DIGITAL DARK MATTER AND THE ECONOMIC CONTRIBUTION OF APACHE

Shane Greenstein
Frank Nagle

Digital Dark Matter and the Economic Contribution of Apache
Shane Greenstein and Frank Nagle
NBER Working Paper No. 19507
October 2013, Revised December 2013
JEL No. O3,O31,O47

ABSTRACT

Researchers have long hypothesized that spillovers from government, university, and private company
R&D contribute to economic growth, but these contributions may be difficult to measure when they
take a non-pecuniary form. The growth of networking devices and the Internet in the 1990s and 2000s
magnified these challenges, as illustrated by the deployment of the descendent of the NCSA HTTPd
server, otherwise known as Apache. This study asks whether this experience could produce measurement
issues in standard productivity analysis, specifically, omission and attribution issues, and, if so, whether
the magnitude is large enough to matter. The study develops and analyzes a novel data set consisting
of a 1% sample of all outward-facing web servers used in the United States. We find that use of Apache
potentially accounts for a mismeasurement of somewhere between $2 billion and $12 billion, which
equates to between1.3 percent and 8.7 percent of the stock of prepackaged software in private fixed
investment in the United States. We argue that these findings point to a large potential undercounting
of the rate or return from IT spillovers from the invention of the Internet, and to a large potential undercounting
of "digital dark matter" in general.

Shane Greenstein
The Elinor and Wendell Hobbs Professor
Kellogg School of Management
Northwestern University
2001 Sheridan Road
Evanston, IL 60208-2013
and NBER
greenstein@kellogg.northwestern.edu

Frank Nagle
Harvard Business School
Soldiers Field
Boston, MA 02163
fnagle@hbs.edu

1. Introduction

Astrophysicists draw on the term "dark matter" to describe the unseen parts of the universe. Many artifacts, such as the rotational speed of galaxies and gravitational effects, indicate the presence of dark matter, although measuring its existence directly can be difficult. Economists need a similar label for some innovative building blocks of the digital economy that standard tools cannot measure. *Digital dark matter* can serve as the phrase for these digital goods and services that are non-pecuniary and effectively limitless, and serve as inputs into production. They are hybrids of public goods and private investments. This study develops an example that illustrates the potential for the growth and importance of these inputs and their impact.

Our study of digital dark matter draws attention to the (mis)measurement of spillovers from knowledge generated by the government, university, or private research and development. It has long been believed that such spillovers are economically important, but are very difficult to measure when they take non-pecuniary form. It is sometimes possible to overcome measurement challenges and assign an economic value to some of these spillovers. We investigate a situation where a spillover became embodied in open source software, and took the form of digital dark matter. By understanding the value of one specific example of digital dark matter, we aim to add evidence to the debate on the contribution of spillovers to economic productivity. Further, we aim to better understand the size of the mismeasurement that occurs due to the presence of digital dark matter.

The growth of networking devices and the Internet in the 1990s and 2000s magnified the challenges affiliated with measuring digital dark matter. After decades of development under the

auspices of the Department of Defense and the National Science Foundation (NSF), the NSF

privatized the Internet backbone in the first half of the 1990s. Software and standards affiliated

with operating TCP/IP networks migrated into widespread commercial use. Additionally, in

1991 Tim Berners-Lee made available the basic building blocks of the World Wide Web,

supporting its use and development by founding the World Wide Web Consortium in 1994. Its

use became common, and formed the basic software infrastructure for a wide range of new forms

of electronic commerce and new media.

This study examines one part of these larger events, the deployment of the descendants of

the National Center for Supercomputing Applications (NCSA)[2] HTTPd server, today known as

Apache. It was one of two notable pieces of NCSA software, the Mosaic browser[3] being the

other one. Both inventions moved into widespread use in the middle of the 1990s, continued to

evolve thereafter, and subsequently became essential for online commercial activities. Apache's

experience deserves academic scrutiny because, in part, it is convenient to examine. Though no

publically available data provides a definitive estimate of the size of the Apache economy, it is

believed to be the second largest open source project after Linux. It is so large that it has left

more observable traces than many other examples of digital dark matter, albeit, such traces are

not easy to find.

This study contains two sections. It initially reviews the practices surrounding Apache's

deployment, and extends existing measurement theory to this setting, showing how Apache's

---

[2] The NCSA is one of the four original supercomputing centers funded jointly by the NSF and state governments. It was founded in 1984 to help address the scientific research needs of the future.

[3] Together, the HTTPd server and the Mosaic browser propelled the World Wide Web forward with the HTTPd server acting as a content publisher and the Mosaic browser acting as a content reader.

experience could produce omission and attribution issues. The paper next develops a quantitative approach to address the open question raised by the first section, namely, whether the attribution and measurement issues are large. This study develops a novel dataset, based on a one-percent sample of all "outward facing" web servers used in the United States (we give a more precise definition below). Our quantitative approach using non-proprietary information is an important innovation in this study. The "best" information is collected for private purposes, is closely guarded (Netcraft, 2012), and, in any event, is not publically available for statistical scrutiny by researchers.

Using principles of GDP measurement (Nordhaus, 2006), the study estimates the monetary value of the stock of servers. The value is compared to different benchmarks, and we conclude that the estimated value is large. We find that Apache potentially accounts for a mismeasurement of somewhere between $2 billion and $12 billion, which equates to between 1.3 percent and 8.7 percent of the stock of prepackaged software in private fixed investment in the United States. We also provide some arguments for why the estimates should tend towards the higher end of this range. We argue that these findings point to a large potential undercounting of the rate or return from IT spillovers affiliated with university and federal funding for the Internet.

The study contributes to one young literature and one mature literature. First, it contributes to the underdeveloped literature on measuring the spillovers from the invention of the Internet. Supporters of federal funding for research often cite the Internet as an example of the best-case scenario, presuming that federal funded research led to public goods with large societal benefit (Greenstein, 2011). Despite much broad interest in measuring the spillovers and economic gains from the invention and deployment of publically funded inventions (See e.g.,

David, Hall, and Toole, 2000), no estimate exists for the benefits the Internet conferred to the economy. Digital dark matter is principally to blame for this gap in knowledge, as there is little appropriate data for distinguishing the contribution of the Internet from contributions from general advances in ICTs (Greenstein, 2012). This is an unfortunate gap in knowledge considering the research on the origins and creation of the Internet (Mowery and Simcoe, 2002) and the contribution of all information technology to productivity gains over the last several decades (Brynjolfsson, 1993, Barua, Kriebel, and Mukhopadhyay, 1995, Barua and Byungtae, 1997, and Brynjolfsson and Hitt, 2003). This is also unfortunate in light of the large body of literature that has examined the important contribution of information technology to productivity growth (Jorgenson, Ho, and Stiroh, 2005, Brynjolfsson and Saunders, 2009, and Tambe and Hitt 2012). The gap is also somewhat inconsistent with other evidence indicating the Internet appears responsible for altering the economic landscape in the late 1990s,[4] and contributed to creating new processes in the economy that had long lasting consequences.[5]

The literature on mismeasurement of economic activity is much more developed, and this study makes a novel contribution, albeit an incremental one to that literature. The potential for mismeasurement in national accounts has received considerable treatment in general discussions (Nordhaus, 2006), and related insights have been applied to a range of long-recognized thorny

---

[4] Forman, Goldfarb and Greenstein (2003) estimate that by the year 2000 approximately 88% of US business establishments with over 100 employees had equipment for basic Internet functions, such as email and browsing, while 12% had evidence of upgrades to enhancing their business processes with Internet functionality. In many industries the former was well over 90%, and the latter was well over 20%. Forman, Goldfarb and Greenstein (2012) find evidence that this upgrade in enterprise use of the Internet was affiliated with major changes in the wage structure across the United States.

