This is because LTPs can reach all destinations using links to their customers or peers, and so they provide the shortest paths. STPs can only attract the few ECs and CPs that use PR-provider selection. When STPs use SEL provider selection, there are no fit STPs (aggregate STP fitness is negative). Further, when STPs use SEL, no form of peering leads to positive aggregate fitness. In the default model, when STPs use SEL, they can be profitable by CB-peering. In the default model, STPs have a large customer base of ECs and CPs, and peering with CPs (or "content-heavy" STPs) can save significant transit expenses. In the EP model, however, STPs have a significantly smaller customer base. Consequently, peering does not increase the aggregate fitness of STPs.

Shorter paths:

When ECs and CPs use PF provider selection, the average path lengths in the network decrease. This is because networks at the edge are attracted to LTPs. End-to-end paths are of the form EC-LTP-EC or EC-LTP-LTP-EC, with no intermediate STPs. This results in shorter end-to-end paths. The unweighted path length is 3.3 for scenario {EP, (SEL,NC), (PR,NC)}, as opposed to 3.9 for scenario {DF, (SEL,NC), (PR,NC)}. *This implies that the*

Table 3: Output metrics for Deviation-1 (P2P), averaged over 20 simulation runs.

STP str	LTP str	wPL	uPL	dia	prof a-STP ($k)	prof a-LTP ($k)	prof f-STP ($k)	prof f-LTP ($k)	num f-STP	num f-LTP	num UA	Traf UA	%PP	Traf PP
PR,NC	PR,NC	4.0	3.9	6.2	266	419	390	540	3.8	1.5	2.0	0.0	2.4	0.1
PR,NC	PR,TR	4.0	3.9	6.2	266	419	390	540	3.8	1.5	2.0	0.0	2.4	0.1
PR,NC	PR,CB	4.1	4.0	6.5	187	405	312	512	4.0	1.6	2.0	0.0	6.1	0.2
PR,NC	SEL,NC	4.0	3.9	6.1	295	388	415	515	4.0	1.4	1.9	0.0	2.9	0.2
PR,NC	SEL,TR	4.0	3.9	6.1	295	388	415	515	4.0	1.4	1.9	0.0	2.9	0.2
PR,NC	SEL,CB	4.1	4.0	6.5	187	405	312	512	4.0	1.6	2.0	0.0	6.1	0.2
PR,TR	PR,NC	4.0	3.9	6.0	295	381	416	508	4.1	1.4	1.9	0.0	2.7	0.2
PR,TR	PR,TR	4.0	3.9	6.1	384	284	506	407	4.0	1.4	2.2	0.0	2.1	0.2
PR,TR	PR,CB	4.1	4.0	6.3	206	404	335	515	3.6	1.5	1.7	0.1	5.2	0.3
PR,TR	SEL,NC	4.0	3.9	5.9	303	373	421	502	4.1	1.4	1.7	0.0	3.2	0.2
PR,TR	SEL,TR	4.0	3.9	6.2	387	273	506	403	4.2	1.2	2.0	0.0	3.2	0.3
PR,TR	SEL,CB	4.1	3.9	6.1	219	395	347	505	3.6	1.5	2.2	0.0	6.1	0.3
PR,CB	PR,NC	4.0	3.9	6.1	379	293	500	421	4.1	1.3	2.3	0.0	2.6	0.2
PR,CB	PR,TR	4.0	3.9	5.9	331	365	455	482	3.8	1.4	2.3	0.0	2.3	0.2
PR,CB	PR,CB	4.0	3.9	6.2	185	433	320	503	3.0	2.2	1.1	0.1	6.7	0.4
PR,CB	SEL,NC	4.0	3.9	6.2	353	311	472	440	4.4	1.3	1.9	0.0	3.7	0.2
PR,CB	SEL,TR	4.0	3.9	5.8	320	356	439	479	4.1	1.3	1.9	0.0	3.4	0.3
PR,CB	SEL,CB	4.0	3.9	6.0	185	436	316	509	3.1	2.2	1.1	0.0	7.5	0.5
SEL,NC	PR,NC	4.1	3.9	5.0	-85	863	59	963	2.0	2.0	2.0	0.3	2.1	0.0
SEL,NC	PR,TR	4.1	3.9	5.0	-85	863	59	963	2.0	2.0	2.0	0.3	2.1	0.0
SEL,NC	PR,CB	4.1	3.9	5.0	-127	835	31	945	2.0	1.6	3.0	0.3	5.0	0.1
SEL,NC	SEL,NC	4.1	3.9	5.0	-85	863	59	963	2.0	2.0	2.0	0.3	2.1	0.0
SEL,NC	SEL,TR	4.1	3.9	5.0	-85	863	59	963	2.0	2.0	2.0	0.3	2.1	0.0
SEL,NC	SEL,CB	4.1	3.9	5.0	-127	835	31	945	2.0	1.6	3.0	0.3	5.0	0.1
SEL,TR	PR,NC	4.0	3.9	5.0	-11	784	125	884	3.0	2.0	1.0	0.0	2.5	0.1
SEL,TR	PR,TR	4.0	3.9	5.0	29	742	163	845	2.9	1.9	1.1	0.0	2.3	0.1
SEL,TR	PR,CB	4.0	3.9	5.0	58	623	199	711	2.8	2.0	1.6	0.0	5.3	0.3
SEL,TR	SEL,NC	4.0	3.9	5.0	-11	784	125	884	3.0	2.0	1.0	0.0	2.5	0.1
SEL,TR	SEL,TR	4.0	3.9	5.0	20	751	155	853	3.0	1.9	1.0	0.0	2.6	0.1
SEL,TR	SEL,CB	4.0	3.9	5.0	58	621	192	712	3.0	2.0	1.7	0.0	6.0	0.4
SEL,CB	PR,NC	4.0	3.9	5.0	1	770	126	870	3.0	2.0	2.0	0.0	3.6	0.1
SEL,CB	PR,TR	4.0	3.9	5.1	66	705	196	809	2.8	1.8	2.3	0.1	2.8	0.2
SEL,CB	PR,CB	4.0	3.9	5.1	92	578	228	652	2.6	2.1	1.6	0.1	5.7	0.4
SEL,CB	SEL,NC	4.0	3.9	5.0	1	770	126	870	3.0	2.0	2.0	0.0	3.6	0.1
SEL,CB	SEL,TR	4.0	3.9	5.0	56	717	183	821	2.8	1.9	2.4	0.1	3.6	0.2
SEL,CB	SEL,CB	4.0	3.9	5.0	91	588	227	666	2.6	1.8	2.0	0.1	6.5	0.4

EP model leads to a situation that is beneficial for the performance seen by edge networks, at the expense of STP profitability.

Weighted paths shorter than unweighted paths:

In scenarios where STPs use (PR,NC), the weighted path lengths are *smaller* than the corresponding unweighted path lengths. For example, the weighted path length for scenario {EP, (PR,NC), (PR,NC)} is 3.3, while the unweighted path length is 3.5. This can be explained as follows: Networks at the edge (ECs and CPs) choose providers based on performance, and are thus attracted to LTPs. STPs, on the other hand, use PR provider selection and connect to other STPs. This creates a hierarchy of STPs, but with *a very small customer base connected to STPs.* Consequently, we see paths that traverse the hierarchy

of STPs and LTPs, but only a small fraction of the total traffic flows on those paths. Most of the traffic flows on the short paths of the form EC-LTP-EC. As a result, *the weighted path length is smaller than the corresponding unweighted path length.*

More traffic over UA providers:

In the default model, significant traffic is carried by UA providers when STPs use PR provider selection and LTPs peer with CPs, *e.g.*, 20% of the total end-to-end traffic is carried by UA providers in scenario {DF, (PR,TR), (PR,CB)}. In other scenarios of the default model, negligible traffic is carried by UA providers. In the EP model, however, *10-20% of the total traffic is carried by UA providers in each scenario.* The reason for this is as follows. When edge networks use PF-provider selection, only a few ECs and CPs connect to STPs. Most of the traffic sourced/consumed by these customers of STPs comes from CPs and ECs that are connected to LTPs. Consequently, STPs send/receive large traffic volumes to their LTP providers, which leads to a larger number of STPs that are unprofitable.

4.8 Deviation-3: CPs replicate their content in every region

In the default model, each CP is present in a single geographical region. A recent trend in the Internet is that CPs increasingly expand their geographical presence [54], either through the use of content distribution networks (CDNs), or by replicating their content at multiple locations. We present a deviation of the default model where CPs are present *in every geographical region.* Geographical expansion by CPs allows them to peer with networks in a larger number of regions, and also them to select providers from a larger number of regions. We call this deviation "GEO".

Larger STP profits:

In GEO, STPs obtain larger as compared to the default model. In GEO, most CPs use PR provider selection, but they are not restricted to choosing the cheapest provider from a single region. Instead, they can choose the cheapest (which are typically STPs) across all regions. This results in larger aggregate profits for STPs. For example, the aggregate STP fitness is \$331k with scenario {DF, (PR,NC), (PR,NC)}, while it is \$450k in scenario

Table 4: Output metrics for Deviation-2 (EP), averaged over 20 simulation runs.

STP str	LTP str	wPL	uPL	dia	prof a-STP ($k)	prof a-LTP ($k)	prof f-STP ($k)	prof f-LTP ($k)	num f-STP	num f-LTP	num UA	Traf UA	%PP	Traf PP
PR,NC	PR,NC	3.3	3.5	6.0	-154	1537	13	1656	0.8	1.6	4.2	0.1	2.7	0.0
PR,NC	PR,TR	3.3	3.5	6.0	-154	1537	13	1656	0.8	1.6	4.2	0.1	2.7	0.0
PR,NC	PR,CB	3.4	3.5	6.5	-204	1512	4	1620	0.4	1.6	5.3	0.2	5.2	0.1
PR,NC	SEL,NC	3.3	3.5	6.0	-154	1537	13	1656	0.8	1.6	4.2	0.1	2.7	0.0
PR,NC	SEL,TR	3.3	3.5	6.0	-154	1537	13	1656	0.8	1.6	4.2	0.1	2.7	0.0
PR,NC	SEL,CB	3.4	3.5	6.5	-204	1512	4	1620	0.4	1.6	5.3	0.2	5.2	0.1
PR,TR	PR,NC	3.3	3.4	5.8	-152	1498	8	1618	0.8	1.5	4.1	0.1	2.8	0.0
PR,TR	PR,TR	3.3	3.4	6.0	-132	1455	24	1583	0.9	1.4	4.0	0.1	2.4	0.0
PR,TR	PR,CB	3.3	3.4	6.2	-154	1469	28	1589	0.8	1.4	4.2	0.1	3.9	0.1
PR,TR	SEL,NC	3.3	3.4	5.8	-152	1498	8	1618	0.8	1.5	4.1	0.1	2.8	0.0
PR,TR	SEL,TR	3.3	3.4	5.9	-132	1456	24	1584	0.9	1.4	3.9	0.1	2.9	0.0
PR,TR	SEL,CB	3.3	3.4	6.3	-154	1468	28	1587	0.9	1.5	4.0	0.1	4.6	0.1
PR,CB	PR,NC	3.3	3.4	5.9	-165	1512	2	1633	0.5	1.4	4.6	0.1	4.3	0.0
PR,CB	PR,TR	3.3	3.4	6.1	-149	1535	19	1653	0.5	1.6	4.2	0.1	3.6	0.1
PR,CB	PR,CB	3.3	3.4	6.0	-175	1475	18	1586	0.4	1.6	4.0	0.2	5.0	0.1
PR,CB	SEL,NC	3.3	3.4	5.9	-165	1512	2	1633	0.5	1.4	4.6	0.1	4.3	0.0
PR,CB	SEL,TR	3.3	3.4	6.1	-143	1526	23	1646	0.7	1.5	4.4	0.1	4.2	0.1
PR,CB	SEL,CB	3.3	3.4	6.1	-174	1474	18	1585	0.5	1.6	4.1	0.2	5.6	0.1
SEL,NC	PR,NC	3.2	3.3	5.0	-183	1519	0	1620	0.0	2.0	3.0	0.2	2.1	0.0
SEL,NC	PR,TR	3.2	3.3	5.0	-183	1519	0	1620	0.0	2.0	3.0	0.2	2.1	0.0
SEL,NC	PR,CB	3.2	3.3	5.0	-193	1517	0	1617	0.0	2.0	4.0	0.2	2.9	0.0
SEL,NC	SEL,NC	3.2	3.3	5.0	-183	1519	0	1620	0.0	2.0	3.0	0.2	2.1	0.0
SEL,NC	SEL,TR	3.2	3.3	5.0	-183	1519	0	1620	0.0	2.0	3.0	0.2	2.1	0.0
SEL,NC	SEL,CB	3.2	3.3	5.0	-193	1517	0	1617	0.0	2.0	4.0	0.2	2.9	0.0
SEL,TR	PR,NC	3.2	3.3	5.0	-180	1514	0	1615	0.0	2.0	3.0	0.2	2.5	0.0
SEL,TR	PR,TR	3.2	3.3	5.0	-178	1511	1	1611	0.1	2.0	2.9	0.2	2.4	0.0
SEL,TR	PR,CB	3.2	3.3	5.0	-192	1510	0	1610	0.0	2.0	4.0	0.2	3.4	0.0
SEL,TR	SEL,NC	3.2	3.3	5.0	-180	1514	0	1615	0.0	2.0	3.0	0.2	2.5	0.0
SEL,TR	SEL,TR	3.2	3.3	5.0	-178	1512	1	1612	0.1	2.0	2.9	0.2	2.6	0.0
SEL,TR	SEL,CB	3.2	3.3	5.0	-192	1510	0	1610	0.0	2.0	4.0	0.2	3.4	0.0
SEL,CB	PR,NC	3.2	3.3	5.0	-192	1514	0	1615	0.0	2.0	3.0	0.2	3.8	0.0
SEL,CB	PR,TR	3.2	3.3	5.0	-186	1500	3	1600	0.2	2.0	2.8	0.2	3.7	0.0
SEL,CB	PR,CB	3.2	3.3	5.0	-198	1496	1	1593	0.1	2.0	3.6	0.2	4.4	0.0
SEL,CB	SEL,NC	3.2	3.3	5.0	-192	1514	0	1615	0.0	2.0	3.0	0.2	3.8	0.0
SEL,CB	SEL,TR	3.2	3.3	5.0	-180	1491	6	1599	0.3	1.9	2.7	0.2	3.9	0.0
SEL,CB	SEL,CB	3.2	3.3	5.0	-197	1495	1	1593	0.1	2.0	3.7	0.2	4.5	0.0

