

1

Introduction

The Internet is a communication network connecting digital devices. It is remarkably versatile, moving information among a vast diversity of desktop computers, supercomputers, personal digital assistants, satellites, spacecraft, cellular phones, sensors, wireless devices, embedded systems, and other systems. Indeed, almost any form of digital electronics can be provided with an Internet connection.

As a result, the Internet has become a globally pervasive ‘nervous system’ for electronic information processing. It connects devices and networks in all continents and in virtually every country, and it permeates many walks of life. The Internet is used for personal communication, for commercial enterprise, for public dissemination of information, and for transmission of secret data. Information on the Internet is exchanged in multiple languages and alphabets. Digital media is distributed through the Internet in text, audio and video formats. Electronic commerce is used daily in many countries to sell virtually any product.

Accordingly, the Internet is a huge system. The first Internet node was installed in 1969 as part of a research project; at the current time (2006) there are hundreds of millions of devices connected to the Internet. The Internet is no longer a research project; current traffic estimates on the Internet range up to several petabytes a day, and the Internet is still growing rapidly. In the near future it will undoubtedly continue to expand in both size and functionality.

One might presume that a system as important as the Internet – with as much impact on society and as much future potential – would be well understood. In fact, while the building blocks of the Internet (its protocols and individual components) are the subject of intensive study and design effort, the immense global entity that is the Internet today has not been precisely characterized. Many quantitative measures of the Internet are simply absent.

There are a number of reasons for this state of affairs. First of all, the physical

structure of the Internet is not the result of any centralized design or plan. The Internet has been built by a large set of independent organizations, often having somewhat different goals. In addition, its construction has taken place (relatively speaking) quite quickly – the vast majority of Internet infrastructure has been installed since the early 1990s.

Another reason for our limited knowledge about the Internet is that the network is dynamic. It is constantly changing in size, configuration, traffic, and application mix. Traffic and application mix vary dramatically in different parts of the network. Devices are moved about, and are frequently connected and disconnected from the network. Even when good measures of some Internet properties are known at one time or in one location, these measures may not apply elsewhere or in the future.

Technical and social factors also affect our ability to quantify the Internet's properties. Internet devices do not always provide the kind of measurements that are most useful for understanding the network. Many useful measures of Internet behavior are hidden from view, in some cases because the architecture of the Internet itself interferes. Collecting measurements of the Internet can result in huge datasets that are difficult to store, transfer, process, and analyze. Commercial service providers often do not share information about the internal details of their networks. Some forms of Internet measurement can violate privacy and raise security concerns.

Finally, the Internet shows a number of unusual statistical properties that complicate measurement attempts. These impose the need for extra care in taking and analyzing Internet measurements, and require the use of nonstandard statistical methods.

All of these factors complicate and impede Internet measurement, and have led to a scarcity of quantitative characterizations of the Internet. Unanswered Internet measurement questions span a diverse range. Some are very general and basic questions, such as “How big is the Internet?” and “How much traffic flows over the Internet?” However, many more practical and immediately useful facts are also unknown: “What is the structure of the Internet? What are the statistical properties of network traffic? What demands do different applications place on the network?” At lower levels of detail, network users and engineers often lack answers to even more specific questions about network properties, such as “What is the capacity of the path to my server?” and “How much peer-to-peer traffic is flowing on my network?”

However, all hope is not lost. There has been an immense amount of effort expended in recent years on various aspects of Internet measurement. Significant progress has been made on many fronts. Important aspects of the Internet's structure have been measured, at least in part, and some general understanding of how the network is organized is starting to emerge. Network traffic has been studied and many aspects of its statistics have been clarified. Applications such as the Web have been characterized and their properties have been documented.

More importantly, there have been considerable achievements in understanding

the specific challenges presented by Internet measurement, and in developing tools and methods to overcome those challenges. Tools are available for a wide range of Internet measurement tasks, and general methods have been developed to overcome many of the challenges presented in Internet measurement. As result, ongoing measurement of the Internet is becoming easier, and our ability to quantify and characterize the Internet is improving.

This book is about the field of Internet measurement. It covers the goals of various kinds of Internet measurements, the challenges that are presented by Internet measurement, the methods and tools that have been developed to facilitate Internet measurement, and the results that have been obtained to date in measuring the Internet.

1.1 Why Measure the Internet?

Why are Internet measurements important? The answer varies; many different people and organizations have an interest in measuring the network or in obtaining Internet measurements. But broadly speaking, interest in Internet measurements arises for three kinds of reasons: commercial, social, and technical.

