Research Policy 43 (2014) 623–631

Contents lists available at ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

Digital dark matter and the economic contribution of Apache

Shane Greenstein[a,*], Frank Nagle[b]

[a] Kellogg School of Management, Northwestern University, 611 Leverone Hall, 2001 Sheridan Road, Evanston, IL 60208, USA
[b] Harvard Business School, Harvard University, Wyss Hall, Soldiers Field, Boston, MA 02163, USA

ARTICLE INFO

Article history:
Received 11 September 2012
Received in revised form 1 January 2014
Accepted 7 January 2014
Available online 5 February 2014

Keywords:
Open source
Apache
Economic measurement
Digital economics

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.

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)[1] HTTPd server, today known as Apache. It was one of two notable pieces of NCSA software, the Mosaic browser[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

[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.

[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.

* Corresponding author. Tel.: +1 847 467 5672.
E-mail addresses: greenstein@kellogg.northwestern.edu (S. Greenstein), fnagle@hbs.edu (F. Nagle).

0048-7333/$ – 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 contains 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,[3] and contributed to creating new processes in the economy that had long lasting consequences.[4]

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,[5] its licensing of the HTTPd server

[3] Forman et al. (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 et al. (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.

[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.