NBER WORKING PAPER SERIES

HIDDEN SOFTWARE AND VEILED VALUE CREATION: ILLUSTRATIONS FROM SERVER SOFTWARE USAGE

Raviv Murciano-Goroff
Ran Zhuo
Shane Greenstein

Working Paper 28738
http://www.nber.org/papers/w28738

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2021

We are grateful to Kenji Nagahashi and Mark Graham from The Internet Archive. We also thank Frank Nagle for providing the data used in Nagle (2019). We thank Audrey Tiew for comments and Alicia Shems for editorial comments. Funding for this research was provided by Harvard Business School, Ewing Marion Kauffman Foundation, the B.F. Haley and E.S. Shaw Fellowship for Economics through a grant to the Stanford Institute for Economic Policy Research, NSF SciSIP Award #1932689, and NSF Education and Human Resources DGE Award #1761008. The authors are pleased to acknowledge that the computational work reported on in this paper was performed on the Shared Computing Cluster which is administered by Boston University's Research Computing Services. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Hidden Software and Veiled Value Creation: Illustrations from Server Software Usage
Raviv Murciano-Goroff, Ran Zhuo, and Shane Greenstein
NBER Working Paper No. 28738
April 2021
JEL No. D22,L17,L86,O3,O33

ABSTRACT

How do you measure the value of a commodity that transacts at a price of zero from an economic standpoint? This study examines the potential for and extent of omission and misattribution in standard approaches to economic accounting with regards to open source software, an unpriced commodity in the digital economy. The study is the first to follow usage and upgrading of unpriced software over a long period of time. It finds evidence that software updates mislead analyses of sources of firm productivity and identifies several mechanisms that create issues for mismeasurement. To illustrate these mechanisms, this study closely examines one asset that plays a critical role in the digital economic activity, web server software. We analyze the largest dataset ever compiled on web server use in the United States and link it to disaggregated information on over 200,000 medium to large organizations in the United States between 2001 and 2018. In our sample, we find that the omission of economic value created by web server software is substantial and that this omission indicates there is over $4.5 billion dollars of mismeasurement of server software across organizations in the United States. This mismeasurement varies by organization age, geography, industry and size. We also find that dynamic behavior, such as improvements of server technology and entry of new products, further exacerbates economic mismeasurement.

Raviv Murciano-Goroff
Boston University
Questrom School of Business
595 Commonwealth Ave
Boston, MA 02215
USA
ravivmg@bu.edu

Shane Greenstein
Technology Operation and Management
Morgan Hall 439
Harvard Business School
Soldiers Field
Boston, MA 02163
and NBER
sgreenstein@hbs.edu

Ran Zhuo
Department of Economics
Harvard University
Cambridge, MA 02138
USA
rzhuo@g.harvard.edu

A data appendix is available at http://www.nber.org/data-appendix/w28738

1. Introduction

First deployed in 1991, the World Wide Web became an essential part of digital activity by the mid-1990s, and it has evolved with the digital economy ever since. Due to the Internet's origins as academic shareware, a large fraction of the Internet's software usage and upgrade activities remain unpriced, or, in other words, open source. Standard economic accounting measures the value of assets at their transactional price, but with software improvements often priced at zero and additions to software created without licensing fees, the operations and improvements to the commercial web have remained largely invisible to national economic analysis and accounting. The lack of visibility into the value created by unpriced software and software updates poses a troubling question: Does it interfere with understanding the sources of national productivity from the digital economy?

Every year the Bureau of Economic Analysis makes estimates based on enormous surveys, but software has long been a difficult commodity to identify economically. In contrast, typical capital goods are tangible items with a transaction cost—such as property, plants, and equipment—and are items that one business uses to produce goods or services for other businesses to use for creating consumer goods. For producers in most markets, capital goods deliver a flow of rentable services, and the price of capital goods reflects the anticipation of those services. The capital stock aggregates across those valuations, as users invest, upgrade, and retire pieces of capital goods. These notions have served as the foundation for the widely deployed neo-classical approach to the economic measurement of capital (Jorgenson 1963; Hall and Jorgenson 1967).

Yet, processes based on such reasoning make little sense in a setting where the capital good remains unpriced or its improvements do not generate any indication about the rental value of the capital. While transactions supporting proprietary products create dollar amounts that enter into economic accounting, specific functionally equivalent open source software upgrades, or *vintages,* do not leave any traceable transaction, and, thus, their value remains hidden, or veiled, to standard economic accounting. Stated broadly, unpriced software creates the potential for what we term *veiled input*. With no input to observe, nor any trace of improvement, if that input grows or improves, analysts will lack explanation for productivity gains, observing gain as if "manna from heaven" (Griliches 2000, Ch. 5).

The contrast between open source software and typical capital services also goes to the heart of the distinction between measurement and explanation. Economic accounting seeks to attribute the fraction of value created by an input by observing its changes over time and associating it with resulting changes in output. Even while major assets due to software grow in numbers or increase in quality, economic accounting may not recognize that an investment and maintenance process determines the spread between leaders and laggards, nor can it trace the improvements to organizations' behavior regarding input usage and adoption of improvements. Thus, a veiled input generates the potential for a distorted explanation for growth and creates the potential for misguided policy.

In this study, we analyze omission and misattribution in web server software. Web servers are one or more computer or groups of computers that run websites, which enable users to search and find information on web pages. Web pages are essentially computer documents consisting of text, images, and style sheets; and web server software are the programs that enable the servers to function as they do.

Web server software offers a useful lens for understanding the veiled parts of the digital economy for several reasons. First, web servers are ubiquitous and critical to the modern web-based commercial internet because they connect customers with sellers. Every commercial firm with an internet presence uses web servers and, thus, their software, which means that tens of millions of firms in the United States and hundreds of millions across the globe use web servers to support hundreds of millions of web pages. Second, server improvements relate directly to improvements in the internet user experience. In other words, there are several hundred million users in the United States alone, and their use of the internet necessarily touches different web servers on a daily basis. Third, the rate of improvement in web server software also correlates with, and sometimes directly causes, the rate of improvement in electronic commerce, which supports hundreds of billions of dollars in transactions a year. This is due to technical interdependence and complementarities between servers and browsers. Fourth, web servers serve as a useful proxy for the divide between the observable and the veiled creation of value, especially as businesses use both proprietary and open source software extensively, and therefore that creates strong concerns around mismeasurement. Relatedly, progresses in complementary technologies are also invisible in standard economic accounting, albeit, in ways that vary across users, a variance which, as we show in the text, was largely undocumented until this study.