Commercially, the ability to sell a product or provide information about a product to a large number of people requires a variety of Internet measurements. For example, demographic information is important: what is the reach of the Internet? How many individuals are connected in a given area? What fraction of users have high-speed connectivity and how many are dependent on dial-up connectivity? Where should network access points be placed? Will users with wireless connectivity be able to access the Internet?

Effective commerce on the Internet also requires an understanding of the network's performance properties. How long do Web pages take to download from a vendor's site? What is the capacity of the path from a customer to a vendor's Web server? How often do network problems prevent efficient transfer of information through the network? In fact these questions are so important that a number of companies have been created to provide answers, via services or products.

Socially, the need for statistics is clear: understanding the amount of network activity involving various sites and protocols gives considerable insight into social issues. Governments, scientists, and corporations may desire to have information about social implications of Internet use.

For example, popular Web sites attract a considerable amount of traffic. Given the millions of sites from which information may be available, it is important to have characterizations of popularity and content. In another example, emerging protocols are indicators of popularity of a new application. This is shown by the sudden ap-

pearance of Napster and later KaZaa, which pointed to the imminent explosion in file sharing.

Finally, there are many technical reasons for Internet measurement. The design of network components and protocols is strongly driven by the nature of Internet workloads. For example, router designs depend strongly on the statistical properties of network traffic and packet size distribution. The statistical properties of Web pages influence the performance and design of Web servers and browsers. Understanding the topology of the network helps identify the places where performance problems may arise and how applications may choose to adapt to the network. The popularity of new applications (such as network games) leads to efforts to build variants (new games). The popularity of new applications can also drive improvements to associated protocols – as in the case of the explosion of Web traffic, which motivated the improvement of the basic HTTP/1.0 protocol to yield HTTP/1.1.

1.2 How to Read this Book

Thus, the growing interest in Internet measurement is not surprising. In general, a need for metrics often signals the emergence of a discipline. The ability to measure is crucial to systematically capture a body of knowledge. The goal of this book is to present what is known about the emerging field of Internet measurement.

Although the field of Internet measurement is relatively new, there are already a vast set of results in the area. It is a considerable challenge to organize a body of work as large as that in this book. We have adopted an approach based on a number of principles.

First, we have divided the ‘things being measured’ into three areas:

Infrastructure. This includes things like links and routers, and their interconnection patterns and various properties. It also includes properties that arise due to the interaction of traffic with the network, such as delay, loss, and throughput.

Traffic. This covers traffic measurement of all kinds, including traffic volume and higher-level characterizations.

Applications. This looks at the most important applications: DNS, Web, peer-to-peer, and online games.

Within each of these three areas, we have organized the topics into four parts:

Properties. This reviews the properties and metrics that are important to measure in this area. For example, in the case of traffic one may seek to measure bytes or packets per unit time, among other metrics.

Challenges. This discusses the various difficulties that arise when trying to mea-

sure the above properties. Example challenges include the inability to determine network topology due to lack of accurate measurement methods.

Tools. This covers the methods and tools that have been developed to measure properties and overcome the above challenges, where possible. Example tools are path capacity measurement methods, or inference techniques to supply missing data.

State of the Art. This summarizes what is known about the properties of interest in today's (2006) Internet. For example, the known statistical properties of Internet traffic are covered.

Here we must note a caveat: the Internet changes at a fast pace. Any attempt to present 'State of the Art' in Internet measurements is certain to be quickly outdated, at least in some aspects. In state of the art discussions we have tried to emphasize the properties that seem to be relatively invariant (or slowly changing) in time. However, material in these sections is likely to have a shorter shelf life than the rest of the book.

Organization of the Book. Having discussed the structuring principles we have used, we can now review the organization of the book. The general organization of the book is shown in Figure 1.1.

In the figure, dotted arrows represent various options for reasonable starting points (we discuss these options later). Solid arrows represent how chapters depend on each other. An arrow from Chapter A to Chapter B means that Chapter B assumes familiarity with concepts or topics introduced in Chapter A (and all Chapter A's predecessors).

