Understanding Internet Topology: Principles, Models, and Validation

David Alderson, *Member, IEEE*, Lun Li, *Student Member, IEEE*, Walter Willinger, *Fellow, IEEE*, and John C. Doyle, *Member, IEEE*

***Abstract*—Building on a recent effort that combines a first-principles approach to modeling router-level connectivity with a more pragmatic use of statistics and graph theory, we show in this paper that for the Internet, an improved understanding of its physical infrastructure is possible by viewing the physical connectivity as an annotated graph that delivers raw connectivity and bandwidth to the upper layers in the TCP/IP protocol stack, subject to practical constraints (e.g., router technology) and economic considerations (e.g., link costs). More importantly, by relying on data from Abilene, a Tier-1 ISP, and the Rocketfuel project, we provide empirical evidence in support of the proposed approach and its consistency with networking reality. To illustrate its utility, we: 1) show that our approach provides insight into the origin of high variability in measured or inferred router-level maps; 2) demonstrate that it easily accommodates the incorporation of additional objectives of network design (e.g., robustness to router failure); and 3) discuss how it complements ongoing community efforts to reverse-engineer the Internet.**

***Index Terms*—Degree-based generators, heuristically optimal topology, network design, network topology, router configuration, topology metrics.**

I. Introduction

A DETAILED understanding of the many facets of the Internet's multiscale structure is fundamental to many important network research problems. These include evaluating the performance of networking protocols, assessing the effectiveness of proposed techniques to protect the network from nefarious intrusions and attacks, and developing improved designs for resource provisioning. In each case, there is a need for realistic models of Internet topology at different levels of detail or scale. For example, when evaluating the performance of next-generation congestion-control protocols, annotated models of topology that reflect IP-level connectivity and include link bandwidths and buffer sizes may be sufficient. However, a more detailed description of the underlying physical structure (including, for example, not only link bandwidths but also node/router capacities) may be required for assessing network vulnerability to physical attacks, for planning the expansion of network capacity, or for reassessing the efficiency of an existing infrastructure in light of new technological advances. Other problems may necessitate different levels of abstraction altogether, with higher layers of the TCP/IP protocol stack defining their own and increasingly more virtual network topologies. For instance, when investigating the behavior of inter-domain routing protocols, it is often more important to focus on annotated models of logical Internet connectivity at the level of Autonomous Systems (ASes) that reflect AS peering relationships and routing policies than on the physical connectivity within an AS. At an even higher layer, the connectivity structure defined by documents (nodes) and hyperlinks (connections) in the WWW is essentially unconstrained, largely separate from the physical connectivity, and of practical importance for evaluating the performance of search algorithms or engines.

Since the Internet is a collection of thousands of smaller networks, each under its own administrative control, there is no single place from which one can obtain a complete picture of its topology. Also, the fear of losing competitive advantage has provided a strong disincentive for network owners and operators to share topology information. Thus, because direct inspection of the network is generally not possible, the task of "discovering" the Internet's topology has been left to researchers who have attempted to infer its structure using both empirical and theoretic approaches. Experimentalists have developed more or less sophisticated methods to reverse-engineer topological features and structures from appropriate network measurements [17], [25], [35], [36]. At the same time, a more theoretical approach has focused on exploiting phenomenological and graph-theoretic descriptions of large-scale network structure and evaluating the ability of synthetic topology generators to reproduce them [13], [27], [29], [45]. Section II provides a brief account of these approaches and of related work. However, each of these approaches suffers because the multitude of potentially relevant network properties makes it difficult to assess what features either characterize or are most essential to network structure and behavior. On the one hand, the elaborate nature of the network protocol suite means that there are many possible measurements that can be made, each having its own strengths, weaknesses, and idiosyncrasies, and each resulting in a potentially distinct view of the network topology. On the other hand, there are a variety of distinctly different graph generation methods that might give rise to the same large-scale statistics, but some of the resulting models may be purely descriptive, with no network-intrinsic meaning whatsoever.

Manuscript received January 21, 2005; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor E. Zegura. This work was supported in part by Boeing, the Air Force Office of Scientific Research (AFOSR), and Caltech's Lee Center for Advanced Networking. Parts of this work were done at the Institute of Pure and Applied Mathematics (IPAM) at UCLA, as part of the 2002 annual program on Large-Scale Communication Networks. Partial and preliminary results appeared in the Proceedings of ACM SIGCOMM, Portland, OR, August 2004.

D. Alderson, L. Li, and J. C. Doyle are with the California Institute of Technology, Pasadena, CA 91125 USA (e-mail: alderd@cds.caltech.edu; lun@cds.caltech.edu; doyle@cds.caltech.edu).

W. Willinger is with AT&T Labs–Research, Florham Park, NJ 07932 USA (e-mail: walter@research.att.com).

Digital Object Identifier 10.1109/TNET.2005.861250

In this paper, we discuss a concrete and complementary approach to Internet topology modeling that explicitly accounts for the inevitable tradeoff between model complexity and fidelity. We build on recent work by Li *et al.* [26] (see also Doyle *et al.* [19]) who considered a *first-principles approach* to modeling Internet topology at the *router level*. More precisely, since for router-level related issues such as performance, reliability, and robustness to component loss, the physical connectivity between routers is more important than the virtual connectivity as defined by the higher layers of the protocol stack (e.g., MPLS, IP), when referring in the following to router-level connectivity, we always mean Layer 2, especially when the distinction between Layer 2 versus Layer 3 issues is important for the purpose of illuminating the nature of the actual router-level connectivity (i.e., node degree) and its physical constraints. While in the following our focus is always on single ISPs or ASes, we observe that the latter form the Internet's fundamental building blocks that are designed largely in isolation and then connected according to both engineering and business considerations to yield the Internet's global router-level connectivity. Finally, the notion of "first-principles approach" refers to a systematic attempt at distilling from the seemingly endless list of potentially relevant technological, economic, or other features and design objectives those that are most essential to a solid understanding of the intrinsic fundamentals of network topology. By identifying a set of minimal functional requirements and physical constraints, Li *et al.* [26] demonstrated how to develop simple models of the Internet's router-level topology that are simultaneously illustrative, insightful, and consistent with engineering reality.

Besides summarizing and elaborating in Section III on this first-principles approach to understanding Internet topology structure, the main focus of this paper is on providing additional empirical evidence in support of the engineering aspects of the proposed approach and its consistency with networking reality. To this end, we consider in Section IV carefully designed state-of-the-art educational networks like Abilene as well as a number commercial networks that are operated by Tier-1 ISPs. These have experienced dramatic growth during the last decade, and they typically accommodate a range of different network technologies, from legacy ATM networks to the latest optical networking technology. We rely on measurements that are either publicly available and exact (e.g., Abilene), proprietary and exact but having some level of aggregation/anonymization (i.e., a Tier-1 ISP), or publicly available and only approximate (e.g., Rocketfuel-derived router-level maps [35]). The picture that emerges from analyzing these different types of measurements reveals that the proposed first-principles approach captures key features of the observed or inferred router-level connectivity of today's Internet and is in stark contrast to existing alternative approaches to Internet topology modeling that favor random constructions that reproduce large-scale statistical connectivity features over the functional requirements and physical constraints that dominate networking reality.

The implications for networking theory and practice of the first-principles approach advocated in this paper are far-reaching, and some of them are discussed in detail in Section V. For example, we show that as the name suggests, this first-principles approach can easily incorporate new design objectives, such as robustness to router failures. Indeed, making this objective part of our heuristic network design described in Section III requires minimally adding some link redundancy and incorporating a simple abstraction of IP routing to account for the feedback mechanism by which the real network "sees damage and works around it." Illustrating this in the context of the toy networks considered in Section III reveals the specious nature of the sensational "Achilles' heel" claim for the Internet that originated with the scale-free network modeling approach popular with physicists [3]. We also illustrate how the first-principles approach complements ongoing measurement based efforts to reverse-engineer the Internet [36], thus providing a natural means to integrate these efforts with the development of next-generation topology generators capable of creating "realistic, yet fictitious" representations of annotated Internet connectivity maps at different levels of detail. We conclude in Section VI by discussing the benefits and shortcomings of the proposed first-principles approach, along with an outlook for future topology modeling.

II. Background and Related Work

The development of abstract, yet informed, models for network topology evaluation and generation has been largely empirical in nature. For example, the first popular topology generator to be used for networking simulation was the Waxman model [42], which is a variation of the classical Erdös–Rényi random graph [20]. Motivated by the general observation that long-range links are expensive, this model places nodes at random in a two-dimensional space and adds links probabilistically between each pair of nodes in a manner that is inversely proportional to their distance. However, the use of this type of random graph model was soon abandoned in favor of models that explicitly introduce nonrandom structure, particularly hierarchy and locality, as part of the network design [8], [18], [46]. The argument for this type of approach was based on the fact that an inspection of real networks shows that they are clearly not random but do exhibit certain obvious hierarchical features. This approach further argued that a topology generator should reflect the design principles in common use. For example, in order to achieve desired performance objectives, the network must have certain connectivity and redundancy requirements, properties which are not guaranteed in random network topologies. These principles were integrated into the Georgia Tech Internetwork Topology Models (GT-ITM).