How much do these potential omission and misattribution problems actually appear in web servers? Our principal goal is to examine evidence about behavior over time—that is with regards to upgrades and retirements of server software--and analyze whether omission and misattribution issues are small or large, and if large, characterize the general circumstances that correspond with realization of the problem. For that purpose, we compile the largest dataset ever assembled on sustained business web server usage, improvements, and retiring of software in the United States. We do this through an enormous search of the Internet Archive records of the Web between 2001 and 2018, in which we extract servers and proxies for their functionality from archived server headers of U.S. organization websites. We obtain disaggregated information on the continuing usage of web server software and the installation of software updates and match that data to information about more than 200,000 medium to large organizations. Our sample achieves good representation of organizations by geography and industry.

Because we develop a standardized approach to measurement—both across time and across organizations in different industries and locations, we can compare and quantify the issues of mismeasurement due to omission and misattribution. Our measures cover both the use of open source web servers, such as Apache and Nginx, and proprietary web servers, such as Microsoft's Internet Information Services, or IIS. These measures include market shares of different web servers, aggregate capital stock of servers, and aggregate distance of server software to the technological frontier. We also analyze the micro-behavior behind changes to these aggregates by identifying upgrades and replacement of software.

Our key findings are as follows. First, we show that the omission of open source web servers, such as Apache and Nginx, produces a large bias in measuring the economic value of server software, on the order of billions of dollars, in traditional accounting metrics. By using the price of the most popular proprietary server software as the proxy for the value of open source servers, we estimate the omitted value for our sample to be approximately $66 million in 2000 with an increase to between $125 and $315 million by 2018. As our sample is just under 3% of

total web server software in the United States,[2] we estimate that the total omitted value of open source server software in the United was between $4.5 billion and $11.2 billion by 2018.

We further show that the issues due to mismeasurement vary across organizations by differences in ages, geographies, industries and sizes. We show that the measurement issues arose at the turn of the millennium and became larger over time. They are particularly pronounced among young organizations, small and medium organizations, organizations in the West, and organizations in healthcare, lodging and food. We find dynamic behaviors—namely, upgrading and switching brands—further exacerbate mismeasurement, and are particularly troubling in the most recent decade, where we find that organizations have adopted a new open source server.

The study ends with a demonstration for how unpriced server software and software updates mislead standard analyses of sources of firm productivity. The study's data on server software usage can statistically and meaningfully explain variations in firm value-added levels. That finding aligns with other work that shows unmeasured open source contributes to productivity improvements (Nagle 2019).

Our study makes several contributions to the literature on the economic measurement of information technology (IT). First, our analysis shows that the usage of open source software creates the potential for mismeasurement. This is not widely appreciated even though open source is routinely used in artificial intelligence, electronic commerce, virtually every smartphone, and any big data application. Although usage of open source has been documented within the United States and several developed high-income countries (Lerner and Schankerman 2010; Robbins et al. 2018; Nagle 2019), virtually all research focuses on characterizing a subset of participants of software, and in only one year, at most.[3] This study is among a small set of studies (Robbins et al. 2018; Keller et al. 2018) to try to develop a standardized approach to measuring open source and its impact in anything resembling a census over time. That is an

[2] The total number of servers in the United States comes from estimates in Greenstein and Nagle (2014). In 2014, Microsoft sold IIS under three different licenses. The first, a basic license, was rarely used by businesses. The second tier was called a "Standard" license. Finally, the third tier was called a "Datacenter" license. We describe more in the text.

[3] An exception is Kim (2020). In that manuscript, Kim tracks the availability of open source software and investigates the open source software's effect on the reviews of router hardware over a twenty-year period.

important step for research to take, because, despite open source software's ubiquity, the its lack of standard measurement results in open source playing no role in the standard international indices for designating industries as "IT-intensive" (Calvino et al. 2018). We are the first study to demonstrate how to track web server usage over time and across industries.[4]

Second, we contribute to the many studies that measure the contribution of IT to economic growth (Jorgenson, Ho, and Samuels 2005). Many studies that seek to understand the impact of investments in IT on firms' output (Brynjolfsson 1993; Byrne, Oliner, and Sichel 2013; Tambe and Hitt 2014) have focused on priced IT assets. Productivity estimates for observable assets tend to have "high" coefficients when IT equipment and other tangible assets proxy for intangible assets, which are unobserved in microdata. Gordon and Sayed (2020) find the increase in ICT investment explains most of the US increase in productivity growth during 1995—2005 but the standard growth accounting method was unable to capture that. More recently, studies link the advance of productivity to advanced functionality enabled by frontier IT, such as data analytics (Wu et al. 2020). Our estimates align with Byrne and Corrado's (2019) findings in a study using novel methods for qualitative adjustment as measured through price indices, which examined improvements in a range of IT services. Our findings suggest standard accounting measures underestimate the existing stock of IT due to the omission of unpriced IT, that is, IT acquired at a zero price. This conclusion also aligns with Korkmaz et al. (2019), who examines the prevalence of frontier statistical software, Python and R, and estimates its unmeasured value, and Robins et al. (2018), who estimate over a billion dollars of open source code in the federal government.

Third, we document previously unrecognized patterns of misattribution correlated with the use of server software. As with Greenstein and Nagle (2014), this study indicates that the benefits from federal support for the internet have been underestimated. While Greenstein and

[4] We note the contrast with Netcraft Ltd, which has published a monthly web server survey (https://news.netcraft.com/archives/category/web-server-survey/) for a long period of time. However, the firm has been opaque about their methodology and many believe their most recent data is not reliable. The strength of our approach lies in its transparency. We construct our data based on archived server headers obtained from the Internet Archive, a non-profit organization supportive of academic research. This makes it possible for us to match with other data, and it enables future researchers to replicate our study and to use techniques similar to ours for various future research related to web servers. We elaborate in the body of the paper.

Nagle (2014) suggest that the underestimate could arise from both omission and misattribution, the findings from their empirical analysis focus on the omission problem alone. That is due to the strengths and limits of their sample, which consist of a single year's cross-section of 1% of outward-facing web servers in the United States in 2011.

In contrast, here we examine a sample that captures almost two decades of outward-facing server usage, which also enables the analysis of upgrades and replacement behavior. Moreover, Greenstein and Nagle (2014) do not match their data with any other, while this study includes a substantial number of organization characteristics, enabling us to study how usage interacts with a rich set of organization attributes. We also build on Nagle's (2019) study of the omission problem in the usage of Linux. His data came from the firms covered in Harte Hanks data, and to our knowledge, is the only other study to directly link open source usage to productivity. Compared to Nagle's study, our study surveys both a much wider sample of organizations, and longer time-frame for their usage of proprietary versus open source software; we also match a subset of the study's data with his, recreate his productivity findings, and then improve upon them.