[5] For example, recent industry assessments estimate that approximately 8% of all retail products sold in the United States are now sold via the Internet (Anderson, Reitsma, Evans, Jaddou, 2011).

measurement issues, such as valuing pollution, valuing national security, and valuing leisure time. No research has extended the insights to open source software, or related matters, such as content accumulated under a creative commons license.[6] That gap exists despite the considerable research about the sources and operations of open source software.[7] The extension is, however, not surprising, as it uses existing frameworks for GDP measurement. This study shows that Apache's diffusion creates potential scenarios for standard problems in the literature, namely, omission and attribution errors. Those errors lead to undercounting and problematic inference in productivity studies (Corrado, 2011, Syverson, 2011).

These two contributions together focus attention on a larger unaddressed question. The micro-mechanisms that create measurement issues for economic accounting of open source software are not unique to Apache. They are common to several Internet inventions that diffused into commercial use without formal market transactions and licenses, and where open source institutions supported deployment and use. Other prominent examples from this time period are Linux, software built around TCP/IP, and the World Wide Web (Greenstein, 2010). Further, while Linux and Apache are two of the most recognized open source software projects, there are many others that play an important role in the digital economy but are not accounted for in any productivity measures, such as Perl, PHP, or Firefox, as well as a creative common license in a not-for-profit setting, such as in Wikipedia. While the study offers only a specific estimate of

---

[6] We are aware of only one partial exception to that statement, Ghosh et al, 2006. It offers a thorough assessment of the size of the open source economy in Europe, and calculates an estimate of the size of all open source activity as a fraction of the economy. However, they calculate their estimates of the size of the open source economy by using the labor cost to reproduce the full body of open source software. We take a very different approach.

[7] See, e.g., Von Hippel, 2001, Lerner and Tirole, 2002, O'Mahony, 2003, Hawkins, 2004, Lakhani and Wolf, 2005, Roberts, Hann, and Slaughter, 2006, West and Lakhani, 2008, Lerner and Schankerman, 2010.

digital dark matter in Apache's case, we think it also illustrates a much broader issue with wide applicability. The study shows why the problem is large in one specific instance, and offers one approach for framing vexing measurement issues in general.

Section 2 provides a general framework for thinking about Apache's experience and the affiliated measurement issues. Section 3 describes the novel data and calculations that hint at the scale of the mismeasurement. Section 4 concludes.

2. Digital Dark Matter: Framework

This section discusses the institutional setting that created Apache. It then discusses the omission and attribution issues created for productivity measurement by Apache's widespread diffusion.

2.1. Institutional background

Apache descended from software invented at the NCSA at the University of Illinois, which also was the home of the Mosaic browser. Apache arose from server software that worked with Mosaic. It was called the NCSA HTTPd server. This was the most widely used HTTP (Hypertext Transfer Protocol) server software in the research-oriented "early-days" of the Internet. The server was a collection of technologies that supported browsing and use of Web technologies.

While the University of Illinois successfully licensed the Mosaic browser for millions of dollars,[8] its licensing of the HTTPd server software did not enjoy a similar experience. In part

---

[8] Notably, the University of Illinois did license the Mosaic browser to a third party, who licensed it to over one hundred other firms, including Microsoft. Netscape never licensed it. Many of the programmers involved in the project left the university in April 1994 and founded Netscape, then got into a dispute with the University over

this was because the server software first became available for use as shareware, with the

underlying code available to anyone, without restriction. Many Webmasters took advantage of

the shareware by adding improvements as needed or by communicating with the lead

programmer, Robert McCool. McCool, however, left the University (along with others) to work

at Netscape in the middle of 1994, and thereafter webmasters and web participants lost their

coordinator.

By early 1995 there were eight distinct versions of the server in widespread use, each

with some improvements that the others did not include. These eight teams sought to coordinate

further improvements. They combined their efforts, making it easier to share resources, share

improvements, and build further improvements on top of the (unified) software. The combination

of eight versions was called *Apache* (ostensibly because it was "a patchy web server"[9]), and,

informally at first and more formally over time, the group adopted the practices of open source.

As has been documented elsewhere, Apache grew into a very large open source project,

widely used in private firms to support electronic commerce.[10] Apache became an essential

---

some ownership rights (initially over the ownership of the name "Mosaic"), and they reprogrammed their commercial browser from scratch. They never paid any licensing fees. In its third year Netscape sold over $500 million dollars of software. It is widely agreed that Netscape's entry was a catalyst for Microsoft's accelerated development of a browser. Those events, in conjunction with Apache's diffusion, catalyzed the entry of thousands of new startups in complementary applications. Though there is no doubt that the licensing revenue collected by Mosaic was a tiny fraction of the value created, which is consistent with this study's theme, fully developing that observation would involve a wider array of historical detail and analysis beyond this study's limited scope.

[9] In a later interview Brian Behlendorf, one of the founders of Apache, acknowledges the pun, but claims it did not motivate his initial thoughts about naming the project Apache. He states " It just sort of connoted: 'Take no prisoners. Be kind of aggressive and kick some ass.'" McMillan (2000).

[10] The Apache Software Foundation, which was founded to support the Apache HTTPd project, has since created a wide array of other open source projects that add additional unquantified value to the Internet ecosystem. However, the HTTPd project remains the largest project and therefore is the primary focus of our inquiry.

component in the customer-facing commercial transactions of many firms, as well as in the procurement activities supported by electronic commerce. Further, Apache is used as the base for many other commercial products, such as the IBM HTTP Server, which comes bundled with the IBM WebSphere Application Server. Today it is widely used across the globe, and is regarded as the second most popular open source project used by businesses, after Linux.[11] Additionally, Apache is disproportionately used to host web sites that receive large amounts of traffic. 57% of the million busiest web sites are hosted on Apache. The next closest server is nginx at 15%. [12]

The lack of prices became essential to the operation and success of the project, and, as we show below, this creates potential measurement issues.[13] The absence of pecuniary transactions first arose at the beginning of Apache's existence, when the HTTPd server moved from universities to commercial use without formal commercial licenses. It continued as Apache emerged as an open source project based on the HTTPd server, and relied upon donations and a community of users who provided new features for free. As with other open source software, Apache eschews standard marketing/sales activities, instead relying on word-of-mouth and other non-priced communication online. Like other open source organizations, Apache also does not develop large support and maintenance arms for their software, although users do offer free

---

[11] See http://httpd.apache.org/ABOUT_APACHE.html, accessed March 2011, or the similar account in Mockus, Fielding, and Herbsleb (2005).

[12] See the "Market share of the top million busiest sites" section of http://news.netcraft.com/archives/2013/09/05/september-2013-web-server-survey.html.

[13] The Apache Software Foundation argues that the lack of price encourages the commitment of the community, and this community would likely fall apart if its products were not free. "Why Apache Software Is Free," http://httpd.apache.org/ABOUT_APACHE.html (accessed July 11, 2011).

assistance to each other via mailing lists and discussion boards (Lakhani and von Hippel, 2003, West and Lakhani, 2008, Lerner and Schankerman, 2010).