{GEO, (PR,NC), (PR,NC}.

STPs can be profitable even with SEL provider selection:

In the default model, we observed that if STPs use SEL, then the aggregate STP profits depend on the relative prices of STPs and LTPs. In the default model, STPs are unprofitable with SEL provider selection, unless they use CB-peering. In the GEO model, however, we find that STPs can be profitable even if they use SEL provider selection and NC or TR-peering. This is because most CPs use PR provider selection, and can select the cheapest STPs from all regions. This increases the aggregate profits of STPs.

Larger aggregate profit for STPs by TR-peering:

In the default model, when STPs use PR provider selection, their aggregate profit is larger with TR-peering than with CB-peering. This is because by using CB-peering, STPs peer with CPs, which would otherwise become their customers. We find that this effect is more pronounced when CPs are present in every region. The difference between scenarios {DF, (PR,TR), (PR,TR)} and {DF, (PR,CB), (PR,TR)} is \$88k (38%), while the difference between {GEO, (PR,TR), (PR,TR)} and {GEO, (PR,CB), (PR,TR)} is \$151k (45%).

Shorter paths:

As in the EP model, we find that several scenarios in GEO lead to weighted paths that are shorter than unweighted paths. For example, the weighted path length for scenario {GEO, (PR,NC), (PR,NC)} is 3.7, while the unweighted path length is 3.9. In GEO, the traffic matrix has mostly CS traffic, with large traffic volumes from from CPs to ECs. Further, CPs can connect to STPs in multiple regions, and the number of CP-STP links is larger than in the default model. Consequently, we see a large number of "short" paths of the form CP-STP-EC, which bypass LTPs and also carry significant traffic. This leads to weighted paths that are shorter than unweighted paths.

4.9 Related Work

A major research effort aimed to characterize the AS-level topology during the last decade. One of the most well cited papers, by Faloutsos *et al.* [47], argued that the Internet AS-level topology is "scale-free". The observation that the degree distribution follows a power-law led

Table 5: Output metrics for Deviation-3 (GEO), averaged over 20 simulation runs.

STP str	LTP str	wPL	uPL	dia	prof a-STP ($k)	prof a-LTP ($k)	prof f-STP ($k)	prof f-LTP ($k)	num f-STP	num f-LTP	num UA	Traf UA	%PP	Traf PP
PR,NC	PR,NC	3.7	3.9	6.2	450	207	570	334	4.0	1.4	0.7	0.0	2.4	0.1
PR,NC	PR,TR	3.7	3.9	6.2	450	207	570	334	4.0	1.4	0.7	0.0	2.4	0.1
PR,NC	PR,CB	4.2	4.0	6.3	17	393	149	493	3.6	2.0	1.4	0.2	6.6	0.5
PR,NC	SEL,NC	3.7	3.9	6.2	464	190	583	324	4.2	1.2	0.8	0.0	2.7	0.1
PR,NC	SEL,TR	3.7	3.9	6.2	464	190	583	324	4.2	1.2	0.8	0.0	2.7	0.1
PR,NC	SEL,CB	4.2	4.0	6.3	17	393	149	493	3.6	2.0	1.4	0.2	6.6	0.5
PR,TR	PR,NC	3.7	3.9	6.4	512	132	630	264	4.3	1.3	0.8	0.0	2.4	0.1
PR,TR	PR,TR	3.7	3.9	6.1	515	152	637	290	4.0	1.2	1.3	0.0	2.1	0.2
PR,TR	PR,CB	4.0	3.9	6.2	92	406	236	499	2.7	2.0	1.9	0.2	6.4	0.4
PR,TR	SEL,NC	3.7	3.9	6.4	528	113	645	252	4.5	1.1	1.0	0.0	2.9	0.2
PR,TR	SEL,TR	3.7	3.9	6.1	527	117	645	259	4.2	1.0	1.4	0.0	3.0	0.2
PR,TR	SEL,CB	4.1	3.9	6.0	53	387	184	489	3.2	2.0	1.3	0.2	6.8	0.5
PR,CB	PR,NC	3.8	3.9	6.1	360	196	483	322	4.0	1.4	1.0	0.0	5.2	0.3
PR,CB	PR,TR	3.8	3.9	6.1	354	229	484	356	3.7	1.3	1.0	0.1	5.2	0.3
PR,CB	PR,CB	3.9	3.9	5.9	89	347	229	419	2.3	2.2	1.6	0.2	7.9	0.6
PR,CB	SEL,NC	3.7	3.9	6.2	398	183	518	314	4.1	1.3	1.0	0.0	5.8	0.3
PR,CB	SEL,TR	3.8	3.9	6.1	392	181	514	315	4.0	1.2	0.8	0.0	5.8	0.3
PR,CB	SEL,CB	3.9	3.9	5.9	84	358	220	439	2.5	1.9	1.8	0.1	8.3	0.6
SEL,NC	PR,NC	3.8	3.9	5.0	123	595	257	696	2.4	1.6	1.0	0.1	2.1	0.0
SEL,NC	PR,TR	3.8	3.9	5.0	123	595	257	696	2.4	1.6	1.0	0.1	2.1	0.0
SEL,NC	PR,CB	3.9	3.9	5.0	-152	662	13	761	1.3	1.8	2.2	0.4	6.0	0.5
SEL,NC	SEL,NC	3.8	3.9	5.0	123	595	257	696	2.4	1.6	1.0	0.1	2.1	0.0
SEL,NC	SEL,TR	3.8	3.9	5.0	123	595	257	696	2.4	1.6	1.0	0.1	2.1	0.0
SEL,NC	SEL,CB	4.0	3.9	5.0	-165	693	2	793	1.0	2.0	2.0	0.5	5.8	0.4
SEL,TR	PR,NC	3.8	3.9	5.0	123	595	257	696	2.4	1.6	1.0	0.1	2.1	0.0
SEL,TR	PR,TR	3.8	3.9	5.0	139	588	273	688	2.6	1.4	1.1	0.1	2.1	0.0
SEL,TR	PR,CB	3.9	3.9	5.0	-27	578	141	678	1.5	1.8	2.1	0.4	5.5	0.5
SEL,TR	SEL,NC	3.8	3.9	5.0	123	595	257	696	2.4	1.6	1.0	0.1	2.1	0.0
SEL,TR	SEL,TR	3.8	3.9	5.0	137	590	270	692	2.6	1.4	1.0	0.1	2.2	0.0
SEL,TR	SEL,CB	3.9	3.9	5.0	-35	588	124	690	1.5	1.9	1.7	0.4	6.3	0.5
SEL,CB	PR,NC	3.7	3.8	5.0	235	478	365	579	3.0	1.6	0.4	0.0	2.5	0.1
SEL,CB	PR,TR	3.7	3.9	5.0	247	467	378	568	2.9	1.5	0.7	0.0	2.6	0.1
SEL,CB	PR,CB	3.8	3.9	5.0	28	484	187	575	2.1	1.8	2.2	0.2	5.8	0.5
SEL,CB	SEL,NC	3.7	3.8	5.0	235	478	365	579	3.0	1.6	0.4	0.0	2.5	0.1
SEL,CB	SEL,TR	3.7	3.9	5.0	246	469	377	571	2.9	1.5	0.6	0.0	2.5	0.1
SEL,CB	SEL,CB	3.8	3.9	5.0	11	498	163	586	2.2	1.8	1.9	0.2	6.8	0.5

to several topology generation models that could produce such distributions, starting with the preferential attachment model of Barabasi *et al.* [15]. Several variants and comparisons of preferential attachment models were later proposed [10, 21, 90, 97, 107, 110, 113] The models in this research thread have been mostly descriptive, meaning that they attempt to reproduce certain known structural characteristics of the Internet.

The previous descriptive models received considerable criticism (for instance, see [67, 70]) because they mostly focus on the degree distribution and clustering, ignoring important characteristics of the Internet topology such as hierarchy or the presence of links of different types (transit versus peering). Further, those models do not explain how the Internet topology is evolving. This led to models that view the Internet topology as the effect of optimization-driven activity by individual ASes. These concepts were first introduced by Carlson and Doyle in [23], and later applied in the context of the Internet in [46]. Chang *et al.* [24] model AS interconnection practices, considering the effects of AS geography, AS business models and AS evolution.

The body of work closest in spirit to ours is that of Chang *et al.* [25, 27]. That work focused on developing a first-principles model for the provider and peer selection behavior of ASes, taking into account various AS-specific decisions in the role of an AS as a customer and as a peer. Their model also accounts for practical practical constraints such as geography and considers realistic economics of transit and peering costs and an interdomain traffic matrix derived from measurement data. Their model can account for the heterogeneity in terms of the different types of ASes with differing business objectives that exist in the Internet. The similarity with our work lies in the fact that they too advocate a bottom-up approach for understanding and modeling the Internet and its evolution. The major difference between the work of Chang *et al.* and our work is the context in which the model is applied. The goal of Chang *et al.* was to generate and evolve Internet-like graphs by assigning various decision strategies to ASes, and studying the structure and evolution of the graph over time. Their model is able to reproduce certain properties seen in the evolution of the Internet's AS-level graph over time, and mainly focused on graph-level properties of the Internet and how these evolve over time. The model proposed in this thesis is *not*

an evolutionary model. We focus mainly on studying the properties of the equilibrium that results as each AS uses certain provider and peer selection strategies. Further, we are interested in AS-specific properties (both for individual ASes and classes of ASes) such as profitability, number of customers and traffic share. In this thesis, we thus go beyond simply producing graphs that resemble the real Internet. In other words, our goal is not to present a model and validate its ability to produce graphs that representative of the real Internet. Instead, we attempt to study the properties of the resulting graph as a function of the different strategies used by networks and conditions such as the interdomain traffic matrix, pricing and cost structures and geographical constraints.

The model presented in this thesis relies on *agent-based simulations* to determine the equilibrium that results as each network uses certain provider and peer selection strategies. Holme *et al.* [60] present an agent-based simulation model where the agents are ASes with economic incentives. In their model, each AS attempts to connect in such a way as to maximize its utility under a set of constraints. Their model captures the effects of economics, geography, user population and traffic flow in AS interconnection. They do not, however, model the presence of different classes of ASes with different incentives and business functions, and their model is rather simplistic, ignoring some important domain-specific details about the Internet at the interdomain level, such as the interdependence between provider and peer selection, real-world economics and pricing, and the role of the interdomain traffic matrix. Corbo *et al.* [33] propose an economically-principled model that is able to create the observed structure of the AS graph. Their model considers the economic utility of an AS, and focuses on growing a network where each new AS tries to maximize its utility from connecting to the Internet. In a sense, this model follows a bottom-up approach, modeling ASes as selfish agents concerned with maximizing their utility function. The goal of their work, however, was to model, from first principles, the evolution of the Internet graph, focusing mainly on the growth phenomenon as new ASes join the Internet.

