The Endogenous Modularity of the Internet[*]

Timothy Simcoe

Boston University School of Management and NBER

PRELIMINARY DRAFT

May 29, 2013

Abstract

This chapter is an empirical case study of the Internet architecture from an economic viewpoint. Data collected from the two main Internet standard setting organizations (IETF and W3C), demonstrate the modularity of the Internet architecture, and the specialized division of labor that produces it. Examining citations to Internet standards provides evidence on the diffusion and commercial applications of new protocols. I tie these observations together by arguing that modularity helps the Internet (and perhaps the digital technology more broadly) avoid long-run decreasing returns, by facilitating low-cost adaptation of a shared general-purpose technology to the demands of heterogeneous applications.

[*] This research was funded by the NBER Digitization program with support from the Kauffman Foundation. Useful comments on a pre-conference presentation were provided by Shane Greenstain, Avi Goldfarb and Catherine Tucker. All errors are my own, and comments are welcome: tsimcoe@bu.edu.

1. Introduction

The Internet is a global computer network comprised of many smaller networks, all of which use a common set of communications protocols (TCP/IP). 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 economists have called a General Purpose Technology (GPT).

The economic literature on GPT's highlights the importance of positive feedback between innovations in a GPT-producing sector and the various application sectors that use it.[1] Much of this literature elaborates on the implications of this framework for understanding productivity growth, notably the importance of co-invention for understanding GPT diffusion and the timing of associated productivity impacts. However, the literature on GPTs is 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* GPT architecture can have implications for industrial organization in the GPT-producing sector and perhaps also prevent the onset of decreasing returns to GPT innovation. 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

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

evidence on the diffusion and commercial application of innovations within this system.

The main point of the paper is to propose modularity as a solution to the question of how digital technologies manage the trade-off between generality and specialization, and to highlight the role of SSOs as institutions that help firms internalize the benefits of coordinated innovation within a GPT-producing sector.

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 by modeling educational investments in reaching the "knowledge frontier" as a fixed investment human capital (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 complementary investments, often 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 inter-dependencies, 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 space of potential product architectures.

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 non-decomposable design. For example, a tightly integrated design may also be required to achieve optimal performance. The fixed costs of defining the components and interfaces that characterize a modular system may exceed the expected benefits *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.[2]

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

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

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, Langlois (2002) describes how the original architects of the operating system for the IBM System 360 line of computers adopted a non-decomposable design, wherein "each programmer should see all the material."[3] Bresnahan and Greenstein (1999) describe the mergence of divided technical leadership (which might be either a cause or consequence of modular product architecture) did not emerge in computing until the arrival of personal computing. During the initial diffusion of electricity, the city electric light company supplied generation, distribution and even lights as part of an integrated system.

The evolution or choice of a non-decomposable 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 tele-communications era, AT&T had a long history of opposing efforts by third-party equipment to sell any equipment that would attach to its network.[4] 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 not necessarily reflect purely design considerations in a manner that weighs social costs and benefits.[5]

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

[3] The quote comes from Brooks (1975).

[4] 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).

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

consist of (i) an architecture that describes a set of modules and their functions, (ii) interfaces that describe how the modules will work together, and (iii) standards that 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 the those design rules that will remain visible in advance of the market.

The process of selecting 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 multi-homing. 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 choice and selection pressures pushes us towards the 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.[6] First, consensus standardization within SSOs (specifically, the IETF and W3C, as described below) is arguably the dominant mode of coordinating the design

[6] Rusell (2006) describes the standards war between TCP/IP and the OSI protocols. Simcoe (2012) analyzes the performance of the IETF as 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 smart-phones are multi-homing devices because they select between Wi-Fi (802.11) and cellular protocols to establish a physical layer network connection.

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).

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.

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.[7] 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 initial deployment of the network. Thus, in many ways, the early IETF formalized a set of

[7] 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.

working relationships among academic, government and commercial researchers who designed and managed the ARPANET and its successor 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. At present the organizations 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.[8] 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.