These *structural topology generators* were the standard models until significant efforts by the networking community provided further insight into the large-scale statistical properties of Internet-specific topologies [17], [22], [25], [33], [35], [38]. These studies have typically focused on statistics related to the *connectivity* of network components, whether they be machines in the router-level graph [25] or entire subnetworks or ASes in the AS-level graph [11], [23]. A particular feature of network connectivity that has attracted considerable attention is the prevalence of heavy-tailed distributions in node *degree* (i.e., number of connections) and whether or not these heavy-tailed distributions conform to power-law distributions [13], [22], [33], [29], [30]. For example, power-law node degree distributions figure prominently in the so-called *scale-free* network models [5], which have been a popular theme in the study of

complex networks, particularly among researchers inspired by statistical physics [2], [31], [32]. This macroscopic statistic not only captures in a parsimonious manner a prominent feature of many real-world networks, namely that most nodes have very few connections and a few nodes have lots of connections, but it has also greatly influenced the recent generation and evaluation of network topologies. In the current environment, node degree distributions and other large-scale statistics are popular metrics for evaluating how representative a given topology is [39], and proposed topology generators are often evaluated on the basis of whether or not they can reproduce the same types of macroscopic statistics, especially power law-type node degree distributions [7]. Since the GT-ITM topology generators fail to produce power laws in node degree, they have recently been abandoned in favor of a new class of *degree-based* generators (see [26] for a partial list) that explicitly replicate these observed statistics.[1]

The popularity of these generators notwithstanding, this emphasis on power-law node degree distributions and the resulting efforts to generate and explain them with the help of newly developed models have not gone without criticism [19], [43]–[45]. Nevertheless, in the absence of concrete examples of alternate models, degree-based methods have remained popular representations for large-scale Internet structure. However, recent work reported in [26] has shown that the perspective offered from the recent degree-based models is both incomplete and can sometimes be misleading or even flawed. For one, there exist many different graphs having the *same node degree distribution*, some of which may be considered *opposites* from the viewpoint of network engineering. Furthermore, as discussed in detail in [27], there are a variety of distinctly different random graph models that might give rise to a given degree distribution, and some of these models may have no network-intrinsic meaning whatsoever. In spirit, this work delivered for degree-based networks a similar message as [46] did for the random graph-type models [42] that were popular with networking researchers in the early 1990s. While [46] identified and commented on the inherent limitations of the various constructs involving Erdös–Rényi-type random graphs, [26] points toward similar shortcomings and unrealistic features when working with probabilistic degree-based graphs.

III. A First Principles Approach

In essence, the first-principles approach developed in [26] starts out by asking the question "what really matters when it comes to topology construction?" and concludes that for the Internet, significant progress can be made by thinking of the physical topology as well-modeled by an annotated graph, with the basic functional requirement of delivering raw connectivity and bandwidth to the higher layers in the TCP/IP protocol stack, subject to simple constraints imposed by the available hardware and the economics of network design. Following closely the presentation in [26], we show that when combined with a more pragmatic view of statistics and graph theory, this approach is capable of providing a perspective that is consistent both with observed measurements and the engineering principles at work in network design.

[1]See however a comment by E. Zegura on router-level topology modeling at http://www.caida.org/analysis/topology/router-level-topology.xml

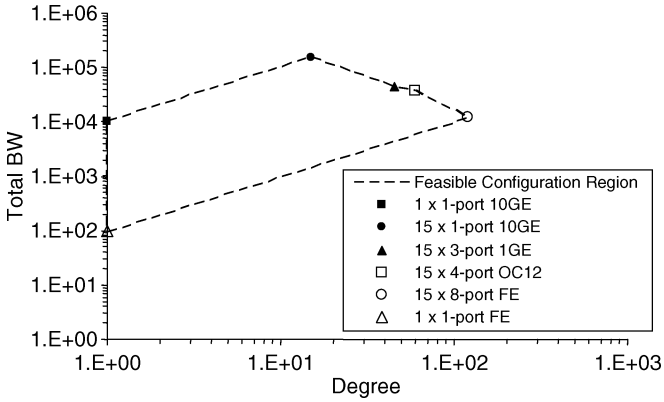


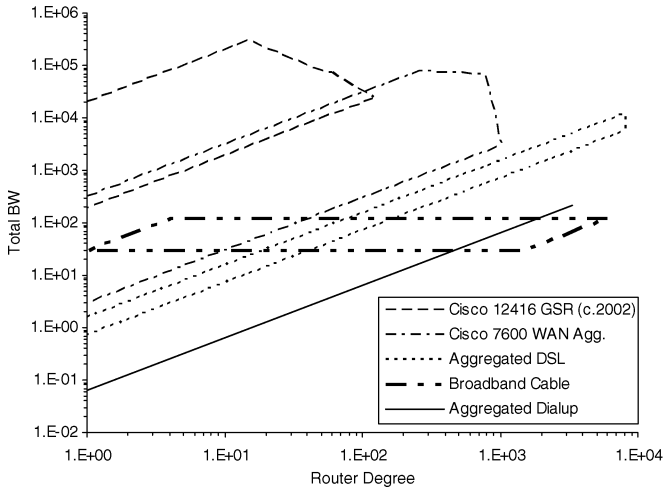
Fig. 1. Technology constraint for Cisco 12416 Gigabit Switch Router (GSR): feasible configuration region as of June 2002. Each point on the plot corresponds to a different combination of physical interface cards and interfaces for the same router. This router has 15 available line card slots. When the router is configured to have less than 15 connections, throughput per degree is limited by the line-card maximum speed (10 Gb/s) and the total bandwidth increases with the number of connections. When the number of connections is greater than 15, the total router bandwidth decreases as the total number of connections increases, up to a maximum of 120 possible connections for this router.

A. Practical Constraints

A key challenge in using large-scale statistical features to characterize something as complex as the topology of an ISP or the Internet as a whole is that it is difficult to understand the extent to which any particular observed feature is "fundamental" to its structure. Here, we consider a complementary approach for thinking about network topology, in which we explore some of the practical constraints and tradeoffs at work in the construction of real networks.

1) Networking Technology: A closer look at the physical topology of the Internet reveals that the underlying *router technology constraints* are a significant force shaping network connectivity. Based on the technology used in the cross-connection fabric of the router itself, a router has a maximum number of packets that can be processed in any unit of time. This means that there is an inherent tradeoff between the number of physical link connections (i.e., node *degree*) and connection speeds (i.e., bandwidth) at each router.

Most high-speed routers are designed in a modular fashion such that the router's chassis (containing the cross-connection fabric) is independent from individual *physical interface cards (PICs)*, each of which must be installed on the chassis via one of a finite number of *slots*. Each PIC will have one or more *ports* that provide the interface for an individual physical cable. In this manner, the *configuration* of a router is defined in terms of the quantity and type of PICs that are installed, and the possible number of physical connections on a router is therefore limited by its number of slots and the *port density* of available PICs. Thus, a router can have a few high bandwidth connections or many low bandwidth connections (or some combination in between), and this type of *conservation law* can be represented



WAN Aggregation: Cisco 7600, 8 slots		
8 x 8-port OC12 channelized into OC3 (155 Mbps)	= 256 x OC3	= 39813 Mbps
8 x 8-port OC12 channelized into DS3 (44.7 Mbps)	= 768 x DS3	= 34357 Mbps
8 x 8-port OC12 channelized into DS1 (1.5 Mbps)	= 1008 x DS1	= 1556 Mbps

Broadband Cable: Cisco uBR700 Cable Module Termination System (CMTS), 4slots	
each card shares 30 Mbps among 1-1500 users	= 80 users @ 1.5 Mbps
	= 160 users @ 768 kbps
	= 1875 users @ 64 kbps

Digital Subscriber Line (DSL): Cisco 6260 Access Concentrator (DSLAM)	
OC3 interface to network backbone	= ~100 users @ 1.5 Mbps
	= ~200 users @ 768 kbps
when used in combination with WAN aggregation	= ~8000 users @ 1.5 Mbps

Dial-up Aggregation	
example: Cisco AS5850 Access Switch	= 3360 users @ 56kbps

Fig. 2. Left: aggregate picture of feasible configuration region for different router technologies. In addition to the Cisco 12416 GSR, the configuration region for the Cisco 7600 wide area network (WAN) aggregation router is also shown along with routing products that support DSL, broadband cable, and dialup aggregation. Right: degree-bandwidth configurations supported by core routers, access routers, and edge technologies. The choice of these particular products is meant to be representative, not comprehensive. Of note for the consumer market, the shared access technology for broadband cable provides service comparable to DSL when the total number of users is about 100, but can only provide service equivalent to dialup when the number of users is about 2000.

in terms of a "feasible configuration region" of possible bandwidth-degree combinations for each router.[2]

Fig. 1 shows the feasible configuration region for the Cisco 12416 GSR, which is one of the most expensive and highest bandwidth routers available from a 2002 Cisco product catalog [41]. Note that because of minimum line speeds (e.g., 100 Mb/s) on available PICs, the feasible configuration region defines both an upper and lower bound on bandwidth-degree combinations. We often refer to the upper bound of the feasible region as the "efficient frontier" since it represents router configurations that have maximized some bandwidth-degree tradeoff. While it is always possible to configure the router so that it falls below the efficient frontier (thereby under-utilizing the router capacity), it is not possible to exceed this frontier (e.g., by having many high bandwidth connections). Although engineers are constantly expanding the efficient frontier with the development of new routing technologies (i.e., higher switching capacity in the chassis, faster ports on individual PICs, or PICs with higher port-density), each particular router model will have a frontier representing its feasible region. Network architects are faced with tradeoffs between capacity and cost in selecting a router chassis and the corresponding interface cards that configure it. Until new technology shifts the frontier, the only way to create throughput beyond the feasible region is to build networks of routers.[3]

The current Internet is populated with many different router models, each using potentially different technologies and each having their own feasible configuration region. However, these technologies are still constrained in their overall ability to tradeoff total bandwidth and number of connections. Fig. 2 presents a simplified picture of the feasible configuration regions for several different types of routing products in common use and provides a summary of the underlying data that support the different configurations. In addition to the Cisco 12416 GSR (intended here to represent the class of *core routers*), we present the possible degree-bandwidth configurations for the Cisco 7600 series aggregation router (intended here to represent the class of *access routers*) as well as several common *edge technologies* (DSL, broadband cable, and dial-up). While we will discuss the implication for each of these product groups in subsequent sections, the message here is that these router products naturally specialize into a few different "roles" which in turn can be understood in terms of their feasible configuration regions. Core routers like the Cisco 12416 GSR tend to specialize in supporting the highest available link speeds (but can only handle a relatively few such connections), while access routers like the Cisco 7600 are designed to support many more connections (but at necessarily lower speeds). Edge technologies are somewhat different in their underlying design, since their intention is to support significantly more end users at fixed (DSL, dialup) or variable (cable) speeds. Details of the technologies in use at the network edge and their impact on network design can be found from [4]. Collectively, these individual constraints form an overall aggregate constraint on available topology design.

2) Economics of Network Design: Besides technological considerations affecting router use and configuration, another key factor affecting network topology concerns the economic considerations of network design and deployment, which are mainly driven by customer demands and ultimately direct the types of technologies that are developed for use by network providers. For example, the cost of installing and operating physical links in a network can dominate the cost of the overall infrastructure, and since these costs tend to increase with link distance, there is strong practical incentive to design wired networks such that they can support traffic using the fewest

[2]Our convention throughout this paper is to compute total router bandwidth as the sum of bandwidth for installed PICs. While router vendors tend to count the bandwidth of PICs twice (since they support duplex traffic), we believe our approach is more transparent, and it does not change the analysis or its implications.