There is a body of work in the area of interdomain network formation that comes from a more theoretical or game-theoretic perspective. A series of papers [74, 75, 73] advocate the use of the Shapley value for revenue distribution between ISPs. They show that if

profits are shared according to the Shapley value, the set of "fair" properties inherent to the Shapley solution exist, and the selfish behavior of ISPs leads to globally optimal routing and interconnection decisions. A body of work known as "network formation games" [13, 14, 64] takes a *game theoretic* approach to the creation of interdomain links between autonomous networks. These papers formulate a game where Autonomous Systems form a graph to route traffic between themselves. Variants of these models assign costs for routing traffic, as well as for a lack of end-to-end connectivity. The goal of each AS is to create the set of links that maximizes its utility. A key difference of these models with ours is that they are *static* in nature; they model one-shot games where an AS knows the payoff obtained from creating a particular link. We consider the realistic case where ASes do not play simultaneously, and are able to observe the moves made by previous players. Also, we assume that an AS cannot predict the payoffs it would obtain by choosing certain providers or peers. Also related is the literature on "potential games" introduced by Monderer and Shapley [79], where the incentive of individual players to change their strategy can be expressed as a single global function called the potential function. The formulation of the potential function is useful, as the incentives of all players can be mapped to a single function. Consequently, finding the equilibrium of such as system amounts to finding the local minima of the potential function. However, it is not clear whether such a formulation can be applied to the case of ASes with differing incentives; For instance, the incentive of a transit provider is to maximize monetary profit, while that of a stub network is to minimize end-to-end path lengths for its traffic. It may not always be possible to formulate a centralized function that can account for the different incentives of individual ASes.

The work of Norton is also related to the general area of Internet economics and peering, but has a more practical and anecdotal flavor. In a series of white papers, Norton discusses, mainly using anecdotal evidence, how economic and competitive interests influence peering and transit connectivity in the Internet [82]. Norton also discusses the "peering playbook" [83], which is a practical guide for ISPs to decide between settlement-free peering and transit connectivity. He also presents anecdotal evidence for evolution trends in the Internet ecosystem [84, 86], focusing on which types of ASes prefer to peer with each other,

and the evolution of transit prices over time [85]. This work gives us many insights into how settlement-free interconnection decisions are made in practice, and also provides real-world data about the relative magnitudes of transit, peering and operational costs for networks. This body of work, however, does not try to determine which strategies for provider and peer selection different types of networks should use to maximize their utility, or the resulting effects on the global Internet.

4.10 Conclusions

In this chapter, we proposed ITER, a detailed first-principles model of interdomain network formation that captures the interdependence between interdomain topology, traffic flow and provider and peer selection strategies of ANs. We present an approach to solve this model using agent-based simulations. As a first practical application of ITER, we evaluate the effect of various strategies for provider selection ("choose cheapest providers", "choose higher-tier providers") and peer selection ("peer by necessity", "peer by traffic ratios" and "peer by cost-benefit analysis") on the profitability of small and large transit providers. We examine the effects of these strategies on the economics, topology and performance of the internetwork at equilibrium.

We find that contrary to conventional wisdom, large transit providers can increase their profits (and decrease those of small transit providers) by peering with content providers. We find that several strategies lead to situations where providers are "unprofitable but active", meaning that these providers carry traffic but are unprofitable. We find that two recent trends in the Internet – an increasing amount of peer-to-peer traffic and the expansion of geographical coverage by large transit providers – can lead to higher profits for small transit providers. We find that if networks at the edge of the Internet become performance-aware, then large transit providers attract most customers. Performance-aware provider selection by edge networks leads to shorter end-to-end paths, at the expense of the profitability of small transit providers.

CHAPTER V

STRATEGIES FOR ACCESS PROVIDERS: THE NETWORK NEUTRALITY DEBATE

5.1 Introduction

A recent trend in the evolution of the Internet is that residential and SOHO (Small Office, Home Office) users download increasingly more content. Several causes are attributed to this phenomenon, in particular the increasing penetration of broadband access, faster last-mile links, the rise of Internet video and peer-to-peer file sharing. This content is delivered to users by Internet Service Providers (ISPs) that are known as Access Providers (APs). APs earn their revenues mostly from their users, and they incur costs to operate their network and to purchase upstream connectivity from transit providers. A trend in recent times is that APs are often not profitable. The increasing volume of traffic that APs need to deliver to their customers leads to escalating operational costs and transit fees. Meanwhile, revenues decline as intense competition in the access market and the commoditization of Internet access leads to falling prices, typically in the form of a flat monthly fee [43, 50, 84]. On the other hand, content providers (CPs) such as Google, Yahoo! etc. are seen as being quite profitable, given their large revenues from advertising and other sources. These CPs must use the AP's infrastructure to reach end users (often without directly connecting to or paying APs), they are often viewed by APs as "free-riders", which has led to significant tension between APs and CPs. This is often referred to as the "network neutrality" debate, as APs have threatened to charge CPs directly or to throttle the traffic from the largest CPs. Despite the many articles in the popular press, articles written by economists and telecommunication policy experts [18, 44, 49, 58, 98, 104, 108], and by computer scientists [35, 19, 109], this debate is still highly misunderstood. We believe that this debate is primarily driven by the profitability of APs. To understand the network neutrality debate, we need to understand both the economic structure and

the traffic characteristics that APs need to work with. We also believe that this debate necessitates a re-think of the strategies that APs need to employ to remain profitable. Which pricing strategies should APs use to improve their profitability? Can they do so without using risking a loss of their customer base? Can APs benefit by peering selectively with content providers or caching their content locally?

In this part of the thesis, we answer the above questions using a simple quantitative model that captures the interactions between an AP, a transit provider, and a number of CPs. The model captures the per-user heavy-tailed traffic distribution, the highly skewed popularity distribution among CPs, and realistic functions for the transit, peering and operating costs incurred by the AP. We first examine a "baseline strategy" that follows current practice, in which the AP charges the same flat rate to all users. Further, the AP does not establish peering sessions with CPs. We then compare this baseline strategy to some strategies that an AP could use to increase its profitability. We focus mainly on strategies that are "network neutral", meaning that the AP does not differentiate between sources of content. These strategies are: usage-based pricing for heavy-hitters, limiting the traffic of heavy-hitters, selectively peering with some CPs, and caching content from selected providers. We also investigate a "non-network neutral" strategy in which an AP charges CPs directly. Our results show that certain strategies are rarely profitable or they are sensitive to factors that are not controlled by the AP (e.g., how would users react to heavy-hitter usage-based pricing?). On the other hand, the strategy of selective peering with CPs is non-disruptive and it can lead to a profit increase, relative to the baseline strategy, for the AP. To increase the effectiveness of such peering, it is important that the AP is co-located with the most popular CPs so that it can reduce peering costs. Caching can also help, even though the profit increase with that strategy depends significantly on the fraction of traffic that can be cached.

5.2 The Network Model

We consider the interactions between three distinct species in the Internet ecosystem.

Access provider (AP): We focus on a single AP that sells Internet access to N paying users.

Content providers (CP): Content providers are the sources of content on the Internet. They do not provide access or transit service to any customers. Instead, CPs earn revenue from sources such as advertisements (out of band revenue). In this work, we do not model the costs and revenues of the CPs, instead focusing on the AP. Further, we take into account only the traffic flow from CPs to the AP, ignoring the requests from customers which are assumed to be small. Also, we assume that the AP does not receive any traffic from other APs, e.g. due to p2p applications. We intend to account for p2p traffic in the extended version of this paper.

Transit providers(TP): TPs provide transit for their customers, which are other ASes. TPs earn revenue by charging their customers for the volume of traffic sent and received. For simplicity, we consider a single TP that can provide transit to any other AS.

All ASes have a certain geographical scope, which is determined by the locations of their points-of-presence (PoPs). Large TPs are typically present globally, while APs and CPs could have only regional presence. As a result, an AP cannot always connect directly to a CP, and the TP is needed to provide reachability. An alternative is for the AP to establish a point of presence in remote locations (using a leased line to that location, for instance) or for the CPs to come closer to the AP by using a content distribution network.

We model two distinct types of inter-AS connections. In a transit (CP) relation, the customer "buys" transit service from the provider. The customer typically pays the provider for traffic sent in both directions on the customer-provider link. In a peering connection, two ASes agree to exchange traffic for free. If the CP and AP have a peering relationship, then the traffic flows directly between the two networks. If both the CP and the AP are customers of the tier-1 provider, then the tier-1 transits traffic that flows from the CP to the AP. Figure 43 illustrates this network model.

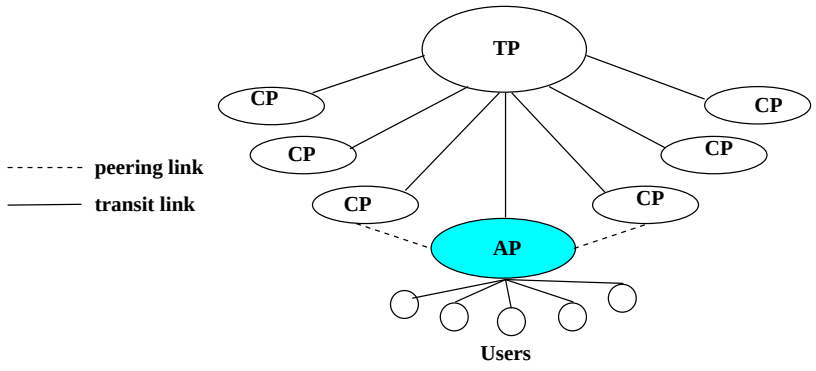

Figure 43: The network model

5.3 The baseline model

This section describes and evaluates a "baseline" model. We believe this model represents the most common current practices, and it captures pricing, connectivity and traffic distribution among the users of the AP and among CPs.

Connectivity and pricing: We consider a situation where both the AP and CP connect to the TP as transit customers, and there are no peering links between the AP and CPs. The AP has N users and charges each of those users based on a flat monthly rate R, giving it a revenue $N * R$. Based on common prices for Internet access in North America, we set the flat rate charged by the AP to \$20/month. The TP charges both the AP and the CPs using the volumes of traffic sent in both directions. The transit pricing function we use for the TP is a concave increasing function of the form $c_t = m_t * V^{0.75}$, where m_t is the transit pricing multiplier used by the TP, V is the charging traffic volume (in Mbps), and c_t gives the monthly price for transit. This pricing function was used in [25] based on pricing data obtained from ISPs, and m_t was around 100 for transit ISPs in North America. Here, we use a transit multiplier $m_t = 100$ for the TP. Using this pricing function, a charging volume of 10Mbps costs \$560 (\$56/Mbps), while 10Gbps costs \$100,000 (\$10/Mbps). This illustrates the well known "economies of scale" in transit prices, i.e., the per-Mbps price decreases as the total charged capacity increases.

The TP typically calculates the charging volume V by dividing the month into 5 minute intervals, and V is the 95th percentile of the load on the customer link over all such intervals.

Norton [86] notes that the 95th percentile charging model is based on the rule of thumb that the ratio of the 95th percentile to the average load is around 2:1 for web traffic. With the increase of video traffic, however, that ratio could be as high as 4:1. In this paper, we assume that the ratio of 95th percentile to average load is 3:1.

Local costs: The local cost of an AP consists of expenses to lease bandwidth for its network, purchase routers and other equipment, and to hire personnel to operate the network. This local cost is modeled as traffic independent and traffic dependent components of the form $c_l = f_l + m_l * V^{0.5}$. f_l is the traffic independent fixed cost component, and we set f_l =\$250000/month for the AP. The local cost multiplier m_l is set to 500. In the absence of data about ISP operational costs, the local cost parameters are chosen to yield a net profit margin of approximately 20% for the AP, which is similar to what was seen in the balance sheet of a large North American access ISP. The local cost exponent is 0.5, which means that the cost incurred to carry traffic scales slower than the transit costs paid by the AP, while also showing economies of scale.

AP profit: The profit of the AP in the baseline model is the total revenues minus the transit and local costs, i.e.,

$$\mathcal{P} = NR - m_t * V^{0.75} - f_l - m_l * V^{0.5} \tag{21}$$

AP users: Each of the N users of the AP downloads a certain amount of traffic every month. To model the user traffic demand, we refer to a study of residential broadband access networks in Japan [30]. That study found that the distribution of the amount downloaded by a user is heavy tailed. In their measurements, approximately 4% of users download more than 75 GB/month (heavy hitters), while the remaining download less than 75 GB/month (normal users). We estimate the average of the normal and heavy hitter users as 300 MB/month and 10GB/month respectively, which gives an overall average of approximately 8GB/month.

