Digital dark matter and the economic contribution of Apache

Shane Greenstein\textsuperscript{a,1}, Frank Nagle\textsuperscript{b}

\textsuperscript{a} Kellogg School of Management, Northwestern University, 611leverone Hall, 2001 Sheridan Road, Evanston, IL 60208, USA
\textsuperscript{b} Harvard Business School, Harvard University, Wyss Hall, Soldiers Field, Boston, MA 02163, USA

\begin{abstract}
Researchers have long hypothesized that research outputs 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 between 1.3% and 8.7% of the stock of prepackaged software in private fixed investment in the United States and a very high rate of return to the original federal investment in the Internet. We argue that these findings point to a large potential undercounting of the rate of return from IT spillovers from the invention of the Internet. The findings also suggest a large potential undercounting of “digital dark matter” in general.

\end{abstract}

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. By understanding the value of one specific example of digital dark matter, 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)\textsuperscript{1} HTTPd server, today known as Apache. It was one of two notable pieces of NCSA software, the Mosaic browser\textsuperscript{2} 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

\footnotesize{\textsuperscript{1} 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.

\textsuperscript{2} 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.}

0048-7333/5 – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.respol.2014.01.003
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 consists of two sections. It initially reviews the practices surrounding Apache’s deployment, and extends existing measurement theory to this setting, showing how Apache’s 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% and 8.7% 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 toward the higher end of this range. After estimating the value of Apache, we calculate the rate of return for federal investments in the technologies that led to the creation of the Internet. By using our value of Apache as the only output from these investments, we are necessarily underestimating the true rate of return. However, even with this significant underestimation, we still find a rate of return between 10.5% and 19%. We argue that these findings point to a large potential undercounting of the rate or return from research output affiliated with university and federal funding for the Internet.

The study contributes to two literatures. 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 economic gains from the invention and deployment of publically funded inventions (see e.g., David et al., 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 et al., 1995, 1997; 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 et al., 2005; Brynjolfsson and Saunders, 2009; 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, and contributed to creating new processes in the economy that had long lasting consequences.

We also contribute to an extensive literature on mismeasurement of economic activity and productivity growth (Nordhaus, 2006; Corrado, 2011; Syverson, 2011). Our study contributes to this literature by showing that mismeasurement of Apache has reduced the estimated contribution of IT to productivity growth. For instance, were it measured like other software Apache should be regarded as an important contributor to economic growth, large enough to have merited investing in the research to create it.

These two contributions together focus attention on a larger unaddressed topic. 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 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, its licensing of the HTTPd server

---

4 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 et al., 2011).

5 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 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.
software did not enjoy a similar experience. In part 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”), 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. Apache became an essential 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. 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%.6

The lack of prices became essential to the operation and success of the project, and, as we show below, this creates potential measurement issues. 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 assistance to each other via mailing lists and discussion boards.

(10) The presence 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 assistance to each other via mailing lists and discussion boards.

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 Mozilla 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.

6 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 (2008).

7 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.

8 See http://httpd.apache.org/ABOUT_APACHE.html, accessed March 2011, or the similar account in Mockus et al. (2002).

9 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.

10 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).

11 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).

12 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 Leontief 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...
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 et al. 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 et al. 2013) could be undercounting investments in IT that are unpriced. Our analysis below (Section 3.5) shows that Apache alone produces a large omission bias, on the order of billions of dollars. 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:

\[ Y_{it} = A_{it} \cdot f(k_{it}, l_{it}, IT_{it}) \]

where \( Y \) is output for firm \( i \) at time \( t \), \( f \) 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

\[ \ln(Y_{it}) = \ln(A_{it}) + \alpha \cdot \ln(l_{it}) + \beta \cdot \ln(k_{it}) + \gamma \cdot \ln(IT_{it}) \]

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, Byrne et al., 2013). 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. 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. Syverson notes that the gains could appear to be disembodied technical change, not attributable to any specific input.
- **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. 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.
- **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 functionality equivalent software from a commercial firm. Both firms will 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. All firms experience an increase in \( Y \) from the software. The commercial software users experience a larger increase in \( IT \) because they paid for the software. Normal productivity analysis will then attribute some of the part of the growth to \( L \) and \( IT \) and some growth to \( A \) for the firm using Apache.

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.

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 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% 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.\(^\text{22}\)

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%) were running a web server.\(^\text{23}\) Of these 195,885 web servers, 44,211 (22.57%) were running Apache and 24,222 (12.37%) were running Microsoft IIS.\(^\text{24}\) 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.\(^\text{25}\) Appendix A gives an analysis of the servers in our sample set, including geographic location and top-level domain distribution.