Our analysis also provides new insights into how errors from veiled inputs accumulate over time, which no previous study could analyze because none had such a long time series. We observe how both the market goods (proprietary web servers) and near-market goods (open source web servers) improve over time, and how usage adapts accordingly. We also observe behaviors that suggest users treat server software like other capital goods. Moreover, we observe this setting long enough to see the set of options change drastically over time. In our sample period, Nginx, one of the near-market options, did not exist for years, and then began to improve and become more widely adopted. To our knowledge, our study is the first study to give attention to switches and substitution in measuring economic activities related to the use of open source software.

The rest of our paper is organized as follows. In Section 2, we elaborate on the problems associated with measuring gains from software, specifically omission and misattribution. In Section 3, we explain the origins of the major open source web servers, Apache and Nginx, and the major proprietary web server IIS. Section 4 explains dataset construction. In Section 5, we

propose a few measures of the economic value of web servers. Section 6 presents our empirical results. In Section 7, we offer conclusions.

2. Theory of Measuring and Mis-measuring Gains from Software

We study what potential economic measurement issues arise from the deployment of open source software. The use of open source server software may create mismeasurement due to omission in the typical accounting measures. Normal economic measurement focuses on transactions taking place in markets and presumes that those transactions involve a positive price. Open source software, such as Apache and Nginx, are freely distributed and do not directly generate revenue, even though they perform functions similar to those performed by proprietary software, such as IIS. Without explicit attention, normal procedures will treat unpriced activities as nonmarket activities, which creates two types of issues that we label *omission* and *misattribution.*

2.1 Omission

Nordhaus (2006) presents a general review of the methods for measuring inputs in many circumstances, including some guidance for settings that lack prices. Although open source is *not* singled out as an example by Nordhaus (2006), this setting partially fits what Nordhaus labels a "near-market good." In his discussion, omission errors arise when standard procedures presume a zero price is affiliated with non-market activity, while real economic activity creates valuable goods with no price. There are also important differences between this setting and the examples discussed in Nordhaus' study, since some of the activities affiliated with open source software can be measured. For example, in the setting of web servers, 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. Complementary activity and participants are key parts of the open source ecosystem (West 2003; Lerner and Schankerman 2010). Most activity will involve market transactions and positive prices. Organizations might also purchase hardware for deployment, as well as additional services in order to accommodate large-scale use. Such expenditure would appear as an operating expense. In practice, measuring the total value created by open source software requires approaches that account for multi-factor productivity (Syverson 2011).

Like Byrne and Corrado (2019), we seek to measure what the standard procedures overlook with regards to unpriced activity. In this setting the mismeasurement due to omission arises from dynamic behavior, such as upgrading and replacement. This limits the insights that cross-sectional analysis can provide for three reasons. First, both the market goods and near-market goods improve over time. As a result, organizations lagging in one time period may leapfrog into leading positions by adopting the latest products. Second, the set of alternative products from which an organization selects its technology stacks changes over time. When option change, organizations may switch extensively between proprietary and open source web servers as different products' features improve. In particular, and especially relevant to our setting, an open source server software, Nginx, did not exist as a viable alternative until at least 2004. Nginx's entry created substitution patterns unseen in cross-sectional data. Third and important for our setting, many organizations combine software from different sources, complicating estimations of the omitted value. Those changes over time can produce scenarios that previous works, such as Nordhaus' (2006), did not examine.

2.2 Misattribution

Economic growth may also be misattributed to observable assets instead of to the use of open source software and the installation of unpriced software updates. Specifically, this misattribution biases the coefficient on observable IT in standard productivity analysis. Prior estimates of the productivity impact of IT have found estimates "too high," which suggests that the presence of unobserved inputs correlated with observed inputs (Brynjolfsson 1993; Byrne, Oliner, and Sichel 2013; Tambe and Hitt 2014), but prior research has suggested a variety of potential mechanisms. This study investigates whether a specific mechanism -- the presence of open source software -- could create a bias. In this section we show that it is possible and plausible.

To understand the potential for 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,[5] 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

$$\ln(Y_{it}) = A_{it} + \beta_L \ln(L_{it}) + \beta_K \ln(K_{it}) + \beta_{IT} \ln(IT_{it}),$$

where Y is revenue, and this equation 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, Oliner, and Sichel 2013; Nagle 2019). Because usage of open source software by an organization 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. Relatedly, all inputs are measured with error. In particular, only proxies are available for measuring open source IT usage.

The intuition for misattribution can be illustrated in a few scenarios. We can illustrate scenarios that could create an upward or downward bias, and we will hypothesize that the scenario behind the downward bias is more plausible. Analysis of server data will be consistent with that hypothesis.

The true amount of IT is not fully observed. Call the observable asset IT^O, and unobservable asset IT^U. All firms have some of both assets, and a fraction of them are observable $IT_{it}^O/(IT^O + IT_{it}^U) = h_i$, where that fraction varies across firms. Now consider a cross-sectional regression in time t. Even though the true level of IT is actually IT_{it}^O/h_i, the econometrician typically estimates:

$$\ln(Y_{it}) = A_{it} + \beta_L \ln(L_{it}) + \beta_K \ln(K_{it}) + \beta_{IT} \ln(IT_{it}^O) + e_{it},$$

[5] This type of analysis can be implemented at either the industry level (Stiroh 2002) or at the firm level (Nagle 2019). For simplicity, we carry it through at the firm/organization.

where $e_{it} = -\beta ln(h_i) + e'_{it}$. While e'_{it} is distributed i.i.d, part of the error is potentially not independent. Because h is always less than one, $-\beta ln(h_i)$ is always positive, and smaller h leads to larger $-\beta ln(h_i)$. That creates the potential for biased estimates, a potential which declines as more IT becomes observable, that is, as h approaches 1.

Three scenarios can arise, and create different "directions" for the bias. They depend on the correlation between h_i and IT_{it}^{O}. Consider the following scenarios that arise when there is no correlation, positive correlations, and negative correlation:

No interdependence. There is no correlation between h_i and IT_{it}^{O}. This scenario occurs when investment in the observable and unobservable asset have no relationship with one another. This leads to *growth without cause*. This can happen when open source code improves or when users receive software updates at no expense. In this case, some firms produce more output without appearing to change their inputs. 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. Growth will be attributed to *A*, because of the appearance of more productivity that cannot be attributed to growth in inputs. This scenario resembles one discussed by 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.[6] Syverson argues that the gains could appear to be disembodied technical change, not attributable to any specific input.[7] By analogy, the greater h is, the larger the disembodied technical change.