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 Laboratory for Particle Physics (CERN). Berners-Lee originally sought to standardize the core web protocols, such at the Hypertext Markup Language (HTLML) and 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 email listserv. The W3C must approve new members, who are typically invited experts, or engineers from dues-paying member companies. The fundamental organizational

[8] http://www.ietf.org/documents/IETF-Regional-Attendance-00.pdf

unit within both SSOs is the Working Group (WG), and the goal of working groups is to publish technical documents.

Working groups IETF and W3C 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 to ensure that independently designed products work together well. The IETF calls these normative documents standards-track RFCs, and the W3C calls them Recommendations.[9] In both organizations, new standards must be approved by consensus, which generally means a substantial super-majority, and in practice is determined by a WG chair, subject to formal appeal and review by the IESG or W3C director.[10]

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.

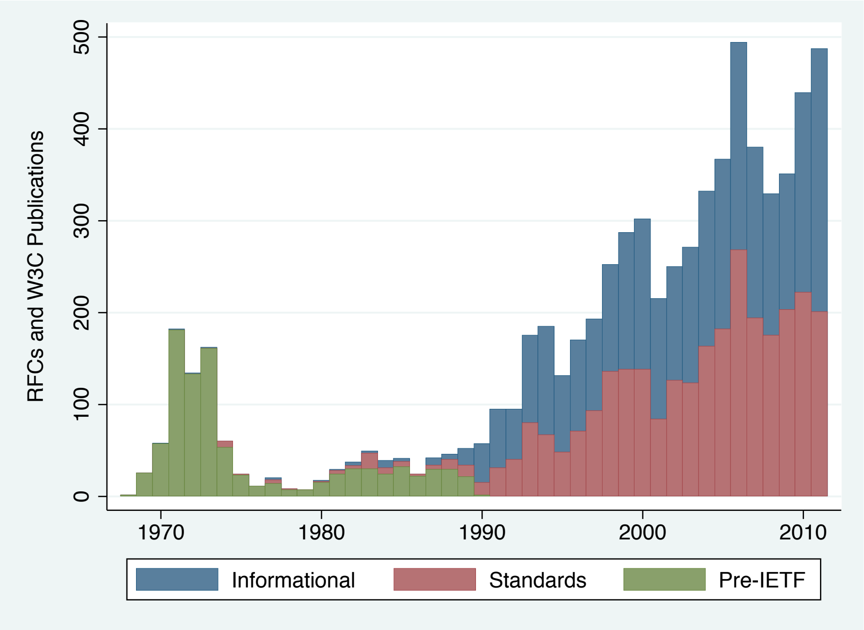
Figure 1 illustrates the annual volume of RFCs and W3C publications between 1969 and 2011. The chart shows a large volume of RFCs published during the early 1970s, followed by a dry spell of almost 15 years, and then a steady increase in output beginning around 1990. This pattern coincides with a burst of inventive activity during the initial development of ARPANET, followed by a long period of

[9] 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.

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

experimentation with various networking protocols – including a standards war between TCP/IP and various proprietary implementations of the OSI protocol suite, as described by 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.

Figure 1: Total RFCs and W3C Publications (1969-2011)



If we interpret the publication counts in Figure 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 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 smart-phones); or the gradual replacement of prior technology that is embedded in substantial capital investments. With respect to replacement effects, in 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 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 R&D. It also illustrates how, at least for "open" standards, much of the information about how to implement a particular module or functionality that is nominally hidden behind the layer of abstraction provided by a standardized interface is actually broadly available.

To provide a sense of better what is actually being counted in Figure 1, Table 1 lists some of the most important IETF standards, as measured by the number of times they have been cited in IETF and W3C publications (Table 1.1) or as non-patent prior art in a US patent (Table 1.2).

Table 1.1: Most Cited Internet Standards (IETF and W3C Citations)[11]

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

[11] 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.

Table 1.2: 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

Both tables contain a number of standards that one might expect to see on such a list, such as the core routing protocols that arguably define the Internet (Transmission Control Protocol and Internet Protocol), 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. All of the documents listed in Table 1 are standards-track publications of the IETF. 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.