[3]Recent product announcements from router manufacturers such as Juniper Networks, Avici Systems, and Cisco Systems suggest that the latest trend in technology development is to build scalable multi-rack routers that do exactly this. However, it remains to be seen whether or not the economics (including configuration and management) for these products will enable their wide deployment in the Internet.

number of links. The ability to share costs via multiplexing is a fundamental driver underlying the design of networking technologies, and the availability of these technologies enables a network topology in which traffic is aggregated at all levels of network hierarchy, from its periphery up to its core.

The development of these technologies has similarly followed the demands of customers, for whom there is high variability in the willingness to pay for network bandwidths. As of 2004, nearly half of all users of the Internet in North America still had dial-up connections (generally 56 kb/s), a similar number had broadband access (256 kb/s–6 Mb/s), and only a small number of users had large ($\geq$10 Gb/s) bandwidth requirements [4]. Again, the cost-effective handling of such diverse end user traffic requires that aggregation take place as close to the edge as possible and is explicitly supported by a common feature that these edge technologies have, namely a special ability to support high connectivity in order to aggregate end user traffic before sending it toward the core. Based on the high variability in population density, it is not only plausible but somewhat expected that there exists high variability in network connectivity. However, the latter is by and large due to the connectivity observed at the network's edge and cannot possibly be the result of the connectivity pattern typically encountered in the network core. Note that all that matters for this discussion is the notion of high variability in network connectivity, and whether or not the latter satisfies some power-law rank-size relationship is irrelevant, though it will be interesting to verify this as well by analyzing accurate ISP backbone networks.

B. Functional Requirements

The primary purpose for building a network is to provide connectivity and to carry effectively a projected overall traffic demand. This observation yields at once two very different metrics for comparing networks that are the same in some respects (e.g., number of nodes and/or links, node degree sequence), but possibly very different otherwise. On the one hand, we define a metric for *network performance* that allows for an engineering-based comparison for how different networks handle one and the same traffic demand. On the other hand, we consider *network likelihood*, a strictly graph-theoretic metric that requires no annotated graph but can be computed for any network for which connectivity information is available.

1) A Performance-Inspired Topology Metric: Consider a network g that is simple (i.e., having no loops or parallel edges), connected, and whose links and nodes are annotated and specify link bandwidth and router type. We define *network performance* as the maximum throughput of g under a gravity model of end user traffic demands [47], subject to some router degree-bandwidth constraints. That is, we consider flows on all source-destination pairs of edge routers, such that the amount of flow X_{ij} between source i and destination j is proportional to the product of the traffic demand x_i, x_j at end points i, j, $X_{ij} = \rho x_i x_j$, where ρ is some constant, and is otherwise uncorrelated from all other flows. Our performance measure $Perf(g)$ for a given network g is then its maximum throughput with gravity flows, computed as

$$Perf(g) = \max_{\rho} \sum_{ij} X_{ij} \quad \text{s.t. } RX \leq B$$

where X is a vector obtained by stacking all the flows $X_{ij} = \rho x_i x_j$; R is the routing matrix obtained using standard shortest path routing and defined such that $R_{kl} = 1$ or 0 depending on whether or not flow l passes through router k; and B is the vector consisting of all router degree-bandwidths constraints as given, for example, in Fig. 1.

For each network g, as a by-product of computing $Perf(g)$, we also obtain the total traffic flow through each router, which we term *router utilization*, as well as the set of bandwidths that are actually delivered to the end users of the network, which we call the *end user bandwidth distribution*. Both quantities can be viewed as auxiliary engineering-related performance metrics, indicating on the one hand how efficient or inefficient each router in the network is used and measuring on the other hand the ability of a network to support "realistic" end user demands. For real ISPs, the objective is clearly not to maximize throughput or router utilization, but to provide some service level guarantees (e.g., reliability). However, our intent here is to evaluate the raw carrying capacity of selected topologies under reasonable global traffic demands, and we avoid modeling the more detailed traffic patterns needed to formulate the objectives of real ISPs. To this end, assuming that the bandwidth demand at a router is proportional to the aggregated demand of any end hosts connected to it, the utilization of the high-end routers tends to be high. While other choices (e.g., choosing the traffic demand between routers as the product of their degrees as in [24]) yield different router utilization, the resulting performance values are qualitatively similar.

2) A Likelihood-Related Topology Metric: Here, we consider a graph-theoretic metric in order to differentiate between raw connectivity structures modeled by simple and connected graphs g having the same vertex set V and the same degree distribution, or equivalently, the same vertex degree sequence $\omega = (\omega_1, \ldots, \omega_n)$, where ω_k denotes the degree of vertex k. Consider the metric $s(g) = \sum_{(i,j)\in E(g)} \omega_i \omega_j$, where $E(g)$ represents the set of edges (with $(i,j) \in E(g)$ if there is an edge between vertices i and j). We define the *network likelihood* of g as the normalized metric

$$S(g) = \frac{s(g)}{s_{\max}}$$

where $s_{\max}$ is the maximum value of $s(g)$ among all $g \in G(\omega)$, the set of all simple connected graphs (i.e., no self-loops or parallel edges) with vertex set V and given node degree sequence ω. Note that graphs $g \in G(\omega)$ with high $s(g)$ values are those with high-degree vertices connected to other high-degree vertices and low-degree vertices connected to low-degree vertices. Conversely, graphs g with high-degree vertices connected to low-degree vertices have necessarily lower $s(g)$ values. Thus, there is an explicit relationship between graphs with high $s(g)$ values and graphs having a "hub-like" core (i.e., high connectivity vertices forming a cluster in the center of the network). By exploiting this connection, it can be shown that the exact $s_{\max}$ graph can be explicitly constructed (for details of this construction and a formal proof that it yields an $s_{\max}$ graph, see [27]).

A particularly appealing feature of the $S(g)$ metric is that it allows for a more traditional interpretation as (relative) likelihood associated with the *general model of random graphs (GRG) with a given expected degree sequence* considered in [14]. In fact,

defining the *(relative) likelihood* of a graph $g \in G(\omega)$ as the logarithm of the probability of that graph under the GRG model, the latter is shown in [27] to be proportional to $S(g)$, which in turn justifies our interpretation of the $S(g)$ metric as *relative likelihood* of $g \in G(\omega)$. However, for the purpose of this paper, we simply use the $S(g)$ metric to differentiate between networks having one and the same degree sequence and refer to [27] for a more detailed account of how this metric relates to graph properties such as "self-similar" or "self-dissimilar," "assortative" or "disassortative," and "scale-free" or "scale-rich."

C. Heuristically Optimal Topologies (HOT)

By combining the technological and economic design issues that apply to the network core and the network edge, respectively, we obtain a consistent story with regard to the forces that appear to govern the build-out and provisioning of the ISPs' core networks: market demands, link costs, and hardware constraints. The tradeoffs that an ISP has to make between what is technologically feasible versus economically sensible can be expected to yield router-level connectivity maps where individual link bandwidths tend to increase while the degree of connectivity tends to decrease as one moves from the network edge to its core. To a first approximation, core routers tend to be fast (have high capacity), but have only a few high-speed connections; and edge routers are typically slower overall, but have many low-speed connections. Put differently, long-haul links within the core tend to be relatively few in numbers but their capacity is typically high.

Thus, the proposed first-principles approach suggests that a reasonably "good" design for an ISP network is one in which the core is constructed as a sparsely connected mesh of high-speed, low-connectivity routers which carry heavily aggregated traffic over high-bandwidth links. Accordingly, this mesh-like core is supported by a hierarchical tree-like structure at the edges whose purpose is to aggregate traffic through high connectivity. We refer to this design as *heuristically optimal topology (HOT)* to reflect its consistency with real design considerations as well as its direct relationship with the *Highly Optimized Tolerance* approach proposed by Carlson and Doyle [9], or its close relative, the *Heuristically Optimized Tradeoffs* approach considered by Fabrikant *et al.* [21]. For the purposes of this paper, such heuristic HOT constructions are appropriate, as it is important to underscore that our results do not depend on designs being formally optimal, which is unlikely to occur in practice. Instead, we argue that any sensible network design process with minimally realistic assumptions would produce something qualitatively similar.

D. Comparing Different Network Topologies

To illustrate that the first-principles approach to modeling the Internet's router-level topology yields networks that are qualitatively as well as quantitatively different from currently considered models and capable of capturing key engineering objectives underlying network design, we consider five toy networks. They represent different methods for generating models of physical Internet connectivity, and they are similar in the sense that they all have one and the same node degree sequence, which happens to be of the power-law type. We show that while some of them appear deceivingly similar from a view that considers only graph theoretic properties, their vastly different structural features become all but too apparent when viewed from a performance-oriented perspective.

1) Five Representative Toy Networks: Fig. 3 depicts five networks constructed explicitly to have *the same* node degree sequence:[4]

(a) The power-law type degree sequence of all five networks.
(b) A graph constructed from *Preferential Attachment (PA)*: nodes are added successively and connected to the existing graph with probability proportional to each existing node's current degree.
(c) A construction based on the *General Random Graph (GRG) method*: we use the degree sequence of the PA network as the expected node degree to generate a random graph using the GRG method [14] and then fine-tune it (e.g., we add additional degree-one edge nodes) in the sense of [39] to obtain the proper degree distribution.
(d) *Heuristically Optimal Topology (HOT)*: we use a heuristic, nonrandom, degree-preserving rewiring of the links and routers in the PA graph to produce a network having a mesh-like core with hierarchical aggregation from the edge to the core.
(e) *Abilene-Inspired Topology.* Inspired by the actual Abilene network (to be discussed in Section IV-A), we use its core and obtain a HOT Abilene-like network by replacing each of its customer and peer networks with a single gateway router supporting the right number of end hosts to yield the desired node degree distribution.
(f) *Sub-optimal Topology.* For the purposes of comparison, we include a heuristically designed network that has been *intentionally constructed to have poor performance*.

Details about the construction of each graph are available in [26], however, what is more important here are the networks and their different features, not the process or particular algorithm that generated them.