A coincidence of assets. The correlation between h_i and IT_{it}^{O} is positive. Theoretically, this scenario arises when the unobservable asset is used more frequently in larger installations. That could occur because the observable and unobservable assets are used together, and the presence of the observable asset makes the unobservable asset more productive in larger installations of the observable asset. This leads to an *overestimate of the contribution of assets* in

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

[7] Or, as in Tambe and Hitt 's (2014) study, problems could arise from mismeasurement of labor, which lacks adjustments for the human capital affiliated with supporting the software, or for the extent to which labor relies on the community to enhance their productivity. They also point out that measurement error may occur due to the differences between labor-based and capital-based estimates of IT productivity.

a cross-sectional regression (where t is fixed and only variance across i is observed). *More* unobserved IT is present in firms with *more* observed IT, and these firms produce more Y. That scenario will bias the estimate for a coefficient upward. The firms with higher IT will seem to get an even larger gain from their observed IT than plausible. Estimation on Y_{it} - $Y_{i,t-1}$ in a first differences estimate does not resolve issues if growth in IT^O_{it} and IT^u_{it} leaves h_i unchanged, which we expect if the underlying cause behind the level of h remains unchanged. Then overestimates again arise, and the coefficients are biased upward.

Substitution of assets. The correlation between h_i and IT_{it}^O is negative. Theoretically, this scenario occurs when an observable asset substitutes for the unobservable asset as installations become larger. This would occur, for example, if the observable asset were more productive than the unobservable asset in large installations, as well as when the opposite was so in small installations. The presence of more of one leads to less of the other, and that results in an underestimate of the contribution of assets. When *more* unobserved IT is present in firms with *less* observed IT, the estimate of the coefficient will be biased downward. That is, the firms with higher IT will *not* seem to get much gain from their higher IT. Once again, estimation of growth does not become resolved issues if h_i remains unchanged. Once again, the coefficients are biased downward.

There is a fourth possibility, *growth attributed to the wrong input.* Greenstein and Nagle (2014) point out that another scenario for misattribution arises if a fraction of firms have a high h_i and another fraction a low h_i, and the former invest in labor to support a new release or upgrade.[8] In that case, the firms using open source software will experience an increase in output, Y, and an increase in L, while showing no measured change in IT capital. Firms using proprietary software, 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. Though not the focus of this study, for analytical completeness we note this is also a possibility.

[8] Higher labor expenditure could arise either from the need to hire more workers or compensate workers more for their efforts. 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). The evidence is consistent with the existence of the premium, but cannot serve as an estimate of its size.

Finally, an important caveat applies. The scenarios above only consider the spillovers from direct usage of open source software as an input into production. They do not account for the spillovers that occur when a competing product adds a feature by imitating a similar feature developed for the product in use. Nor does this include further gains from enabling the entry of complementary applications or the enabling of more productivity business processes. We speculate that such a sequence of imitation and enabling activities would be measured as improvement in intangible capital. Below we provide direct evidence for the first steps in that mechanism -- improvement and replacement of server software. We regard the next behaviors as widespread, albeit, difficult to measure with precision, and the measurement of that behavior as an exercise for further research.

The above scenarios frame what can be learned if one could observe what has previously not been visible. How do the observable and unobservable assets correlate as firms and IT installations become larger or small, if at all? Where is the potential for mismeasurement, as evidence about variance in correlations between the observable and previously unobservable assets—across industries, regions and time?

3. Origins and Pricing of Web Server Software

While every online business uses web servers, the three most popular web server software packages—Apache, IIS, and Nginx—had very different development histories, pricing strategies, and visibility. Most notably, while Apache and Nginx emerged as open and freely available, Microsoft's IIS is proprietary software. These differences mean that these servers contribute to traditional economic accounting measures in different ways. In this section, we recount how these web servers emerged, as well as how their pricing affects the way they enter aggregate valuation measures.

Apache descended from the very first web server, which was developed in an academic setting and was freely distributed as shareware. In 1993, the National Center for Supercomputing Applications (NCSA) at the University of Illinois developed a computer program known as the NCSA HTTPd server. The HTTPd server software supported the sharing of content on the web

through the Hypertext Transfer Protocol (HTTP). NCSA made HTTPd available as shareware within academic and research settings, along with the underlying code. The original developers did not place any restrictions on the usage or modification of the software. Many webmasters took advantage by adding improvements as needed or by communicating with the lead programmer, Robert McCool, who coordinated the addition and releases of new extensions. When in the spring of 1994, McCool left the University to become one of the first ten employees of the newly founded Netscape, the development of the web server software fragmented with eight distinct teams working on eight distinct vintages of the software. In 1995, the eight teams decided to coordinate and unify their efforts into one server to be known as Apache (ostensibly because it was "a patchy web server"). The University of Illinois then fully transferred the development of the server software to the Apache organization without any licensing or restrictions. After 1995, Apache grew in popularity as the commercial internet grew, becoming widely used in customer facing and procurement activities of many organizations. It is regarded as the second most popular open source project used by businesses, after Linux.

The growth and continued deployment of Apache, therefore, has largely taken place outside the visibility of standard economic measurements. Traditional economic accounting measures aggregate the cost of inputs and value outputs at their transaction price. In contrast, the Apache server has never had a price affiliated with either its inputs or outputs. The Apache Foundation relies upon donations and a community of technically skilled users who provide new features at no charge, motivated both by the intrinsic and extrinsic rewards. Further improvements in Apache code relied on the equivalent of donations for support. These came in the form of explicit donations from organizations who provide personnel time and firm capital, or it came from programmers devoting leisure time to open source activity. It also may have come in the form of in-kind or unacknowledged donations of capital or services, such as computer time and hosting facilities.

As with other open source software (Lakhani and von Hippel 2003; Lerner and Schankerman 2010), Apache eschews standard marketing/sales activities, instead relying on word-of-mouth and other non-priced communication online. 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. While Apache itself continues to be developed by volunteers and distributed without a price, and is therefore invisible to traditional economic

accounting, Apache is affiliated with a plethora of revenue-generating economic activity. For example, there is a large labor market for Apache programmers, administrators, and third-party consultants.

Around the same time that Apache emerged as open source web server, Microsoft developed a proprietary web server and intended it as a substitute for Apache. Beginning in 1995, Microsoft provided their server software, known as the Internet Information Services or IIS, as part of the Windows NT suite. That software is proprietary and a major revenue generating product for Microsoft. The Windows server became popular during the dot-com era in large part because Microsoft put its sale and support behind the product, which fostered the growth of many supporting documents and tools. Users of IIS say it possesses appealing features, including its compatibility with other Microsoft products, as well as its certification, documentation, and ease-of-use in enterprises with routine requirements. Many believe that Microsoft benefited from suspicion and security concerns among some large organizations about using open source code they had not vetted. Similar to Apache, a large labor market for IIS programmers, administrators, and third-party consultants exists.