Differences between the two lists are also suggestive. IETF and W3C publications frequently cite the Structure of Management Information (SMIv2) protocol, which defines a language and database used to manage objects in a communications network, such as switches or routers. US patents, on the other hand, are more likely to cite protocols for reserving network resources (DHCP and RSVP) and security standards. These differences hint at the idea that citations from the IETF and W3C measure technical interdependencies or knowledge flows within the GPT producing sector, whereas patent cites reflect complementary innovation linked to particular applications of the GPT.[12] I return to this idea below when examining diffusion.

2.2 The Protocol Stack

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, and Baldwin and Colfer 2010), the IETF and W3C clearly conform to the basic cross-sectional prediction. In particular, both organizations assign individual Working Groups to broad technical areas that correspond to distinct modules within the TCP/IP protocol stack.

The protocol stack is the 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 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 that the details of how other layers are implemented is hidden behind a standardized interface. Salzer, Reed and Clark (1984) provide 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 simpler terms, application designers need not worry about the details of how a packet travels from A to B, and router manufacturers can safely ignore the contents of the packets they transmit.

[12] Examining citations to Informational publications reinforces this intuition: the nonstandards-track RFCs most cited by other RFCs describe IETF processes and procedures, whereas the nonstandards-track RFCs most cited by US patent 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.

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 paper, though most engineers would view the organization as a developer of application-layer protocols.[13]

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 corresponding to layers in the traditional protocol stack, the IETF has created a Realtime 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 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. In the mid to late 1990s, the emergence of the web was associated with an increased number of high-level protocols, including the early IETF work on HTML/HTTP, and the first standards from the W3C and Realtime areas. 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 in the late 2000s was driven by a combination of upgrades to legacy technology and the creation of new standards, such as label-switching protocols

[13] 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 stdanrds, and standards for web services (SOAP and WSDL).

(MPLS) that allow IP-based routed networks to function more like a switched network that maintains a specific path between source and destination devices.

Figure 2: Evolution of the Internet Protocol Stack[14]

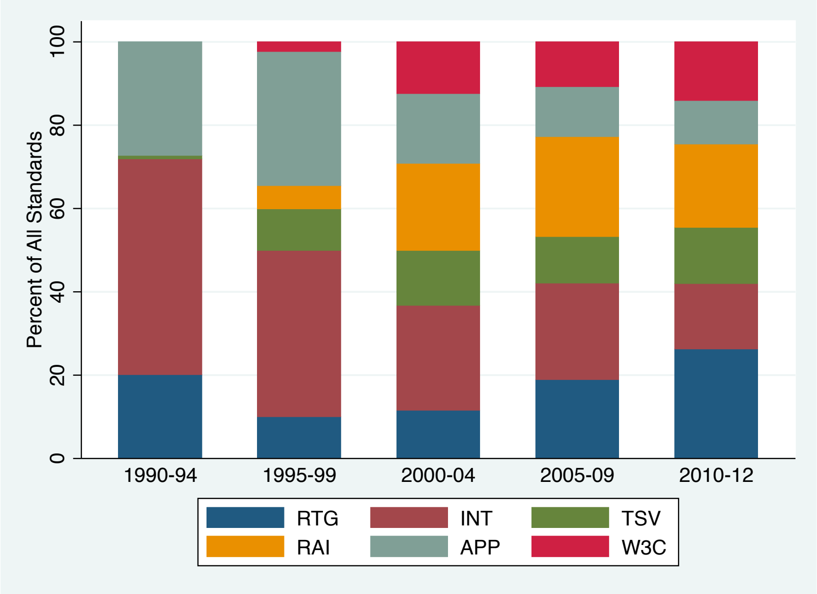


Figure 2 illustrates several points about the Internet's modular architecture that are linked to the literature on GPTs. First, if one views the web as a technology that enables complementary inventions across a wide variety of Application Sectors (e.g. e-commerce, digital media, voice-over IP, online advertising or cloud services), it is not surprising to see initial growth in Application layer protocol development, followed by the emergence of a new Realtime layer, followed by a resurgence of lower layer routing technology. This evolution is broadly consistent with the notion of innovation complementarities between the application sectors and the GPT. Unfortunately, like many papers in the GPT literature, this chapter lacks detailed data on Internet-related inventive activity across the full range of application

[14] These figures are based on the author's calculations using data from IETF and W3C, and include *only* standards-track RFCs and W3C Recommendations. RTG = Routing, INT = Internet, TSV = Transport, RAI = Realtime Applications and Infrastructure, and APP = Applications.

sectors, and is thus limited to detailed observations of the innovation process where it directly touches the GPT.[15]

Figure 2 also raises several questions that will be taken up in the remainder of the paper. 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 Realtime 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 bang" technology transitions than alternative modes of platform governance, such as a dominant platform leader?

