Volume Title: Economic Analysis of the Digital Economy

Volume Author/Editor: Avi Goldfarb, Shane M. Greenstein, and Catherine E. Tucker, editors

Volume Publisher: University of Chicago Press

Volume ISBN: 0-226-20684-X; 978-0-226-20684-4

Volume URL: http://www.nber.org/books/gree13-1

Conference Date: June 6–7, 2013

Publication Date: April 2015

Chapter Title: Modularity and the Evolution of the Internet

Chapter Author(s): Timothy Simcoe

Chapter URL: http://www.nber.org/chapters/c13000

Chapter pages in book: (p. 21 – 47)

1

Modularity and the Evolution of the Internet

Timothy Simcoe

1.1 Introduction

The Internet is a global computer network comprised of many smaller networks, all of which use a common set of communications protocols. This network is important not only because it supports a tremendous amount of economic activity, but also as a critical component within a broader constellation of technologies that support the general-purpose activity of digital computing. Given its widespread use and complementary relationship to computing in general, the Internet is arguably a leading contemporary example of what some economists have called a general purpose technology (GPT).

The literature on GPTs highlights the importance of positive feedback between innovations in a GPT-producing sector and the process of "coinvention" (i.e., user experimentation and discovery) in various application sectors that build upon the GPT.[1] Much of this literature elaborates on the implications of coinvention for understanding GPT diffusion and the timing of associated productivity impacts.[2] However, the literature on GPTs is

Timothy Simcoe is associate professor of strategy and innovation at Boston University School of Management and a faculty research fellow of the National Bureau of Economic Research.

This research was funded by the NBER Digitization program with support from the Kauffman Foundation. Useful comments were provided by Tim Bresnahan, Shane Greenstein, Avi Goldfarb, Joachim Henkel, and Catherine Tucker. All errors are my own, and comments are welcome: tsimcoe@bu.edu. For acknowledgments, sources of research support, and disclosure of the author's material financial relationships, if any, please see http://www.nber.org/chapters/c13000.ack.

1. See Bresnahan (2010) for a recent review of this literature.

2. For a historical example, see Paul David (1990) on the role of coinvention in industrial electrification. For a contemporary quantitative application of these ideas, see Dranove et al.'s (2012) analysis of the productivity benefits from adopting health information technology.

less precise about how the supply of a GPT can or should be organized, or what prevents a GPT from encountering decreasing returns as it diffuses to application sectors with disparate needs and requirements.

This chapter provides an empirical case study of the Internet that demonstrates how a *modular* system architecture can have implications for industrial organization in the GPT-producing sector, and perhaps also prevent the onset of decreasing returns to GPT innovation. In this context, the term "architecture" refers to an allocation of computing tasks across various subsystems or components that might either be jointly or independently designed and produced. The term "modularity" refers to the level (and pattern) of technical interdependence among components. I emphasize voluntary cooperative standards development as the critical activity through which firms coordinate complementary innovative activities and create a modular system that facilitates a division of innovative labor. Data collected from the two main Internet standard-setting organizations (SSOs), the Internet Engineering Task Force (IETF), and World Wide Web Consortium (W3C), demonstrate the inherent modularity of the Internet architecture, along with the division of labor it enables. Examining citations to Internet standards provides evidence on the diffusion and commercial application of innovations within this system.

The chapter has two main points. First, architectural choices are multidimensional, and can play an essential role in the supply of digital goods. In particular, choices over modularity can shape trade-offs between generality and specialization among innovators and producers. Second, SSOs play a crucial role in designing modular systems, and can help firms internalize the benefits of coordinating innovation within a GPT-producing sector. While these points are quite general, it is not possible to show how they apply to all digital goods. Instead, I will focus on a very specific and important case, showing how modularity and SSOs played a key role in fostering design and deployment of the Internet.

The argument proceeds in three steps. First, after reviewing some general points about the economics of modularity and standards, I describe the IETF, the W3C, and the Transmission Control Protocol/Internet Protocol (TCP/IP) "protocol stack" that engineers use to characterize the Internet's architecture. Next, I use data from the IETF and W3C to illustrate the modularity of the system and the specialized division of labor in Internet standard setting. In this second step, I present results from two empirical analyses. The first analysis demonstrates the modular nature of the Internet by showing that citations among technical standards are highly concentrated within "layers" or modules in the Internet Protocol stack. The second analysis demonstrates that firms contributing to Internet standards development also specialize at particular layers in the protocol stack, suggesting that the technical modularity of the Internet architecture closely corresponds to the division of labor in standards production. The final step in the chapter's

broader argument is to consider how components within a modular system evolve and are utilized through time. To illustrate how these ideas apply to the Internet, I return to citation analysis and show that intermodule citations between standards occur later than intramodule citations. Similarly, citations from patents (which I use as a proxy for commercial application of Internet standards) occur later than citations from other standards. These patterns suggest that modularity facilitates asynchronous coinvention and application of the core GPT, in contrast to the contemporaneous and tightly coupled design process that occur within layers.

1.1.1 Modularity in General

Modularity is a general strategy for designing complex systems. The components in a modular system interact with one another through a limited number of standardized interfaces.

Economists often associate modularity with increasing returns to a finer division of labor. For example, Adam Smith's famous description of the pin factory illustrates the idea that system-level performance is enhanced if specialization allows individual workers to become more proficient at each individual step in a production process. Limitations to such increasing returns in production may be imposed by the size of the market (Smith 1776; Stigler and Sherwin 1985) or through increasing costs of coordination, such as the cost of "modularizing" products and production processes (Becker and Murphy 1992). The same idea has been applied to innovation processes by modeling educational investments in reaching the "knowledge frontier" as a fixed investment in human capital that is complementary to similar investments made by other workers (Jones 2008). For both production and innovation, creating a modular division of labor is inherently a coordination problem, since the ex post value of investments in designing a module or acquiring specialized human capital necessarily depend upon choices and investments made by others.

A substantial literature on technology design describes alternative benefits to modularity that have received less attention from economists. Herb Simon (1962) emphasizes that modular design isolates technological interdependencies, leading to a more robust system, wherein the external effects of a design change or component failure are limited to other components within the same module. Thus, Simon highlights the idea that upgrades and repairs can be accomplished by swapping out a single module instead of rebuilding a system from scratch. Baldwin and Clark (2000) develop the idea that by minimizing "externalities" across the parts of a system, modularity multiplies the set of options available to component designers (since design constraints are specified ex ante through standardized interfaces, as opposed to being embedded in ad hoc interdependencies), and thereby facilitates decentralized search of the entire design space.

Economists often treat the modular division of labor as a more or less

inevitable outcome of the search for productive efficiency, and focus on the potential limits to increasing returns through specialization. However, the literature on technology design is more engaged with trade-offs that arise when selecting between a modular and a tightly integrated design. For example, a tightly integrated or nondecomposable design may be required to achieve optimal performance. The fixed costs of defining components and interfaces could also exceed the expected benefits of a modular design that allow greater specialization and less costly ex post adaptation. Thus, modularity is not particularly useful for a disposable single-purpose design. A more subtle cost of modularity is the loss of flexibility at intensively utilized interfaces. In a sense, modular systems "build in" coordination costs, since modifying an interface technology typically requires a coordinated switch to some new standard.[3]

The virtues of modular design for GPTs may seem self-evident. A technology that will be used as a shared input across many different application sectors clearly benefits from an architecture that enables decentralized end-user customization and a method for upgrading "core" functionality without having to overhaul the installed base. However, this may not be so clear to designers at the outset, particularly if tight integration holds out the promise of rapid development or superior short-run performance. For example, during the initial diffusion of electricity, the city electric light company supplied generation, distribution, and even lights as part of an integrated system. Langlois (2002) describes how the original architects of the operating system for the IBM System 360 line of computers adopted a nondecomposable design, wherein "each programmer should see all the material."[4] Similarly, Bresnahan and Greenstein (1999) describe how divided technical leadership—which might be either a cause or a consequence of product modularity—did not emerge in computing until the personal computer era.