Here, we draw the amount downloaded by each user from a truncated Pareto distribution with shape parameter 1.1 and mean 8 GB/month, in which case 2% of the users download more than 75 GB/month. The distribution is truncated from above at a point

corresponding to the the access link speed. For example, a user behind a 1.5Mbps connection cannot download more than 486 GB/month. We consider different values of the cutoff point corresponding to the various common access speeds: 300kbps (97 GB/month), 700kbps (226 GB/month), 1.5Mbps (486 GB/month) and 10Mbps (3240 GB/month). Figure 44 shows the complementary CDF (CCDF) of the amount downloaded by each user. Unless noted otherwise, we use a cutoff point corresponding to the 1.5Mbps access speed in the rest of this paper.

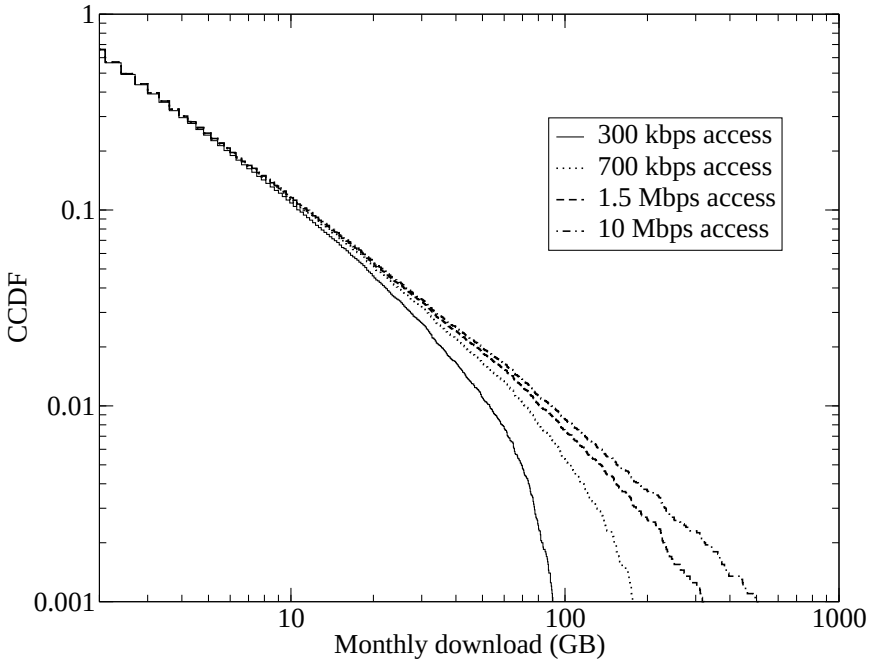

Figure 44: CCDF of the amount downloaded by users (GB/month).

Inter-AS traffic matrix: After generating the traffic demands for AP users, we create the distribution of the traffic among CPs as follows. The total incoming traffic for the AP is calculated as the sum of the incoming traffic demands for each of its users. This total is used to obtain the *average* incoming traffic, in Mbps, for the AP. Chang et al. [28] found that the traffic distribution from the top content providers follows a Zipf-like distribution, with shape parameter ranging from 0.9 to 1.1. We assign the actual traffic volumes from each CP i to the AP using a Zipf distribution with shape parameter 1. This produces an effect where certain CPs are "popular" sources of content for the AP.

5.3.1 Evaluation of the baseline scheme

Here, we evaluate the performance of the baseline model used by the AP. We examine how the profit of the AP varies with the number of users, the different random samples of users, and the increasing amount of video traffic.

Variability in the set of users: We first examine what happens when the number of AP users increases. Recall that the traffic demand of each user is drawn from the heavy-tailed truncated Pareto distribution described earlier. The large variability in the traffic demand of individual users leads to also large variability in the costs incurred by the AP. To demonstrate this effect, we draw 1000 samples (corresponding to different samples of the user population) from a truncated Pareto distribution with shape parameter 1.1 and different cut-off points corresponding to different access speeds. We then create the inter-AS traffic matrices and calculate the costs incurred by the AP. Figure 45 shows the median and the min-max range of the AP costs across 1000 simulation runs. We find that the costs of the AP can vary significantly depending on the amount downloaded by its set of users. Moreover, as the user access speed increases, both the sample mean and the variance increase. This means that the increasing access speeds that users enjoy in the last few years will lead to increasing variability in the AP costs, making it harder for access providers to guarantee their profitability.

The impact of video traffic: A recent trend is that a large fraction of the traffic from content providers is streaming video. Norton [86] notes that video traffic is fundamentally different from web traffic, as the ratio of the 95th percentile to the average load due to video is 4:1, while for web traffic it is roughly 2:1. Consider, for example, that the users download web content at an average rate of V Mbps. Using the 2:1 ratio for web traffic, the AP provisions its network and gets charged by the TP for a charging volume $2V$. If the traffic is video, the AP must provision its network and purchase transit capacity for $4V$. This leads to a significant increase in the costs incurred by the AP, as shown in Figure 45.

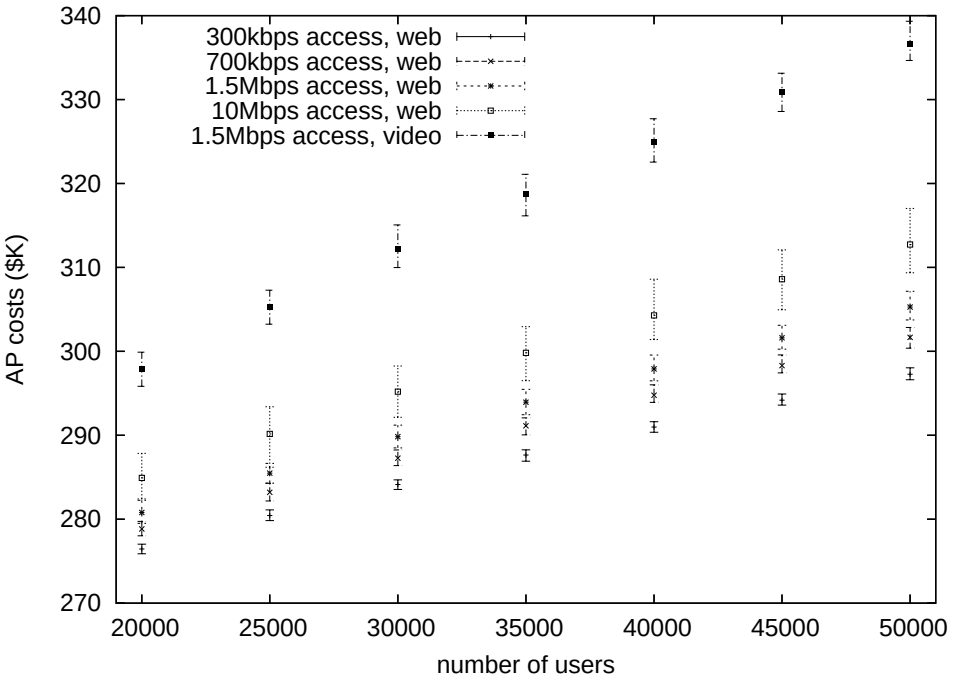

Figure 45: Variability of AP costs with the number of users, access speeds and type of traffic.

5.4 ISP strategies

There are various strategies that an AP could deploy to increase its profits. In this section, we evaluate some strategies that are anecdotally mentioned in discussions about network neutrality and ISP economics. We attempt to gain a deeper, quantitative understanding of the pros and cons of these strategies. Further, we compare each strategy with the baseline, and evaluate the conditions under which the AP is able to achieve better profits than the baseline.

5.4.1 AP charges heavy hitters

In this charging strategy, the AP sets a threshold $\mathcal{T}$ to identify the users that download the largest amounts of traffic. These users are called the "heavy hitters", and the AP uses a volume-based pricing scheme for these users, rather than the flat rate. The price charged to a heavy hitter that downloads an amount of traffic $D > \mathcal{T}$ is given by: $c_v(D) = \frac{D*R}{\mathcal{T}}$, i.e., a heavy hitter is charged proportional to the amount of traffic downloaded.

A volume-based charging strategy is likely to be unpopular with the AP's users. In the

presence of sufficient competition in the AP market, customers would switch from an AP that uses volume-based charging to an AP that offers flat-rate, "all you can eat" service. We model the unpopularity of volume-based charging with a probability that a user leaves this AP, referred to as *departure probability*. The departure probability depends on the threshold $\mathcal{T}$ set by the AP and is calculated as follows. For a value of $\mathcal{T}$ set by the AP, it is possible to calculate the number of users N_h that would be classified as heavy hitters. The number of users N_d that are expected to depart at this threshold is assumed to be proportional to N_h, $N_d = d * N_h$, where d is a positive parameter. The departure probability is then set to N_d/N (as long as $N_d \leq N$). The departure probability is applied to all users, not only those that are classified as heavy hitters. This captures the pragmatic fact that users are uncertain about their monthly usage and so they may leave the AP to avoid the possibility of extra fees if they get classified as heavy hitters. The parameter d determines the shape of the departure probability curve, as shown in Figure 46. The parameter d is also related to the degree of competition in the Internet access market. Without competition, users would be bound to a particular AP and d would be quite low as long as users need to have Internet access.

We evaluate the heavy hitter charging strategy by calculating the profit of the AP for different values of the threshold $\mathcal{T}$ and the parameter d. Figure 47 shows the profit of the AP as a function of the threshold $\mathcal{T}$. In the case of $d = 0.1$ and $d = 1$, the user departure probability decreases quickly with $\mathcal{T}$. In this case, even for low values of $\mathcal{T}$, the AP retains a significant fraction of users, and also charges them according to the downloaded traffic. Consequently, it can achieve higher profits than the baseline scheme. In the case of $d = 2$ and $d = 10$, the user departure probability decreases slowly, and the optimal value of $\mathcal{T}$ shifts higher. In the most extreme case of $d = 10$, the optimal value of $\mathcal{T}$ occurs when the AP is able to keep all its users. The AP's profit in that case is similar to that of the baseline scheme. The curves for different access speeds are qualitatively similar. As expected, the benefit of heavy hitter charging is smaller if the users are limited by a smaller access speed. This is simply because there are fewer heavy hitters at any given value of $\mathcal{T}$ that the AP would be able to charge.

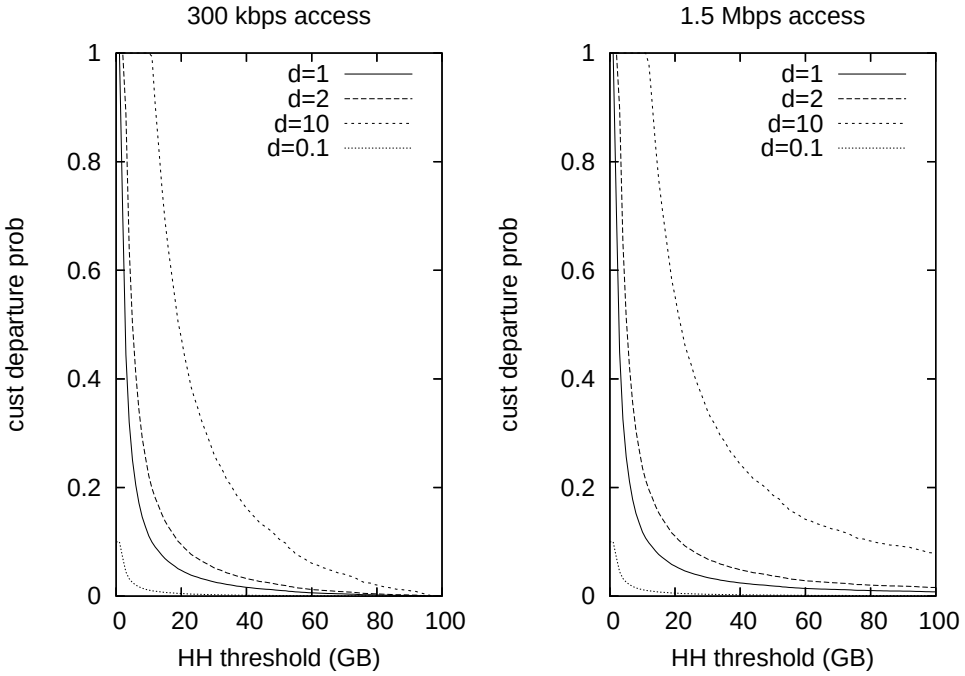

Figure 46: User departure probability as a function of $\mathcal{T}$, N=20000.

The previous results illustrate that a volume-based charging strategy is quite sensitive to the user departure probability, which is not controlled by the AP. Even if the departure probability is low, it is difficult to determine the optimal value of $\mathcal{T}$, and hence this strategy is not robust to the selection of this threshold. If the AP sets $\mathcal{T}$ to a sub-optimal point, it could end up with even lower profit than in the baseline scheme.