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.

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

[15] 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 SSOs that develop structured information standards for particular commercial applications on top of the general-purpose W3C specifications (OASIS).

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.

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 references (or citations).[16] 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.

Clearly, Normative references are an attractive measure of inter-dependency. Unfortunately, the distinction between normative and informative cites was not clear in 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.

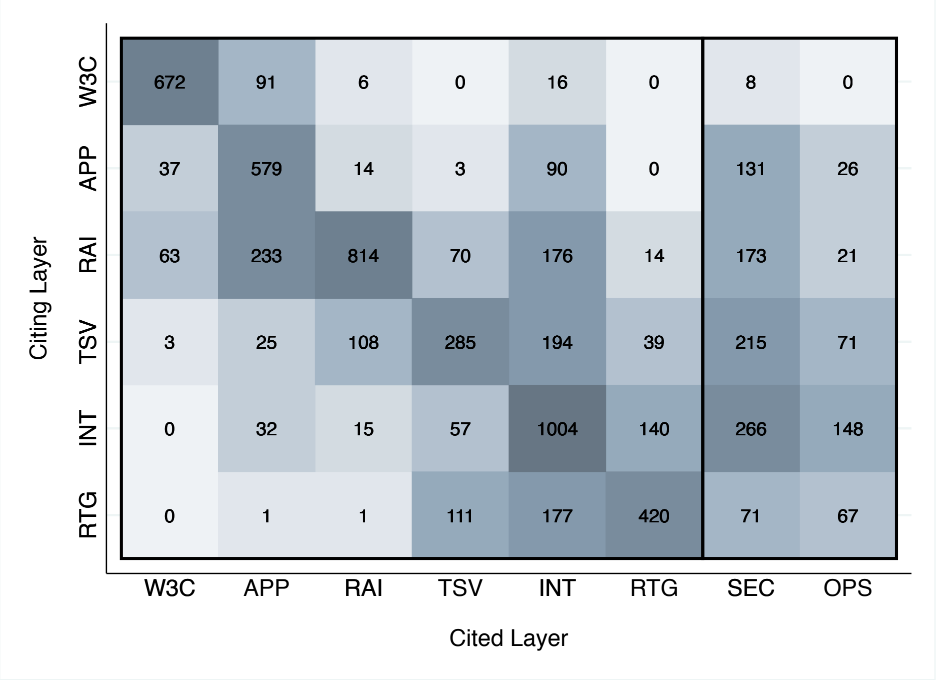
Figure 3 is directed graph of citations among all standards produced by the IETF and W3C, with citing Layers/Areas arranged on the Y axis and cited layers/Areas arranged on the X axis.[17] Shading is based on each cells' decile in the cumulative citation distribution. Twenty-seven percent of all citations link two documents produced by the same Working Group, and I exclude these from the analysis.

[16] The following quotes come from an official IESG statement on citations and referencing, see http://www.ietf.org/iesg/statement/normative-informative.html.

[17] This approach to measuring modularity is closely related to the Deesign Structure Matrix (DSM) in Baldwin and Clark (2000).

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

Figure 3: Citations in the Internet Protocol Stack



	W3C	APP	RAI	TSV	INT	RTG	SEC	OPS
W3C	672	91	6	0	16	0	8	0
APP	37	579	14	3	90	0	131	26
RAI	63	233	814	70	176	14	173	21
TSV	3	25	108	285	194	39	215	71
INT	0	32	15	57	1004	140	266	148
RTG	0	1	1	111	177	420	71	67

In a completely decomposable system, all citations would be contained with the cells along the main diagonal. Figure 3 suggests that the Internet more closely resembles a nearly decomposable system, with the majority of technical inter-dependencies and information flows occurring either within a module, or between a module and its adjacent neighbor in the protocol stack.[18]

The exceptions to near-decomposability illustrated in Figure 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

[18] An alternative non-modular and non inter-dependent design configuration would be a hierarchy, with all cites either above or below the main diagonal.

encounters decreasing returns to scale due to the costs of adapting a shared input to serve heterogeneous application sectors.