The evolution or choice of a modular architecture may also reflect expectations about the impact of modularity on the division of rents in the GPT-producing sector. For example, during the monopoly telecommunications era, AT&T had a long history of opposing third-party efforts to sell equipment that would attach to its network.[5] While the impact of compatibility on competition and the distribution of rents is a complex topic that goes beyond the scope of this chapter, the salient point is that the choice of a modular architecture—or at a lower level, the design of a specific interface—will

3. A substantial economics literature explores such dynamic coordination problems in technology adoption, starting from Arthur (1989), David (1985), and Farrell and Saloner (1986).

4. The quote comes from Brooks (1975).

5. Notable challenges to this arrangement occurred in the 1956 "Hush-a-Phone" court case (238 F.2d 266, D.C. Cir., 1956) and the Federal Communication Commission's 1968 Carterphone ruling (13 F.C.C.2d 420).

not necessarily reflect purely design considerations in a manner that weighs social costs and benefits.[6]

It is difficult to say what a less modular Internet would look like. Comparisons to the large closed systems of earlier eras (e.g., the IBM mainframe and the AT&T telecommunications network) suggest that there would be less innovation and commercialization by independent users of the network, in part because of the greater costs of achieving interoperability. However, centralized design and governance could also have benefits in areas such as improved security. Instead of pursuing this difficult counterfactual question, the remainder of this chapter will focus on documenting the modularity of the Internet architecture and showing how that modularity is related to the division of labor in standardization and the dynamics of complementary innovation.

1.1.2 Setting Standards

If the key social trade-off in selecting a modular design involves up-front fixed costs versus ex post flexibility, it is important to have a sense of what is being specified up front. Baldwin and Clark (2000) argue that a modular system partitions design information into visible design rules and hidden parameters. The visible rules consist of (a) an architecture that describes a set of modules and their functions, (b) interfaces that describe how the modules will work together, and (c) standards that can be used to test a module's performance and conformity to design rules. Broadly speaking, the benefits of modularity flow from hiding many design parameters in order to facilitate entry and lower the fixed costs of component innovation, while its costs come from having to specify and commit to those design rules before the market emerges.

The process of selecting globally visible design parameters is fundamentally a coordination problem, and there are several possible ways of dealing with it. Farrell and Simcoe (2012) discuss trade-offs among four broad paths to compatibility: decentralized technology adoption (or "standards wars"); voluntary consensus standard setting; taking cues from a dominant "platform leader" (such as a government agency or the monopoly supplier of a key input); and ex post efforts to achieve compatibility through converters and multihoming. In the GPT setting, each path to compatibility provides an alternative institutional environment for solving the fundamental contracting problem among GPT suppliers, potential inventors in various applications sectors, and consumers. That is, different modes of standardization imply alternative methods of distributing the ex post rents from complementary inventions, and one can hope that some combination of conscious

6. See Farrell (2007) on the general point and MacKie-Mason and Netz (2007) for one example of how designers could manipulate a specific interface.

choice and selection pressures pushes us toward a standardization process that promotes efficient ex ante investments in innovation.

While all four modes of standardization have played a role in the evolution of the Internet, this chapter will focus on consensus standardization for two reasons.[7] First, consensus standardization within SSOs (specifically, the IETF and W3C, as described below) is arguably the dominant mode of coordinating the design decisions and the supply of new interfaces on the modern Internet. And second, the institutions for Internet standard setting have remarkably transparent processes that provide a window onto the architecture of the underlying system, as well as the division of innovative labor among participants who collectively manage the shared technology platform. If one views the Internet as a general purpose technology, these standard-setting organizations may provide a forum where GPT-producers can interact with application-sector innovators in an effort to internalize the vertical (from GPT to application) and horizontal (among applications) externalities implied by complementarities in innovation across sectors, as modeled in Bresnahan and Trajtenberg (1995).

1.2 Internet Standardization

There are two main organizations that define standards and interfaces for the Internet: the Internet Engineering Task Force (IETF) and World Wide Web Consortium (W3C). This section describes how these two SSOs are organized and explains their relationship to the protocol stack that engineers use to describe the modular structure of the network.

1.2.1 History and Process

The IETF was established in 1986. However, the organization has roots that can be traced back to the earliest days of the Internet. For example, all of the IETF's official publications are called "Requests for Comments" (RFCs), making them part of a continuous series that dates back to the very first technical notes on packet-based computer networking.[8] Similarly, the first two chairs of the IETF's key governance committee, called the Internet Architecture Board (IAB), were David Clark of MIT and Vint Cerf, who worked on the original IP protocols with Clark before moving to the Defense Advanced Research Projects Agency (DARPA) and funding the

7. For example, Russell (2006) describes the standards war between TCP/IP and the OSI protocols. Simcoe (2012) analyzes the performance of the IETF as a voluntary SSO. Greenstein (1996) describes the NSF's role as a platform leader in the transition to a commercial Internet. Translators are expected to play a key role in the transition to IPv6, and smartphones are multihoming devices because they select between Wi-Fi (802.11) and cellular protocols to establish a physical layer network connection.

8. RFC 1 "Host Software" was published by Steve Crocker of UCLA in 1969 (http://www.rfc-editor.org/rfc/rfc1.txt). The first RFC editor, Jon Postel of UCLA, held the post from 1969 until his death in 1998.

initial deployment of the network. Thus, in many ways, the early IETF formalized a set of working relationships among academic, government, and commercial researchers who designed and managed the Advanced Research Projects Agency Network (ARPANET) and its successor, the National Science Foundation Network (NSFNET).

Starting in the early 1990s, the IETF evolved from its quasi-academic roots into a venue for coordinating critical design decisions for a commercially significant piece of shared computing infrastructure.[9] At present the organization has roughly 120 active technical working groups, and its meetings draw roughly 1,200 attendees from a wide range of equipment vendors, network operators, application developers, and academic researchers.[10]

The W3C was founded by Tim Berners-Lee in 1994 to develop standards for the rapidly growing World Wide Web, which he invented while working at the European Organization for Nuclear Research (CERN). Berners-Lee originally sought to standardize the core web protocols, such as the Hypertext Markup Language (HTML) and Hypertext Transfer Protocol (HTTP), through the IETF. However, he quickly grew frustrated with the pace of the IETF process, which required addressing every possible technical objection before declaring a consensus, and decided to establish a separate consortium, with support from CERN and MIT, that would promote faster standardization, in part through a more centralized organization structure (Berners-Lee and Fischetti 1999).

The IETF and W3C have many similar features and a few salient differences. Both SSOs are broadly open to interested participants. However, anyone can "join" the IETF merely by showing up at a meeting or participating on the relevant e-mail listserv. The W3C must approve new members, who are typically invited experts or engineers from dues-paying member companies. The fundamental organizational unit within both SSOs is the working group (WG), and the goal of working groups is to publish technical documents.

The IETF and W3C working groups publish two types of documents. The first type of document is what most engineers and economists would call a standard: it describes a set of visible design rules that implementations should comply with to ensure that independently designed products work together well. The IETF calls this type of document a standards-track RFC, and the W3C calls them Recommendations.[11] At both SSOs, new standards must be approved by consensus, which generally means a substantial supermajority, and in practice is determined by a WG chair, subject

9. Simcoe (2012) studies the rapid commercialization of the IETF during the 1990s, and provides evidence that it produced a measurable slowdown in the pace of standards development.

10. http://www.ietf.org/documents/IETF-Regional-Attendance-00.pdf.

11. Standard-track RFCs are further defined as proposed standards, draft standards, or Internet standards to reflect their maturity level. However, at any given time, much of the Internet runs on proposed standards.