2.2 Measuring the gains: Omission

What potential economic measurement issues could result from this invention's deployment? If any major issues arise, they arise from the measurement of the software's contribution to production. Two categories of issues need attention, a problem affiliated with *omission* and another affiliated with *attribution*.

Normal procedures of economic accounting omit Apache as input into production or into stocks of capital. Normal economic measurement focuses on measuring transactions taking place in markets, and presumes that transactions involve a positive price (Nordhaus, 2006). Without explicit attention, normal procedures presume that unpriced activities are nonmarket activities. In sum, like other open source software, the prices and revenue for Apache are zero.

Though open source is *not* singled out as an example by Nordhaus (2006), this setting fits one of the settings he outlines as problematic, namely what Nordhaus labels a "near-market good." He discusses omission errors that arise when standard procedures presume that a zero price is affiliated with non-market activity, but real economic activity creates goods that have a value, but no price. This setting fits Nordhaus' description in many respects. Creating Apache code relied on the equivalent of donations for support. These may come in the form of explicit donations from firms who provide personnel time and firm capital, or it may come from programmers devoting leisure time to open source activity. It also may come in the form of in-kind or unacknowledged donations of capital or services, such as computer time and hosting

facilities. Further, the software also contributes to producing more or better output that may appear unaccounted for.

There are also important differences with the examples discussed in Nordhaus. In this case, some of the activities affiliated with Apache can be measured. Like other widely used open source software, third party firms perform many complementary support functions. This activity typically involves consultants, independent programmers, and providers of bridging software between open source software and commonly used proprietary software.[14] This activity of complementary actors is a key part of the open source ecosystem (West, 2003). Most of that activity will involve market transactions and positive prices. In addition, to obtain service from Apache a firm might have to make considerable investments, using paid personnel, including training personnel to install Apache and conduct ongoing operations, and customizing and adapting Apache to the unique needs of the enterprise. Finally, firms also might purchase hardware for deployment, and potentially additional hardware to accommodate large-scale use.[15] Such expenditure would appear as an operating expense.

---

[14] We also note that similar issues pertain to licensed software, though a considerable variety applies there as well. Licensing can be on a per-CPU, per-employee, or per-copy basis. In most other respects, investment activities with personnel and customization and a complementary ecosystem remain the same. A key difference may be the size and operations of the network that has grown up around the standardized commercial software, especially when proprietary firms subsidize those operations with tools and technical support. See Lerner and Schankerman (2010).

[15] While at any point in time there must be a strong association between the number of Apache web servers in use and the number of hardware machines acting as servers, that association does not imply a fixed or constant Leontiff production function over time between the number of Apache servers and the amount of hardware in a firm or industry. There need not be as strong an association between the number of web pages and number of Web servers deployed, for example. One Apache web server can support many web pages, and that has grown over time. In addition, the software improves through software upgrades after new version releases, yielding improvement with no hardware expenditure. Improvement also may arise from better practices at complementary processes within the network, such as mirror servers. Hence, many users have enjoyed functional upgrades without any change in their own hardware.

We will argue that the presence of open source software, specifically, and digital dark matter, more broadly, raises the potential for attribution and omission biases in productivity analysis. The problem with omission bias is readily transparent. For example, studies that measure the importance of IT to economic growth (e.g. Jorgenson, Ho, and Samuels 2013) could be underestimating the existing stock of IT due to the non-pecuniary nature of digital dark matter. Further, productivity studies that seek to understand the impact of investments in IT on a firm's output (e.g. Brynjolfsson, 1993; Byrne, Oliner, and Sichel, 2013) could be undercounting investments in IT that are unpriced. The issues with attribution bias are subtler, however, and merit a deeper discussion.

2.3. Measuring the gains: Attribution

To understand the mechanisms behind omission and misattribution, consider the standard productivity model.

Begin with this representation:

$$Y_{it} = A_{it} * f(L_{it}, K_{it}, IT_{it}),$$

where Y is output for firm i at time t[16], which results from a production function with arguments for (L) labor, (K) capital stock, and (IT) information technology capital stock, and A is an unmeasured contributor to firm efficiency. In the standard Cobb-Douglas production model this becomes

$$\ln(Y_{it}) = A_{it} + \alpha * \ln(L_{it}) + \beta * \ln(K_{it}) + \gamma * \ln(IT_{it}).$$

---

[16] This type of analysis can be implemented at the industry level (Stiroh, 2002), but for simplicity, we carry it through at the firm level.

where, typically, the natural log of each side is taken. This results in an equation that can be used

for regression estimates. In typical analyses, growth is measured by improvement over time,

namely, $Y_{it} - Y_{i,t-1}$, and productivity is measured as multifactor productivity (Corrado, 2011,

Syverson, 2011). Because usage of open source software by a firm does not have a specific

pecuniary measure, there is no mechanism for such usage to enter the equation as an input

variable on the right hand side. This results in several possible scenarios of misattribution:

- *Growth without cause.* One scenario for misattribution arises if firms experience growth

  without hiring more labor, and seemingly without paying for more IT capital or L or K or, for

  that matter, any visible service. This can happen when Apache code improves and users

  receive updates at no expense. In this case some firms grow without appearing to change

  their inputs. Growth will be attributed to A, because of the appearance of more productivity

  that cannot be attributed to growth in inputs.[17] This scenario resembles a scenario discussed

  in Syverson (2011), misattribution due to externalities from the local environment, which is

  analogous to firms relying on the quasi-public goods created by the open source

  community.[18] Syverson argues that the gains could appear to be disembodied technical

  change, not attributable to any specific input.[19]

---

[17] A similar scenario arises when donations by firms lead to an increase in output prices at many firms. If the price increase eventually leads to an increase in revenue, this would lead to a growth in Y improperly attributed to A.

[18] The mismeasurement is analogous to mismeasuring an improving public good. In her analysis of the various types of protections used in OSS, for example, O'Mahony (2003) highlights this analogy and finds it is an important driver of legal efforts of OSS projects to protect their work.

[19] Or, as in Tambe and Hitt (2012), problems could arise from mismeasurement of labor, which lacks adjustments for human capital affiliated with supporting the software, or for the extent to which labor relies on the community to enhance their productivity. Tambe and Hitt (2012) also points out that measurement error may occur due to the differences between labor-based and capital-based estimates of IT productivity.

14

- *Growth attributed to the wrong input.* Another scenario for misattribution arises if a large

  fraction of firms employ Apache software and another fraction makes no investment in

  Apache, and those investing in Apache invest in labor to support a new release or upgrade.[20]

  In that case, the firms using open source software will experience an increase in output, Y,

  and an increase in L. They will show no measured change in IT capital. Non-Apache users do

  not show any change in Y, L, or IT. Normal productivity analysis will then attribute output

  growth to the growth of L, even though it is due to increases in unmeasured IT capital. This

  can cause particular problems in cross-sectional analysis since growth may be measured

  accurately for some firms and inaccurately for other firms. An interesting variant in this

  scenario arises from deploying a new web server, which generates purchase of hardware

  upon which to run Apache. That generates an increase in Y, L and IT among Apache users,

  but the real measure of IT will be lower than the actual level. The growth will be attributed to

  both the L and IT. In such an instance, IT expenditure will appear especially productive due

  to the unmeasured complementary software input.