The second noticeable departure from near-decompsability in Figure 3 is the relatively high number of inter-layer 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. Of course, the vintage effects themselves are interesting to the extent that they highlight potential "lock in" to early design choices made for an important interface, such as TCP/IP.

Finally, Figure 3 shows that Realtime and Transport-layer protocols have a somewhat greater inter-module citation propensity than standards from other layers. Recall that these layers arrived somewhat later than the original Applications Internet and Routing Areas (see Figure 2). Thus, this observation suggests that when a new module is added to an existing system (perhaps to enable or complement co-invention 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.

3.2 Division of Labor

While Figure 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.[19] The data for this analysis are extracted from actual RFCs by identifying all email addresses in the section listing each author's contact information, and parsing those addresses to

[19] 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.

obtain an author's organizational affiliation. The analysis is limited to the IETF, as it was not possible to reliably extract author information from W3C publications.[20] 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 each document's layer in the 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 2 lists the 15 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.

Table 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

One way to assess whether there is a specialized division of labor in standards creation is to ask whether firms are more concentrated within particular layers of the protocol stack than would be predicted by random choice (i.e. where the probability that a given RFC is associated with a particular layer equals the marginal

[20] 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.

probability of that layer). Greenstein and Rysman (2005) propose a test statistic based on the multinomial distribution that compares the actual distribution of RFCs across layers to a simulated distribution based on random choice. Applying their likelihood-based Multinomial Test for Agglomeration or Dispersion reveals that organizations participating in the IETF are highly concentrated within particular layers. Specifically, the observed value of the multinomial test statistic was -7.1, as compared to a simulated value of -5.3 under the null hypothesis of random assignment.[21] 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).

In order to place a bit more structure on our analysis of the division of labor within the IETF, it helps to have in mind a simplistic model of an organization's decision to contribute to drafting an RFC. Suppose that several firms (indexed by *i*) must decide whether to join a new Working Group *w* in Layer *j*. Each firm either creates a new RFC within that Working Group, or does not: a_i = 0,1. Let us further assume that each firm receives a gross public benefit B_w if the Working Group produces a new protocol, a gross private benefit S_{iw} if the firm helps draft the protocol, and incurs a fixed participation cost F_{ij} per Working Group that varies across layers. In this setup, public benefits flow from increasing the functionality of the network and growing the installed base of users. Private benefits might reflect a variety of factors, such as intellectual property in the underlying technology or improved interoperability with proprietary complements. The fixed costs are assumed constant within each layer to reflect the idea of a layer-specific human capital investment.

To derive a firm's WG-joining decision, let Φ represent the probability that at least one other firm joins the Working Group. Thus, firm *i*'s benefits if it joins the Working Group are $B_w + S_{iw} - F_{ij}$,, while the expected benefits of not joining are ΦB_w. If all firms

[21] 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

have private knowledge of S_{iw} and make simultaneous WG participation decisions, the optimal rule is to join the committee if and only if $(1-\Phi)B_w + S_{iw} > F_{ij}$.

While dramatically over-simplified, 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, the main point here is that firms derive private benefits from participation, and are likely to join a WG if they expect S_{iw} to be large. 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 with the log of one plus the actual number of other WG participants.[22] I parameterize F_{ij} as a linear function of two dummies variables that measure prior participation in WG's at the same layer of the protocol stack, or at an adjacent layer condition on not having participated at the same layer as the focal WG. These dummies for past participation at the same layer should pick up much of the specialization in protocol development observed using the multinomial test statistic.