5.4.2 AP caps heavy hitters

In this strategy, the AP imposes "download caps" on its users, i.e., users are not permitted to download more than $\mathcal{T}$ GB/month. If a user reaches that threshold, her account is blocked for the remainder of the month[1]. In this strategy, the AP charges each access customer with the same flat rate R. As with the strategy of charging heavy hitters, capping the amount that a user is allowed to download can be an unpopular strategy. We assume that the departure probability with this strategy is modeled using the same function as in 46. In practice, the departure probability in this model may be higher or lower than in

[1]In practice, the AP may choose to seriously rate-limit a user that exceeds her threshold. For simplicity, we consider the more extreme measure where the user is blocked.

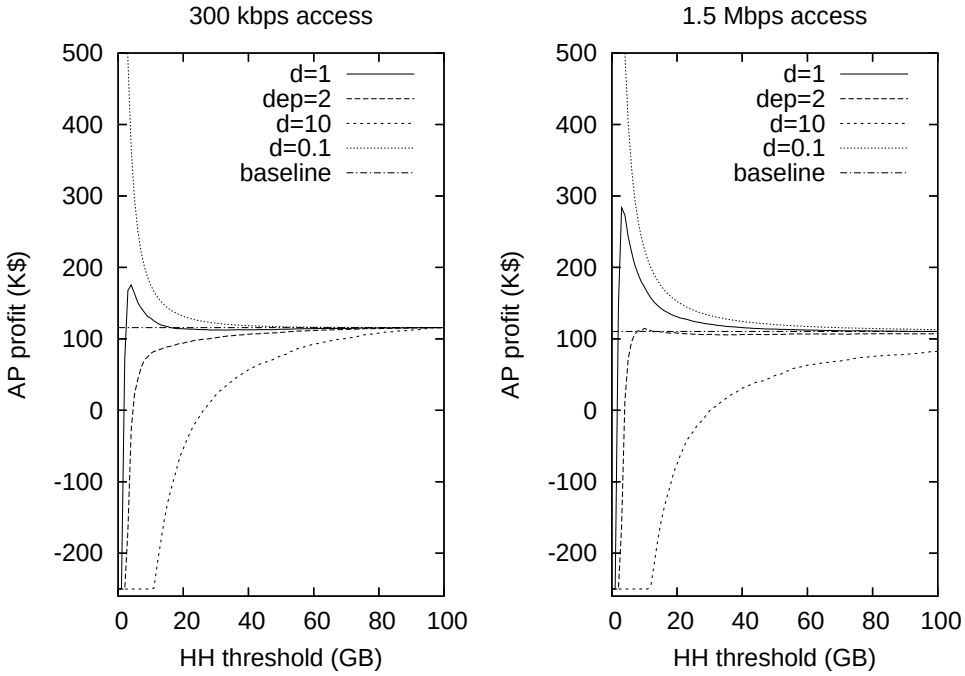

Figure 47: AP profit as a function of $\mathcal{T}$ when the AP charges heavy hitters, N=20000.

the heavy hitter charging scheme, depending on the user population, the available pricing plans and policies of competing APs, and how APs justify/present these policies to their users.

Figure 48 shows the profit of the AP as a function of the threshold $\mathcal{T}$ used by the AP to cap customers. We find similar trends as in the case of heavy hitter charging. The strategy of capping heavy hitters performs worse than heavy hitter charging, even when the customer departure probability drops quickly (curves marked "d=1" and "d=0.1"). By capping heavy users, the AP is only able to save on its operating costs, and does not gain any additional revenue. With the same user departure profile as in the case of heavy-hitter charging, this strategy would be less profitable for the AP.

5.4.3 AP charges CPs

There has been much debate on whether an AP should be able to discriminate between CPs. To recover the costs due to increasing traffic volumes, APs would like to charge the CPs that produce most traffic. The AP could rank CPs in decreasing order of traffic volume,

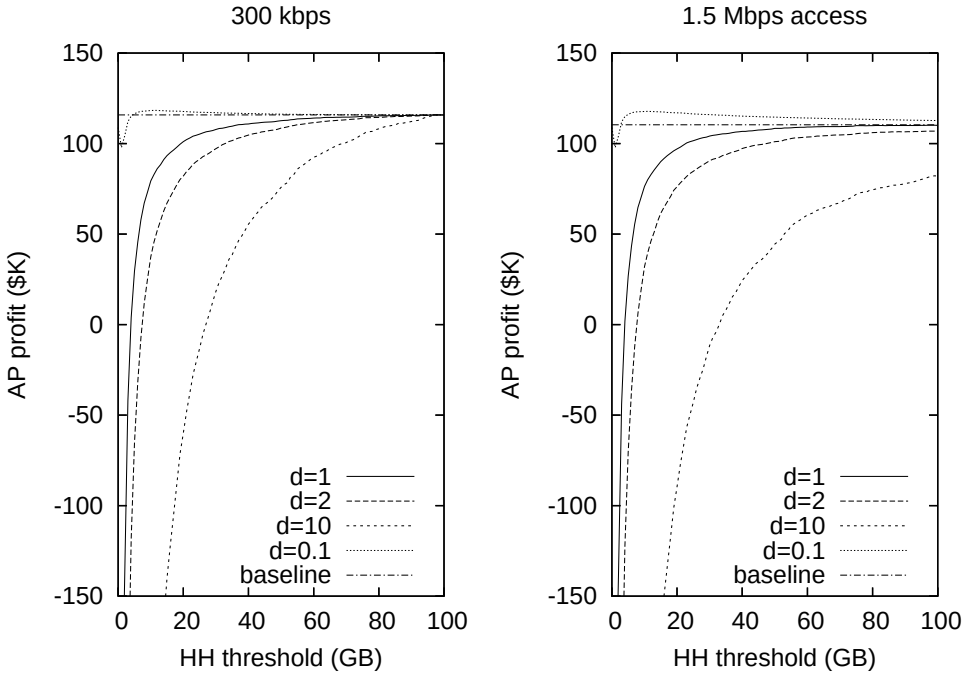

Figure 48: AP profit as a function of $\mathcal{T}$ when the AP caps heavy hitters, N=20000.

and charge a certain fraction of the top providers. We assume that the AP would use its transit pricing function to charge those CPs. This strategy is again likely to be unpopular, and a fraction of the AP's customers may choose to switch to another AP. We model this by making the customer departure probability dependent on the fraction of CPs charged by the AP using a function of the form $y = ax^b + c$. The parameter b determines the shape for the departure probability curve as shown in Figure 49. The values of a and c are adjusted to give a departure probability of 0 when no CPs are charged and 1 when all CPs are charged. We find that the profit of the AP depends stongly on the customer departure probability.

The trends in all the three previous strategies highlight an important tradeoff involved with strategies that can compromise the customer base of the AP. If the AP charges or throttles heavy hitters, or tries to charge CPs instead, it may lose some of its customers. Whether such a charging strategy increases the profitability of the ISP depends heavily on the customer departure probability. As such, the fate of an AP that deploys such a strategy would be highly dependent on user behavior. In the following, we investigate alternate, non-disruptive strategies that the AP could use to increase its profits.

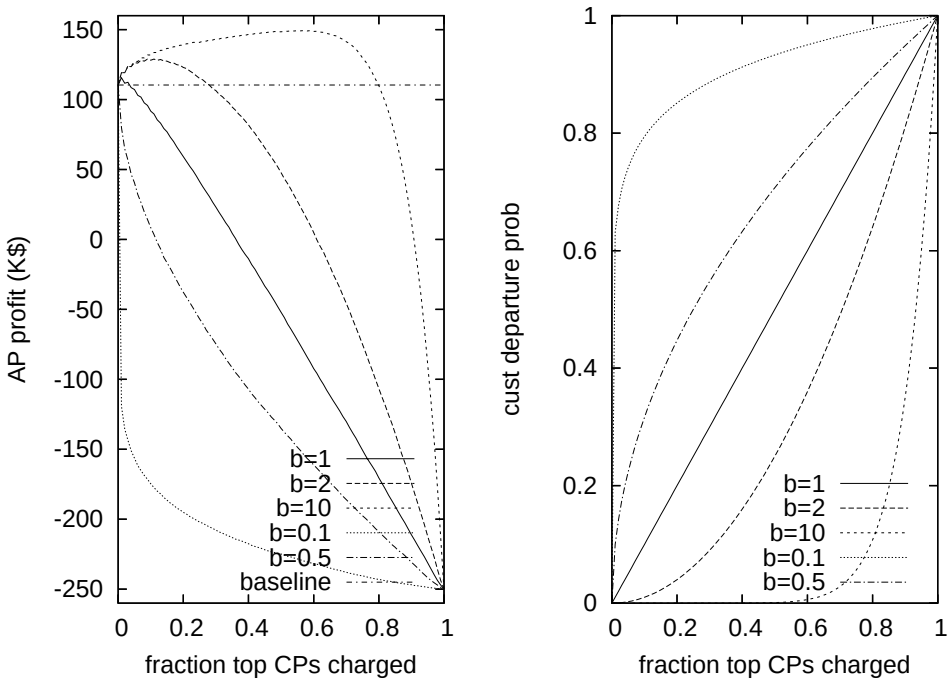

Figure 49: AP profits by charging CPs, as a function of the fraction of CPs charged, N=20000, 1.5Mbps access.

5.4.4 Selective peering with CPs

So far, we have considered the baseline model in which the AP and CPs are customers of the TP, and there is no direct peering between the AP and CPs. Here, we study a strategy where the AP follows a selective peering policy, peering with a CP depending on the potential benefits and costs associated with peering.

Chang et al. [25] studied the fixed and traffic dependent costs associated with peering. To model the traffic dependent peering costs, we use the function $c_p = f_p + m_p * V^{0.25}$, which is the function used in [25]. The parameters f_p and m_p are different for each CP, and they indicate the difficulty of peering with that CP. For example, some CPs may be colocated in the same city as the AP, in which case the peering costs are low. On the other hand, some CPs may be in entirely different continents, in which case it costs much more (or it may even be impossible) to peer with that CP. We assume that content providers fall into different classes depending on the ease of peering with that content provider. The peering cost multiplier is different for each class of CPs and the values are 10 (easiest),

100 (medium) and 1000 (hardest) peering. The fixed peering costs for these classes are \$500/month (easy), \$5000/month (medium) and \$50000/month (hard). These classes of peering costs are meant to capture the fact that it may be practically impossible for the AP to peer with certain CPs (the "hard" class). For CPs in the "medium" and "easy" classes, it makes sense for the AP to peer, if the traffic volume is sufficiently large. The figures we use for the fixed costs of the easy and medium classes are in the same range as those quoted by Norton [86]. The fixed cost of the "hard" class is very large, to model the fact that it does not make sense for the AP to peer with a CP in that class.

We investigate two distinct divisions of the CPs into the three peering cost classes. In the first, a CP is equally likely to be in any of the three classes. In the second, a CP is in the "easy" and "medium" peering class with probability 0.1 each, and with probability 0.8 is in the "hard" class. We also vary the assignment of CPs to these classes. In one case, the set of CPs in each class is determined randomly. In the second case, the most popular CPs are also the easiest to peer with. This scenario is likely in the case that the popular CPs expand their networks and are thus present in multiple peering points. A recent study gives evidence that some content providers are indeed expanding their networks in recent times [54].

To determine the set of CPs with which to peer, the AP uses the following procedure. The AP considers separately each CP i, and decides whether to peer with CP i based on a simple rule-of-thumb. Let $V(i)$ be the traffic from CP i. The AP calculates the estimated benefit of peering (saving in transit costs) as the amount that would be paid to the TP, assuming a charging volume $V(i)$. This is an approximation, as it does not account for the economies of scale when multiple CPs send traffic to the AP through the same TP.

The AP decides to peer with the CP if the following condition is satisfied:

$$\frac{m_t * V(i)^{0.75}}{f_{p_i} + m_{p_i} * V(i)^{0.25}} > \mathcal{R}$$

The estimated cost of sending the traffic $V(i)$ through the TP is given by $m_t * V(i)^{0.75}$. The cost of peering with CP i is given by $f_{p_i} + m_{p_i} * V(i)^{0.25}$.