2) The $Perf(g)$ Versus $S(g)$ Plane: Before projecting each of these networks onto the $Perf(g)$ versus $S(g)$ plane, note that while computing their $S(g)$ values is trivial, evaluating their network performance $Perf(g)$ requires further specifications. In particular, for each of our five networks, we impose the same router technological constraint on the non-edge routers, and for simplicity, we use a fictitious router based on the Cisco 12410 GSR model,[5] but modified so that the maximum number of ports it can handle coincides with the maximum degree generated above. As a result, each of the networks has the same number of non-edge nodes and links, as well as the same degree distribution among non-edge nodes, and collectively, these observations guarantee the same total "cost" (measured in routers) for each network.

A striking contrast is observed by simultaneously plotting network performance versus network likelihood for all five models in Fig. 4. The HOT networks have high performance and

[4]The degree sequences of these networks match exactly in the tail but differ slightly in the degree-one and degree-two nodes as a result of different generation mechanisms. These minor differences do not affect the comparison.

[5]The Cisco 12410 GSR is similar to the 12416 except that it has only nine available PIC slots instead of 15.

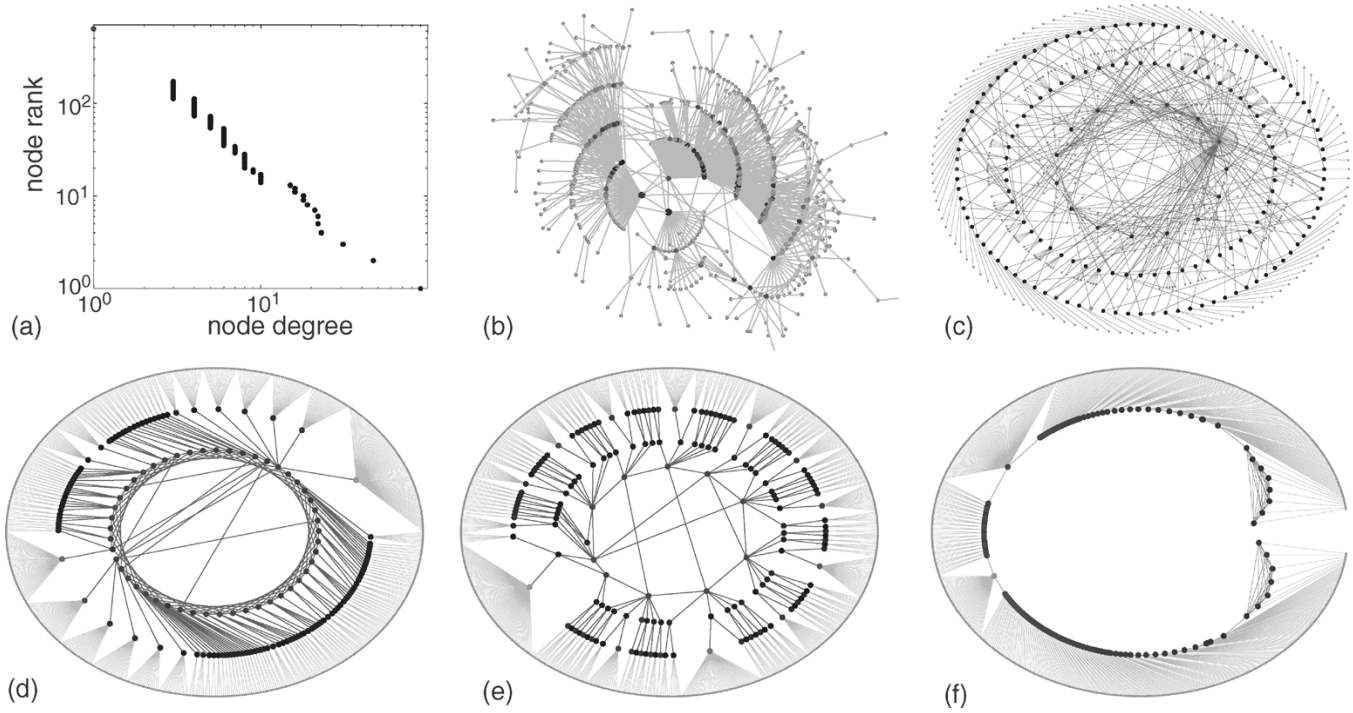


Fig. 3. Five networks having the same node degree distribution: identical from a degree-based perspective, but opposites in terms of engineering performance.

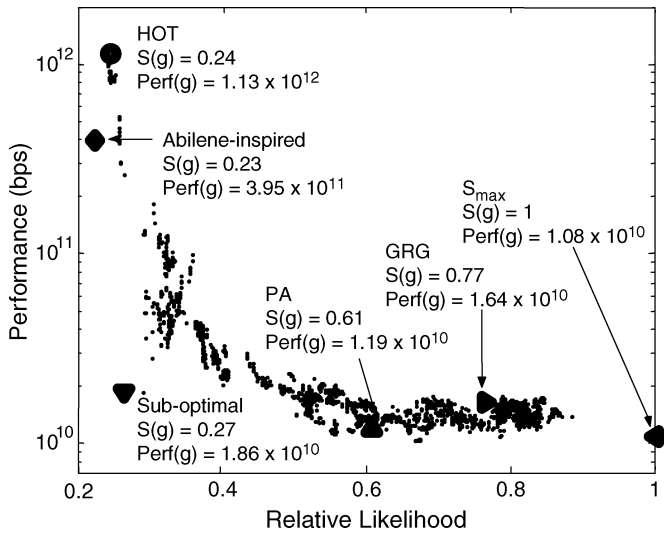


Fig. 4. Performance versus Likelihood for each network in Fig. 3. Additional points are other networks having the same node degree sequence obtained by pairwise random rewiring of links (see [26] and [27] for details).

low likelihood while the degree-based PA and GRG networks have high likelihood but low performance. The main reason for the degree-based models to have such poor performance is exactly the presence of the highly connected "hubs" that create low-bandwidth bottlenecks. The HOT models' mesh-like cores, like the real Internet, aggregate traffic and disperse it across multiple high-bandwidth routers. The interpretation of this picture is that a careful design process explicitly incorporating technological constraints can yield high-performance topologies, but these are extremely rare from a probabilistic graph point of view. In contrast, equivalent power-law degree distribution networks constructed by generic degree-based probabilistic constructions result in more likely, but poorly-performing topologies. Consistent with this, the "most likely" $s_{\max}$ network (included in Fig. 4) has also sub-par performance. This picture is further enhanced when considering the two auxiliary performance measures mentioned earlier, namely the distribution of end user bandwidths and router utilization. As detailed in [26], the heuristically optimal networks [Fig. 3(d) and (e)] achieve high utilization in their core routers and support a wide range of end-user bandwidth requirements. In contrast, the degree-based networks [Fig. 3(b) and (c)] saturate only their "hub" nodes and leave all other routers severely underutilized, thus providing uniformly low bandwidth and poor performance to their end-users.

IV. Empirical Evidence

As evidence that technological and economic forces described above are relevant to the real Internet, we consider in detail the router-level topology of several ISP networks. While no single data set is conclusive, we argue that the consistency across networks provides collective evidence that these forces are real and provide a reasonable starting point for topology generation models.

A. Exact and Publicly Available Data: Abilene and CENIC

While commercial ISPs tend to consider network topology information proprietary and a source of competitive advantage, public networks supporting higher education are less sensitive to such pressures and often willing to provide exact information about their design, configuration, and operation. Here, we consider educational backbone networks at the national and regional level.

The Abilene Network [1] is the national Internet backbone for higher education. It is comprised of high-speed connections between core routers located in 11 U.S. cities and carries approx-

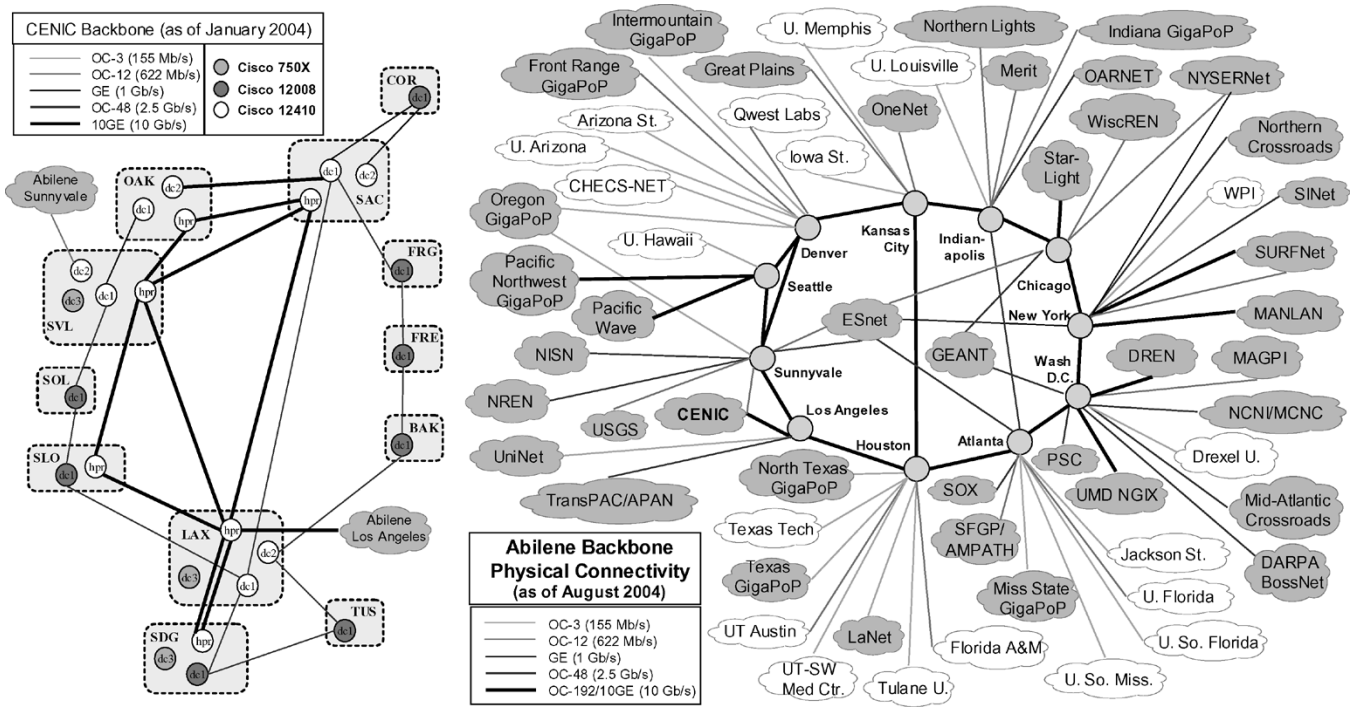


Fig. 5. CENIC and Abilene networks. Each node represents a router, and each link represents a physical connection. Left: The CENIC backbone is comprised of two backbone networks in parallel—a high performance (HPR) network supporting the University of California system and other universities, and the digital California (DC) network supporting K-12 educational initiatives and local governments. Connectivity within each POP is provided by Layer-2 technologies, and connectivity to the network edge is not shown. Right: The Abilene network is a sparsely connected mesh. End user networks are represented in white, while peer networks (other backbones and exchange points) are represented in gray. For both networks, each router has only a few high bandwidth connections, but each physical connection can support many virtual connections that give the appearance of greater connectivity to higher levels of the Internet protocol stack.

imately 1% of all traffic in North America. Fig. 5 depicts the exact physical connectivity of Abilene (as obtained from router configuration files). Abilene maintains peering connections with other higher educational networks (domestic and international) but does not connect directly to the commercial Internet.