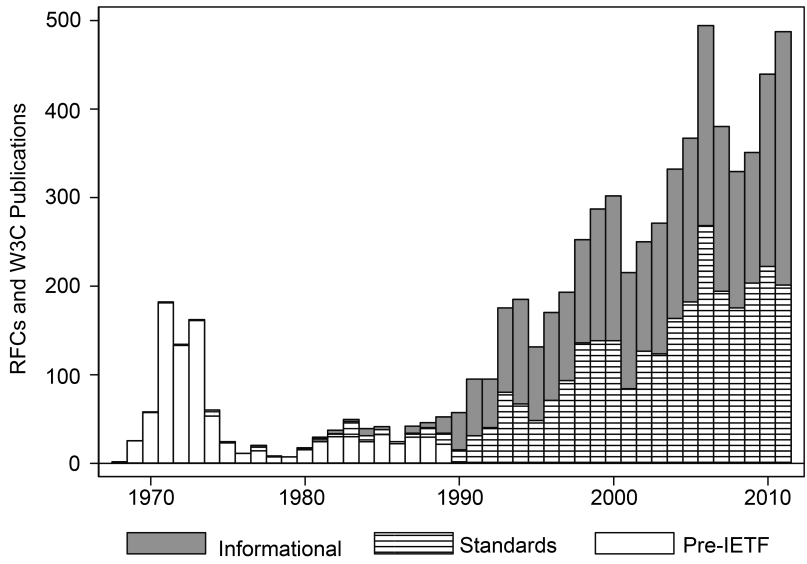

Fig. 1.1 Total RFCs and W3C publications (1969–2011)

Notes: Figure 1.1 plots a count of publications by the IETF and W3C. Pre-IETF publications refer to Request for Comments (RFCs) published prior to the formation of the IETF as a formal organization. Standards are standards-track RFCs published by IETF and W3C Recommendations. Informational publications are nonstandards-track IETF RFCs and W3C notes.

to formal appeal and review by the Internet Engineering Steering Group (IESG) or W3C director.[12]

The IETF and W3C working groups also publish documents that provide useful information without specifying design parameters. These informational publications are called nonstandards-track RFCs at the IETF and Notes at the W3C. They are typically used to disseminate ideas that are too preliminary or controversial to standardize, or information that complements new standards, such as "lessons learned" in the standardization process or proposed guidelines for implementation and deployment.

Figure 1.1 illustrates the annual volume of RFCs and W3C publications between 1969 and 2011. The chart shows a large volume of RFCs published during the early 1970s, followed by a dry spell of almost fifteen years, and then a steady increase in output beginning around 1990. This pattern coincides with a burst of inventive activity during the initial development of ARPANET, followed by a long period of experimentation with various

12. For an overview of standards-setting procedures at IETF, see RFC 2026 "The Internet Standards Process" (http://www.ietf.org/rfc/rfc2026.txt). The W3C procedures are described at http://www.w3.org/2005/10/Process-20051014/tr.

networking protocols—including a standards war between TCP/IP and various proprietary implementations of the open systems interconnection (OSI) protocol suite (Russell 2006). Finally, there is a second wave of sustained innovation associated with the emergence of TCP/IP as the de facto standard, commercialization of the Internet infrastructure and widespread adoption.

If we interpret the publication counts in figure 1.1 as a proxy for innovation investments, the pattern is remarkably consistent with a core feature of the literature on GPTs. In particular, there is a considerable time lag between the initial invention and the eventual sustained wave of complementary innovation that accompanies diffusion across various application sectors. There are multiple explanations for these adoption lags, which can reflect coordination delays such as the OSI versus TCP/IP standards war; the time required to develop and upgrade complementary inputs (e.g., routers, computers, browsers, and smartphones); or the gradual replacement of prior technology that is embedded in substantial capital investments. With respect to replacement effects, it is interesting to note that the share of IETF standards-track publications that upgrade or replace prior standards has averaged roughly 20 percent since 1990, when it becomes possible to calculate such statistics.

Another notable feature of figure 1.1 is the substantial volume of purely informational documents produced at IETF and W3C. This partly reflects the academic origins and affiliations of both SSOs, and highlights the relationship between standards development and collaborative research and development (R&D). It also illustrates how, at least for "open" standards, much of the information about how to implement a particular module or function is broadly available, even if it is nominally hidden behind the layer of abstraction provided by a standardized interface.

To provide a better sense of what is actually being counted in figure 1.1, table 1.1A lists some of the most important IETF standards, as measured by the number of times they have been cited in IETF and W3C publications, or as nonpatent prior art in a US patent in table 1.1B.

All of the documents listed in tables 1.1A and 1.1B are standards-track publications of the IETF.[13] Both tables contain a number of standards that one might expect to see on such a list, including Transmission Control Protocol (TCP) and Internet Protocol (IP), the core routing protocols that arguably define the Internet; the HTTP specification used to address resources on the Web; and the Session Initiation Protocol (SIP) used to control multimedia sessions, such as voice and video calls over IP networks.

Several differences between the two lists in tables 1.1A and 1.1B are also noteworthy. For example, table 1.1A shows that IETF and W3C publica-

13. I was not able to collect patent cites for W3C documents, and the W3C Recommendation that received the most SSO citations was a part of the XML protocol that received 100 cites.

Table 1.1A **Most cited Internet standards (IETF and W3C citations)**

Document	Year	IETF & W3C citations	Title
RFC 822	1982	346	Standard for the format of ARPA Internet text messages
RFC 3261	2002	341	SIP: Session Initiation Protocol
RFC 791	1981	328	Internet Protocol
RFC 2578	1999	281	Structure of Management Information Version 2 (SMIv2)
RFC 2616	1999	281	Hypertext Transfer Protocol—HTTP/1.1
RFC 793	1981	267	Transmission Control Protocol
RFC 2579	1999	262	Textual conventions for SMIv2
RFC 3986	2005	261	Uniform Resource Identifier (URI): Generic syntax
RFC 1035	1987	254	Domain names—implementation and specification
RFC 1034	1987	254	Domain names—concepts and facilities

Note: This list excludes the most cited IETF publication, RFC 2119 "Key Words for Use in RFCs to Indicate Requirement Levels," which is an informational document that provides a standard for writing IETF standards, and is therefore cited by nearly every standards-track RFC.

tions frequently cite the Structure of Management Information Version 2 (SMIv2) protocol, which defines a language and database used to manage individual "objects" in a larger communications network (e.g., switches or routers). On the other hand, table 1.1B shows that US patents are more likely to cite security standards and protocols for reserving network resources (e.g., Dynamic Host Configuration Protocol [DHCP] and Resource Reservation Protocol [RSVP]). These differences hint at the idea that citations from the IETF and W3C measure technical interdependencies or knowledge flows within the computer-networking sector, whereas patent cites measure complementary innovation linked to specific applications of the larger GPT.[14] I return to this idea below when examining diffusion.

1.2.2 The Protocol Stack

The protocol stack is a metaphor used by engineers to describe the multiple layers of abstraction in a packet-switched computer network. In principle, each layer handles a different set of tasks associated with networked communications (e.g., assigning addresses, routing and forwarding packets, session management, or congestion control). Engineers working at a particular layer need only be concerned with implementation details at that layer, since the functions or services provided by other layers are described in a set of standardized interfaces. Saltzer, Reed, and Clark (1984) provide

14. Examining citations to informational publications reinforces this interpretation: The nonstandards-track RFCs most cited by other RFCs describe IETF processes and procedures, whereas the nonstandards-track RFCs most cited by US patents describe technologies that were too preliminary or controversial to standardize, such as Network Address Translation (NAT) and Cisco's Hot-Standby Router Protocol (HSRP). On average, standards receive many more SSO and patent citations than informational publications.

Table 1.1B Most cited Internet standards (US patent citations)

Document	Year	US Patent citations	Title
RFC 2543	1999	508	SIP: Session Initiation Protocol
RFC 791	1981	452	Internet Protocol
RFC 793	1981	416	Transmission Control Protocol
RFC 2002	1996	406	IP mobility support
RFC 3261	2002	371	SIP: Session Initiation Protocol
RFC 2131	1997	337	Dynamic Host Configuration Protocol
RFC 2205	1997	332	Resource ReSerVation Protocol (RSVP)—Version 1
RFC 1889	1996	299	RTP: A transport protocol for real-time applications
RFC 2401	1998	284	Security architecture for the Internet Protocol
RFC 768	1980	261	User Datagram Protocol

an early description of this modular or "end-to-end" network architecture that assigns complex application-layer tasks to "host" computers at the edge of the network, thereby allowing routers and switches to focus on efficiently forwarding undifferentiated packets from one device to another. In practical (but oversimplified) terms, the protocol stack allows application designers to ignore the details of transmitting a packet from one machine to another, and router manufacturers to ignore the contents of the packets they transmit.