The value that organizations get from adopting Microsoft IIS is visible in standard economic metrics. Because Microsoft charges for the Windows software that IIS is packaged inside, the transaction price of the software can represent the value that organizations place on this software and its features. Though Microsoft does not separately charge for IIS software updates, the additional features and fixes contained in those updates should also be priced in the software's original transaction price.

Nginx, the third most popular web server package and the most recent entrant, was developed through a different path than Apache or IIS, and also is freely shared as open source software. Compared to Apache and IIS, Nginx is a latecomer. Programmer Igor Sysoev started the initial work in 2002 when he sought to scale a server for a large online media company, optimizing it to handle at least 10,000 concurrent connections. In 2004, Sysoev opened Nginx to the public as an open source project, using a permissive Berkeley Software Distribution (BSD) license. Steady improvement from many contributors turned the software into a viable web server around 2007. Soon it was widely believed to become the most popular web server for streaming and video. Nginx performs well on large volumes of traffic, and that performance

gives it a foothold in media and entertainment enterprises with high peak loads, which comes at the cost of sacrificing some of the adaptability found in Apache. With a different appeal than the ease-of-management of IIS, which comes with considerable support, Nginx partially compensates by being interoperable with other web servers and by facilitating load-balancing[9] between multiple applications. Over time the software community around Nginx added extensions and modifications in an attempt to grow the capabilities of this niche. In other words, Nginx began as a niche product that complemented IIS and Apache, and over time became a substitute for a range of applications.

Like Apache, Nginx is freely available for anyone to use and modify as they see fit. Therefore, like Apache, the decision of organizations to adopt Nginx is not captured in transactional exchanges. As many of the users of Nginx require enterprise-level features, Nginx has spawned a number of complimentary revenue-generating activities. In July, 2011 in San Francisco, Sysoev and his business partners founded a company, also named Nginx (see nginx.org or nginx.com), and. Sysoev serves as the CTO. Over the course of the data set used in this study, Nginx the company supports "Nginx Plus," which is server software that includes enterprise-level services and offers a set of paid extensions on top of the open source Nginx. These commercial extensions target long-time users who desire commercial-grade features that are not normally available in any existing open source product. They also target enterprises that require both technical support and license payments. These features help maintain Nginx's use as an "edge web server" for the cloud, hosting, and content delivery network (CDN)[10] service providers. This aspect of Nginx-related activities does fall into standard economic accounting.

4. Data

To study the prevalence of the mismeasurement scenarios discussed in the previous section, we compiled the largest data set ever on business web server use between 2001 and

[9] A load balancer mediates all the requests coming in to the server and makes sure that no one server is overworked, all of which maintain speed and reliability. If and when a server crashes, the load balancer redirects web traffic to other servers.

[10] A CDN is a network of servers that are geographically dispersed so that different servers can be closer and therefore faster and more reliable to users in different locations.

2018, covering over 200,000 medium to large organizations in the United States. We make use of multiple data sources for data construction and we summarize these sources in Table 1. In this section, we discuss in more detail our sample construction.

We define our sample of organizations as the organizations listed in Bureau van Dyke's Orbis database with at least 50 employees at some point between 2000 and 2018. The Orbis database contains 230,611 organization homepages for organizations matching these criteria. We then extract additional information about these organizations from the Orbis database. For each organization homepage in each year, we find the Orbis organization records with the associated website. We use the Orbis record that contains consolidated financial information, or when such data is not available, we use unconsolidated data for the largest subsidiary. Among the variables that we capture from the Orbis data are the number of employees that an organization has in that year, the main industry the organization operates in, the location of organization headquarters, the organization's year of incorporation (and, if applicable, year of disincorporation), as well as the organization's capital expenditures. We supplement the organization data with CompuStat data for public firms. In 0, we describe the overlap in the coverage and provide more details on the data construction.

The data on server software vendor and vintage in this paper comes from analyzing the server headers of U.S. organization websites collected by the Internet Archive (IA). When an individual visits a website, her computer connects with the web server that hosts that particular site. Upon connecting, the web server responds with both the content of the webpage as well as metadata known as the server header. By default, server headers contain the name of the responding web server's vendor as well as the vintage of web server software being used to host the requested webpage. The IA is a nonprofit organization that stores snapshots of websites for archival purposes. IA's computers regularly connect to large numbers of websites and record both the content as well as the server headers.

Out of the 230,611 organization homepages we initially identified, the IA holds server header data for 213,956 of those sites. To get a sense of scale, we compared this data with Greenstein and Nagle (2014)'s sample of U.S. web servers in 2011. Greenstein and Nagle (2014) estimated that 4.28 million total servers were running Apache in the United States, and 2.35 million servers were running IIS. Our sample in 2011 examines 102,376 Apache servers and

82,776 IIS servers, which, taking both estimates at face value, is approximately 2.79% of total servers. In spite of seeing only a fraction, the dataset is able to identify the identity of the organization, an exercise which Greenstein and Nagle (2014) did not perform. We can take advantage of matching information about organizations and can look for indications that we sampled representative U.S organizations. The data in 0 indicate that our sample appears representative in terms of dimensions we can measure, such as region of headquarters and industry among medium and large organizations.

We then transform the server headers into a panel dataset. In the panel, an observation is an organization website in a month. Each observation in the panel includes the vendor of the server software used by that organization in that month as well as an indication of the vintage of server software, which indicates something about when the user updated the software.[11] The definition of vintage takes advantage of the practice, common among software producers, to number their software in ordinal sequences. Improved software takes higher numbers.[12] We refer to improved software as a later vintage of software. We observe three features of servers: The vintage number, its date of installation at each organization, and the date at which the vintage is first introduced at any organization. We define a software update as any time that the server vendor used by a homepage stays the same between two months but the server vintage changes.

There are some limitations of our server header data. First, the IA does not capture every website each month. On average, a homepage in the dataset appears in approximately half of the months between its first observation in the dataset and its last observation. Therefore, the panel is unbalanced. Second, some savvy server administrators turn off the display of information in server headers. By modifying a server's configuration files, the server can be made to only respond with limited information about the server vendor and vintage being run.[13]

To handle missing data due to the intermittent scanning of websites and server administrators turning off the display of server information, we impute some missing observations. Specifically, if a website's server header does not contain the server vendor and

[11] The parsing of the raw server headers into this information is described in Appendix A.2.

[12] Major improvements increase the first digit from 1.0 to 2.0 to 3.0, and so on. Minor improvements improve smaller digits, going from 1.1 to 1.2 to 1.3, and so on.