- *Competition between open source and commercial software leading to misattribution:* The

  third scenario is related to the two scenarios described above. Consider a situation – observed

  in the data below – where a large fraction of firms invest in Apache software while another

  large fraction use functionally equivalent software from a commercial firm. Both firms will

---

[20] Higher labor expenditure could arise either from the need to hire more workers or compensate workers more for their efforts. Though the prevailing view in industry is that open source labor receives higher compensation, there is only limited evidence for this belief. There is some evidence that contributions to open source projects yield increases in pecuniary compensation (see e.g., Hann, Roberts, Slaughter, and Fielding, 2002, Hann, Roberts and Slaughter, 2013). However, the evidence is limited to whether contributors gain monetary rewards, not whether an otherwise equivalent worker gains premiums on their wages for Apache-specific skills in comparison to others. The monetary gains from contributions are consistent with the existence of the premium, but cannot serve as an estimate of its size.

also invest in more labor, with the firms using Apache software making similar or larger increases in expenditure for labor than those investing in commercial software.[21] All firms experience an increase in Y. Both users experience a growth in L, while the commercial software users experience a larger increase in IT because they paid for the software. Normal productivity analysis will then attribute some part of the growth to L and IT and some growth to A for the firm using Apache. An interesting variant arises when Apache labor gets a premium. Then Apache users experience a larger growth in L than commercial software users, but a smaller growth in IT. If most firms are Apache users then standard estimates will attribute much of the gains to L and not enough to IT.

That explanation also illustrates the omission and attribution problems in tracing the gains to the economy from federally funded research if the gains diffuse into the economy as unpriced inventions, as Apache did. Many of the costs to developing Apache were incurred as part of the research to support the development of the Internet at NCSA. Those were monetary costs and real economic costs. Most of the gains, however, were not recorded – either omitted or misattributed – because the software took the form of open source, and the code improved without any explicit costs or transactions.

---

[21] If the labor for open source software cost the same or less, in addition to open source software costing nothing, and yielded outcomes equivalent to the commercial software, then the commercial software would fade from being used at all. This is not what we observe in the data. Though Apache is the largest service software for Web commerce, functionally-equivalent software from commercial firms has achieved substantial market share, especially from Microsoft. For this situation to be sustainable as market equilibrium, labor expenditure for open source software has to be higher than that for commercial software. A related possibility is general resistance to using open source software or some other distaste for it, or, equivalently, a taste for some attribute affiliated with pecuniary products, which would lead some potential users to pecuniary products for reasons other than labor costs.

Further, the scenarios above only consider the spillovers from direct usage of Apache as an input into production. They do not account for the spillovers that occur when a competing product, such as Microsoft's Internet Information Services (IIS), add a feature by imitating a similar feature developed for Apache. Nor does this include further gains from enabling the entry of complementary applications.

While the omission and attribution issues discussed above are possible and likely, that does not settle whether they are large and important. The next section addresses the question: Is the evidence about unmeasured value of Apache software large enough to suggest the attribution and measurement issues are important economic issues?

3. The shadow value of Apache HTTP Server

To demonstrate the potential impact of digital dark matter, we will calculate the shadow value of the Apache HTTP Server market by considering the price of substituting the non-pecuniary Apache HTTP Server with the pecuniary Microsoft IIS. Although we could have also considered the impact of substituting Microsoft IIS for nginx, the second most popular open source web server, as well, we chose to limit our analysis to only one product, as this adequately illustrates the core point.

3.1.   The shape of the server economy

Although data on the number of websites hosted via Apache HTTP Server is readily available in a public manner (Netcraft, 2012), data on the number of actual Apache HTTP Servers used is not. Additionally, existing public data does not clearly identify the location/country for these servers. However, because web servers are primarily used to host

public web pages, and are therefore directly reachable via the Internet, we were able to collect information on the number of Apache HTTP Servers used to serve public web pages in the US. Because Apache HTTP Servers can be used internally by organizations, our calculation of the number of Apache HTTP Servers that serve public web pages can be considered a lower bound on the number of actual Apache HTTP Servers in use. Furthermore, a number of different network architectures –load balancing, elastic/cloud computing, and so on – allow for multiple web servers to run on one IP address, which would also lead to our collection method yielding an underestimate of the true number of Apache HTTP Servers.

We first identified the full list of IPv4[22] addresses registered to U.S. organizations. To do this, we utilized information published by the American Registry for Internet Numbers, the organization responsible for managing the distribution of IPv4 addresses in the United States. As of October 15, 2011, there were 1537.37 million IPv4 addresses allocated in the United States. It was too costly to scan every one of these IPv4 addresses, so we took a random sampling of 15,865,522 addresses, which is just over 1 percent of the entire U.S. IPv4 space. For each IPv4 address in our sample, we checked to see if the system was running a web server. If it was, we determined whether the server ran Apache, Microsoft IIS, or anything else including unidentified servers.[23]

---

[22] IPv4 is version 4 of the Internet Protocol and is currently the most widely used protocol for routing Internet traffic. It is in the process of being replaced by IPv6, but at the time the data was collected all IPv6 addresses also used a backwards-compatible IPv4 address.

[23] The details are straightforward for someone technically skilled in web programming and administration, albeit tedious to report in this context. This method will identify "outward" facing servers, but will systematically undercount any server used entirely for internal purposes. Hence, it is necessarily an underestimate of all Apache HTTP Server software in use. Further details about the process are available from the authors, upon request.

This method will generate a sample of server use and its characteristics, which otherwise is not available. It has one principal drawback. One server may support a large or small number of pages. This method will be proportional to Apache's actual importance in the economy when the size of use is uncorrelated with our measurement strategy (i.e., no selection bias), and our sample size is large. We look for selection issues in the sample, and do not find any symptoms of such issues (Appendix A). This feature of our method also makes us cautious about inference from small sample sizes, as it will be when analysis focuses on narrow geographies or industries.

Of the 15,865,522 addresses in our sample, we found that 195,885 (1.23 percent) were running a web server.[24] Of these 195,885 web servers, 44,211 (22.57 percent) were running Apache and 24,222 (12.37 percent) were running Microsoft IIS.[25] If we extrapolate these numbers to the full U.S. IPv4 space, we estimate that there are 18,981,268 outward-facing web servers in the United States, 4,284,049 of which are running Apache Web Server.[26] Appendix A gives an analysis of the servers in our sample set, including geographic location and top-level domain distribution.

---

[24] The other 98.77% of the IPs scanned were either inactive or were devices that were not web servers on standard TCP ports.

[25] Apache and IIS account for 34.94% of all web servers in our sample. The remaining web servers were either unable to be correctly identified or were running a different web server such as nginx or a proprietary web server. For example, Google has developed its own internal web server that it uses in place of a publicly available web server.

[26] Continuing this extrapolation to the entire range of IP addresses in the world, of which there are 3.706 billion that are not reserved, there would be 10,288,264 Apache servers in the world. Based on Netcraft's publicly released data on websites, (see news.netcraft.com/archives/2011/12/09/december-2011-web-server-survey.html) that translates into 33 websites per Apache server. This is plausible because the number of web pages per web server must be very skewed. While some Apache servers serve only a single website, many are used by hosting facilities and host hundreds of websites.