The canonical TCP/IP protocol stack has five layers: applications, transport, Internet, link (or routing), and physical. The IETF and W3C focus on the four layers at the "top" of the stack, while various physical layer standards are developed by other SSOs, such as the IEEE (Ethernet and Wi-Fi/802.11b), or 3GPP (GSM and LTE). I treat the W3C as a distinct layer in this chapter, though most engineers would view the organization as a developer of application-layer protocols.[15]

In the management literature on modularity, the "mirroring hypothesis" posits that organizational boundaries will correspond to interfaces between modules. While the causality of this relationship has been argued in both directions (e.g., Henderson and Clark 1990; Sanchez and Mahoney 1996; Colfer and Baldwin 2010), the IETF and W3C clearly conform to the basic cross-sectional prediction that there will be a correlation between module and organizational boundaries. In particular, both organizations assign individual working groups to broad technical areas that correspond to distinct modules within the TCP/IP protocol stack.

For each layer, the IETF maintains a technical area comprised of several related working groups overseen by a pair of area directors who sit on the Internet Engineering Steering Group (IESG). In addition to the areas cor-

15. Within the W3C there are also several broad areas of work, including Web design and applications standards (HTML, CSS, Ajax, SVG), Web infrastructure standards (HTTP and URI) that are developed in coordination with IETF, XML standards, and standards for Web services (SOAP and WSDL).

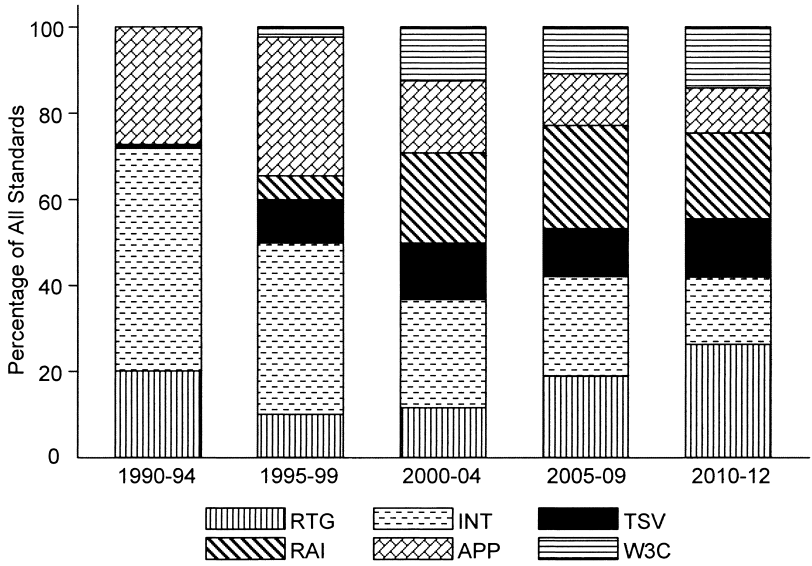

Fig. 1.2 Evolution of the Internet Protocol Stack

Notes: Figure 1.2 plots the share of all IETF and W3C standards-track publications associated with each layer in the Internet Protocol Stack, based on the author's calculations using data from IETF and W3C. The full layer names are: RTG = routing, INT = Internet, TSV = transport, RAI = real-time applications and infrastructure, APP = applications, and W3C = W3C. The figure excludes RFCs from the IETF operations and security areas, which are not generally treated as a "layer" within the protocol stack (see figure 1.3).

responding to layers in the traditional protocol stack, the IETF has created a real-time applications area to develop standards for voice, video, and other multimedia communications sessions. This new layer sits "between" application and transport-layer protocols. Finally, the IETF manages two technical areas—security and operations—that exist outside of the protocol stack and develop protocols that interact with each layer of the system.

Figure 1.2 illustrates the proportion of new IETF and W3C standards from each layer of the protocol stack over time. From 1990 to 1994, protocol development largely conformed to the traditional model of the TCP/IP stack. Between 1995 and 1999, the emergence of the Web was associated with an increased number of higher-level protocols, including the early IETF work on HTML/HTTP, and the first standards from the W3C and real-time applications and infrastructure layers. From 2000 to 2012 there is a balancing out of the share of new standards across the layers of the protocol stack. The resurgence of the routing layer between 2005 and 2012 was based on a combination of upgrades to legacy technology and the creation of new standards, such as label-switching protocols (MPLS) that allow IP networks to function more like a switched network that maintains a specific path between source and destination devices.

Figure 1.2 illustrates several points about the Internet's modular architecture that are linked to the literature on GPTs. If one views the Web as a technology that enables complementary inventions across a wide variety of application sectors (e.g., e-commerce, digital media, voice-over IP, online advertising, or cloud services), it is not surprising to see initial growth in application-layer protocol development, followed by the emergence of a new real-time layer, followed by a resurgence of lower-layer routing technology. This evolution is broadly consistent with the notion of positive feedback from application-sector innovations to extensions of the underlying GPT. Unfortunately, like most papers in the GPT literature, I lack detailed data on Internet-related inventive activity across the full range of application sectors, and I am therefore limited to making detailed observations about the innovation process where it directly touches the GPT. Nevertheless, if one reads the RFCs and W3C Recommendations, links to protocols developed by other SSOs to facilitate application sector innovation are readily apparent. Examples include standards for audio/video compression (ITU/H.264) and for specialized commercial applications of general-purpose W3C tools like the XML language.

Figure 1.2 also raises several questions that will be taken up in the remainder of the chapter. First, how modular is the Internet with respect to the protocol stack? In particular, do we observe that technical interdependencies are greater within than between layers? Is there a specialized division of labor in protocol development? Second, is it possible to preserve the modularity of the entire system when a new set of technologies and protocols is inserted in the middle of the stack, as with the real-time area? Finally, the dwindling share of protocol development at the Internet layer suggests that the network may be increasingly "locked in" to legacy protocols at its key interface. For example, the IETF has long promoted a transition to a set of next generation IP protocols (IPv6) developed in the 1990s, with little success. This raises the question of whether modularity and collective governance render technology platforms less capable of orchestrating "big push" technology transitions than alternative modes of platform governance, such as a dominant platform leader.

1.3 Internet Modularity

Whether the Internet is actually modular in the sense of hiding technical interdependencies and, if so, how that modularity relates to the division of innovative labor, are two separate questions. This section addresses them in turn.

1.3.1 Decomposability

Determining the degree of modularity of a technological system is fundamentally a measurement problem that requires answering two main questions: (1) how to identify interfaces or boundaries between modules, and

(2) how to identify interdependencies across modules. The TCP/IP protocol stack and associated technical areas within the IETF and W3C provide a natural way to group protocols into modules. I use citations among standards-track RFCs and W3C Recommendations to measure interdependencies. The resulting descriptive analysis is similar to the use of design structure matrices, as advocated by Baldwin and Clark (2000) and implemented in MacCormack, Baldwin, and Rusnak (2012), only using stack layers rather than source files to define modules, and citations rather than function calls to measure technical interdependencies.

Citations data were collected directly from the RFCs and W3C publications. Whether these citations are a valid proxy for technical interdependencies will, of course, depend on how authors use them. Officially, the IETF and W3C distinguish between normative and informative citations. Normative references "specify documents that must be read to understand or implement the technology in the new RFC, or whose technology must be present for the technology in the new RFC to work." Informative references provide additional background, but are not required to implement the technology described in a RFC or Recommendation.[16] Normative references are clearly an attractive measure of interdependency. Unfortunately, the distinction between normative and informative cites was not clear for many early RFCs, so I simply use all cites as a proxy. Nevertheless, even if we view informative cites as a measure of knowledge flows (as has become somewhat standard in the economic literature that relies on bibliometrics), the interpretation advanced below would remain apt, since a key benefit of modularity is the "hiding" of information within distinct modules or layers.

Figure 1.3 is a directed graph of citations among all standards produced by the IETF and W3C, with citing layers/technical areas arranged on the Y-axis and cited layers/areas arranged on the X-axis. Shading is based on each cell's decile in the cumulative citation distribution. Twenty-seven percent of all citations link two documents produced by the same working group, and I exclude these from the analysis.[17]

In a completely modular or decomposable system, all citations would be contained with the cells along the main diagonal. Figure 1.3 suggests that the Internet more closely resembles a nearly decomposable system, with the majority of technical interdependencies and information flows occurring either within a module or between a module and an adjacent layer in the protocol stack.[18] If we ignore the security and operations areas, 89 percent of all citations in figure 1.3 are on the main diagonal or an adjacent cell,

16. For the official IESG statement on citations, see http://www.ietf.org/iesg/statement/normative-informative.html.

17. Including within-WG citations would make the Internet architecture appear even more modular.

18. An alternative nonmodular and non-interdependent design configuration would be a hierarchy, with all cites either above or below the main diagonal.