[13] The Apache server configuration files allow for five different levels of information being sent back to visitors in the server headers. These are described in more detail in Appendix D.

vintage number in a particular month—either because IA did not scan that website in that month or because a server administrator had disabled it—and then reappeared in a subsequent month with the exact same vendor/vintage combination as previously used, then we impute the months in between as the same vendor/vintage combination. On average, an imputation consists of adding 2.48 months between observations of the same server vendor and vintage number (with the median imputation inserting just 1 month).

We also quantify the measurement error in the timing of vintage updates and vendor switches. The number of months between when we detect a vintage update and the previous observation is 1.37 months and the 75th percentile is 1 month. The average number of months between an observation in which we detect a vendor switch and the previous observation is 2.63 months with a 75th percentile of 2 months. Therefore, we are able to detect the timing of software updates and vendor switches within a narrow window of time despite some missing data.

One other factor shapes what is possible. Apache and IIS server headers show different levels of details. Apache server headers frequently reveal the precise vintage of software that a website is using, including the major and all minor vintage numbers (e.g., Apache 1.3.6 improves upon 1.3.5). This enables us to know the precise release date of those server vintages being used. In contrast, IIS server headers only reveal the major and first minor vintage number (e.g., IIS 6.5 improves upon IIS 6.4). While IIS has many minor updates and security patches, the server headers do not show them. Thus, our visibility into the vintage used by IIS users is less granular than that for Apache.

We stress that the data examine an important place in virtual space—namely, an organization's central web page on the Internet—where server software necessarily plays an important role for an organization. The servers support the publicly displayed online face of an organization and direct inbound queries: Specifically, the organizations' websites, which are essential for generating online sales, coordinating employees, and performing many operational functions, are supported by the servers. Thus, we do not expect the actions for this server to be "an afterthought," nor an investment decision lightly taken.

Our focus also comes with limits in that the website is not the only virtual location where web servers play an important role in a typical organization, and we have little visibility into

those other roles. For example, we do not see the web servers that support procurement, human resources, some parts of order fulfilment and sales support, or data analysis. Prior work has hypothesized that investment in IT can play a crucial role across the entire organization (Calvano 2018) and can support frontier processes in data analytics that drive productivity improvements (Wu et al. 2020). However, developing exhaustive processes for measurement of hidden inputs across an entire organization remains a large and open research topic.

5. Measurement

Traditional economic accounting mismeasures the value created by web servers, and the dataset helps uncover what was previously veiled. To understand whether this measurement applies systematically, we propose several summary indices.

Our first measure compares the date of a *release* of a vintage of frontier server software with the date associated with the most recent vintage of the *installed* software, which we call the *distance from the technological frontier* (DTF). Inspired by standard approaches to measuring improvements in capital through the benchmarking of them by "how many years the productive capital is out of date," *DTF* measures technological progress by emphasizing the introduction of new features to sequential vintages of software, which traditional economic accounting does not capture. Our measure improves over the traditional measures on multiple dimensions of technological progress in web servers, such as being capable of handling larger numbers of concurrent requests and being able to respond to requests more quickly. It ignores if an organization uses a server with more or less intensity over time, and how much, and is computed as follows:

$$DTF_{v,t} = (t - v) - (t - v_{frontier})$$
$$= (v_{frontier} - v),$$

where v denotes the vintage of server software released in time v, t denotes the time of observation, $v_{frontier}$ is the vintage of the most recently released software.

To illustrate how this index works, let us suppose that we are interested in computing the $DTF_{v,t}$ at $t =$ November 2002 for Apache 2.0.11, which was release at $v =$ February 2001. The

most recently released Apache software in November 2002 was Apache 2.0.43, which was released at $v_{frontier} =$ October 2002. We compute *DTF* as follows:

$$\begin{aligned} DTF_{Feb\ 2001, Nov\ 2002} &= (Nov\ 2002 - Feb\ 2001) - (Nov\ 2002 - Oct\ 2002) \\ &= (Oct\ 2002 - Feb\ 2001) \\ &= 20\ months. \end{aligned}$$

In order to examine if *DTF* is increasing or decreasing over time for a large group of organizations, we construct an aggregate DTF_t in each moment of observation using a weighted average of $DTF_{v,t}$, weighted by the number of users of each vintage:

$$DTF_t = \frac{\sum_{v \in \mathcal{V}_t} Q_{v,t} * (v_{frontier} - v)}{\sum_{v \in \mathcal{V}_t} Q_{v,t}},$$

where $\mathcal{V}_t$ is the subset of all vintages used at time *t* and $Q_{v,t}$ is the number of users of vintage v at time *t*.

For example, suppose we are interested in examining the aggregate DTF_t of organizations using the Apache server software at $t =$ November 2002. For simplicity, let us assume that there were only 100 organizations, among which 45 organizations used Apache 2.0.11 and 55 organizations used Apache 2.0.43. Our aggregate DTF_t for Apache users would be:

$$\begin{aligned} DTF_{Nov\ 2002} &= \frac{45 * 20\ months + 55 * 1\ month}{45 + 55} \\ &= 9.55\ months. \end{aligned}$$

We also define the *quality-adjusted DTF* (QADTF). Inspired by quality adjustments found in the tradition of hedonic price indices, this measure adjusts the baseline DTF by accounting for the quality that can be purchased for a dollar over time.[14] We propose to use the inverse of the CPI as our weight, or *W(v) = 100/CPI(v),* as our measure of *quality per dollar*, where *CPI(v)* is the Consumer Price Index value for software vintage *v*. We use the Consumer Price Index (CPI) for "Computer Software and Accessories" from the U.S. Bureau of Labor

[14] For many years the CPI for packaged software has included a hedonic-estimated adjustment for qualitative change.

Statistics. This CPI is on a December 1997=100 base.[15] The QADTF measure is defined as follows:

$$QADTF_t = \frac{\sum_{v \in V_t} W(v) Q_{v,t} * (v_{frontier} - v)}{\sum_{v \in V_t} W(v) Q_{v,t}} .$$

To illustrate how the quality-adjusted index works, we again make use of the above example. The CPI for computer software and accessories in February 2001 was 80.1 and in October 2002 was 71.2 for a December 1997=100 base, meaning that our *quality per dollar* measure $W(Feb\ 2001) = \frac{100}{80.1} = 1.248$ and $W(Oct\ 2002) = \frac{100}{71.2} = 1.404$. We compute the quality-adjusted DTF for $t =$ November 2002 as follows:

$$QADTF_{Nov\ 2002} = \frac{1.248 * 45 * 20\ months + 1.404 * 55 * 1\ month}{1.248 * 45 + 1.404 * 55}$$
$$= 9\ months.$$

These weights are appropriate for use if the rate of qualitative change in server software resembles the rates of change observed broadly across all widely used software. Relatedly, using such weights introduces the potential for error in short periods, and should deter us from making inferences that depend too critically on a small number of observations. Hence, in the discussion below we favor inferences about trends that manifest over the long term.