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.\(^\text{26}\) Appendix B discusses the substitutability of Apache and IIS.

At the time of this study the price for Windows Server 2008 R2 Standard is $1029 for five licenses, Windows Server 2008 R2 Enterprise is $3999 for twenty-five licenses, and Windows Server 2008 R2 Datacenter Edition is $2999 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

---

\(^\text{22}\) 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.

\(^\text{23}\) The other 98.77% of the IPs scanned were either inactive or were devices that were not web servers on standard TCP ports.

\(^\text{24}\) 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.

\(^\text{25}\) 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.


---
purely for “the development and deployment of Internet-facing Web sites and services.”27 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.

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 $2999 for one license. Finally, we can consider the bare-bones version Windows Web Server 2008 in the middle,28 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.29 With our estimate of the number of Apache Web Servers publicly 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 publicly 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 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.30 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 toward the functionality of the higher end Datacenter version of IIS.

3.3. Economic importance of Apache

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.31 By this yardstick, the stock of Apache software in the United States is as much as 5% (12.8/257) of software sales. However, this compares a stock to a flow, so some readers 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%) was prepackaged.32 If that ratio holds for investment stocks, then the stock of prepackaged software in the United States was $146 billion dollars in 2010.33 By that yardstick, Apache software is bounded by as much as 8.7% (12.8/146) of the measured capital stock of packaged software, or as little as 1.3% (2/146).

This 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. 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. 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,34 and the size of US online advertising revenue in 2009 is approximately $21 billion.35 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.36 Against that, the $2 or $12 billion of Apache software appears quite large; albeit, a reader could

27 “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” is “custom software,” namely, software built by a third party, such as a consultant, or software built by in-house employees.
29 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#NAICS204849 (accessed November 2011).
30 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#NAICS204849 (accessed November 2011).
worry about 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 et al., 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, 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 et al., 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 et al. (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? 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. 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. 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 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.

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. 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. 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, 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.

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.

---

37 Specifically, this uses Eq. (1) in BOS, and the estimates in Table 1. The simulation asks what would happen if Apache had pecuniary value.

38 As shown in Eq. (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.

39 This uses the same source of data, as noted above.

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

41 See Netchcraft.com.

42 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.
Table 1
Contribution of Apache to GDP, simulation, billions of dollars.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.04</td>
<td>0.08</td>
<td>0.13</td>
<td>0.19</td>
<td>0.25</td>
<td>0.30</td>
<td>0.36</td>
<td>0.43</td>
</tr>
<tr>
<td>B</td>
<td>0.07</td>
<td>0.14</td>
<td>0.23</td>
<td>0.32</td>
<td>0.43</td>
<td>0.53</td>
<td>0.63</td>
<td>0.75</td>
</tr>
<tr>
<td>C</td>
<td>0.03</td>
<td>0.10</td>
<td>0.20</td>
<td>0.32</td>
<td>0.46</td>
<td>0.63</td>
<td>0.80</td>
<td>0.99</td>
</tr>
<tr>
<td>D</td>
<td>0.06</td>
<td>0.17</td>
<td>0.34</td>
<td>0.55</td>
<td>0.80</td>
<td>1.08</td>
<td>1.38</td>
<td>1.71</td>
</tr>
<tr>
<td>2004</td>
<td>0.52</td>
<td>0.59</td>
<td>0.65</td>
<td>0.71</td>
<td>0.74</td>
<td>0.82</td>
<td>0.93</td>
<td>1.04</td>
</tr>
<tr>
<td>2005</td>
<td>0.90</td>
<td>0.86</td>
<td>1.00</td>
<td>1.13</td>
<td>1.23</td>
<td>1.28</td>
<td>1.42</td>
<td>1.61</td>
</tr>
<tr>
<td>2006</td>
<td>1.19</td>
<td>1.36</td>
<td>1.54</td>
<td>1.73</td>
<td>2.08</td>
<td>2.25</td>
<td>2.42</td>
<td>2.59</td>
</tr>
<tr>
<td>2007</td>
<td>2.06</td>
<td>2.36</td>
<td>2.67</td>
<td>2.99</td>
<td>3.30</td>
<td>3.59</td>
<td>3.89</td>
<td>4.18</td>
</tr>
<tr>
<td>2008</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 et al. (2013). See text.

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.43

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 et al., 2014). 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.

43 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.
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.

Acknowledgements

We thank Justin Ng for research assistance. We are grateful for the many useful comments we received from audiences at the NBER, Kristina McElheran, Dan Sichel, Ashish Arora, and two anonymous referees. All errors are our responsibility.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.respol.2014.01.003.

References