Part I consists of Chapters 1 through 4. These chapters provide background material that is necessary to understand before the body of the book can begin. The first chapter after this one, **Chapter 2**, provides an introduction to the architecture of the Internet. It covers the organizational principles of the Internet at an introductory level, providing the reader who has little previous knowledge of the Internet enough background to understand what follows in the rest of the book. **Chapter 3** covers the analytic tools needed to discuss Internet measurements in a formal way, covering basics in linear algebra, probability models, statistics, and graph theory. As with Chapter 2, the coverage is not thorough, but provides sufficient foundation for topics later in the book. Note however that the sections on probability, statistics, and graph theory each include subsections entitled "Special Issues in the Internet." These cover topics that arise particularly in Internet measurement and should be read even if one is already familiar with the basics in this chapter. **Chapter 4** covers the pragmatics of Internet measurement: where and how measurements can be taken, how time is measured, what existing sources of data already exist, and how measurements are taken at different layers.

Part II represents the body of the book, which is organized according to the principles mentioned above. It consists of Chapters 5 to 7. **Chapter 5** covers infras-

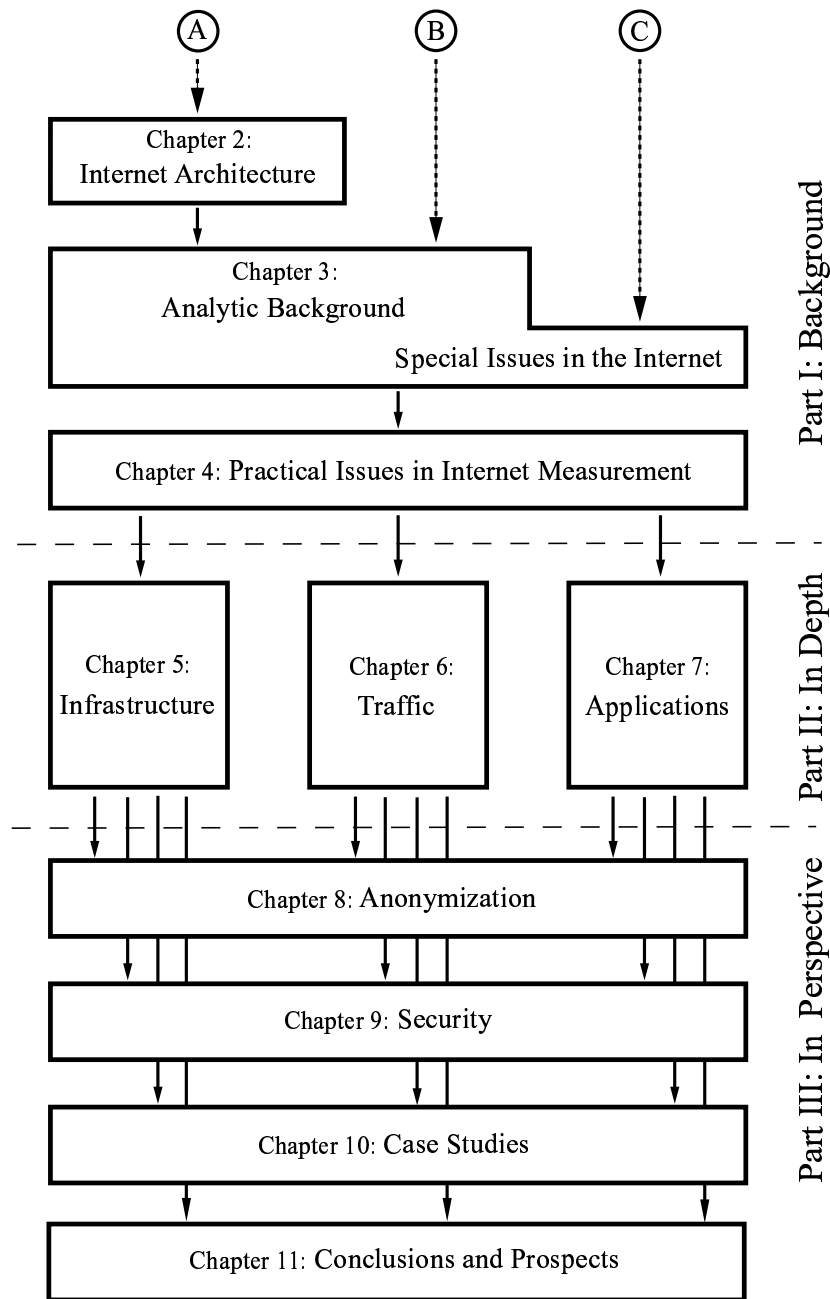


Figure 1.1 Organization of the Book.

structure, **Chapter 6** covers traffic, and **Chapter 7** covers applications. Within each of these chapters subject matter is organized according to Properties, Challenges, Tools, and State of the Art.