To be clear, the models presented below ignore the potential simultaneity of organizations' decisions to enter a given WG. However, the model suggests that the main strategic interaction is related to free-riding, which will push firms to disperse across various Working groups if the public benefits of protocol development (B) are large relative to the private rents. Thus, if we observe a positive correlation

[22] An alternative approach would be to estimate the model as a static game of incomplete information following Bajari et al (2010). Unfortunately, I lack instruments that produce plausibly exogenous variation in Φ, as required by that approach.

between firms' entry decisions, we should conclude that the evidence points towards the relative importance of rent-seeking, and towards a large shared WG-level component in S_{iw}. 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-seeking (e.g. because they specify how products will actually be implemented), there should be a stronger positive correlation between a focal firm's entry decision and that of other firms when "entry" is measured as standards-track RFC production than 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 10 or more RFCs and is assumed to be at risk of participating in every WG. I report results from linear probability models, which are nearly identical to the average marginal effect from a logistic specification. Table 3 presents summary statistics for the sample and Table 4 presents the regression results.

Table 3: Summary Statistics

Variable	Mean	SD	Min	Max
Stds-Track WG Entry	0.06	0.24	0	1
Nonstds-track Entry	0.05	0.22	0	1
Past WG (This Layer)	0.34	0.47	0	1
Past WG (Next Layer)	0.17	0.38	0	1
log(Other Participants)	2.11	0.86	0	4.51

The first four columns in Table 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 layer.

Table 4: Linear Probability Models of IETF Working Group Participation

Outcome	Stds-Track WG Entry	Stds-Track WG Entry	Stds-Track WG Entry	Stds-Track WG Entry	Stds-Track WG Entry	Nonstds-track Entry
Past WG (This Layer)	0.06	0.07	0.07	0.05	0.06	0.05
	[0.01]**	[0.01]**	[0.01]**	[0.01]**	[0.01]**	[0.01]**
Past WG (Next Layer)	0.02	0.02	0.02	0.01	0.02	0.01
	[0.01]**	[0.01]**	[0.01]**	[0.01]	[0.01]**	[0.01]*
log(Other Participants)					0.06	0.04
					[0.00]**	[0.00]**
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

*p<0.05; **p<0.01; SEs clustered at WG (except RE models)

The fifth column in Table 4 shows that the number of other WG participants has a strong positive correlation with the focal firm's participation decision. A one 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 evidence that private benefits from contributing to specification development are highly correlated across firms at the WG-level, and that the cost 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 4 changes the outcome to an indicator of entry through publication of nonstandards-track RFCs. In this model, the partial correlation between a focal firms entry 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 larger standards-track correlation suggests that the unobserved private-interest component of joining decisions is either weaker (relative to the free-riding incentive) for nonstandards, or less correlated across firms for the same WG.[23]

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 various 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 the standard appear to be correlated across Working Groups, suggesting that participants know when a particular technical standard is likely to be important.

Stepping back from the analysis, it is important to note that modularity and the division of innovative labor offer a potential answer to the question of what prevents the GPT from ultimately hitting diminishing returns as it is expanded and

[23] 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 Realtime, Internet, Routing and Security.

applied across many different application sectors. Thus, while this analysis of WG participation focused on firms that produce at least 10 RFCs in order to disentangle their motivations for contributing, it is important to recognize that these firms are a minority of IETF participants. I identified 1,299 unique organizations that supplied an author on one or more RFCs. Although the largest organizations do a great deal of protocol development, eighty percent of participating organizations contribute four RFCs or less. By hiding many of the details of what happens within any given layer of the protocol stack, the Internet's modular architecture may lower the costs of entry and component innovation for this large group of small participants.

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. This exercise can provide a window onto the diffusion and utilization of the underlying technology. As described above, lags in diffusion and co-invention 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 macro-economic productivity effects, and (2) they highlight the role of positive innovation externalities between and among application sectors and the GPT-producing sector.

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 inter-dependency 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.