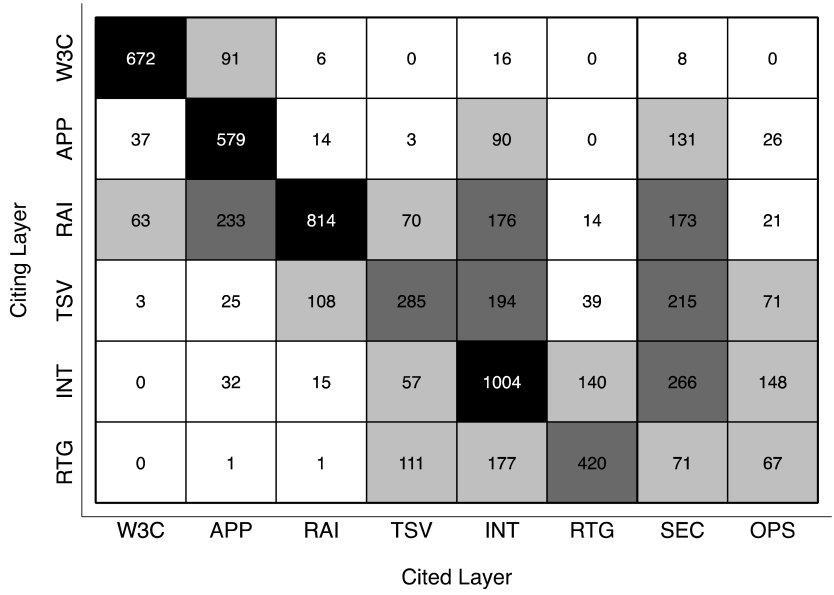

Fig. 1.3 Citations in the Internet Protocol Stack

Notes: Figure 1.3 is a matrix containing cumulative counts of citations from citing layer standards-track publications to cited layer standards-track publications based on the author's calculations using data from IETF and W3C. Layer names are: RTG = routing, INT = Internet, TSV = transport, RAI = real-time applications and infrastructure, APP = applications, W3C = W3C, SEC = security, and OPS = operations.

whereas a uniformly random citation probability would lead to just 44 percent of all citations on or adjacent to the main diagonal.

The exceptions to near-decomposability illustrated in figure 1.3 are also interesting. First, it is fairly obvious that security and operations protocols interface with all layers of the protocol stack: apparently there are some system attributes that are simply not amenable to modularization. While straightforward, this observation may have important implications for determining the point at which a GPT encounters decreasing returns to scale due to the costs of adapting a shared input to serve heterogeneous application sectors.

The second notable departure from near-decomposability in figure 1.3 is the relatively high number of interlayer citations to Internet layer protocols. This turns out to be a function of vintage effects. Controlling for publication-year effects in a Poisson regression framework reveals that Internet layer specifications are no more likely to receive between-layer citations than other standards.[19] Of course, the vintage effects themselves are inter-

19. These regression results are not reported here, but are available from the author upon request.

esting to the extent that they highlight potential "lock in" to early design choices made for an important interface, such as TCP/IP.

Finally, figure 1.3 shows that real-time and transport-layer protocols have a somewhat greater intermodule citation propensity than standards from other layers. Recall that these layers emerged later than the original applications, Internet, and routing areas (see figure 1.2). Thus, this observation suggests that when a new module is added to an existing system (perhaps to enable or complement coinvention in key application areas), it may be hard to preserve a modular architecture, particularly if that module is not located at the "edges" of the stack, as with the W3C.

1.3.2 Division of Labor

While figure 1.3 clearly illustrates the modular nature of the Internet's technical architecture, it does not reveal whether that modularity is associated with a specialized division of labor. This section will examine the division of labor among organizations involved in IETF standards development by examining their participation at various layers of the TCP/IP protocol stack.[20] The data for this analysis are extracted from actual RFCs by identifying all e-mail addresses in the section listing each author's contact information, and parsing those addresses to obtain an author's organizational affiliation.[21] The analysis is limited to the IETF, as it was not possible to reliably extract author information from W3C publications. On average, IETF RFCs have 2.3 authors with 1.9 unique institutional affiliations.

Because each RFC in this analysis is published by an IETF working group, I can use that WG to determine that document's layer in the protocol stack. In total, I use data from 3,433 RFCs published by 328 different WGs, and whose authors are affiliated with 1,299 unique organizations. Table 1.2 lists the fifteen organizations that participated (i.e., authored at least one standard) in the most working groups, along with the total number of standards-track RFCs published by that organization.

One way to assess whether there is a specialized division of labor in standards creation is to ask whether firms' RFCs are more concentrated within particular layers of the protocol stack than would occur under random assignment of RFCs to layers (where the exogenous assignment probabilities equal the observed marginal probabilities of an RFC occupying each layer in the stack). Comparing the actual distribution of RFCs across layers to a simulated distribution based on random choice reveals that organizations participating in the IETF are highly concentrated within particular

20. In principle, one might focus on specialization at the level of the individual participant. However, since many authors write a single RFC, aggregating to the firm level provides more variation in the scope of activities across modules.

21. In practice, this is a difficult exercise, and I combined the tools developed by Jari Arkko (http://www.arkko.com/tools/docstats.html) with my own software to extract and parse addresses.

Table 1.2 Major IETF participants

Sponsor	Unique WGs	Total standards
Cisco	122	590
Microsoft	65	130
Ericsson	42	147
IBM	40	102
Nortel	38	78
Sun	35	76
Nokia	31	83
Huawei	28	49
AT&T	27	50
Alcatel	26	64
Juniper	25	109
Motorola	24	42
MIT	24	42
Lucent	23	41
Intel	23	33

layers. Specifically, I compute the likelihood-based multinomial test statistic proposed by Greenstein and Rysman (2005) and find a value of –7.1 for the true data, as compared to a simulated value of –5.3 under the null hypothesis of random assignment.[22] The smaller value of the test statistic for the true data indicates agglomeration, and the test strongly rejects the null of random choice (SE = 0.17, $p = 0.00$).

To better understand this pattern of agglomeration in working group participation, it is helpful to consider a simplistic model of the decision to contribute to drafting an RFC. To that end, suppose that firm i must decide whether to draft an RFC for working group w in layer j. Each firm either participates in the working group or does not: $a_i = 0,1$. Let us further assume that *all* firms receive a gross public benefit B_w if working group w produces a new protocol. Firms that participate in the drafting process also receive a private benefit S_{iw} that varies across working groups, and incur a participation cost F_{ij} that varies across layers. In this toy model, public benefits flow from increasing the functionality of the network and growing the installed base of users. Private benefits could reflect a variety of idiosyncratic factors, such as intellectual property in the underlying technology or improved interoperability with proprietary complements. Participation costs are assumed constant within-layer to reflect the idea that there is a fixed cost to develop the technical expertise needed to innovate within a new module. If firms were all equally capable of innovating at any layer ($F_{ij} = F_{ik}$, for all $i, j \neq k$), there would be no specialized division of labor in standards production within this model.

22. Code for performing this test in Stata has been developed by the author and is available at http://econpapers.repec.org/software/bocbocode/s457205.htm.

To derive a firm's WG-participation decision, let Φ_w represent the endogenous probability that at least one other firm joins the working group. Thus, firm *i*'s payoff from working group participation are $B_w + S_{iw} - F_{ij}$, while the expected benefits of not joining are ΦB_w. If all firms have private knowledge of S_{iw}, and make simultaneous WG participation decisions, the optimal rule is to join the committee if and only if $(1 - \Phi_w)B_w + S_{iw} > F_{ij}$.