3.2. Substitution with pecuniary goods

We seek to put a monetary value on the Apache HTTP Server by comparing it with the most widely used proprietary and pecuniary choice. We follow Nordhaus (2006), who states that (p. 146) "…the price of market and nonmarket goods and services should be imputed on the basis of the comparable market goods and services," and (p. 151) valuation "…should rely on available market and behavioral data wherever and whenever possible." At the time of this study a number of proprietary source web servers exist, the most prevalent of which is Microsoft's IIS. IIS's most obvious cost as a substitute for Apache HTTP Server is pecuniary. IIS is shipped for free with Microsoft's Windows Server 2008 operating system, the price of which varies greatly.[27] Appendix B discusses the substitutability of Apache and IIS.

At the time of this study the price for Windows Server 2008 R2 Standard is $1,029 for five licenses, Windows Server 2008 R2 Enterprise is $3,999 for twenty-five licenses, and Windows Server 2008 R2 Datacenter Edition is $2,999 for one license. The most bare-bones version of Windows Server 2008, called the Windows Web Server 2008, is priced at $469. This version of Server 2008 is intended purely for "the development and deployment of Internet-facing Web sites and services."[28] Finally, IIS also comes installed with Windows 7, which can be purchased for as low as $119.99. However, Windows 7 is not designed to be used as a production scale web server and it is unlikely that any company hosting a public website would use this version of Windows.

---

[27] http://www.microsoft.com/windowsserver2008/en/us/pricing.aspx (accessed July 11, 2011).

[28] http://www.microsoft.com/windowsserver2008/en/us/pricing.aspx (accessed July 11, 2011).

What is a representative price for IIS? We utilize three of the above price points to understand the range of possible prices. On the cheap end, we consider Windows 7, which can cost as low as $119.99, albeit, it also possesses too little functionality to be of practical use. On the high end, we consider Windows Server 2008 R2 Datacenter Edition, which costs $2,999 for one license. Finally, we can consider the bare-bones version Windows Web Server 2008 in the middle,[29] and is currently priced at $469. These three price points allow us to construct a range of possible values for the shadow value of Apache HTTP Server. [30]

With our estimate of the number of Apache Web Servers publically reachable in the United States, we can compute a pecuniary cost of replacing all of these Apache Web Servers with Microsoft IIS. Based on the valuations of Microsoft license fees as mentioned above, the cost of replacing all publically reachable Apache Web Servers in the United States would be between $514 million and $12.8 billion, with a middle estimate of $2 billion.

As previously mentioned, this middle number should be considered a lower bound, because it is based solely on web servers that are attached to the public Internet and does not account for web servers on corporate Intranets or in private use, servers that are behind load balancers or other configurations where multiple servers may exist on one IP address. In

---

[29]We consider Windows Web Server 2008 a comparable match with Apache HTTP Server because it exhibits the closest functionality set and the other versions have additional functions that Apache HTTP Server does not provide. However, it should be noted that most, if not all, of these additional functions can be replicated by free open source software. For example, the operating system functionality is equivalent to the Linux operating system, which is open source and free.

[30] This procedure follows standard GDP measurement principles. It is not a valuation of user gains from employing Apache. Standard revealed preference suggests, for example, that the valuation of IIS by the infra-marginal IIS users would be higher than the market price, and similarly, valuation by infra-marginal Apache users should be higher as well. Conventional GDP measurement does not use consumer surplus. Rather, it uses the marginal valuation.

addition, the valuation we employ – namely, the present price of IIS – reflects the presence of this differentiated competition. In the absence of any other price, we have to presume that this price reflects the marginal value of the software. [31] Further, although we consider the Windows Web Server 2008 to be the most similar to Apache, Apache is disproportionately used to host the busiest websites (Netcraft, 2012). Hence, there are reasons to think the functionality of Apache tends towards the functionality of the higher end Datacenter version of IIS.

3.3. The potential for productivity miscalculation

Is the estimate of the value of Apache a large or a small number? It depends on whether it is compared to sales or investment. First, consider sales. Of the $357 billion (2010 dollars) in software sales by U.S. firms in 2010, $257 billion went to private fixed investment. [32] By this yardstick, the stock of Apache software in the United States is as much as 5 percent (12.8/257) of software sales. However, this compares a stock to a flow, so some might consider it like comparing apples with oranges.

Consider a benchmark against investment. Of the $295 billion of software invested in by U.S. firms in 2010, $81 billion (or just over 27 percent) was prepackaged. [33] If that ratio holds for

---

[31] These prices reflect the current state of the market, where the market leading good (Apache) is unpriced. We do recognize that the presence of differentiated competition lends doubt to the assumption that prices reflect marginal value. Microsoft may not have pricing power in setting the price for IIS. It seems possible and plausible that the price of IIS would be higher if Apache was a priced good. Nonetheless, we follow Nordhaus' dictum to use observed prices, and not counterfactual prices. This is another reason why our calculations could be considered an underestimate of the value of a single Apache server.

[32] "GDP and Final Sales of Software," Bureau of Economic Analysis, http://www.bea.gov/national/info_comm_tech.htm (accessed October 2011).

[33] "Software Investment and Prices, by Type," Bureau of Economic Analysis, http://www.bea.gov/national/info_comm_tech.htm (accessed October 2011). The vast majority of software

investment stocks, then the stock of prepackaged software in the United States was $146 billion dollars in 2010.[34] By that yardstick, Apache software is bounded by as much as 8.7 percent (12.8/146) of the measured capital stock of packaged software, or as little as 1.3 percent (2/146).

This single illustration suggests the scale of the issue is more than merely a rounding error, particularly when one considers the ubiquity of other widely used, free open source software, such as the software that supports the World Wide Web, Linux, Firefox, PHP, software implementing TCP/IP, and many other widely-used open source programs in Internet-based services. We conclude it is likely that the sum of a few of these cases reaches a significant fraction of the total value of the packaged software capital stock and in turn results in a significant impact on overall U.S. GDP.

3.4. The economic size of the ecosystem supported

Apache can be viewed through another lens, as a part of the large ecosystem that supports Internet activity. Are our estimates large or small in relation to the value of Internet activities?

Apache is one of several complementary components that together provide Internet services. Indeed, Apache is one of several important non-pecuniary complements. As noted, others include a wealth of standards and software built around the TCP/IP protocol stack, maintained by the Internet Engineering Task Force; a wealth of standards and software built

---

investment is "custom software" or "own-account," namely, software built by a third party, such as a consultant, or software built by in-house employees.

[34] This is 27 percent of the total stock of software in the United States (under nonresidential equipment and software), which was $533 billion in 2010. See "Fixed Assets and Consumer Durable Goods for 1997-2010," http://www.bea.gov/scb/pdf/2011/09%20September/0911_fixed-assets.pdf (accessed October 2011).

around the World Wide Web, maintained by the World Wide Web Consortium; and important components of Linux, which embed a range of functionality to enable Internet activities.

How important a component is Apache? Consider these comparisons. The size of Internet access revenue in the United States in 2009 (the last year of reliable data) is $59.6 billion,[35] and the size of US online advertising revenue in 2009 is approximately $21 billion.[36] That number combines access revenue from both households and businesses, and it includes $10.1 billion of wireless Internet access revenue. Compared to the revenue it helps produce, the $2 billion to $12 billion of Apache software appears significant.