4.1 Diffusion Across Modules

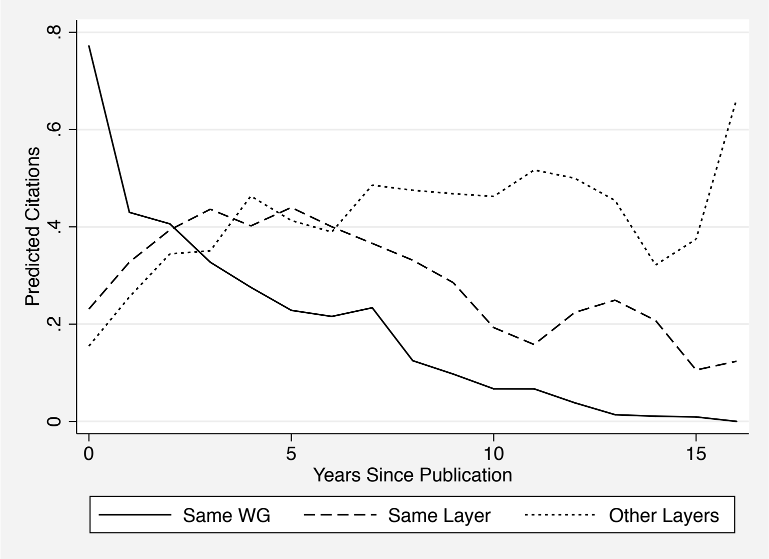
I begin by examining citation flows across different modules and layers within the IETF and the TCP/IP protocol stack. If the level of technical inter-dependency between any two standards increases as we move inwards from protocols in different layers, to protocols in the same layer, to protocols in the same Working Group, we would 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. (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.)

To test the idea that innovation diffuse within and beteen modules at different rates, I create a panel of annual citations to standards-track RFCs for 16 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: $E[\text{Cites}_{iy}] = \exp\{\lambda_{age} + f(\text{Citing-year})\}$.

To summarize these regression results, I set the citing-year to 2000 for all observations and generate the predicted number of citations at each age. These values are plotted and used to calculate A hypothetical mean citation age, along with its standard error (using the delta method). The result is a normalized distribution of predicted citations over the first 16 years of RFC-life that I call the citation-age profile.

Figure 4 illustrates the citation age profile for standards-track RFCs conditional 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 (excluding Operations and Security). The pattern is consistent with the idea that more inter-connected 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.

Figure 4: Age Profiles for RFC-to-RFC Citations



The main lesson contained in Figure 4 is that even within a GPT, innovations diffuse "outward" slowly. I argue that much of this pattern is 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 interface. And before a GPT becomes useful in various application sectors, many interface layers may need to be specified. 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 inter-related interfaces is likely a general pattern, the Internet and digital technology more broadly seems especially suited to modular architecture that facilitates low-cost re-use and time-shifting.

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 citations is based on its 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 inventors monopoly, and is therefore carefully assessed for its relevance to the claimed invention. For this paper, the key assumption is simply that citing patents are more likely to reflect inventions that enable applications of the GPT than RFCs.

Figure 5: Age Profiles for RFC-to-RFC and US Patent-to-RFC Citations

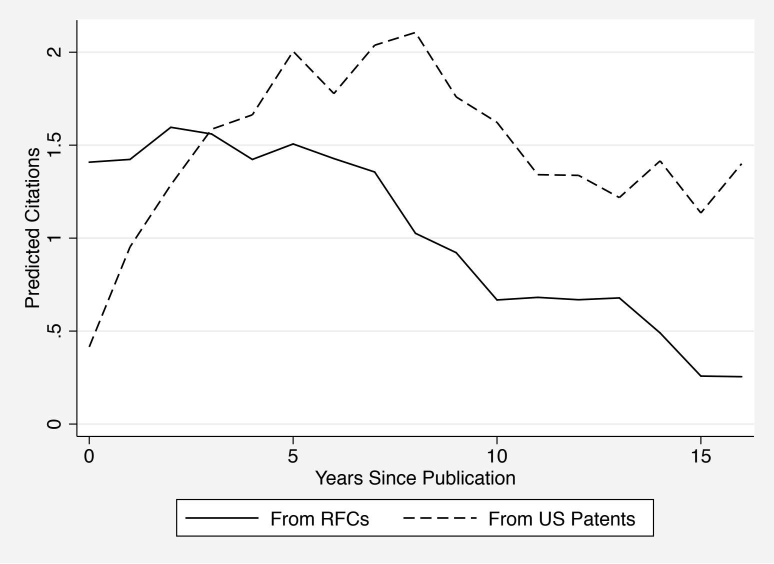


Figure 5 graphs the age profiles for all RFC cites and all patent cites. The RFC age profile represents a cite-weighted average of the three lines in Figure 4, and the average age of an RFC citation is 5.9 years (SE=0.5). Patent citations clearly take longer to arrive, and are more persistent in later years than RFC cites. The average age of a US patent non-prior citation to an RFC is 8.2 years (SE=0.51), which is close to the age as cites from RFCs at other layers of the protocol stack.