Focusing on a regional network, we consider California, where the Corporation for Education Network Initiatives in California (CENIC) acts as ISP for the state's colleges and universities [15]. Its backbone is similarly comprised of a sparse mesh of routers connected by high-speed links (Fig. 5). Here, routing policies, redundant physical links, and the use of virtual private networks support robust delivery of traffic to edge campus networks. We argue that similar observations about network structure are found when examining (where available) topology information of global, national, or regional commercial ISPs.

These two networks provide evidence that the heuristic design presented in Section III-A shares similar qualitative features with the real Internet, namely the presence of a relatively sparse backbone with high connectivity at the periphery for aggregation purposes. However, this topology information also provides an opportunity to examine the extent to which router technology constraints are in effect for each network. That is, by aggregating the number and capacity of the line cards installed on each router, we obtain the total bandwidth and degree for each machine, which we can then locate within the feasible region for each router model. Fig. 6 shows this router configuration information for both Abilene and CENIC. One common feature of these relatively new networks is that only a subset of

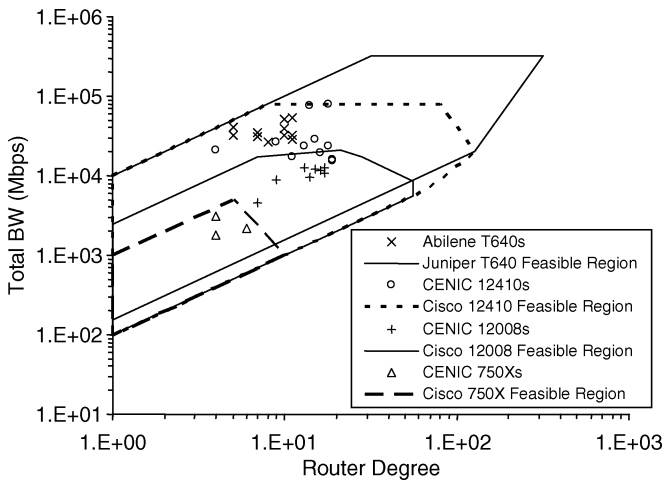


Fig. 6. Configuration of Abilene routers (as of August 2004) and CENIC routers (as of January 2004). Abilene is comprised of Juniper T640 routers, which are another type of core router similar to the Cisco GSR series. CENIC is comprised of Cisco 12410, 12008, and 7500 series routers. Note that in the time since the Cisco catalog [41] was published, the introduction of a new line card (supporting 10×1 GE interfaces) has shifted the feasible region for Cisco GSR routers. Thus, the Cisco 12410 GSR (which has nine available slots) can therefore achieve a maximum of 90 Gb/s with either nine 10 GE line cards or nine 10×1GE line cards.

available slots or ports on routers is currently populated. By allowing for significant future growth in capacity, the core routers in these networks are currently configured to be far away from the efficient frontier.

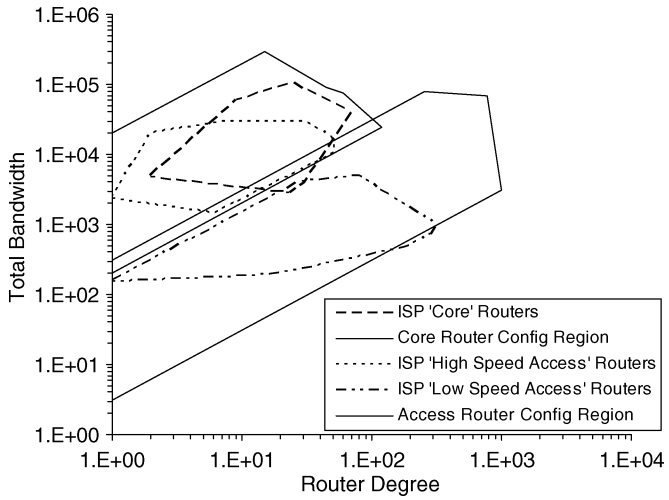


Fig. 7. Configuration of a Tier-1 commerical ISP as of 2003. Routers are grouped into three different types: high-speed access routers, low-speed access routers, and core routers. For each group, we show the convex hull surrounding the points corresponding to the bandwidth-degree configuration for each router. Also shown is the feasible configuration region for a typical core router (i.e., the Cisco 12416 GSR) and a typical access router (i.e., the Cisco 7600).

B. Exact but Proprietary Data: A Tier-1 ISP

One of the overriding concerns of commerical ISPs in sharing topology data is that it will reveal information about its customers, thereby putting them at risk to competition. However, in cases where topology information is sufficiently anonymized and aggregated, we have found ISPs more willing to share and publish connectivity data. Here, we present aggregate router configuation information for AS 7018 (AT&T), as it existed during the second half of 2003. This Tier-1 ISP has hundreds of routers across the U.S. and Fig. 7 shows aggregate router configuration data for "core" and "access" routers. Here, "core routers" can be understood as those that provide long-haul connectivity between individual *points of presence (POPs)* for the ISP. Conversely, "access routers" can be understood as those that provide aggregation connectivity between the ISP and its customers within a POP. For this ISP, access routers are further categorized according to whether they facilitate high-speed or low-speed connections.

Fig. 7 depicts the convex hull containing the bandwidth-degree configuration for the routers of each type. This aggregated information obscures individual router configurations as well as the total number of routers in each group, but it provides useful information nonetheless. First, the maximum number of connections to a core router is 68, while the maximum number of connections to a low-speed access router is 313. The maximum number of connections to a high-speed access router is less than that for both low-speed and core routers. Also, the relative position of these convex hulls reinforces the notion that routers are specialized according to their role (again, Fig. 2). The core routers in this AS tend to have higher overall capacity than access routers, and they also tend to have fewer connections than many low-speed access routers. The high-speed access routers tend to have higher overall capacity but fewer connections than low-speed access routers. Also shown in Fig. 7 is the feasible region for representative core and access routers.[6] While certainly not all of the routers deployed in this AS were of these specific router models, it is likely that some of them were. Nonetheless, a striking feature of Fig. 7 is the way in which the core routers conform rather tightly to the feasible configuration regions.

C. Indirect Measurements: Rocketfuel-Derived Evidence

While measurement based approaches to inferring router-level topology are fraught with ambiguity and subject to interpretation, recent efforts by the community have yielded topology maps of increasing fidelity [17], [25], [35]. The Rocketfuel Project [35] has provided the most detailed and accurate ISP-based router-level maps to date (however, see [40] for a partial accounting of errors and ambiguities in the Rocketfuel data). Using BGP and other publicly available information to design traceroute experiments targeted at individual ISPs, Rocketfuel maps combine basic connectivity data with DNS information and other domain-specific knowledge to generate realistic topologies of ISP backbone and POP structure.

Our first principles approach suggests that, when using measurement-based connectivity maps like Rocketfuel, one should leverage domain-specific information whenever possible to validate and augment each ISP topology, with the ultimate objective of producing an annotated graph that contains realistic bandwidths/capacities and is consistent with engineering reality. While some annotation is available from the Rocketfuel data itself (i.e., routing weights, delays, see [28]), additional insight comes from a closer look at router configuration data.

Each physical interface on a router generally has its own *IP address* and corresponding *domain name service (dns) name*, which are fundamental to communication within and across subnetworks. Often, dns naming conventions are used to embed information for management simplicity on the part of network operators, as evidenced by the published naming convention for AS 1239 (Sprintlink) [37]. As in the case of AS 1239, dns names are typically chosen to reflect the location, role, physical interface, and business relationship associated with the router. For example, the dns name `sl-bb21-chi-1-0.sprintlink.net` refers to an interface on backbone (bb) router number 21 in Sprintlink's Chicago POP. Similarly, the dns name `ar13-a300s1.attga.ip.att.net` for AS 7018 refers to an interface on access router (ar) number 13 in the Atlanta PoP. Rocketfuel makes use of dns information where available to help determine the geographic location of individual routers. However, there is additional information contained in these dns names that is relevant to our discussion. For example, we believe that the `-1-0` portion of the above dns name for AS 1239 and `-a300s1` in the dns name for AS 7018 contain information about the physical interface (i.e., slot and port number) on the router itself. Thus, for ISPs that make consistent use of dns naming conventions, one obtains insight into the underlying structure of individual POPs and the overall engineering design of the ISP network.

One additional piece of information available from these naming conventions is the *dns prefix associated with the role of*

[6]While the technology represented in the 2002 catalog [41] is now outdated, we argue that the product deployment lifecycle for routers makes it reasonable to believe that the production network of our ISP in 2003 is adequately reflected using this older technology.

TABLE I
(LEFT) ASSUMED MAPPING BETWEEN THE DNS PREFIX AND ROUTER ROLE FOR TWO TIER-1 ISPS (RIGHT) ASSUMED LINK SPEEDS (IN Mb/s) FOR ANNOTATING ROCKETFUEL DATA

	AS 1239	AS 7018
cust		
slow-access	gw	ar
fast-access		gar, sar
core	bb	br, gbr, tbr gr, ggr, gw
other	st, dr	dc, nr

	cust	slow-access	fast-access	core	other
cust		10	100	1000	1000
slow-access	10	1000	1000	1000	1000
fast-access	100	1000	1000	1000	1000
core	1000	1000	1000	2500	1000
other	1000	1000	1000	2500	1000

a router, which in the two cases above is `bb21` (for AS 1239) and `ar13` for (AS 7018). Recall that we previously defined two different roles for a router within an ISP, namely a *core router* and an *access router*. We now expand that taxonomy slightly to include *customer routers* (representing the connection with a customer subnetwork) and *other routers* (e.g., data servers). While each ISP may use its own set of dns prefixes, they can most always be mapped to these four roles. In what follows, we use the mapping in Table I to associate dns prefix and router role.