While dramatically oversimplified, this model yields several useful insights. First, there is a trade-off between free riding and rent seeking in the decision to join a technical committee. While a more realistic model might allow for some dissipation of rents as more firms join a working group, the main point here is that firms derive private benefits from participation, and are likely to join when S_{iw} is larger. Likewise, when S_{iw} is small, there is an incentive to let others develop the standard, and that free-riding incentive increases with the probability (Φ) that at least one other firm staffs the committee. Moreover, because Φ depends on the strategies of other prospective standards developers, this model illustrates the main challenge for empirical estimation: firms' decisions to join a given WG are simultaneously determined.

To estimate this model of WG participation I treat S_{iw} as an unobserved stochastic term, treat B_w as an intercept or WG random effect, and replace Φ_w with the log of one plus the actual number of other WG participants.[23] I parameterize F_{ij} as a linear function of two dummy variables—prior RFC (this layer) and prior RFC (adjacent layer)—that measure prior participation in WGs at the same layer of the protocol stack, or at an adjacent layer conditional on the same-layer dummy being equal to zero. These two dummies for prior RFC publication at "nearby" locations in the protocol stack provide an alternative measure of the division of labor in protocol development that may be easier to interpret than the multinomial test statistic reported above.

The regression results presented below ignore the potential simultaneity of WG participation decisions. However, if the main strategic interaction involves a trade-off between free riding and rent seeking, the model suggests that firms will be increasingly dispersed across working groups when the public benefits of protocol development (B_w) are large relative to the private rents (S_{iw}). Conversely, if we observe a strong positive correlation among participation decisions, the model suggests that private benefits of exerting some influence over the standard are relatively large and/or positively correlated across firms. It is also possible to explore the rent-seeking hypothesis by exploiting the difference between standards and nonstandards-track RFCs, an idea developed in Simcoe (2012). Specifically, if the normative aspects of standards-track documents provide greater opportunities for rent seek-

23. An alternative approach would be to estimate the model as a static game of incomplete information following Bajari et al. (2010). However, I lack instrumental variables that produce plausibly exogenous variation in Φ_w, as required for that approach.

Table 1.3 **Summary statistics**

Variable	Mean	SD	Min.	Max.
Stds.—track WG participation	0.06	0.24	0	1
Nonstds.—track participation	0.05	0.22	0	1
Prior RFC (this layer)	0.34	0.47	0	1
Prior RFC (adjacent layer)	0.17	0.38	0	1
log(1 + other participants)	2.11	0.86	0	4.51

ing (e.g., because they specify how products will actually be implemented), there should be a stronger positive correlation among firms' WG participation decisions, leading to more agglomeration when "participation" is measured as standards-track RFC production than when it is measured as nonstandards-track RFC publication.

The data used for this exercise come from a balanced panel of 43 organizations and 328 WGs where each organization contributed to ten or more RFCs and is assumed to be at risk of participating in every WG.[24] Table 1.3 presents summary statistics for the estimation sample and table 1.4 presents coefficient estimates from a set of linear probability models.[25]

The first four columns in table 1.4 establish that there is a strong positive correlation between past experience at a particular layer of the protocol stack and subsequent decisions to join a new WG at the same layer. Having previously published a standards-track RFC in a WG in a given layer is associated with a 5 to 7 percentage-point increase in the probability of joining a new WG at the same layer. There is a smaller but still significant positive association between prior participation at an adjacent layer and joining a new WG. Both results are robust to adding fixed or random effects for the WG and focal firm. Given the baseline probability of standards-track entry is 6 percent, the "same layer" coefficient corresponds to a marginal effect of 100 percent, and is consistent with the earlier observation that participation in the IETF by individual firms is concentrated within layers.

The fifth column in table 1.4 shows that the number of other WG participants has a strong positive correlation with the focal firm's participation decision. A 1 standard deviation increase in participation by other organizations, or roughly doubling the size of a working group, produces a 5 percentage-point increase in the probability of joining and is therefore roughly equivalent to prior experience at the same layer. I interpret this as

24. Increasing the number of firms in the estimation sample mechanically reduces the magnitude of the coefficient estimates (since firms that draft fewer RFCs participate in fewer working groups, and therefore exhibit less variation in the outcome) but does not qualitatively alter the results.

25. The linear probability model coefficients are nearly identical to average marginal effects from a set of unreported logistic regressions.

Table 1.4 **Linear probability models of IETF working group participation**

	Stds.—track particip.					
Outcome	(1)	(2)	(3)	(4)	(5)	(6)
Prior RFC (this layer)	0.06	0.07	0.07	0.05	0.06	0.06
	[6.87]***	[11.98]***	[9.64]***	[6.25]***	[11.24]***	[11.19]***
Prior RFC (adjacent layer)	0.02	0.02	0.02	0.01	0.02	0.01
	[3.27]***	[3.12]***	[2.72]***	[1.54]	[3.49]***	[2.36]**
log(other WG participants)					0.06	0.04
					[23.70]***	[17.82]***
WG random effects	N	Y	N	N	N	N
WG fixed effects	N	N	Y	Y	N	N
Firm fixed effects	N	N	N	Y	N	N
Observations	14,104	14,104	14,104	14,104	14,104	14,104

Notes: Unit of analysis is a firm-WG. Robust standard errors clustered by WG (except random effects model). *T*-statistics in brackets.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

evidence that private benefits from contributing to specification development are highly correlated across firms at the WG level, and that the costs of WG participation are low enough for these benefits to generally outweigh temptations to free ride when an organization perceives a WG to be important.

The last column in table 1.4 changes the outcome to an indicator for publishing a nonstandards-track RFC in a given WG. In this model, the partial correlation between a focal firm's participation decision and the number of other organizations in the WG falls by roughly one-third, to 0.04. A chi-square test rejects the hypothesis that the coefficient on log(other participants) is equal across the two models in columns (5) and (6) ($\chi^2(1) = 6.22$, $p = 0.01$). The stronger association among firms' WG participation decisions for standards-track RFCs than for nonstandards-track RFCs suggests that the benefits of exerting some influence over the standards process are large (relative to the participation costs and/or the public-good benefits of the standard) and positively correlated across firms.[26]

In summary, data from the IETF show that the division of labor in protocol development does conform to the boundaries established by the modular protocol stack. This specialized division of labor emerges through firms' decentralized decisions to participate in specification development in vari-

26. In unreported regressions, I allowed the standards/nonstandards difference to vary by layer, and found that standards was larger at all layers except applications and operations, with statistically significant differences for real-time, Internet, and routing and security.

ous working groups. The incentive to join a particular WG reflects both the standard economic story of amortizing sunk investments in developing expertise at a given layer, and idiosyncratic opportunities to obtain private benefits from shaping the standard. The results of a simple empirical exercise show that forces for agglomeration are strong, and suggests that incentives to participate for private benefit are typically stronger than free-riding incentives (perhaps because the fixed cost of joining a given committee are small). Moreover, firms' idiosyncratic opportunities to obtain private benefits from shaping a standard appear to be correlated across working groups, suggesting that participants know when a particular technical standard is likely to be important.

Finally, it is important to note that while this analysis focused on firms that produce at least ten RFCs in order to disentangle their motivations for working group participation, those forty-three firms are only a small part of the total population of 1,299 unique organizations that supplied an author on one or more RFCs. Large active organizations do a great deal of overall protocol development. However, the organizations that only contribute to one or two RFCs are also significant. By hiding many of the details of what happens within any given layer of the protocol stack, the Internet's modular architecture lowers the costs of entry and component innovation for this large group of small participants.

1.4 Diffusion across Modules and Sectors

The final step in this chapter's exploration of Internet modularity is to examine the distribution of citations to RFCs over time. As described above, lags in diffusion and coinvention occupy center stage in much of the literature on GPTs for two reasons: (1) they help explain the otherwise puzzling gap between the spread of seminal technologies and the appearance of macroeconomic productivity effects, and (2) they highlight the role of positive innovation externalities between and among application sectors and the GPT-producing sector.

Analyzing the age distribution of citations to standards can provide a window onto the diffusion and utilization of the underlying technology. However, it is important to keep in mind the limitations of citations as a proxy for standards utilization in the following analysis. In particular, we do not know whether any given citation represents a normative technical interdependency or an informative reference to the general knowledge embedded in an RFC. One might also wish to know whether citations come from implementers of the specification, or from producers of complements, who reference the interface in a "black box" fashion. While such fine-grained interpretation of citations between RFC are not possible in the data I use here, examining the origin and rate of citations does reveal some interesting patterns that hint at the role of modularity in the utilization of Internet standards.