At one level, the results illustrated in Figures 4 and 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 (as illustrated in Figure 1). However, it is likely that patented technology represents only a first step 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.

5. Conclusion

The main goal of this chapter is descriptive. It illustrates the modular 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 macro-economic impacts.

At a broader level, this chapter argues 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 "co-invention" 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 re-usability required to sustain growth at the extensive margin, and delivers the familiar benefits of a specialized division of labor in GPT production.

It is tempting to conclude by asking whether modularity promotes innovation. However, a better question is whether the marginal returns to a modular product or organizational design are higher under certain types of governance. Modularity is ultimately a general strategy for reducing coordination costs in complex systems. The Internet is a shared technology platform with divided technical leadership where modular design principles seem to work well. (Though even on the Internet, we sometimes observe what seem to be significant coordination problems, such as the long delayed upgrade to IPv6.) The degree of modular specialization within platforms managed by a single firm, and their impact on innovation outcomes, are topics for additional research.

References

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

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

Becker, G. S. and K. M. Murphy (1992). The division-of-labor, coordination costs and knowledge. Quarterly Journal of Economics 107(4), 1137–1160.

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

Bresnahan, T. (2010). "General Purpose Technologies" Ch. 18 in the Handbook of the Economics of Innovation, Volume 2, Pages 761–791

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 (1995), p. 83

Brooks, F. (1975) *The Mythical Man-Month*. Addison-Wesley.

Colfer, L. and C. Baldwin. (2010) "The mirroring hypothesis: Theory, evidence and exceptions." Working Paper 10-058, Harvard Business School, Boston.

David, Paul A. (1985) "Clio and the Economics of QWERTY." American Economic Review, 77(2): 332-337

Farrell, J. (2007) "Should competition policy favor compatibility?" in Standards and Public Policy, Greenstein, S. and V. Stango, eds. Cambridge Univ. Press.

Farrell, J. and G. Saloner (1986). Installed base and compatibility - innovation, product preannouncements, and predation. American Economic Review 76(5), 940–955.

Farrell, J. and T. Simcoe (2012) "Four Paths to Compatibility." pages 34-58 in the Oxford Handbook of the Digital Economy. Oxford University Press.

Goolsbee, A. and P. Klenow (2006) "Valuing Consumer Products By The Time Spend Using Them: An Application To The Internet," American Economic Review, 96(2): 108-113.

Greenstein, S. and M. Rysman (2005). "Testing for Agglomeration and Dispersion." Economics Letters 86(3): 405-411.

Greenstein, S. (1996) "Invisible Hand versus Invisible Advisors." in *Private Networks, Public Objectives,* ed. Noam, Eli; Amsterdam: Elsevier, 1996.

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 #14138.

Langlois, R. (2002) "Modularity in technology and organization," Journal of Economic Behavior & Organization, 49(1): 19-37.

Mackie-Mason, J. and J. Netz, (2007) "manipulating Interface Standards as an Anticompetitive Strategy" in Standards and Public Policy, Greenstein, S. and V. Stango, eds. Cambridge Univ. 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), 1920–1934.

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), 277–288.

Sanchez, R., 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): 305-336.

Simon, H. A. (1962). "The architecture of complexity" Proceedings of the American Philosophical Society, 106(6) 467–482.

Smith, A. (1776) Wealth of Nations, edited by C. J. Bullock. Vol. X. The Harvard Classics. New York: P.F. Collier & Son.

Stigler, G. and R. Sherwin (1985). "The Extent of the Market," Journal of Law and Economics, University of Chicago Press, vol. 28(3), pages 555-85.