Part III covers material that spans multiple areas. It consists of Chapters 8 to 11. **Chapter 8** covers methods for removing sensitive information from measured data: anonymization. Anonymization can be needed when any kind of measurements are shared – infrastructure, traffic, or application measurements. **Chapter 9** covers ways in which Internet measurements can be used in the area of security. The need for security oriented analysis of Internet measurements has grown as the importance of the Internet has increased, and as the volume and variety of attacks has grown. Next, **Chapter 10** looks at some examples of Internet measurement in practice. It looks at some specific software tools that have been developed for Internet measurement, and some specific large-scale projects that are focused on Internet measurement. Finally, **Chapter 11** concludes with an attempt to review where the field of Internet measurement has come from, and (to the extent possible) where it is going.

Paths Through the Book. This book can be read in different ways depending on the reader’s background and goals. We expect this book to be useful to networking practitioners, to new and seasoned researchers in Internet measurement, and to outside experts such as statisticians and physicists with an interest in the field.

First of all, readers coming from outside the field of networking can start at **A** in Figure 1.1. This provides the necessary background in networking for the non-specialist.

Readers with a basic understanding of networking (equivalent to an undergraduate course in computer networks) can start at **B** in Figure 1.1, i.e., skip Chapter 2 and begin at Chapter 3.

Furthermore, if the reader is grounded in the analytic methods covered in Chapter 3, most of the chapter can be skipped; however, the three subsections entitled “Special Issues in the Internet” should be read regardless. This is entry point **C**. Graduate students in computer science will probably want to start here.

Leaving the “Background” chapters, the “In Depth” chapters are relatively independent of each other. Readers having a special interest in one topic (for example, a reader looking for a review of Internet traffic measurement) can move directly to the chapter of interest. Furthermore, readers already having a background in Internet measurement can turn directly to the sections of interest in Chapters 5 through 11.

Finally the “In Perspective” chapters cover topics that relate to multiple areas. Anonymization and security connect in various ways to infrastructure, traffic, and applications. The case studies involve tools and projects addressing multiple areas as well. These chapters are best read after most or all of Part II.

1.3 Resources for More Information

At many places throughout this book we note texts and references to the research literature that provide more information about topics at hand. As already mentioned, this literature is extensive; we have attempted to restrict citations to key papers in each area, but nonetheless our bibliography runs to around 900 entries. This large set of references is needed because almost all of the literature in this area is in the form of RFCs and research papers in journals, conferences, and workshops. There are few or no prior books that extensively cover our topics. As such, there are many topics in the book which can only be covered completely by combining information from a large number of papers or other publications.

Where possible we have tried to cite journal papers rather than conference papers, and conference papers rather than workshop papers. This is in the expectation that journal papers represent the most polished results in an area, and that they should have the longest shelf life. Likewise, conference papers are often more rigorously reviewed than workshop papers. We also reference many RFCs, which represent relatively durable sources of information and are often needed for in-depth understanding of certain protocols.

It is clear that Internet measurement is a rapidly advancing field. New developments on most of the topics in this book are appearing regularly. Here we note where to look for ongoing and future work in the field.

The *Internet Measurement Conference* (IMC) has a purview most closely related to the focus of this book. It began as a small workshop in 2001 and 2002, became a larger conference in 2003, and has taken place annually since then. This is a high-quality conference; in recent years it has selected its program of about 35 papers from about 150 submissions. The conference has grown dramatically in submissions and attendance since its inception; the growth of IMC is indicative of the increasing level of research interest in Internet measurement.

Another venue that has this topic as its primary focus is the *Passive and Active Measurement* (PAM) workshop. The focus of PAM is to present up-and-coming work and serve as an early testbed to discuss ideas. There has been considerable work relating to measurement infrastructure hardware that has primarily been presented in this venue. There has been significant growth in submissions and attendance in PAM as well.

Internet measurement topics have been addressed in the broader networking literature as well. The Infocom conference is broad and generally has a more analytic bent than other networking conferences. It has featured measurement-related research work in several areas of link-level, flow-level, and bandwidth measurements, as well as topology modeling. The ACM SIGMETRICS conference has a focus on measurement issues in computer systems in general, and as such covers topics in In-

ternet measurement (as well as a number of topics outside the scope of this book). Finally, ACM SIGCOMM focuses on all aspects related to network communication and protocols, and has presented papers that include an Internet measurement component.

The flagship journal in computer networking is *IEEE/ACM Transactions on Networking*, which has published many significant papers on Internet measurement. Other journals sometimes carrying results in Internet measurement are *Computer Communication Review* and *Computer Networks*.