1.4.1 Diffusion across Modules

I begin by examining citation flows across different modules and layers within the IETF and the TCP/IP protocol stack. If the level of technical interdependency between any two standards increases as we move inward from protocols in different layers, to protocols in the same layer, to protocols in the same working group, we should expect to see shorter citation lags. The intuition is straightforward: tightly coupled technologies need to be designed at the same time to avoid mistakes that emerge from unanticipated interactions. Two technologies that interact only through a stable interface need not be contemporaneously designed, since a well-specified interface defines a clear division of labor.[27]

To test the idea that innovations diffuse within and between modules at different rates, I created a panel of annual citations to standards-track RFCs for sixteen years following their publication. Citation dates are based on the publication year of the citing RFC. The econometric strategy is adapted from Rysman and Simcoe (2008). Specifically, I estimate a Poisson regression of citations to RFC i in citing year y that contains a complete set of age effects (where age equals citing year minus publication year) and a third order polynomial for citing years to control for time trends and truncation: $E[\text{Cites}_{iy}] = \exp\{\lambda_{\text{age}} + f(\text{Citing year})\}$.

To summarize these regression results, I set the citing year equal to 2000 and generate the predicted number of citations at each age. Dividing by the predicted cumulative cites over all sixteen years of RFC life yields a probability distribution that I call the citation-age profile. These probabilities are plotted and used to calculate a hypothetical mean citation age, along with its standard error (using the delta method).

Figure 1.4 illustrates the citation-age profile for standards-track RFCs using three different outcomes: citations originating in the same WG, citations originating in the same layer of the protocol stack, and citations from other layers of the protocol stack.[28] The pattern is consistent with the idea that more interconnected protocols are created closer together in time. Specifically, I find that the average age of citations within a working group is 3.5 years (SE = 0.75), compared to 6.7 years (SE = 0.56) for cites from the same layer and 8.9 years (SE = 0.59) for other layers.

The main lesson contained in figure 1.4 is that even within a GPT, innovations diffuse faster within than between modules. This pattern is arguably driven by the need for tightly interconnected aspects of the system to coordinate on design features simultaneously, whereas follow-on innovations can rely on the abstraction and information hiding provided by a well-defined

27. The costs of time shifting when the division of labor is nor clearly defined ex ante will be familiar to anyone who has worked on a poorly organized team project.

28. For this analysis, I exclude all cites originating in the security and operations layers (see figure 1.3).

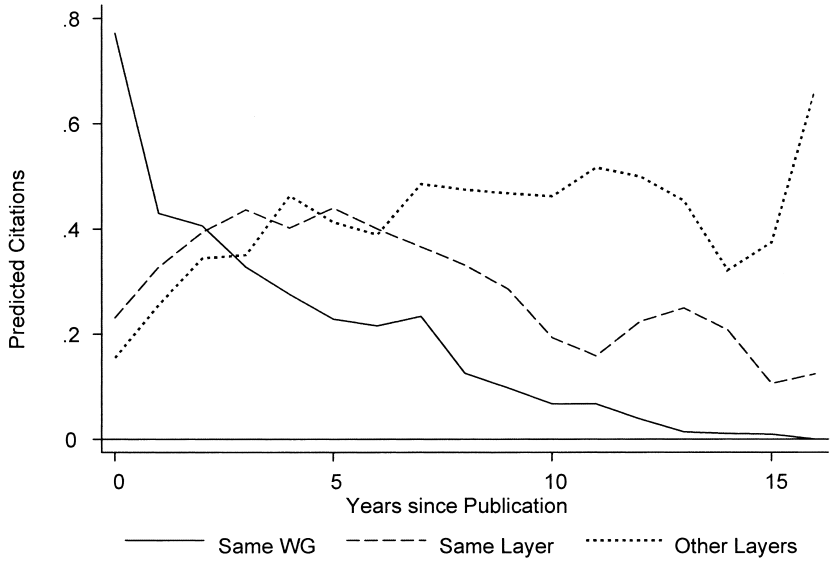

Fig. 1.4 Age profiles for RFC-to-RFC citations

interface. The importance of contemporaneous design for tightly coupled components may be compounded by the fact that many interface layers may need to be specified before a GPT becomes useful in specific application sectors. For example, in the case of electricity, the alternating versus direct current standards war preceded widespread agreement on standardized voltage requirements, which preceded the ubiquitous three-pronged outlet that works with most consumer devices (at least within the United States). While this accretion of interrelated interfaces is likely a general pattern, the Internet and digital technology seems particularly well suited to the use of a modular architecture to reduce the rate at which technical knowledge depreciates and to facilitate low-cost reuse and time shifting.

1.4.2 Diffusion across Sectors

To provide a sense of how the innovations embedded in Internet standards diffuse out into application sectors, I repeat the empirical exercise described above, only comparing citations among all RFCs to citations from US patents to RFCs. The citing year for a patent-to-RFC citation is based on the patent's application date. While there are many drawbacks to patent citations, there is also a substantial literature that argues for their usefulness as a measure of cumulative innovation based on the idea that each cite limits the scope of the inventor's monopoly and is therefore carefully assessed for its relevance to the claimed invention. For this chapter, the key assumption is simply that citing patents are more likely to reflect inventions that enable applications of the GPT than citations from other RFCs.

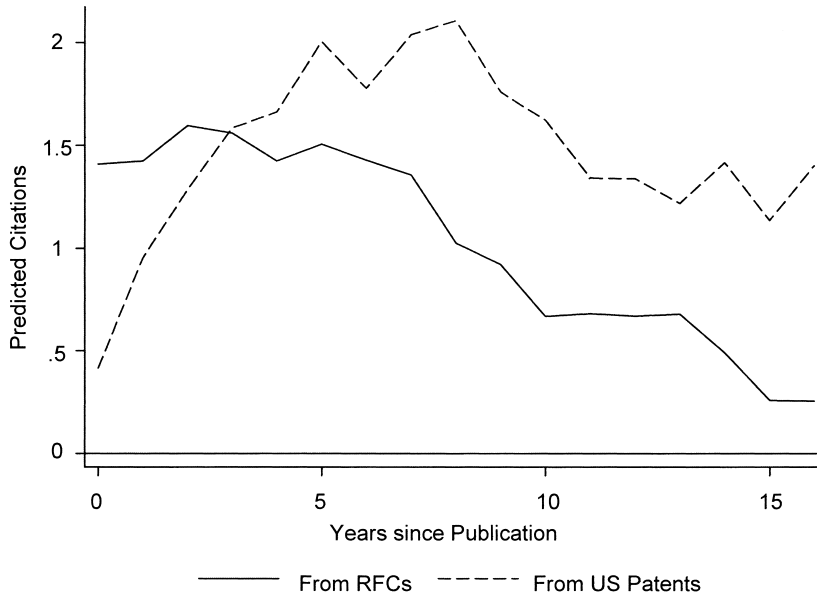

Fig. 1.5 Age profiles for RFC-to-RFC and US patent-to-RFC citations

Figure 1.5 graphs the age profiles for all RFC cites and all patent cites. The RFC age profile represents a cite-weighted average of the three lines in figure 1.4, and the average age of an RFC citation is 5.9 years (SE = 0.5). Patent citations clearly take longer to arrive, and are more persistent in later years than RFC cites. The average age of a US patent nonprior citation to an RFC is 8.2 years (SE = 0.51), which is quite close to the mean age for a citation from RFCs at other layers of the protocol stack.

At one level, the results illustrated in figures 1.4 and 1.5 are not especially surprising. However, these figures highlight the idea that a GPT evolves over time, partly in response to the complementarities between GPT-sector and application sector innovative activities. The citation lags illustrated in these figures are relatively short compared to the long delay between the invention of packet-switched networking and the emergence of the commercial Internet illustrated in figure 1.1. Nevertheless, it is likely that filing a patent represents only a first step in the process of developing application-sector-specific complementary innovations. Replacing embedded capital and changing organizational routines may also be critical, but are harder to measure, and presumably occur on a much longer time frame.