Extracting the role of a router is important for two reasons. First, by recognizing that different types of hardware are typically used in different roles (as described in Section III-A-1), knowing the role of a router provides a rough guide for how to annotate the router with capacity and link speed information. Before explaining this in more detail, we note that a second use of dns prefix information provides indirect validation of the way in which raw traceroute data is incorporated into topology maps. Specifically, one challenge for interpreting traceroute data is to decide which IP addresses (and corresponding dns names) refer to the same router, a process known as *alias resolution* [34]. One of the contributing factors to the high fidelity of the Rocketfuel maps was an improvement in the way that alias resolution was performed [35]. Nonetheless, making sense of traceroute data to obtain connectivity information within a large POP is a nontrivial task, and here again knowledge about router configuration is useful. Assuming that the dns prefix (e.g., `bb21`, `ar13`) provides a unique identifier for routers within the same POP, we found many instances where alias resolution in Rocketfuel resulted in seemingly duplicate nodes.[7] More importantly, designating router by role provides insight into the connectivity structure within the ISP, as evidenced by Table II.

While consistent with our understanding of the feasible configuration of individual routers as well as the reasonable design of a POP, these results are not conclusive and should be viewed with skepticism for several reasons. First, dns responses can be unreliable and are not guaranteed to be correct. Indeed, many of the nodes in the Rocketfuel data set identified as backbone routers did not have dns information at all, so this procedure could not be applied. Second, the ability to leverage dns information in this manner depends entirely on the presence and interpretation of rational naming conventions. Early attempts to apply this procedure to other Rocketfuel data sets have met with mixed success. While some networks such as AS 3967 (Exodus) followed conventions that could be leveraged, others such as AS 1755 (Ebone) and AS 6461 (AboveNet) did not. At the same time, AS 1221 (Telstra) includes link speed information in some of its dns naming (e.g., `gigabiteth-ernet1-1.fli2.adelaide.telstra.net`), and the extent to which this information can be used to improve the fidelity of existing router-level maps remains an open question.

[7]For AS 1239, our approach suggests that 215 of 673 (32%) are duplicates, while for AS 7018 the number of duplicates is 156 out of 640 (24%). These duplicates were found exclusively within the largest POPs, again where alias resolution is the most difficult.

TABLE II
ROCKETFUEL-DERIVED DATA SUMMARIZING CONNECTIVITY AND NUMBER OF NODES FOR AS 7018 (TOP) AND AS 1239 (BOTTOM)

AS 7018 link connections	cust	slow-access	fast-access	core	other	total routers
cust	0	7888	622	131	8	8649
slow-access	7888	96	1	662	9	299
fast-access	622	1	24	161	4	71
core	131	622	161	503	12	145
other	8	9	4	12	0	26

AS 1239	cust	access	core	other	total routers
cust	0	6653	288	24	6965
access	6653	152	868	2	317
core	288	868	584	23	122
other	24	2	23	8	23

While a detailed discussion of the implications of our first principles approach to router-level mapping techniques is beyond the scope of this paper, we consider one simple experiment that provides additional indirect evidence for our story of heuristically optimal ISP design and also indicates one potential direction for the annotation of connectivity-only maps. Starting with Rocketfuel connectivity maps, we use information about the role of individual routers to assist in the annotation of link bandwidths (and router capacities). Specifically, we assume that links between backbone routers are used for long-haul aggregated traffic between POPs and access routers are used primarily for customer traffic aggregation, and we annotate bandwidth according to the link speeds defined by Table I.

While direct validation of the assumed link bandwidths is difficult, the resulting bandwidth-degree combinations for each router provide an initial check for qualitative consistency between the annotated network maps and what is known about Tier-1 ISPs. Fig. 8(a) presents the inferred bandwidth-degree combinations for routers in AS 7018, differentiated by router type. We again superimpose the feasible configuration regions for typical core and access routers, and we note that the resulting picture is qualitatively similar to what we observe in Fig. 7. The core routers tend to have higher overall bandwidth, which is not surprising, but what is striking is that none of the bandwidth-degree combinations for backbone routers fall outside the corresponding feasible configuration. Similar information for AS 1239 is presented in Fig. 8(b), and a qualitatively similar picture can be found there as well. However, we make two cautionary remarks regarding the interpretation of these figures. First, the similarity between Figs. 7 and 8(a) does not necessarily mean that the routers in AS 7018 are configured as represented in Fig. 8(a). Second, Fig. 8(b) does not necessarily reflect what is really in AS 1239. Rather, the point here is that even a heuristic process informed by a detailed understanding of router role and technology constraints can go a long way toward generating realistic annotated router-level maps. While not conclusive, what

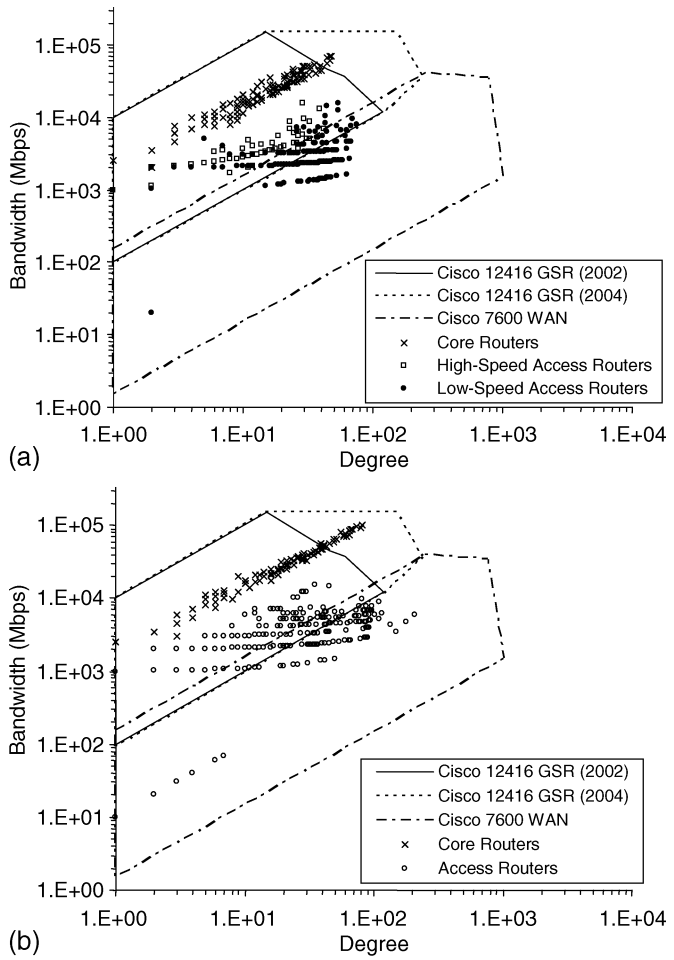


Fig. 8. Rocketfuel-derived router configurations based on assumed link annotations: (a) AS 7018; (b) AS 1239.

is remarkable about these results is that the simple assumptions for (fixed) link bandwidths in Table I result in bandwidth-degree combinations that are not inconsistent with our understanding of heuristically optimal network design.

While the general problem of reverse-engineering router-level topologies [36] remains an open problem (see Section V-C), this heuristic procedure inspired by our first-principles approach provides a reasonable "strawman" from which additional information may be leveraged to generate even higher fidelity maps. For example, one could include higher granularity in link speeds by incorporating typical *packet over SONET (POS)* speeds: DS1 (1.544 Mb/s), DS3 (45 Mb/s), OC3 (155 Mb/s), OC12 (622 Mb/s), and OC48 (2.5 Gb/s). Adding a mixture of such speeds can be exptected to result in bandwidth-degree combinations that are more dispersed within the feasible configuration region. Similarly, the incorporation of links at 10 Gb/s (and higher) will result in core routers configured much closer to the efficient frontier. In addition, adding equipment prices such as those found in [41] would provide an approximate means for calculating network infrastructure cost.

V. Insights and Implications

In addition to providing an understanding of key factors driving network design, the first-principles approach offers several implications for the explanation, validation, and ongoing investigation of network structure.

A. On the Origin of High Variability in Node Degrees

One of the more striking statistical features in measured or inferred models of the Internet is the commonly observed high variability in network connectivity (i.e., node degrees). The first principles approach provides new insight into the causes or origins behind this seemingly ubiquitous network characteristic. As argued previously, the technological and economic forces shaping the buildout of router-level topologies by ISPs suggest that high variability in connectivity, if present, only makes sense at the network edge. The need for high-performance in network throughput precludes the possibility of high connectivity in the network core, while the presence of high variability in terms of end-user connection speeds and population density [4] suggest that this is the source of empirically observed variability or power laws. This notion is further supported by a realization that the same core network design can support many different end-user bandwidth distributions and that by and large, the variability in end-user bandwidth demands determines the variability of the node degrees in the resulting network (see [26, Fig. 9] for details). At the same time, it is worth noting that there exists considerable evidence suggesting that claims of power laws in the router-level Internet may be the result of misinterpretation of available measurements and/or their naive and inappropriate statistical analysis (see [27] for a detailed discussion).

B. Demystifying the Achilles' Heel Claim for the Internet

A topic of increasing importance to complex network researchers across disciplines has been the extent to which the connectivity structure of a network is responsible for its *robustness to random failures* or *fragility to targeted attacks*. In particular, following the work of Albert *et al.* [3], the literature on "scale-free networks" has advocated the claim that the presence of high-connectivity nodes in the core of the network is a hallmark of networks having power laws in the distribution of node degree, and furthermore that attacks on them can destroy network connectivity as a whole. The basic argument underlying this claim is that while typical nodes in a scale-free network have small degree and hence contribute little or nothing to overall network connectivity, the presence of hubs is critical in that their removal literally fragments the network. In the context of the Internet, this discovery has been touted "the Achilles' heel of the Internet" [3], a vulnerability that has presumably been overlooked by networking engineers. If true, this finding would certainly be startling and profound, as it directly contradicts the Internet's legendary and most clearly understood robustness property, namely its ability, in the presence of router or link failures, to "see damage and work around it" [16].