Finally, we propose a measure of the quality-adjusted server software capital stock over time. The measure is defined as follows:

$$QACAP_t = \sum_{v \in \mathcal{V}_t} W(v) * Q_{t,v} .$$

[15] We were concerned about the robustness of this estimate, so we also compute an adjustment using the quality adjustments developed by Bryne and Corrado (2019), which they estimate at intervals of one year. The CPI is estimated at monthly intervals, so we display that. Using the Byrne and Corrado estimates as weights gives qualitatively similar results over the long term. When we have multiple observations for one website in a year, we take the average CPI of the servers that they used in that year.

This measure captures the number of server software used by organizations in each moment of observation, taking into account that more recent server software has higher quality.[16] Note that the unit of this measure is the quality-adjusted quantity of server software, on a base where vintages of server software released in December 1997 have a $QACAP_t$ of 1. As in standard capital measurement, this adjustment enables us to make an estimate of the increase or decline in quality-adjusted server capital stock—an estimate that permits us to see what previously lay hidden. Major changes over time or major differences across industries create the potential for large measurement problems.

To illustrate how this index works, we compute the $QACAP_t$ for our above example as follows:

$$QACAP_{Nov\,2002} = 1.248 * 45 + 1.404 * 55 = 133.4$$

6. Results

In this section, we document the patterns of software usage that mislead traditional economic accounting. Using the organizations in our sample, we attempt to quantify the amount of value derived from web server use, which has been omitted from previous economic measurements. For convenience of the readers, we summarize the questions, methods, and results in this section in Table 2.

Figure 1 displays the fraction of organizations in the data set using server software from the major vendors between 2000 and 2018. In the left-hand figure, we show the unweighted market shares, and in the right-hand figure we show the quality-adjusted market shares (based on the vintage of the software).[17] In both plots, Apache and Microsoft IIS had similar market share

[16] Because this measure sums over the inferred quality of the server software and the quality of server software is based on the server version number, hidden server numbers pose a challenge for this measure. If hidden version numbers are ignored then QACAP numbers would be underestimated. We therefore provide lower and upper bounds on the QACAP. The lower bound is developed by interpolating hidden version numbers with the last visible server version number used by a website. The upper bound is developed by interpolating with the most recently released server version given the observed server vendor. In the main text, we show the lower bound as that is the most conservative measure of QACAP. In Appendix I, we show both the lower and upper bound. Our qualitative results are consistent and robust regardless of which interpolation method is utilized.

[17] Because Apache and Nginx users are more likely to hide their version numbers in recent years (see Figure 15), computing the QACAP using only the observations in which software vintages are visible would undervalue the Apache and Nginx servers. For Figure 1(b), we treat servers that turned off their server version number as staying at the vintage of their last observation. While this attenuates the bias caused by organizations changing the visibility of

during much of the early 2000s. Beginning in 2010 and accelerating after that, Nginx began to capture market share from both Microsoft and Apache.

The patterns displayed in Figure 1 indicate that until 2010, calculating the software capital stock on the basis of just Microsoft IIS, the leading proprietary web server during this time period, would result in proportionally undervaluing the total stock. For each organization paying for Microsoft IIS, another organization captured similar value by using open source Apache software.

Although the results in Figure 1 suggest a periodic survey of organizations' software usage might have been sufficient for aggregate estimates, the patterns change after 2010. Beginning in 2010, Nginx, an open source solution, began to take significant market share both incumbent proprietary and open source software. Hence, after 2010, a more appropriate approach would have had to account for changes taking place. A periodic survey would have missed the extent of change.

Figure 2 shows the quality-adjusted capital stock based on the vintages of server software (*QACAP*).[18] The pattern shown in Figure 2 demonstrates a similar finding about usage prior to 2010 as indicated by Figure 1, and, if anything, suggests Apache and IIS were similarly ranked in usage prior to 2010. Use of Apache and IIS grew with little interruption until approximately 2013, when their growth flattened and then declined, with IIS declining earlier than shown in Figure 1. In the dozen years between 2001 and 2013, IIS usage tripled and Apache's quadrupled. As with Figure 1, Nginx usage grew quickly after 2013, but Figure 2 suggests this growth was even faster than indicated by Figure 1, in part because Nginx was so young and close to the frontier. These plots reveal that mismeasurement due to omission may be economically substantial during the ascendancy of Nginx use.

How large is the mismeasurement due to omission? Figure 3 provides an estimate of the value based on the shadow value of the open source servers used by organizations to host their homepages. Following Nordhaus' (2006) reasoning, for each year that an organization used an

---

their server software version numbers, given that organizations that hide their server versions are typically closer to the technological frontier than organizations that leave their version numbers visible (see Figure 16), the quality-adjusted market share of Apache and Nginx are likely even higher in recent years. See Appendix I for alternative estimates.

[18] See Appendix I for alternative estimates.

open source server, we find the "nearest market good," in this case, the most popular Windows IIS vintage used in that same year. The prices of these Windows IIS vintages represent the best proxy for the shadow value of the open source server used. By adding these up, we get the omitted value due to open source server software.[19] The omitted value shown in Figure 3 is large, starting at approximately $66 million in 2000 and increasing to between $125 and $315 million by 2018.[20] The increase is both a reflection of the increasing usage of open source server software, as well as the expanding number of features and value created by this software. If our sample represents 2.79% of all open source server software, as we estimated for 2011, and if that persisted, then the total value in 2018 would be between $4.48 billion and $11.29 billion.[21] While only an approximation, this suggests the scale of mismeasurement for just servers reaches many billions of dollars.

In Figure 4, we plot the heterogeneity of omission by organizations' industry, size, geography, and age. The magnitude of omission is both a reflection of the usage of open source server software, as well as the size of the relevant subsample. Due to the large proportion of manufacturing firms in our sample, we find that the manufacturing industry has a particularly large number of missing dollars. Similarly, we find that the omitted value to be the highest in the South, partly because the South has more organizations than any other region in our sample.[22] As our sample achieves good representation of medium to large organizations in the United States by industry and geography, we believe our finding is representative of heterogeneity in omitted value due to open source web servers in the U.S. economy. We also find that the omitted value is

[19] In Appendix E, we show the price series of Windows IIS vintages as well as the most popular Windows IIS vintage by year.

[20] Starting in 2014, Microsoft IIS could be purchased under a "standard" license or under a more expensive "datacenter" license. Organizations would typically opt for different licenses depending on the size of their deployment. We provide a range of estimates for the omitted value of open source servers based on where the low end uses the "standard" license to approximate the shadow value of open source servers and where the high end uses the "datacenter" license price.