Figure 50 shows the profit of the AP as the ratio $\mathcal{R}$ is changed. The left plot is for

the case where the CPs are distributed randomly in the three cost classes ("rand"). The number of CPs in each class is either the same (marked "eq"), or is skewed towards "hard" peering (marked "sk"). We repeat the simulations for different number of content providers ("N 50" and "N 200"). All curves show qualitatively similar behavior. If the ratio $\mathcal{R}$ is set too low, the AP forms peering relationships that incur more cost than the transit savings. On the other hand, if the ratio is too large, the AP does not peer with certain content providers that would have reduced the transit costs for the AP. We see that the optimal point for the ratio $\mathcal{R}$ occurs *after* $\mathcal{R} = 1$. This is because of the fact that the AP uses an estimate of the transit savings. Due to the economies of scale in the TP's transit pricing function, the AP *over-estimates* the potential savings in transit. An interesting trend is that above a certain value of $\mathcal{R}$, the profit is fairly robust to changes in $\mathcal{R}$. Also, the profit from peering is larger when the incoming traffic is split into a smaller number of CPs (N=50 vs N=200).

In Figure 50, the profit increase from peering is only 5% over the baseline scheme. Note that the absolute value of the profit (and the improvement over the baseline) depends on certain parameter values, such as the number of AP customers and the fixed local costs. We stress that with an appropriate choice of $\mathcal{R}$, the AP's profit with peering is guaranteed to be equal to or greater than that with the baseline scheme. The peering strategy does not increase the revenues of the AP or affect the fixed local costs. Instead, it reduces the traffic-dependent costs incurred by the AP. For the case of ("N 50 sk sort") in Figure 50, the traffic-dependent cost is \$33,000 with peering and \$40,000 for the baseline scheme, i.e., peering reduces those costs by 17%.

The strategy of selective peering appears to be quite attractive because the parameter $\mathcal{R}$ is controlled solely by the AP. The right graph shows that the benefit from selective peering is larger for the peering cost structure where the CPs with the largest traffic volume fall into the "easy" or "medium" classes. This could happen if the largest CPs expand their networks, and are thus easy to peer with. Given that such expansion by CPs is already happening [54], selective peering could be a profitable strategy for many APs.

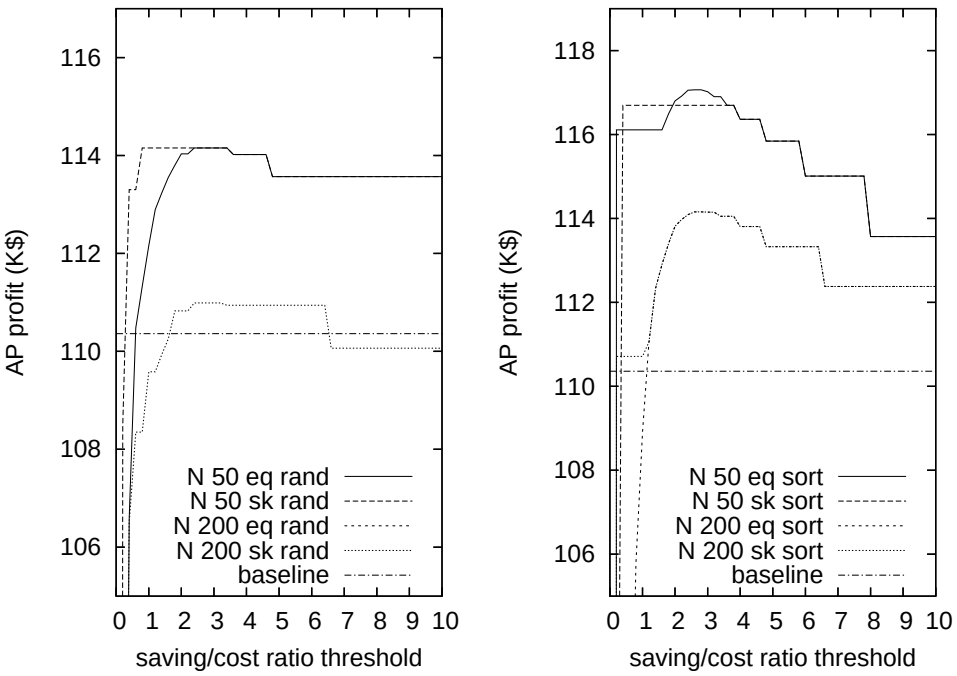

Figure 50: AP profits with selective peering as a function of $\mathcal{R}$. N=20000, 1.5Mbps access.

5.4.5 AP caches CP content

We also consider the case where the AP chooses to cache the content that it receives from the major content providers. By caching content, most requests for content by the access customers of the AP are handled locally, and hence can save transit costs for the AP. Here, we assume that there exists a certain fraction h of content from each CP that is "cacheable". When the AP caches content from CP i, the traffic $h * V(i)$ is served locally, while $(1 - h) * V(i)$ has to be downloaded through the transit provider. The fraction h captures the fact that all content from the CP may not be cacheable, e.g. dynamically generated content or live video streams. By caching content locally the AP saves transit costs. We model the costs of caching CP content, which involve purchasing servers and bandwidth to serve the content locally. We assume that caching adds to the fixed local cost of the AP according to the relation $f_c = s * f_l$, where s is a parameter that determines how the caching cost relates to the local cost. The AP must decide how many of the largest CPs to cache. The CPs are considered in decreasing order of traffic volume, because caching the largest CPs can lead to the largest potential savings in the transit costs.

Figure 51 shows the profitability of the AP as a function of the fraction of CPs cached. We simulate two cases corresponding to $s = 0.01$ and $s = 0.5$. First, we observe that the profit of the AP increases with the fraction of CPs that it caches, following a concave function. The figure on the left shows the case where the caching cost is large (equal to half of the fixed local cost). In this case, the AP is not able to do better than the baseline scheme, even if it caches all CPs. The right graph shows the case where caching costs are very low in comparison with the fixed local costs. In this case, the AP is able to achieve higher profits than the baseline scheme, depending on how much traffic is cacheable.

This analysis indicates that the attractiveness of content caching depends on the additional local cost incurred by the AP. The AP may be able to optimize its network in such a way that caching costs are small in relation to fixed local costs. In that case, the amount of CP traffic that is cacheable determines whether the AP can obtain higher profits than the baseline scheme. Note that this is again a parameter that is out of the AP's control. The previous scenario represents the case where a CP allows the AP to freely cache its content. It is possible that the CP does not allow the AP to do so due to copyright or privacy concerns. As such, our analysis of this strategy evaluates the *best case* scenario for the AP.

5.5 Conclusions

We took a quantitative approach towards understanding the network neutrality issue from the point of view of an access provider. We examined a baseline scheme that follows current practices, and some variants of charging and connection schemes. Our results show that AP strategies based on charging are rarely profitable or are highly sensitive to factors out of the control of the AP. On the other hand, the AP can obtain substantial additional profit by engaging in selective peering with CPs or caching CP content locally.

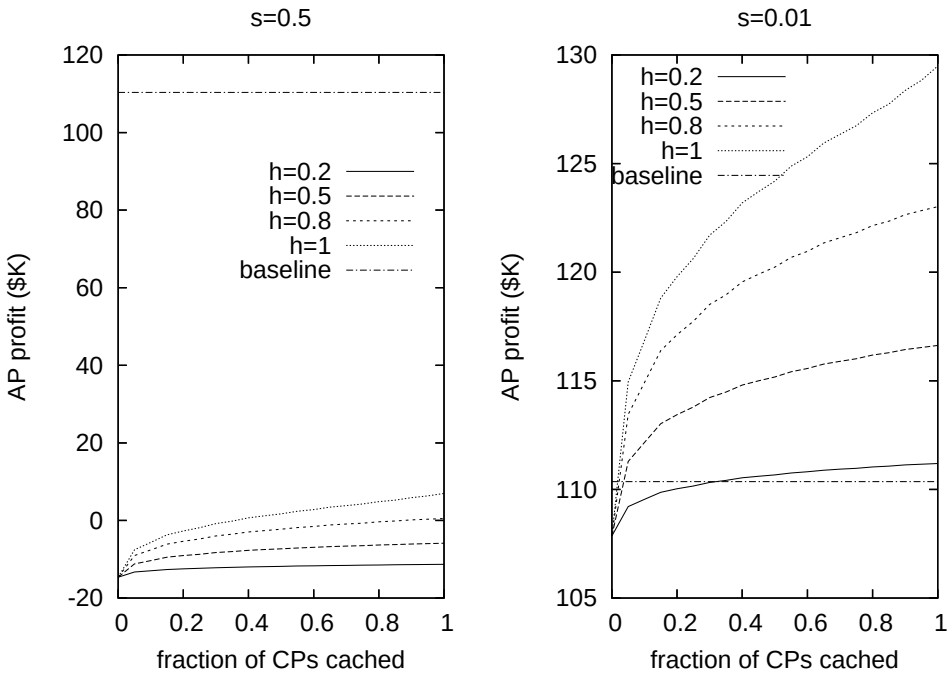

Figure 51: AP profits from caching CP content. n_A=20000, 1.5Mbps access

CHAPTER VI

CONTRIBUTIONS AND FUTURE WORK

The Internet at the interdomain level is a system of interacting, selfish networks (ASes). The Internet is dynamic, as ASes are born and die, and interdomain links appear and disappear. A plausible reason for these dynamics is that these networks try to optimize a certain objective function (either monetary cost or performance) in a distributed manner. In this thesis, we took some steps towards understanding the evolution of the Internet ecosystem, focusing on economics, traffic flow, and the implications of various provider and peer selection strategies of autonomous networks. We focused on provider and peer selection by different types of networks, and the effects of these local changes on the properties of the global Internet. The work in this thesis has potential impact for the research community, as well as more practical applications for network operators and peering coordinators at ISPs. The results of the measurement study can serve as inputs to future research that attempts to study the performance of network architectures or protocols in the future, or create synthetic topologies for simulation studies. The results from the other parts of the thesis can provide insights to network operators about how to to choose their providers and peers in order to achieve desirable properties, and the effects of their decisions on the global Internet. Presented next is a summary of the main contributions from each part of this thesis.

- **Measurement study of the evolution of the Internet ecosystem**

  We studied the dynamic properties of the Internet graph at the Autonomous System (AS)-level, focusing on provider and peer selection by ASes [39]. We found important trends in the growth of the Internet, indicating that the Internet now grows linearly in terms of both ASes and inter-AS links, following the exponential growth of the late 90s. We found that the average path lengths in the Internet stay almost constant, due to a densification process arising from aggressive multihoming by transit providers in

the core of the Internet. We emphasized the need to classify ASes into different types based on their business function, as these types differ significantly with respect to their activity (how often they make a change to their upstream connectivity), multihoming degrees, and types of providers. We also found significant geographical differences, with Europe increasingly dominant in the Internet ecosystem. There are several areas where the results of this study can be applied. First, our results can help in generating realistic synthetic topologies for evaluating new network architectures and protocols. Second, the growth and rewiring trends that we have measured can help studies that attempt to predict the performance of new or existing protocols in the future.

- **Algorithms for provider selection by edge networks**

  We proposed algorithms for provider selection by Enterprise Customers (EC) and Content Providers (CP) that at the edge of the Internet [36]. ECs are mostly concerned with minimizing their monetary costs, while Content Providers (CP) try to optimize the performance of their egress traffic. In this part of the thesis, we proposed algorithms using which these networks can select the best set of upstream ISPs. The optimization objective is to minimize the monetary cost incurred while achieving good performance (short AS-level paths and high path diversity) for the major destinations of the egress traffic. We also proposed an algorithm for egress path selection (once the best set of ISPs has been chosen) that determines a congestion-free allocation of egress flows to upstream providers (if it exists) with minimum cost for the source network. This work provides a systematic framework for network operators to make decisions about how to choose their upstream providers to minimize monetary cost and obtain good performance, using information that can be obtained *offline* (without actually connecting to a provider). We showed that by using the proposed algorithms, edge networks can find a set of providers that is close to optimal in terms of the monetary cost, path length and path diversity that these providers can offer.

- **A model for interdomain network formation**

  We proposed a first-principles model for interdomain network formation [37] that

accounts for the interdependence between topology, traffic flow, and the utility of different types of networks in the Internet (*e.g.,* monetary cost or performance for edge networks and monetary profit for transit providers.). Our model also includes geographical constraints, realistic pricing/cost structures, and BGP-like interdomain routing. We used this model to evaluate the effects of various provider and peer selection strategies, the interdomain traffic matrix, pricing/cost structures and customer preferences on the economics, performance and topology of the Internet. This work has several applications, particularly for Internet Service Providers (ISPs). Our model provides a framework for ISPs to make decisions about the provider and peer selection strategies that they should use to maximize their utility (profit, monetary cost or performance). This framework can also be used to reason about the effects of network strategies on global metrics such as interdomain path lengths or economic efficiency. Finally, the from this model could also be of interest to policy makers and regulators who would like to know what to expect from the Internet in the future, in terms of the profitability of various entities, and the risk of emerging oligopolies or monopolies.