While a comprehensive study of large-scale network robustness is well beyond the scope of this article, an initial account of the "robust yet fragile" nature of the Internet's actual router-level topology is provided in [19], where we use a scale-free model like the one in Fig. 3(b) and a HOT model like the one in Fig. 3(d) to compare network performance in the presence of successive router loss. For each network, we

target the worst-case node that has not yet been deleted, and after each router loss we compute the amount of original traffic (as measured by our previously defined notion of performance) that can still be served by the remaining network, possibly after some re-routing, but with routers that remain constrained to their original feasible region. Consistent with the claims in [3] (and further illustrated in [19, Fig. (3c)]), the scale-free network is indeed fragile to the deletion of worst-case nodes (here, worse-case means highest-degree); after removing the hubs, the performance drops by more than one order of magnitude. In contrast, the HOT network is not only more robust to worst-case deletions (here, worst-case are low-connectivity core nodes), but also shows high tolerance to deleting other nodes, particularly high-degree edge routers. In fact, because the scale-free network has such poor nominal performance to start with, it is worse intact than the HOT network after the latter has sustained substantial damage.

This example illustrates two appealing features of the proposed first-principles approach. First, our detailed study of the technological and economic forces shaping the router-level topology of a single ISP provides convincing evidence that in today's Internet, the existence of highly connected routers in the core of the network is a myth. Size issues not notwithstanding, the real Internet is nothing like Fig. 3(b), but is qualitatively more like the network shown in Fig. 3(d): it cannot possibly have a hub-like core, and the highly connected nodes, if they exist, must be situated at the periphery of the network. Second, when trying to answer the question "What really matters when it comes to the ability of the Internet to perform in the presence of router or link losses?" we note that modeling router-level robustness requires at a minimum adding some link redundancy (e.g., multi-homing) and incorporating a simple abstraction of IP routing that accounts for the feedback mechanisms that react to the loss or failure of a network component. In particular, our approach makes it clear why the type of connectivity-only perspective pursued in [3] (i.e., one that completely ignores the existence of routing protocols sitting on top of the raw router-level connectivity) is bound to provide an overly simplistic and even misleading view of network robustness. Indeed, it is well-known that the Internet's actual fragilities are not to physical attacks on routers or links, but to perturbations that were not part of the Internet's original design objectives [16], particularly misbehaving components or hijacked services.

C. Reverse-Engineering the Internet

Reverse-engineering the Internet typically refers to the process of learning about its design (i.e., how the different network components are assembled and configured) by studying its implementation. Spring *et al.* [36] describe it as a community effort that produces annotated maps of the Internet, complete with features such as client population, traffic patterns and workloads; network ownership, capacity, connectivity, geography and routing policies; patterns of packet loss, congestion, bottlenecks, failure, growth; etc. The argument in [36] is that reverse-engineering the Internet is feasible based on: 1) continuing improvements in measurement techniques; 2) ongoing refinements of methods to infer network-internal details from external measurements; and 3) a more focused accounting of the resources required to complete the process. However, much of the ongoing efforts in this area remain heuristic/empirical in nature, reflect a piecemeal effort with little quality control, and lack by and large a coherent framework for a realistic cost-benefit assessment of a given measurement effort (as considered for example in [6]). Given the essentially unlimited opportunities for Internet measurements, questions like *What to measure?, Which measurements are more important/informative than others?*, and *How to infer network-internal features that cannot be measured directly from what sort of external measurements?* are left either unanswered or to the experimentalist.

Here we argue that our proposed first-principles approach provides a prototype framework for a systematic and informed effort to reverse-engineer the Internet that complements present efforts and assists the experimentalists in answering some of the above questions. Starting with the connectivity-only graphs provided by projects such as Rocketfuel, we contend that network annotations such as link bandwidth and router capacity should be consistent, both internally and with the technological and economic considerations at work in router-level network design. The type of annotations considered in Section IV-C represents an initial attempt to do exactly this, and trying to re-produce the original input that led to summary plots like Fig. 7 represents a concrete example of such an informed reverse-engineering effort. However, additional measurements on the part of empiricists would significantly help to reduce the uncertainty inherent in such annotated maps. For example, is it conceivable to design measurement experiments for the purpose of inferring the model/make of a given router from external measurements? Similarly, is it feasible to perform large-scale active measurement studies intended to infer link speeds on an ISP-wide level? Also, to what extent can additional information (e.g., metropolitan populations, population densities, market prices for network services, ISP market share, ISP revenues) be leveraged to generate more accurate annotated router-level maps?

VI. Conclusions and Outlook

We have shown that a first-principles approach to modeling Internet connectivity at the router-level provides direct and immediate insight into current approaches based on topology generation and measurement-based reverse engineering to understand the large-scale structure of the Internet as a whole. Connectivity data from real ISP networks provides empirical evidence that these constraints are consistent with engineering reality, and it also supports our explanation for the origin of high variability in measured or inferred router-level maps. Perhaps more importantly, this approach easily accommodates the incorporation of additional objectives of network design, such as network robustness to router failures as discussed here. It is also ideally suited for investigating the sensitivity of network design to deviations from the assumed gravity traffic demand model (e.g., a trend toward more localized traffic demands), to revolutionary advances in router technologies (e.g., all-optical Terabit routers), or to radical changes affecting the economics of network design (e.g., the large-scale deployment of ultralong-haul (ULH) fiber links). In particular, studying a possible evolution from today's networks to future design scenarios (e.g., see the discussion in [48]) and understanding the main drivers behind such a transition loom as intriguing open problems.

Although the emphasis in this paper has been on a reasonably good design of the physical network connectivity at the level of a single ISP, we recognize that the broader Internet is a collection of thousands of ASes that interconnect at select locations. While the important issue of understanding how the large-scale structure of the Internet relates to the heuristically optimal network design of single ISPs is not addressed in this paper, we speculate that similar technology constraints and economic drivers will exist at peering points between ISPs, but that the complexity of routing management may emerge as an additional consideration. As a result, we fully expect border routers again to have a few relatively high bandwidth physical connections supporting large amounts of aggregated traffic. In turn, high physical connectivity at the router level is again expected to be confined to the network edge.

With a detailed annotated map of global Internet connectivity at the physical level (Layer 2) now within reach, there are natural ways of coarse-graining such a map to obtain less detailed, yet physically or logically meaningful representations of the Internet's topology. For example, one natural coarsification of the physical connectivity could represent Internet connectivity at the IP layer (Layer 3) as seen by traceroute. Coarsifying yet further could result in a POP-level view of Internet connectivity. Finally, coarse-graining even further by collapsing all POPs within an AS, combined with an adequate accounting and annotation of all physical links, would result in annotated AS-level maps that include such details as network ownership, capacity, POP-level connectivity and geography, routing policies, etc. While inferred connectivity-only AS-maps of the Internet and their properties have been studied extensively in the recent past, with the exception of work by Chang *et al.* [12], there have been no attempts at relating them to or explaining some of their particularly striking features in terms of POP-level, IP-level, let alone physical layer Internet connectivity. However, the picture that has emerged of the large-scale statistical properties of a number of different inferred AS maps and that is based on the combined efforts of the networking community [17], [22], [25], [33], [35], [38] strongly suggests that Internet connectivity at the various levels are shaped by very different forces and factors. While these level-specific connectivity maps may appear deceivingly similar when viewed from the perspective of certain large-scale statistical properties (i.e., node degree distribution), their structures are often completely different, especially of those at the two opposite ends of the multi-scale spectrum, i.e., the physical-level and AS-level maps. A more detailed understanding of the different forces at work in shaping the different topologies at the different levels or scales remains an open problem. Also, the possibility to exploit this networking-specific, multi-scale view of Internet topology that reflects key aspects of the Internet's architecture for the purpose of network visualization looms as a promising open research problem, especially when trying to combine it with estimating or inferring the different traffic matrices associated with the different "scales" and studying their multi-scale properties (e.g., see [10]).

ACKNOWLEDGMENT

The authors are indebted to M. Roughan for generating Fig. 8 and negotiating its suitability for "public consumption," N. Spring for assistance in using the Rocketfuel data, S. Shalunov for data on the Abilene network, H. Sherman for help with the CENIC backbone, and R. Govindan and S. Low for fruitful discussions of router-level topologies. The authors also thank the four anonymous SIGCOMM reviewers for constructive comments on an earlier version of this paper.

REFERENCES

[1] Abilene Network. Detailed Information About the Objectives, Organization, and Development of the Abilene Network. [Online]. Available: http://www.Internet2.edu/abilene

[2] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Rev. Modern Phys.*, vol. 74, no. 1, pp. 47–97, Jan. 2002.

[3] A. Albert, H. Jeong, and A.-L. Barabási, "Attack and error tolerance of complex networks," *Nature*, vol. 406, pp. 378–382, 2000.

[4] D. Alderson, "Technological and economic drivers and constraints in the Internet's "Last Mile"," California Inst. Technol., Pasadena, CA, Tech. Rep. CIT-CDS-04-004, 2004.

[5] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509–512, 1999.

[6] P. Barford, A. Bestavros, J. Byers, and M. Crovella, "On the marginal utility of deploying measurement infrastructure," in *Proc. 1st ACM SIGCOMM Internet Measurement Workshop*, San Francisco, CA, Nov. 2001, pp. 5–17.

[7] T. Bu and D. Towsley, "On distinguishing between Internet power law topology generators," in *Proc. IEEE INFOCOM*, New York, 2002, pp. 638–647.

[8] K. L. Calvert, M. Doar, and E. Zegura, "Modeling Internet topology," *IEEE Commun. Mag.*, vol. 35, no. 6, pp. 160–163, Jun. 1997.

[9] J. M. Carlson and J. Doyle, "Complexity and robustness," *Proc. Nat. Acad. Sci.*, vol. 99, no. 1, pp. 2539–2545, Feb. 2002.

[10] H. Chang, S. Jamin, Z. M. Mao, and W. Willinger, "An empirical approach to modeling inter-AS traffic matrices," in *Proc. 5th ACM SIGCOMM Internet Measurement Conf.*, Berkeley, CA, 2005, pp. 139–152.