1.5 Conclusion

The chapter provides a case study of modularity and its economic consequences for the technical architecture of the Internet. It illustrates the modu-

lar design of the Internet architecture, the specialized division of innovative labor in Internet standards development, and the gradual diffusion of new ideas and technologies across interfaces within that system. These observations are limited to a single technology, albeit one that can plausibly claim to be a GPT with significant macroeconomic impacts.

At a broader level, this chapter suggests that modularity and specialization in the supply of a GPT may help explain its long-run trajectory. In the standard model of a GPT, the system-level trade-off between generality and specialization is overcome through "coinvention" within application sectors. These complementary innovations raise the returns to GPT innovation by expanding the installed base, and also by expanding the set of potential applications. A modular architecture facilitates the sort of decentralized experimentation and low-cost reusability required to sustain growth at the extensive margin, and delivers the familiar benefits of a specialized division of labor in GPT production.

Finally, this chapter highlights a variety of topics that can provide grist for future research on the economics of modularity, standard setting, and general-purpose technologies. For example, while modularity clearly facilitates an interfirm division of labor, even proprietary systems can utilize modular design principles. This raises a variety of questions about the interaction between modular design and "open" systems, such as the Internet, which are characterized by publicly accessible interfaces and particular forms of platform governance. The microeconomic foundations of coordination costs that limit the division of innovative labor within a modular system are another broad topic for future research. For example, we know little about whether or why the benefits of a modular product architecture are greater inside or outside the boundaries of a firm, or conversely, whether firm boundaries change in response to architectural decisions. Finally, in keeping with the theme of this volume, future research might ask whether there is something special about digital technology that renders it particularly amenable to the application of modular design principles. Answers to this final question will have important implications for our efforts to extrapolate lessons learned from studying digitization to other settings, such as life sciences or the energy sector.

References

Arthur, W. Brian. 1989. "Competing Technologies, Increasing Returns, and Lock-In by Historical Events." *Economic Journal* 97:642–65.

Bajari, P., H. Hong, J. Krainer, and D. Nekipelov. 2010. "Estimating Static Models of Strategic Interactions" *Journal of Business and Economic Statistics* 28 (4): 469–82.

Baldwin, C. Y., and K. B. Clark. 2000. *Design Rules: The Power of Modularity*, vol. 1. Boston: MIT Press.

Becker, G. S., and K. M. Murphy. 1992. "The Division-of-Labor, Coordination Costs and Knowledge." *Quarterly Journal of Economics* 107 (4): 1137–60.

Berners-Lee. T., and M. Fischetti. 1999. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its Inventor*. San Francisco: Harper.

Bresnahan, T. 2010. "General Purpose Technologies." In *Handbook of the Economics of Innovation*, vol. 2, edited by B. Hall and N. Rosenberg, 761–91. Amsterdam: Elsevier.

Bresnahan, T. F., and S. Greenstein. 1999. "Technological Competition and the Structure of the Computer Industry." *Journal of Industrial Economics* 47 (1): 1–40.

Bresnahan, T., and M. Trajtenberg. 1995. "General Purpose Technologies: Engines of Growth?" *Journal of Econometrics* 65:83.

Brooks, F. 1975. *The Mythical Man-Month*. Boston: Addison-Wesley.

Colfer, L., and C. Baldwin. 2010. "The Mirroring Hypothesis: Theory, Evidence and Exceptions." Working Paper no. 10–058, Harvard Business School, Harvard University.

David, Paul A. 1985. "Clio and the Economics of QWERTY." *American Economic Review* 77 (2): 332–37.

David, Paul A. 1990. "The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox." *American Economic Review Papers and Proceedings* 80 (2): 355–61.

Dranove, D., C. Forman, A. Goldfarb, and S. Greenstein. 2012. "The Trillion Dollar Conundrum: Complementarities and Health Information Technology." NBER Working Paper no. 18281, Cambridge, MA.

Farrell, J. 2007. "Should Competition Policy Favor Compatibility?" In *Standards and Public Policy*, edited by S. Greenstein and V. Stango. Cambridge: Cambridge University Press.

Farrell, J., and G. Saloner. 1986. "Installed Base and Compatibility—Innovation, Product Preannouncements, and Predation." *American Economic Review* 76 (5): 940–55.

Farrell, J., and T. Simcoe. 2012. "Four Paths to Compatibility." In *Oxford Handbook of the Digital Economy*, edited by M. Peitz and J. Waldfogel, 34–58. Oxford: Oxford University Press.

Greenstein, S. 1996. "Invisible Hand versus Invisible Advisors." In *Private Networks, Public Objectives*, edited by Eli Noam. Amsterdam: Elsevier.

Greenstein, S., and M. Rysman. 2005. "Testing for Agglomeration and Dispersion." *Economics Letters* 86 (3): 405–11.

Henderson, R., and K. B. Clark. 1990. "Architectural Innovation: The Reconfiguration of Existing Product Technologies and the Failure of Established Firms." *Administrative Science Quarterly* 35 (1): 9–30.

Jones, B. F. 2008. "The Knowledge Trap: Human Capital and Development Reconsidered." NBER Working Paper no. 14138, Cambridge, MA.

Langlois, R. 2002. "Modularity in Technology and Organization." *Journal of Economic Behavior & Organization* 49 (1): 19–37.

MacCormack, A., C. Baldwin, and J. Rusnak. 2012. "Exploring the Duality between Product and Organizational Architectures: A Test of the 'Mirroring' Hypothesis." *Research Policy* 41:1309–24.

MacKie-Mason, J., and J. Netz. 2007. "Manipulating Interface Standards as an Anticompetitive Strategy." In *Standards and Public Policy*, edited by S. Greenstein and V. Stango, 231–59. Cambridge: Cambridge University Press.

Russell, A. 2006. "'Rough Consensus and Running Code' and the Internet-OSI Standards War." *Annals of the History of Computing, IEEE* 28 (3): 48–61.

Rysman, M., and T. Simcoe. 2008. "Patents and the Performance of Voluntary Standard Setting Organizations." *Management Science* 54 (11): 1920–34.
Saltzer, J. H., D. P. Reed, and D. D. Clark. 1984. "End-to-End Arguments in System Design." *ACM Transactions on Computer Systems* 2 (4): 277–88.
Sanchez, R., and J. T. Mahoney. 1996. "Modularity, Flexibility, and Knowledge Management in Product and Organization Design." *Strategic Management Journal* 17:63–76.
Simcoe, T. 2012. "Standard Setting Committees: Consensus Governance for Shared Technology Platforms." *American Economic Review* 102 (1): 305–36.
Simon, H. A. 1962. "The Architecture of Complexity." *Proceedings of the American Philosophical Society* 106 (6): 467–82.
Smith, A. 1776. *Wealth of Nations*, vol. 10, Harvard Classics, edited by C. J. Bullock. New York: P. F. Collier & Son.
Stigler, G., and R. Sherwin. 1985. "The Extent of the Market." *Journal of Law and Economics* 28 (3): 555–85.

Comment

Timothy F. Bresnahan

In "Modularity and the Evolution of the Internet" Tim Simcoe brings valuable empirical evidence to bear on the structure and governance of the Internet's more technical, less customer-facing, layers. His main empirical results are about the Internet's protocol stack, that is, the structure of the technical layers' modular architecture and of the division of labor in invention of improvements.

To organize my discussion, I will follow Simcoe's main results. There are, however, three distinctions that I want to draw before proceeding: (1) modularity is not the same as openness; (2) one can say that an architecture is modular (or open), which is not the same as saying the process by which the architecture changes is modular (or open); and (3) the Internet, like most ICT platforms, includes both purely technical standards and de facto standards in customer-facing products.

1. Modularity is related to, but not the same as, openness. Modularity is an engineering design concept. A large, complex problem can be broken up into pieces, and engineers working on one piece need know only a small amount about all the other pieces. They *do* need to know how their piece can interact with the other pieces—for which they (ideally) need know only the information contained in the interface standards described in the IETF (and preceding) and W3C documents analyzed by Simcoe. In contrast, openness is an economic organization concept. It refers to the availability and control

Timothy F. Bresnahan is the Landau Professor in Technology and the Economy at Stanford University and a member of the board of directors of the National Bureau of Economic Research.

For acknowledgments, sources of research support, and disclosure of the author's material financial relationships, if any, please see http://www.nber.org/chapters/c13056.ack.