- **Strategies for access providers – A technical view of the "network neutrality" debate**

  We approached the debate over "network neutrality" from the point of view of access providers [38]. An increasing amount of video traffic in the Internet has threatened the profitability of access providers, who mostly charge their customers a flat rate. In this work, we used a simple model to study the possible reasons for the non-profitability of access providers. We further evaluated the effectiveness of different pricing and connection strategies that the AP can use to remain profitable. We showed that AP strategies that rely on differential pricing or violations of network neutrality are not likely to be successful due to competition in the market for Internet access. We showed that strategies based on connection (peering with content providers or caching content from content providers) are more promising, and can improve the profitability of APs without risking the loss of customers. The main contribution of this part of the

thesis is a quantitative approach to the debate over network neutrality. The results showed that co-operation between access networks and content providers in the form of content caching or strategic peering is likely to be the most profitable strategy for access providers. We also showed that some strategies such as differential pricing or non-neutral charging are "unsafe", meaning that they are highly sensitive to the behavior of end users and competition in the access market.

6.0.1 Future work

The work in this thesis represents a step towards understanding the evolution and dynamics of the Internet ecosystem. Several directions for future work are natural extensions of the work in this thesis:

- **Better Measurements**

  The results of the measurement study (chapter 2), indicate a poor visibility of settlement-free peering links in the Internet. In particular, some peering links are visible in publicly available BGP data only if a route monitor is present at either endpoint of that link. Consequently, we are not able to reliably study the evolution of peering links. The number of route monitors in publicly available data has increased steadily over the years, and there are now close to 500 active monitors. Route monitors are important, because if an AS provides a route monitor, then *we can detect all the links of that AS, including peering links.* We propose to study the evolution of the set of customer-provider and peering links for ASes that are route monitors. Even though this is a small subset of the ASes in the Internet, these monitors could serve as valuable case studies. In particular, we propose to classify the monitor ASes into different types (enterprise customers, content providers and transit providers), and study the changes in the set of customer-provider and peering links for these monitor ASes. The goal is to study how frequently customer-provider links change to peering links and vice versa. Doing so could give us further insights into the inter-dependence between provider and peer selection, and the dynamics of the relationship between two ASes. We expect that this study will augment the model described in chapter 4.

- **Distributed solution of a centralized problem?**

  The work presented in this thesis considers the Internet at the interdomain level as a system of selfish, interacting agents, each trying to optimize a certain utility function. The distributed nature of these optimizations leads to the question: What objective function does the Internet, as a whole, try to optimize via these distributed optimizations? Does such a global objective function even exist? We can approach this question by formulating a centralized optimization problem, where a single entity creates the interdomain topology to optimize that objective function (*e.g.*, the AS-level path lengths, economic efficiency, or the profitability of transit providers). How does the topology that results from this centralized optimization compare with that formed from the distributed optimizations by ASes (*i.e.*, the result of ITER)? Is there a set of strategies for ASes that results in an internetwork that is close to the global optimal with respect to a metric of interest? Can we design mechanisms that operate in a distributed manner, implemented by autonomous networks, that lead the Internet towards "favorable" global states (by some definition of favorable)?

- **Interdomain traffic matrix estimation**

  In chapter 4, we described the interdependence between interdomain topology, traffic flow, and the provider and peer selection strategies of ASes. Further, we observed in chapter 4 that the nature of the interdomain traffic matrix has significant impact on the resulting steady-state network and the best strategies for STPs and LTPs. In fact, much of the work on modeling the Internet's AS level topology dynamics relies on an estimate of the interdomain traffic matrix. Unfortunately, however, there is little understanding of how the interdomain traffic matrix looks like, and how it evolves over time. We propose to measure the interdomain traffic matrix using data collected from a large tier-1 ISP. Assuming that a subset end-to-end traffic still flows through tier-1 ISPs, we would be able to directly measure a part of the interdomain traffic matrix. Further, the availability of data from multiple ISPs would improve our coverage of the entire interdomain traffic matrix. We will also develop techniques to improve the inference of some traffic matrix entries using AS topology data. This

measurement study can give important insights into the nature of this traffic matrix and how it evolves over time. We expect that this study will be an important input for future efforts to model the dynamics of the Internet.

- **Best response strategies**

  In chapter 4, we studied the effects of various provider and peer selection strategies by small and large transit providers. We studied the properties of the steady-state when all providers in a class (STPs or LTPs) used the same strategies for provider and peer selection. We did not, however, attempt to determine the best-response strategy that each class of providers should use to maximize their profitability, given the expected strategies that other networks would use. In future work, we plan to derive the best strategies that providers from different classes should adopt. We will also consider the case where each provider can use a different strategy. We also plan to extend the ITER model to scenarios where providers change their transit prices.

REFERENCES

[1] "Peering Database." http://www.peeringdb.com, Jan 2009.

[2] "Planetlab." http://planet-lab.org, Apr 2009.

[3] "Renesys Market Intelligence." http://www.renesys.com/products_services/market_intel/, Jan 2009.

[4] "TeleGeography: IP Transit Pricing Service." http://www.telegeography.com/cu/article.php?article_id=25445, Jan 2009.

[5] Akella, A., Seshan, S., and Shaikh, A., "Multihoming Performance Benefits: An Experimental Evaluation of Practical Enterprise Strategies," in *Proceedings of USENIX Annual Technical Symposium*, 2004.

[6] Akella, A., Maggs, B., Seshan, S., Shaikh, A., and Sitaraman, R., "A Measurement-Based Analysis of Multihoming," in *Proceedings of ACM SIGCOMM*, pp. 353–364, 2003.

[7] Akella, A., Pang, J., Maggs, B., Seshan, S., and Shaikh, A., "A Comparison of Overlay Routing and Multihoming Route Control," in *Proceedings of ACM SIGCOMM*, pp. 93–106, 2004.

[8] Akella, A., Seshan, S., and Shaikh, A., "An Empirical Evaluation of Wide-Area Internet Bottlenecks," in *Proceedings of Internet Measurement Conference (IMC)*, pp. 101–114, 2003.

[9] Akella, A., Seshan, S., and Shaikh, A., "An Empirical Evaluation of Wide-area Internet Bottlenecks," in *Proceedings of ACM SIGMETRICS*, pp. 316–317, 2003.

[10] Albert, R. and Barabasi, A. L., "Topology of Evolving Networks: Local Events and Universality," *Physical Review Letters 85, 5234*, 2000.

[11] Alderson, D., Doyle, J. C., Li, L., and Willinger, W., "Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications," *Internet Math, Vol. 2, Number 4*, 2005.

[12] Anagnostopoulos, A., Michel, L., Hentenryck, P. V., and Vergados, Y., "A Simulated Annealing Approach to the Traveling Tournament Problem," in *Proceedings of CPAIOR'03*, 2003.

[13] Anshelevich, E., Dasgupta, A., Tardos, E., and Wexler, T., "Near-optimal Network Design with Selfish Agents," in *STOC '03: Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pp. 511–520, 2003.

[14] Anshelevich, E., Shepherd, B., and Wilfong, G., "Strategic Network Formation through Peering and Service Agreements," in *FOCS '06: Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, (Washington, DC, USA), pp. 77–86, IEEE Computer Society, 2006.

[15] Barabasi, A. L. and Albert, R., "Emergence of Scaling in Random Networks," *Science 286 509512*, 1999.

[16] Bates, T. and Rekhter, Y., "Internet RFC 2260: Scalable Support for Multi-homed Multi-provider Connectivity," January 1998.

[17] Battista, G. D., Patrignani, M., and Pizzonia, M., "Computing the Types of Relationships Between Autonomous Systems," in *Proceedings of IEEE Infocom*, 2003.

[18] Bauer, J. M., "Dynamic Effects of Network Neutrality," *International Journal of Communication Vol.1, pp. 531-547, 2007.*

[19] Beverly, R., Bauer, S., and Berger, A., "The Internet Is Not a Big Truck: Toward Quantifying Network Neutrality," in *Proceedings of Passive and Active Measurement Conference (PAM)*, 2007.

[20] Bollobas, B. and Riordan, O., "The Diameter of a Scale-Free Random Graph," *Combinatorica*, vol. 24, no. 1, 2004.

[21] Bu, T. and Towsley, D., "On Distinguishing Between Internet Power Law Topology Generators," in *Proceedings of IEEE Infocom*, 2002.

[22] CAIDA, "The Skitter Project." `http://www.caida.org/tools/measurement/skitter/`, Jul 2005.

[23] Carlson, J. M. and Doyle, J., "Highly Optimized Tolerance: A Mechanism for Power Laws in Designed Systems," *Physical Review E 60*, 1999.

[24] Chang, H., Jamin, S., and Willinger, W., "Internet Connectivity at the AS-level: An Optimization-Driven Modeling Approach," in *Proceedings of ACM SIGCOMM Workshop on MoMeTools*, 2003.

[25] Chang, H., Jamin, S., and Willinger, W., "To Peer or Not to Peer: Modeling the Evolution of the Internet's AS-Level Topology," in *Proceedings of IEEE Infocom*, 2006.

[26] Chang, H. and Willinger, W., "Difficulties Measuring the Internet's AS-Level Ecosystem," in *Proceedings of the 40th Annual Conference on Information Sciences and Systems*, 2006.

[27] Chang, H., "An Economic-based Empirical Approach to Modeling the Internet's Interdomain Topology and Traffic Matrix," *Ph.D. Thesis*, 2006. Advisor Sugih Jamin.

[28] Chang, H., Jamin, S., Mao, Z. M., and Willinger, W., "An Empirical Approach to Modeling Inter-AS Traffic Matrices," in *Proceedings of the Internet Measurement Conference (IMC)*, pp. 12–12, 2005.

[29] Chen, Q., Chang, H., Govindan, R., Jamin, S., Shenker, S., and Willinger, W., "The Origin of Power-Laws in Internet Topologies Revisited," in *Proceedings of IEEE Infocom*, 2002.

[30] Cho, K., Fukuda, K., Esaki, H., and Kato, A., "The Impact and Implications of the Growth in Residential User-to-user Traffic," in *Proceedings of ACM SIGCOMM*, pp. 207–218, 2006.

[31] Cisco Systems, "Optimized Edge Routing (OER)."

[32] Cohen, R. and Raz, D., "The Internet Dark Matter - On the Missing Links in the AS Connectivity Map," in *Proceedings of IEEE Infocom*, 2006.

[33] Corbo, J., Jain, S., Mitzenmacher, M., and Parkes, D., "An Economically Principled Generative Model of AS Graph Connectivity," in *Proceedings of the International Joint Workshop on The Economics of Networked Systems and Incentive-Based Computing*, 2007.

[34] Crovella, M. E. and Taqqu, M. S., "aest: A Tool For Estimating the Heavy Tail Index from Scaling Properties." `http://www.cs.bu.edu/faculty/crovella/aest.html`, Jul 2005.

[35] Crowcroft, J., "Net Neutrality: The Technical Side of the Debate: A White Paper," *ACM SIGCOMM Computer Communication Review (CCR)*, 2007.

[36] Dhamdhere, A. and Dovrolis, C., "ISP and Egress Path Selection for Multihomed Networks," in *Proceedings of IEEE Infocom*, 2006.

[37] Dhamdhere, A. and Dovrolis, C., "A Model for Interdomain Network Formation, Economics and Routing," *under review at ACM SIGCOMM*, 2008.

[38] Dhamdhere, A. and Dovrolis, C., "Can ISPs be Profitable without Violating Network Neutrality?," *Proceedings of ACN Sigcomm Workshop on the Economics of Networks (NetEcon)*, 2008.

[39] Dhamdhere, A. and Dovrolis, C., "Ten Years in the Evolution of the Internet Ecosystem," in *In the Proceedings of ACM SIGCOMM Internet Measurement Conference (IMC)*, 2008.

[40] Dimitropoulos, X., Krioukov, D., Fomenkov, M., Hyun, Y., Claffy, K., and Riley, G., "AS Relationships: Inference and Validation," *ACM SIGCOMM Computer Communication Review (CCR)*, 2007.

[41] Dimitropoulos, X., Krioukov, D., Riley, G., and Claffy, K., "Revealing the Autonomous System Taxonomy: The Machine Learning Approach," in *Proceedings of Passive and Active Measurement Conference (PAM)*, 2006.

[42] Dimitropoulos, X. and Riley, G., "Modeling Autonomous-System Relationships," in *PADS '06: Proceedings of the 20th Workshop on Principles of Advanced and Distributed Simulation*, 2006.

[43] Economides, N., "The Economics of the Internet Backbone," *Handbook of Telecommunications Economics Ed. S. Majumdar, I. Vogelsang, M. Cave. Amsterdam: Elsevier Publishers*, 2006.

[44] Economides, N. and Tag, J., "Net Neutrality on the Internet: A Two-Sided Market Analysis," *SSRN eLibrary*, 2007.