[11] H. Chang, R. Govindan, S. Jamin, S. Shenker, and W. Willinger, "Toward capturing representative AS-level Internet topologies," in *Proc. ACM SIGMETRICS*, Marina Del Rey, CA, Jun. 2002, pp. 280–281.

[12] H. Chang, S. Jamin, and W. Willinger, "Internet connectivity at the AS-level: an optimization-driven modeling approach," presented at the ACM SIGCOMM Workshop on Models, Methods and Tools for Reproducible Network Research (MoMeTools) 2003, (extended version, Tech. Rep. UM-CSE-475-03, 2003, Univ. Michigan).

[13] Q. Chen, H. Chang, R. Govindan, S. Jamin, S. Shenker, and W. Willinger, "The origin of power laws in Internet topologies revisited," in *Proc. IEEE INFOCOM*, New York, 2002, pp. 608–617.

[14] F. Chung and L. Lu, "The average distance in a random graph with given expected degrees," *Internet Math.*, vol. 1, pp. 91–113, 2003.

[15] Corporation for Education Network Intitiatives in California (CENIC) [Online]. Available: http://www.cenic.org

[16] D. D. Clark, "The design philosophy of the DARPA Internet protocols," *Proc. ACM SIGCOMM'88, ACM Comput. Commun. Rev.*, vol. 18, no. 4, pp. 106–114, 1988.

[17] Skitter, Cooperative Association for Internet Data Analysis (CAIDA). [Online]. Available: http://www.caida.org/tools/measurement/skitter/

[18] M. B. Doar, "A better model for generating test networks," in *Proc. IEEE GLOBECOM*, London, U.K., Nov. 1996, pp. 86–93.

[19] J. C. Doyle, D. Alderson, L. Li, S. Low, M. Roughan, S. Shalunov, R. Tanaka, and W. Willinger, "The "robust yet fragile" nature of the Internet," *Proc. Nat. Acad. Sci.*, vol. 102, no. 41, pp. 14497–14502, 2005.

[20] P. Erdos and A. Renyi, "On random graphs I," *Publ. Math. Debrecen*, vol. 6, pp. 290–297, 1959.

[21] A. Fabrikant, E. Koutsoupias, and C. Papadimitriou, "Heuristically optimized trade-offs: a new paradigm for power-laws in the Internet," in *Proc. 29th Int. Colloq. Automata, Languages and Programming (ICALP 2002)*, Jul. 2002, pp. 110–122.

[22] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the Internet topology," *Proc. ACM SIGCOMM 1999, ACM Comput. Commun. Rev.*, vol. 29, pp. 251–262, 1999.

[23] L. Gao, "On inferring autonomous system relationships in the Internet," in *Proc. IEEE GLOBECOM*, San Francisco, CA, Nov. 2000, pp. 387–396.

[24] C. Gkantsidis, M. Mihail, and A. Saberi, "Conductance and congestion in power law graphs," in *Proc. ACM SIGMETRICS*, San Diego, CA, 2003, pp. 148–159.

[25] R. Govindan and H. Tangmunarunkit, "Heuristics for Internet map discovery," in *Proc. IEEE INFOCOM*, Tel Aviv, Israel, 2000, pp. 1371–1380.
[26] L. Li, D. Alderson, J. Doyle, and W. Willinger, "A first-principles approach to understanding the Internet's router-level topology," *Proc. ACM SIGCOMM 2004, ACM Comput. Commun. Rev.*, vol. 34, pp. 3–14, 2004.
[27] L. Li, D. Alderson, J. C. Doyle, and W. Willinger, "Toward a theory of scale-free graphs: definition, properties, and implications," *Internet Math.*, to be published.
[28] R. Mahajan, N. Spring, D. Wetherall, and T. Anderson, "Inferring link weights using end-to-end measurements," in *Proc. 2nd ACM SIGCOMM Internet Measurement Workshop*, Marseille, France, 2002, pp. 231–236.
[29] A. Medina, I. Matta, and J. Byers, "On the origin of power laws in Internet topologies," *Comput. Commun. Rev.*, vol. 30, pp. 18–28, 2000.
[30] M. Mitzenmacher, "A brief history of generative models for power law and lognormal distributions," *Internet Math.*, vol. 1, pp. 226–249, 2003.
[31] M. E. J. Newman, "The structure and function of complex networks," *SIAM Rev.*, vol. 45, pp. 167–256, 2003.
[32] R. Pastor-Satorras and A. Vespignani, *Evolution and Structure of the Internet: a Statistical Physics Approach*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[33] G. Siganos, M. Faloutsos, P. Faloutsos, and C. Faloutsos, "Power laws and the AS-level Internet topology," *IEEE/ACM Trans. Netw.*, vol. 11, no. 4, pp. 514–524, Aug. 2003.
[34] N. Spring, M. Dontcheva, M. Rodrig, and D. Wetherall, "How to resolve IP aliases," Univ. Michigan, UW CSE Tech. Rep. 04-05-04, 2004.
[35] N. Spring, R. Mahajan, D. Wetherall, and T. Anderson, "Measuring ISP topologies with Rocketfuel," *IEEE/ACM Trans. Netw.*, vol. 12, no. 1, pp. 2–16, Feb. 2004.
[36] N. Spring, D. Wetherall, and T. Anderson, "Reverse-engineering the Internet," in *Proc. ACM SIGCOMM 2nd Workshop on Hot Topics in Networks (HotNets-II)*, Boston, MA, Nov. 2003, pp. 3–8.
[37] Sprintlink Router Naming Conventions [Online]. Available: http://www.sprint.net/faq/namingconvention_sl.html
[38] L. Subramanian, S. Agarwal, J. Rexford, and R. Katz, "Characterizing the Internet hierarchy from multiple vantage points," in *Proc. IEEE INFOCOM*, New York, 2002, pp. 618–627.
[39] H. Tangmunarunkit, R. Govindan, S. Jamin, S. Shenker, and W. Willinger, "Network topology generators: degree-based versus structural," in *Proc. ACM SIGCOMM 2002, Comput. Commun. Rev.*, vol. 32, 2002, pp. 147–159.
[40] R. Teixeira, K. Marzullo, S. Savage, and G. M. Voelker, "In search of path diversity in ISP networks," in *Proc. 3rd ACM SIGCOMM Internet Measurement Conf.*, Miami, FL, Oct. 2003, pp. 313–318.
[41] State of Washington Master Contract for Cisco Products (2002). [Online]. Available: http://techmall.dis.wa.gov/master_contracts/intranet/routers_switches.asp
[42] B. M. Waxman, "Routing of multipoint connections," *IEEE J. Sel. Areas Commun.*, vol. 6, no. 9, pp. 1617–1622, Dec. 1988.
[43] W. Willinger, D. Alderson, and L. Li, "A pragmatic approach to dealing with high variability in network measurements," in *Proc. 4th ACM SIGCOMM Internet Measurement Conf.*, Taormina, Sicily, Italy, Oct. 2004, pp. 88–100.
[44] W. Willinger, D. Alderson, J. C. Doyle, and L. Li, "More "normal" than normal: Scaling distributions and complex systems," in *Proc. 2004 Winter Simulation Conf.*, Washington, DC, pp. 130–141.
[45] W. Willinger, R. Govindan, S. Jamin, V. Paxson, and S. Shenker, "Scaling phenomena in the Internet: critically examining criticality," in *Proc. Nat. Acad. Sci.*, vol. 99, Feb. 19, 2002, pp. 2573–2580.
[46] E. Zegura, K. L. Calvert, and M. J. Donahoo, "A quantitative comparison of graph-based models for Internet topology," *IEEE/ACM Trans. Netw.*, vol. 5, no. 6, pp. 770–783, Dec. 1997.
[47] Y. Zhang, M. Roughan, C. Lund, and D. Donoho, "An information-theoretic approach to traffic matrix estimation," *Proc. ACM SIGCOMM 2003, Comput. Commun. Rev.*, vol. 33, pp. 301–312, 2003.
[48] R. Zhang-Shen and N. McKeown, "Designing a predictable Internet backbone network," in *Proc. ACM SIGCOMM 3rd Workshop on Hot Topics in Networks (HotNets-III)*, San Diego, CA, Nov. 2004.

David Alderson (M'05) received the B.S.E. degree in civil engineering and operations research from Princeton University, Princeton, NJ, and the M.S. and Ph.D. degrees from the Department of Management Science and Engineering, Stanford University, Stanford, CA.

He has been a Research Fellow at the Institute for Pure and Applied Mathematics at the University of California at Los Angeles and at the Santa Fe Institute. He is currently a Postdoctoral Scholar at the California Institute of Technology, Pasadena. His research interests include network design, operations, and management.

Dr. Alderson has been a member of the ACM since 2004.

Lun Li (S'05) received B.S. degrees in optics and automatic control from Tsinghua University, Beijing, China, in 1999, and the M.S. degree in mechanical engineering from the University of California at Berkeley in 2001. She has been working towards the Ph.D. degree in electrical engineering at the California Institute of Technology, Pasadena, since 2001.

Her research interests include network congestion control, routing and Internet topology.

Ms. Li received the Best Student Paper Award from ACM SIGCOMM 2004.

Walter Willinger (F'05) received the Diplom (Dipl. Math.) degree from the ETH Zurich, Switzerland, and the M.S. and Ph.D. degrees from the School of ORIE, Cornell University, Ithaca, NY.

He is currently a member of the Information and Software Systems Research Center at AT&T Labs–Research, Florham Park, NJ. He was a Member of Technical Staff at Bellcore Applied Research from 1986 to 1996.

Dr. Willinger was a co-recipient of the 1996 IEEE W. R. G. Baker Prize Award and the 1994 W. R. Bennett Prize Paper Award. He has been a member of the ACM since 2000.

John C. Doyle (M'96) received the B.S. and M.S. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, in 1977, and the Ph.D. degree in mathematics from the University of California at Berkeley in 1984.

He is the John G. Braun Professor of Control and Dynamical Systems, Electrical Engineering, and Bioengineering at the California Institute of Technology, Pasadena.

Dr. Doyle received the 2004 IEEE Control Systems Award, the 1984 IEEE Centennial Outstanding Young Engineer Award, the 1984 Bernard Friedman Award, the 1983 American Automatic Control Council (AACC) Eckman Award, and the 1976 IEEE Hickernell Award. His best paper awards include the 1991 IEEE W. R. G. Baker Prize, the 1994 AACC O. Hugo Schuck Award, and the 1990 IEEE G. S. Axelby Award (twice).