Now consider another benchmark: the value of Apache in comparison to the size of the software market. The size of system software revenue in 2009 was $48 billion, though personal computer software comprised the largest category, and it was not comparable to Apache. The enterprise and mainframe software revenue together amounted to $26 billion.[37] Against that, the $2 or $12 billion of Apache software appears quite large, albeit, a reader could worry about

---

[35] 2009 Service Annual Survey Data, Information Sector Services-NAICS 51, does not provide a direct estimate of online access revenue, but it lists four categories of access revenue in four tables: Table 3.3.6. Wired telecommunications carriers (NAICS 5171); Table 3.3.9. Wireless and other telecommunications carriers (NAICS 517212); Table 3.3.12. Cable and other program distribution (NAICS 5175); and 3.4.1. Internet service providers (NAICS 518111), http://www.census.gov/services/sas_data.html#NAICS%2048/49 (accessed November 2011).

[36] 2009 Service Annual Survey Data, Information Sector Services - NAICS 51, does not provide a direct estimate of online advertising revenue, but it lists three categories in three tables: Table 3.3.5. Internet publishing and broadcasting (NAICS 516); Table 3.4.1. Internet service providers (NAICS 518111); and Table 3.4.2. Web search portals (NAICS 518112). The latter table does not provide an estimate for 2009, but it does provide sufficient data for 2008 and other categories in 2009 to make an educated guess at its level. For the estimate above, that guess was $14 billion. http://www.census.gov/services/sas_data.html#NAICS%2048/49 (accessed November 2011).

[37] 2009 Service Annual Survey Data, Information Sector Services - NAICS 51, Table 3.1.6. Software Publishers (NAICS 5112), http://www.census.gov/services/sas_data.html#NAICS%2048/49 (accessed November 2011).

comparing apples with oranges, once again. This comparison mixes different time scales, as we are comparing sales of one year to replacing the entire stock of Apache.

Of course, neither of these comparisons is precise. With estimates of the replacement cycle for Apache it would be possible to translate the stock into service flows, or their equivalent.

3.5. The Rate of Return

Another way to calculate the value of Apache is to place it in a cost benefit framework. We provide two different calculations on the rate of return. In each case we use the same estimate for the cost of creating the Internet, but different methods for estimating the benefits. The second calculation uses productivity estimates, much like those discussed above. Due to data limitations, at best we can make an estimate in the "ballpark." The exercise offers evidence of the large value of Apache specifically, and digital dark matter more broadly. It also provides evidence that is useful for thinking about misattribution.

During NSF's management (approximately 1985–1995) the agency invested $200 million dollars in creating Internet related technology (Leiner, Cerf, Clark, Kahn, Kleinrock, Lynch, Postel, Roberts, Wolff, 2003). That budget covered the costs of the Internet backbone, operational expenses, and the supercomputer centers, of which NCSA was one. Apache originated out of NCSA around 1995, so the rate of return calculations will start counting benefits after 1995.

Before the NSF programs, DARPA funded most of the early invention related to the Internet in the 1970s and early 1980s. While DARPA's financial commitment was considerable,

25

no historian has made a precise estimate of the size of the commitment to networking, which was one of several programs it funded. Government secrecy about DARPA's specific projects and budgets prevents any historian from uncovering further details. Nonetheless, we do know something useful. The entire expenditure for the IPTO, the agency within DARPA that funded most of the Internet, did not exceed approximately $500 million over its entire existence (1963–1986), and the funding for what became the Internet was but one of many IPTO projects (Norberg, O'Neill, Freedman, 1996). To be conservative we add another $200 million to the costs of creating the Internet, which is much more than likely.

This calculation includes the most direct costs for creating the Internet, and excludes numerous other costs for research. For example, this estimate of costs does not include a range of other experiments in computer science that NSF paid for (out of different budgets) and from which the general community of researchers learned. We are comfortable with this because it provides a number that is likely more than the actual amount of the direct costs.

That sets up the first cost-benefit calculation. Above we estimated that Apache is worth between 2 and 12.2 billion dollars in 2012, seventeen years later. As noted above we consider 2 and 12.2 to be implausibly low and high. So for this cost/benefit calculation we consider two more plausible levels, $7.1B and $10B, where the former is the midpoint between the two numbers and the latter is as high as we consider plausible. What constant rate of growth would generate such levels of benefit after $400M investment seventeen years earlier? Using the current dollar numbers, the former generates a rate of growth of 17% and the latter 19%. While inflation (which averaged 2-3% a year in this period) would diminish some of that gain, it is quite high for only one output from NCSA, and there were many more benefits than just this one.

A second cost-benefit calculation examines the scale of economic growth attributable to Apache, using productivity gains from investment in software, much as discussed above. For such a calculation we use the model of Byrne, Oliner and Sichel (2013) (hereafter BOS), because it distinguishes between the gains to labor productivity from distinct IT inputs in the US economy – IT hardware, software and communications. Using this model we ask the question: if additional software were added to the capital stock of software, how much economic growth would it generate?[38] In each year it generates more income, and that continues to accumulate after 1995. We consider the total through 2012. This estimate must make assumptions about the rate of growth of Apache, which we do not observe directly.

To illustrate how this calculation works we begin with one year, the estimate for 2012. In 2012 BOS estimate that software accounts for 0.16 of labor productivity growth (out of 1.56 total labor productivity growth per year). Software accounts for only 3.75% of the income share in that year, but software makes a big contribution to productivity growth in comparison to the size of software in use.[39] Our low and high estimates for the value of Apache would place it at 1.3% and 1.8% of the stock of software in 2012.[40] While the increase in labor productivity due to Apache has to be comparatively small, the US economy is so large that even a small improvement can yield a substantial economic gain. In this case, we assume that labor does not

---

[38] Specifically, this uses Equation 1 in BOS, and the estimates in Table 1. The simulation asks what would happen if Apache had pecuniary value.

[39] As shown in Equation 1 in BOS, each component's contribution to labor productivity arises from multiplication of the income share and the growth in that input. Our simulation is equivalent to asking how the contribution would change if the income share for software increased without any increase in labor. Thanks to Dan Sichel for patiently walking us through the steps of the calculation so it remained consistent with the estimates in Table 1.

[40] This uses the same source of data, as noted above.

increase, but only software does, and simulate how much GDP would grow with more software. This model estimates that additional software stock would generate a low/high estimate of $1.0B and $1.8B in additional income in 2012.[41]

Apache also generated income in all the years between 1995 and 2012. For those estimates we must make an assumption about the rate of growth in Apache servers over this period, which we did not observe directly. We make an estimate from public data on the growth of web pages supported by Apache over this time period.[42] It shows web pages grow at one rate over the 1990s, and then grow at a slower rate in the last decade, with Apache generally supporting 60% of all web pages. This history suggests that a constant rate of growth over all seventeen years is probably slower than the true rate of growth, so we make the assumption of linear growth in order to be conservative.

Next we estimate the incremental contribution of Apache for each year, just as we did for 2012, but now we ask a slightly different question. A rate of return calculation addresses the question: "How much economic activity did the investment in the Internet generate by 2012?" Such a calculation requires aggregating all future benefits into the same dollar units. We assume a 10% discount rate on the future from 1995, so we can add up the contemporary dollars.[43] This assumption will weight short-term gains against those that come many years later, as Table 1 shows (comparing line A to line C, or comparing line B to line D). Then we add up those gains,

---

[41] This estimate follows BOS and estimates the labor productivity gain on non-farm private income.