[45] f5 networks, "BIG-IP Link Controller." `http://www.f5.com/solutions/tech/multi_homing.html`, Jul 2005.

[46] Fabrikant, A., Koutsoupias, E., and Papadimitriou, C. H., "Heuristically Optimized Trade-Offs: A New Paradigm for Power Laws in the Internet," in *Proceedings of ICALP*, 2002.

[47] Faloutsos, M., Faloutsos, P., and Faloutsos, C., "On Power-law Relationships of the Internet Topology," in *Proceedings of ACM SIGCOMM*, 1999.

[48] FatPipe, "WARP." `http://www.fatpipeinc.com/warp/index.htm`, Jul 2005.

[49] Felten, E. W., "Nuts and Bolts of Network Neutrality," tech. rep., 2006.

[50] Frieden, R., "Network Neutrality or Bias? Handicapping the Odds for a Tiered and Branded Internet," *Bepress Legal Series. Working Paper 1755.*, 2006.

[51] Gao, L., "On Inferring Autonomous System Relationships in the Internet," *IEEE/ACM Transactions on Networking*, vol. 9, no. 6, 2001.

[52] Gao, L. and Wang, F., "The Extent of AS Path Inflation by Routing Policies," in *Proceedings of IEEE Global Telecommunications Conference, GLOBECOM*, 2002.

[53] Garey, M. R. and Johnson, D. S., *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., 1979.

[54] Gill, P., Arlit, M., Li, Z., and Mahanti, A., "The Flattening Internet Topology: Natural Evolution, Unsightly Barnacles or Contrived Collapse?," in *Proceedings of Passive and Active Measurement Conference (PAM)*, 2008.

[55] Goldenberg, D. K., Qiuy, L., Xie, H., Yang, Y. R., and Zhang, Y., "Optimizing Cost and Performance for Multihoming," in *Proceedings of ACM SIGCOMM*, pp. 79–92, 2004.

[56] Guo, F., Chen, J., Li, W., and Cker, T., "Experiences in Building a Multihoming Load Balancing System," in *Proceedings of IEEE Infocom*, 2004.

[57] Haddadi, H., Uhlig, S., Moore, A., Mortier, R., and Rio, M., "Modeling Internet Topology Dynamics," *ACM SIGCOMM Computer Communication Review (CCR)*, 2008.

[58] Hahn, R. W. and Litan, R. E., "The Myth of Network Neutrality and What We Should Do About It," tech. rep., 2006.

[59] He, Y., Siganos, G., Faloutsos, M., and Krishnamurthy, S. V., "A Systematic Framework for Unearthing the Missing Links: Measurements and Impact," in *Proceedings of 4th USENIX/SIGCOMM NSDI*, 2007.

[60] Holme, P., Karlin, J., and Forrest, S., "An Integrated Model of Traffic, Geography and Economy in the Internet," *ACM SIGCOMM Computer Communication Review (CCR)*, vol. 38, no. 3, pp. 5–16, 2008.

[61] Huston, G., "The 32-bit AS Number Report." `http://www.potaroo.net/tools/asn32`, Apr 2009.

[62] Ingber, L., "Simulated annealing: Practice versus Theory," Tech. Rep. 93sa, Lester Ingber, 1993. available at http://ideas.repec.org/p/lei/ingber/93sa.html.

[63] InterNAP, "Premise-Based Route Optimization." `http://www.internap.com/products/route-optimization.htm`, Jul 2005.

[64] Johari, R., Mannor, S., and Tsitsiklis, J., "A Contract-based Model for Directed Network Formation," *Games and Economic Behavior*, vol. 56, pp. 201–224, August 2006.

[65] Johnson, D., "Approximation Algorithms For Combinatorial Problems," in *Jnl. of Comp. and System Sciences*, 1974.

[66] Jr., E. G. C., Garey, M. R., and Johnson, D. S., "Approximation Algorithms for Bin Packing: A Survey," *Approximation Algorithms for NP-Hard Problems*, pp. 46–93, 1997.

[67] Keller, E. F., "Revisiting "Scale-free" Networks," *BioEssays 27, Wiley Periodicals Inc.*, 2005.

[68] Kirkpatrick, S., "Optimization by Simulated Annealing: Quantitative Studies," *Journal of Statistical Physics 34:975–986*, 1984.

[69] Kirkpatrick, S., Gelatt, C., and Vecchi, M., "Optimization by Simulated Annealing," *Science, Number 4598, 13 May 1983*, vol. 220, 4598, pp. 671–680, 1983.

[70] Krioukov, D., kc claffy, Fomenkov, M., Chung, F., Vespignani, A., and Willinger, W., "The Workshop on Internet Topology (wit) Report," *ACM SIGCOMM Computer Communication Review (CCR)*, vol. 37, no. 1, 2007.

[71] Kumar, R., Novak, J., and Tomkins, A., "Structure and Evolution of Online Social Networks," in *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006.

[72] Leskovec, J., Kleinberg, J., and Faloutsos, C., "Graph Evolution: Densification and Shrinking Diameters," *ACM Transactions on Knowledge Discovery from Data (ACM TKDD)*, 2007.

[73] Ma, R. T. B., Chiu, D., Lui, J. C. S., Misra, V., and Rubenstein, D., "Internet Economics: the use of Shapley Value for ISP Settlement," in *Proceedings of the ACM Conference on Emerging network experiment and technology (CoNEXT)*, 2007.

[74] Ma, R. T., Chiu, D., Lui, J. C., Misra, V., and Rubenstein, D., "Interconnecting Eyeballs to Content: A Shapley Value Perspective on ISP Peering and Settlement," in *Proceedings of the ACM SIGCOMM 2008 Workshop on Economics of Networked Systems (NetEcon)*, 2008.

[75] Ma, R. T., Chiu, D., Lui, J. C., Misra, V., and Rubenstein, D., "On Cooperative Settlement Between Content, Transit and Eyeball Internet Service Providers," in *Proceedings of the ACM Conference on Emerging network experiment and technology (CoNEXT)*, 2008.

[76] Magoni, D. and Pansiot, J. J., "Analysis of the Autonomous System Network Topology," *ACM SIGCOMM Computer Communication Review (CCR)*, 2001.

[77] Mahadevan, P., Krioukov, D., Fomenkov, M., Huffaker, B., Dimitropoulos, X., kc claffy, and Vahdat, A., "The Internet AS-Level Topology: Three Data Sources and One Definitive Metric," *ACM SIGCOMM Computer Communication Review (CCR)*, 2005.

[78] Metropolis, R., Rosenbluth, A., Teller, A., and Teller, E., "Simulated Annealing," in *Journal of Chemical Physics, 21:1087–1092*, 1953.

[79] Monderer, D. and Shapley, L. S., "Potential games," in *Games and Economic Behavior*, vol. 14, 1996.

[80] Nahar, S., Sahni, S., and Shragowitz, E., "Simulated Annealing and Combinatorial Optimization," in *DAC '86: Proceedings of the 23rd ACM/IEEE conference on Design automation*, (Piscataway, NJ, USA), pp. 293–299, IEEE Press, 1986.

[81] Nortel networks, "Alteon Link Optimizer." `http://www.nortelnetworks.com/products/01/alteon/optimizer/`, Jul 2005.

[82] Norton, W. B., "A Business Case for ISP Peering," *Equinix white papers*, 2002.

[83] Norton, W. B., "The Art of Peering: The Peering Playbook," *Equinix white papers*, 2002.

[84] Norton, W. B., "The Evolution of the U.S. Internet Peering Ecosystem," *Equinix white papers*, 2004.

[85] Norton, W. B., "Transit Cost Survey." `www.nanog.org/mtg-0606/pdf/bill.norton.2.pdf`, Jul 2006.

[86] Norton, W. B., "Video Internet: The Next Wave of Massive Disruption to the U.S. Peering Ecosystem," *Equinix white papers*, 2007.

[87] Oliveira, R., Pei, D., Willinger, W., Zhang, B., and Zhang, L., "In Search of the Elusive Ground Truth: The Internet's AS-level Connectivity Structure," in *Proceedings of ACM SIGMETRICS*, 2008.

[88] Oliveira, R. V., Zhang, B., and Zhang, L., "Observing the Evolution of Internet AS Topology," in *Proceedings of ACM SIGCOMM*, 2007.

[89] Orda, A. and Rom, R., "Multihoming in Computer Networks: a Topology-design Approach," *Comput. Netw. ISDN Syst.*, vol. 18, no. 2, pp. 133–141, 9/90.

[90] Park, S., Pennock, D. M., and Giles, C. L., "Comparing Static and Dynamic Measurements and Models of the Internet's AS Topology," in *Proceedings of IEEE Infocom*, 2004.

[91] Radware, "Peer Director." `http://www.radware.com/content/products/pd/`, Jul 2005.

[92] Rainfinity, "RainConnect." `http://www.rainfinity.com/products/rainconnect.html`, Jul 2005.

[93] Rether networks, "Internet Service Management Device." `http://rether.com/ISMD.htm`, Jul 2005.

[94] RIPE, "RIPE Network Coordination Centre." http://www.ripe.net, Apr 2009.

[95] Route Science, "Adaptive Networking Software." http://www.avaya.com/gcm/master-usa/en-us/products/offers/adaptive_networking_software.htm, Jul 2005.

[96] Routeviews, "University of Oregon Route Views Project." http://www.routeviews.org, Apr 2009.

[97] Serrano, M. A., Boguna, M., and Guilera, A. D., "Modeling the Internet," *The European Physics Journal B*, 2006.

[98] Sidak, J. G., "A Consumer-Welfare Approach To Network Neutrality Regulation of the Internet," *Journal of Competition Law and Economics, Vol. 2, pp. 349-474, 2006.*

[99] Siganos, G., Faloutsos, M., and Faloutsos, C., "The Evolution of the Internet: Topology and Routing," *University of California, Riverside technical report*, 2002.

[100] Stonesoft, "StoneGate Multi-Link Technology." http://www.stonesoft.com/products/ISP_Multi-homing, Jul 2005.

[101] Subramanian, L., Agarwal, S., Rexford, J., and Katz, R., "Characterizing the Internet hierarchy from Multiple Vantage Points," in *Proceedings of IEEE Infocom*, 2002.

[102] Tangmunarunkit, H., Doyle, J., Govindan, R., Willinger, W., Jamin, S., and Shenker, S., "Does AS Size Determine Degree in AS Topology?," *ACM SIGCOMM Computer Communication Review (CCR)*, 2001.

[103] Tao, S., Xu, K., Xu, Y., Fei, T., Gao, L., Guerin, R., and D. Towsley, J. K., and Zhang, Z., "Exploring the Performance Benefits of End-to-End Path Switching," in *Proceedings of IEEE ICNP*, 2004.

[104] Van Schewick, B., "Towards an Economic Framework for Network Neutrality Regulation," *Journal on Telecommunications and High Technology Law, Vol. 5, pp. 329-391, 2007.*

[105] Vazquez, A., Pastor-Satorras, R., and Vespignani, A., "Large-scale Topological and Dynamical Properties of the Internet," *Physical Review E, vol. 65*, 2002.

[106] Wang, H., Xie, H., Qiu, L., Silberschatz, A., and Yang, Y. R., "Optimal ISP Subscription for Internet Multihoming: Algorithm Design and Implication Analysis," in *Proceedings of IEEE Infocom*, 2005.

[107] Wang, X. and Loguinov, D., "Wealth-Based Evolution Model for the Internet AS-Level Topology," in *Proceedings of IEEE Infocom*, 2006.

[108] Wu, T., "Network Neutrality, Broadband Discrimination," *Journal of Telecommunications and High Technology Law, Vol. 2, p. 141, 2003.*

[109] Yang, X., Tsudik, G., and Liu, X., "A Technical Approach to Net Neutrality," in *Proceedings of Fifth Workshop on Hot Topics in Networks (Hotnets-V)*, 2006.

[110] Yook, S. H., Jeong, H., and Barabasi, A. L., "Modeling the Internet's Large-scale Topology," *Proceedings of the National Academy of Sciences*, 2002.

[111] Zhang, B., Liu, R., Massey, D., and Zhang, L., "Collecting the Internet AS-level Topology," *ACM SIGCOMM Computer Communication Review (CCR)*, 2005.

[112] Zhang, Y., Zhang, Z., Mao, Z. M., Hu, C., and Maggs, B. M., "On the Impact of Route Monitor Selection," in *Proceedings of Internet Measurement Conference (IMC)*, 2007.

[113] Zhou, S., "Understanding the Evolution Dynamics of Internet Topology," *Physical Review E, vol. 74*, 2006.

[114] Zhou, S. and Mondragon, R., "Accurately Modeling the Internet Topology," *Physical Review E, vol. 70*, 2004.