[42] See Netcraft.com

[43] Note that this calculation takes all values in contemporary values and discounts from that value by 10% per year, blending all price level corrections and forecasting into one value in 1995 terms.

and calculate that Apache generated economic activity equivalent to between $2.6B and $4.5B by the end of 2012. On an investment of $400M, that is a rate of return between 10.5% and 14%. Table 1 shows the components that went into that calculation. Once again, that is quite high for only one output from NCSA.

Table 1: Contribution of Apache to GDP, simulation, Billions of Dollars

Year	1996	1997	1998	1999	2000	2001	2002	2003
A	0.04	0.08	0.13	0.19	0.25	0.30	0.36	0.43
B	0.07	0.14	0.23	0.32	0.43	0.53	0.63	0.75
C	0.03	0.10	0.20	0.32	0.46	0.63	0.80	0.99
D	0.06	0.17	0.34	0.55	0.80	1.08	1.38	1.71
2004	2005	2006	2007	2008	2009	2010	2011	2012
0.52	0.50	0.58	0.65	0.71	0.74	0.82	0.93	1.04
0.90	0.86	1.00	1.13	1.23	1.28	1.42	1.61	1.80
1.19	1.36	1.54	1.73	1.91	2.08	2.25	2.42	2.59
2.06	2.36	2.67	2.99	3.30	3.59	3.89	4.18	4.48

A: Additional GDP in that year, low estimate, contemporary dollars
B: Additional GDP in that year, high estimate, contemporary dollars
C: Accumulation of GDP up until that year, low estimate, discounted 10% per year
D: Accumulation of GDP up until that year, high estimate, discounted 10% per year
Source: Author's calculation, based on Byrne, Oliner, Sichel (2013). See text.

There are important qualifications to this second set of estimates. First, the estimates in BOS also examine the effect of productivity gains in the production of IT, as distinct from capital deepening, but we did not use those in our calculation. If Apache generated productivity gains to production, the above estimates would not capture those gains.[44]

---

[44] As noted by the literature, to the extent that open source investments were capitalized into the private value of firms, some of this would have been accounted for in other capital. The BOS estimates do not use intangible capital as another form of capital, however, so this is not a concern.

Second, no reader should quote the rate return on Apache as precisely 10-14%. The estimate depends on assumptions, such as of linear growth and 10% discounting. A small increase in the rate of discount would mildly lower the rate of return, as would later growth in Apache. Similarly, a lower rate of discount would mildly increase the rate of return, as would sooner growth in Apache. So why make such an estimate? These estimates provide a good sense of the scale of the gains, and small changes in those assumptions do not alter the quality of the answer. Anything in this range has to yield a large rate of return, which is our point.

Third, these are estimates on a counter-factual in order to illustrate the scale of importance of additional software, had it been accounted for like any other asset. Because it is a small change in the value of assets it is possible to approximate its effect with this simulation. Were Apache measured properly the other estimates of the contribution of other IT capital would change as well. In that light, this exercise yields one other insight. Because Apache is such a small percentage of total capital, the attribution biases associated with mismeasuring this asset, therefore, appear to be small. This implies that Apache alone, as one piece of open source software, produces large omission bias but does not produce attribution bias. That also suggests that proper measurement of open source software would produce omission bias, but it leaves open the question of whether it would produce a major attribution bias.

We would stress that our estimates are necessarily below the true value. Despite this underestimate, these numbers, as well as the estimates from the first cost-benefit analysis, indicate that the return on investment for Apache was quite high. Since this is only one program, this leads us to conclude that the returns from federal R&D invested in the Internet must be underestimated.

4. Concluding thoughts and future research

In this study we argued that digital dark matter is an important issue to consider in the online economy. Like other private assets, digital dark matter acts at times like an input into the production of a pecuniary good, and regular investment extends functionality or delays obsolescence. Like a public good, more than one user can employ digital dark matter nonexclusively. In contrast to many private assets or public goods, something other than market prices shapes the extent of investment and use. Finally, even when visible, digital dark matter is measured indirectly at best. Omission and attribution errors are possible, even likely.

We illustrated these observations by focusing on one prominent case, Apache, which is a key piece of software in the operation of the Internet. We argued that Apache contributes value to the online economy, and that this value could be quite large, and that it is not currently captured through standard GDP measurement. We find evidence that the omission biases are significant, but the attribution biases are not. Our estimates also imply that, were we to add additional open source software, we could reach a significant fraction of the total value of packaged software sales. Once again, we conclude that this evidence suggests that it is likely that open source software significantly contributes to omission bias.

These findings point to a large potential undercounting of "digital dark matter" and related IT spillovers from university and federal funding. Apache's experience focuses attention on a broader set of open source software projects, such as Linux, the software built around IETF standards, the World Wide Web, PERL, or a creative common license in a not-for-profit setting, such as Wikipedia. Every project took a distinct institutional form, but shares similar potential for omission and attribution errors.

While open source software is certainly an important piece of digital dark matter, we speculate that similar concerns about measurement may arise in other activities where digital goods and services are non-pecuniary, effectively limitless, and serve as inputs into production. For example, user contributed content powers websites as diverse as Twitter, Yelp, and YouTube, but these free "inputs" from users go unmeasured by standard productivity measurement. As another example, digitized blueprints, many of which are non-pecuniary, have become widely available for 3D printing, and as that activity grows, these prints will contribute to production, despite their lack of price.

We speculate that the effect of omission biases are likely to increase as information costs approach zero and firms rely more on non-pecuniary digital inputs from communities of users and developers (Altman, Nagle, and Tushman, 2013). Although such quantification may be difficult to attain directly, we have shown that indirect methods of estimating this value are possible. More precise and broad-based estimates may be used to create GDP calculations that more accurately reflect the true production of the U.S. economy, resulting in policies that are more suited to the reality of the online economy. We foresee such studies shedding light on the measurement of the gains from research and development in universities that diffused into commercial use as part of open source software and in many other ways. Such quantification may also lead to a better understanding of the impact of free and open source software on the economy as a whole.

These concerns lead to a number of open questions. If the undercounting of digital dark matter leads to mismeasurement of productivity, does it also lead to underinvestment – both public and private - in projects that create digital dark matter? Would demand for digital dark matter products decrease significantly if they were pecuniary? We also wonder how digital dark

matter shapes a variety of online activities where these and related products are common, such as online news, entertainment, scientific inquiry, educational and reference activities, and business operations.

5. References

Altman, Elizabeth, Frank Nagle, and Michael Tushman. 2013. "Innovation without Information Constraints: Organization, Communities, and Innovation When Information Costs Approach Zero," in Oxford *Handbook of Creativity, Innovation, and Entrepreneurship*, edited by Michael Hitt, Christina Shalley, and Jing Zhou. Oxford University Press.

Anderson, Jacqueline, Reineke Reitsma, Patti Freeman Evans, and Samantha Jaddou. 2011. Understanding Online Shopper Behaviors. *Forrester Research*.

Barua, Anitesh, Charles H. Kriebel, and Tridas Mukhopadhyay. 1995. Information technologies and business value: An analytic and empirical investigation. *Information Systems Research* 6(1): 3-23.

Barua, Anitesh, and Byungtae Lee. 1997. The information technology productivity paradox revisited: A theoretical and empirical investigation in the manufacturing sector. *International Journal of Flexible Manufacturing Systems* 9(2): 145-166.

Brynjolfsson, Erik. 1993. The productivity paradox of information technology. *Communications of the ACM* 36 (12): 66–77.

Brynjolfsson, Erik, and Lorin M. Hitt. 2003. Computing productivity: Firm-level evidence. *Review of Economics and Statistics* 85(4): 793-808.

Brynjolfsson, Erik, and Adam Saunders. 2009. *Wired for innovation: how information technology is reshaping the economy.* MIT Press.

Byrne, David, Stephen Oliner, and Daniel Sichel. 2013. Is the Information Technology Revolution Over? *International Productivity Monitor, Centre for the Study of Living Standards,* Vol 25, pp 20-36. Spring.

Corrado, Carol, 2011. Communications Capital, Metcalfe's Law, and U.S. Productivity Growth, The Conference Board, EPWP #11-01. March.

David, P.A., Bronwyn H. Hall, and Andrew A. Toole. 2000. Is public R&D a complement or substitute for private R&D? A review of the econometric evidence. *Research Policy* 29 (4-5): 497-529.

Forman, Chris, Avi Goldfarb, and Shane Greenstein. 2003. Which Industries use the Internet? in (ed) Michael Baye, *Organizing the New Industrial Economy,* Elsevier. Pages 47-72.

Forman, Chris, Avi Goldfarb, and Shane Greenstein. 2012. The Internet and Local Wages: A Puzzle, *American Economic Review.* February. 102(1), pages 556-575.

Ghosh, R. A., R. Glott, B. Krieger, G. Robles. 2006. Study on the economic impact of open source software on innovation and the competitiveness of the Information and Communication Technologies (ICT) sector in the EU. Technical Report, UNU-MERIT, the Netherlands.

Greenstein, Shane. 2010. Innovative Conduct in U.S. Commercial Computing and Internet Markets. in *Handbook on the Economics of Innovation*, edited by Bronwyn Hall and Nathan Rosenberg. Burlington: Academic Press. Pp. 477-538.

Greenstein, Shane. 2011. Nurturing the Accumulation of Innovations: Lessons from the Internet. in *Accelerating Innovations in Energy. Insights from Multiple Sectors,* edited by Rebecca Henderson and Richard Newell. Chicago: University of Chicago Press. Pp. 189-224.

Greenstein, Shane, 2012. The absence of data for measuring the economic impact of IT in the US. in *Regulation and Performance of Communications and Information Networks,* Edited by Gary Madden, Gerry Faulhaber, and Jeffery Petchey, Edward Elgar Press; Cheltenham, UK. Pp. 328-344.

Hann, I., Roberts, J., Slaughter, S. 2013. All Are Not Equal: An Examination of the Economic Returns to Different Forms of Participation in Open Source Software Communities. *Information Systems Research,* April 2013.

Hann, I., Roberts, J., Slaughter, S., and R. Fielding. 2002. Delayed returns to open source participation: An empirical analysis of the Apache HTTP Server Project. Working Paper.

Hawkins, R. 2004. The economics of open source software for a competitive firm: Why give it away for free? *Netnomics* 6 (2): 103-117.

Jorgenson, Dale, Mun Ho, and Jon Samuels. 2013. Economic Growth in the Information Age: A Prototype Industry-Level Production Account for the United States, 1947-2010. Working Paper.

Jorgenson, Dale, Mun Ho, and Kevin Stiroh. 2005. *Information Technology and the American Growth Resurgence*. MIT Press; Cambridge, MA.

Lakhani, K., and E. von Hippel. 2003. How open source software works: "Free" user-to-user assistance. *Research Policy* 32 (6): 923-943.

Lakhani, K. R., and R. G. Wolf. 2005. Why hackers do what they do : Understanding motivation effort in free/open source software projects. In *Perspectives on free and open source software,* ed. J. Feller, B. Fitzgerald, S. Hissam, and K. Lakhani. Cambridge, MA: MIT Press.

Leiner, Barry, Vint Cerf, David Clark, Robert Kahn, Leonard Kleinrock, Daniel Lynch, Jon Postel, Larry Roberts, Stephen Wolff, 2003, A Brief History of the Internet, Version 3.32. Reston Virginia, The Internet Society, http://www.isoc.org/internet/history/brief.shtml.

Lerner, J., and J. Tirole. 2002. Some simple economics of open source. *The Journal of Industrial Economics* 50 (2): 197-234.

Lerner, J., and M. Schankerman. 2010. *The comingled code, open source and economic development*. Cambridge, MA: MIT Press.

McMillan, Robert. 2000. Apache Power. *Linux Magazine.* April 15. http://www.linux-mag.com/id/472/, accessed September, 2013.

Mockus, A., Fielding, R. T., and J. D. Herbsleb. 2002. Two case studies of open source software development: Apache and Mozilla. *ACM Transactions on Software Engineering and Methodology* 11 (3): 309-346.

D. Mowery and T. Simcoe. 2002. Is the Internet a US Invention? An Economic and Technological History of Computer Networking. *Research Policy*, 31(8-9): 1369-87.

Netcraft. 2012. June 2012 Web Server Survey. *Netcraft*. Retrieved from http://news.netcraft.com/archives/2012/07/03/june-2012-web-server-survey.html, accessed September, 2013

Norberg, Arthur, Judy O'Neill, and Kerry Freedman. 1996. *Transforming Computer Technology, Information Processing for the Pentagon, 1962-1986.* Baltimore, MD., Johns Hopkins University Press.

Nordhaus, William D., 2006, "Principles of National Accounting for Nonmarket Accounts," in editors, Dale W. Jorgenson, J. Steven Landefeld, and William D. Nordhaus, *A new Architecture for the US National Accounts,* University of Chicago Press.

O'Mahony, S. 2003. Guarding the commons: how community managed software projects protect their work. *Research Policy* 32 (7) 1179-1198.

Roberts, J., Hann, I., and S. Slaughter. 2006**.** Understanding the motivations, participation, and performance of open source software developers: A longitudinal study of the Apache projects. *Management Science* 52 (7): 984-999.

Scacchi, W. 2002. A case study in electronic commerce and open source software development. Institute for Software Research, University of California, Irvine. Working Paper.

Stiroh, Kevin J., 2002. "Information Technology and the U.S. Productivity Revival: What do the Industry Data Say?" *American Economic Review,* 92 (5), pp 1559-1576. December.

Syverson, C. 2011. What Determines Productivity? *Journal of Economic Literature* 49(2): 326-365.

Tambe, P., and L. M. Hitt. 2012. The productivity of information technology investments: New evidence from IT labor data. *Information Systems Research* 23(3-1): 599-617.

Von Hippel, E. 2001. Innovation by user communities: Learning from open source software. *Sloan Management Review* 42 (4): 82-86.

West, Joel. 2003. How open is open enough? Melding proprietary and open source platform strategies. *Research Policy* 32 (7): 1259-1285.

West, J., and K. R. Lakhani. 2008. Getting clear about communities in open innovation. *Industry & Innovation* 15 (2): 223-231.

Zhu, K., Kraemer, K. L., Gurbaxani, V., and S. Xu. 2006. Migration to open-standard interorganizational systems: Network effects, switching costs and path dependency. *Management Information Systems Quarterly* 30 (1): 515-538.