[21] This fraction from 2011 is 2.79%, and assumes a similar fraction in 2018. That is probably an underestimate due to the increase in the use of Nginx and the increasing fraction of open source software in wide use, which came at the expense of proprietary software use. Hence, the estimate is conservative.

[22] In our sample, the South region has 61,285 organizations while the West has 38,572, the Northeast has 34,051, and the Midwest has 37,075. The Census Bureau's South region had a population of approximately 126 million in 2019, while the Northeast had 56 million. https://www.census.gov/newsroom/press-releases/2019/popest-nation.html In Appendix C, we show that the number of organizations in our sample is highly correlated with the total number of firms in each of these regions as reported by the Census Bureau Statistics of US Businesses data.

the largest for organizations with fewer than 250 employees and organizations more than ten years old, due to the fact that the majority of our sample are these organizations.

Figure 5 supplements Figure 4 by showing how the potential for mismeasurement due to omission correlates with organization characteristics. For Figure 5, we plot the fraction of organizations using open source server software by industry, size, geography, and age. The graphs show that younger organizations, West Coast organizations, organizations operating in accommodation, and smaller organizations are more inclined toward using open source software, while organizations in the Midwest, and organizations in finance and (eventually) retail are the slowest. This finding implies that mismeasurement due to omission may vary. It also suggests that mismeasurement may be larger for the software capital stock of the former types of organizations.

A descriptive logistic regression confirms the previous results and provides more nuance, while showing that mismeasurement is not driven by coincidental correlations. The marginal effect estimates at the means of all organization attributes are included in Table 3. The qualitative inference does not differ from those drawn with the figures, but these estimates do provide some statistical grounding for describing the heterogeneity. At mean values for all other variables, a large organization is 14.38% less likely to use open source than a medium sized organization, a young organization 3.91% more likely than an older organization, and a organization headquartered in the West is 10.43% more likely than one in the Midwest.[23] Organizations in the accommodation and food and information industries are, respectively, 39.51 % and 21.44% more likely than organizations in finance, which has the lowest probability.[24] The not-for-profit sector especially eschews open source in the data, as represented by the low coefficients on education and public administration, which means any organization in these

[23] These numbers are derived by comparing the average predicted probability of an organization with a feature being on open source software versus those without the feature. For example, 46.35% of large organizations are predicted to use open source software, while 54.14% of smaller organizations are predicted to use that open source software.

[24] One might wonder if accommodation and restaurant websites are more likely to be hosted on services like Squarespace and Wix. If so, could the hosting company be making the decision of which server software to use rather than the organizations themselves. For approximately 8,000 organizations in our sample, we analyze the server headers, cookies, and scripts on the organizations' websites in 2016 to ascertain if there are indications that the organization used WordPress, Wix, or Squarespace hosting. We find that 18.2% of the restaurant and accommodation websites versus 16.02% of information industry organizations use one of these hosted website services.

industries are 13.82% and 18.75% less likely to adopt open source software than the average organization.[25]

To illustrate substantial potential for mismeasurement in productivity analysis, consider the estimates altogether. A healthcare organization headquartered in the West would be 13.58% more likely to use open source than a finance organization headquartered in the Midwest. Since the dataset covers the entire range of industries and geographies of medium to large organizations operating in the United States, these relatively popular categories represent 1.21% and 2.84% of the organizations in our sample. In other words, the mismeasurement does not arise solely from one or two subsamples, but correlates with some of the most basic features of an organization, such as its location, industry, and size. Potential for mismeasurement, therefore, is prevalent throughout the economy.

In light of our motivation, the differences in results between the large and medium-sized organizations in Table 3 and Figure 5 are striking. Almost from the beginning, organizations with fewer than 250 employees are substantially more frequent users of open source web servers than large organizations. That is evidence that standard productivity mismeasurement makes its greatest errors with the smaller and younger organizations.

As organizations install software updates that advance the capabilities of the software they are already using, this dynamic factor further exacerbates mismeasurement due to omission. In Figure 6, we show that the dispersion in technology age of organizations' choices of server software increases over time. As shown in the figure, the interquartile range in the distance from the technical frontier grows over time. The panel on the left shows this for Apache, the panel on the right shows this for IIS, and panel on the bottom shows this for Nginx. A key difference between the figures arises from updating by Apache and Nginx in contrast with that by Microsoft. Apache and Nginx users initiate the update, while many IIS users were automatically updated (within a given vintage of 1.0, 2.0, and so on). Until recently, Microsoft IIS users tended to utilize the same vintage of IIS. In contrast, Apache and Nginx users rely on a wide range of different vintages. In the single year of 2010, some organizations used Apache vintage that were

[25] An organization at the mean of all attributes would be predicted to use open source with a probability of 52.75%. Organizations in the education industry are predicted to use open source at a rate of 45.46% and public administration organizations use open source with a predicted rate of 42.86%. That finding contrasts with Robins et al. (2018), which estimates that there is over a billion dollars of open source software in the Federal government.

at the technical frontier, while 50% of organizations used Apache vintages that were more than four years old.

Several insights emerge from these plots. Foremost, they demonstrate that simply assuming that the majority of server software reflects similar functionality is an incorrect assumption. In fact, updates of software like Apache create different functionality, which creates different amounts of value for different users, depending on their tendency towards installing these updates. Second, the variance in vintages, especially for Apache users, can be substantial because some organizations regularly update to stay close to the frontier while others do not take much action with much frequency, and some take none after their first installation. If server software were treated as any other capital model, such behavior would motivate the need for quality adjustments (linked to vintage), such as those we presented in Figure 2 and Figure 3. Third, and relatedly, the results in Figure 6 show that simply adding servers to produce a capital aggregate would lead to enormous errors. In fact, Figure 6 reinforces the need to properly adjust for quality and capacity during assembly of a capital aggregate. Finally, the Figure 6 also suggests that a one-size-method does not fit all servers. Different servers require adjustment procedures to reflect the practices of users.

In Table 4, we show the heterogeneity in organizations' updating behavior by organization characteristics. The table displays the estimated coefficients from a regression of the distance from the technological frontier of the web server software used by organizations on covariates representing the characteristics of those organizations.

The estimates show that geography is predictive of distance to the frontier, with many of the organizations that are closer to the frontier being located in traditional technology hubs. Organizations headquartered in Massachusetts using open source server are 20.19% and 35.63% closer to the technological frontier than organization headquartered in Mississippi and Wyoming respectively. Also, organizations less than five years old use open source server software that is 9.39% closer to the frontier than organizations ten years and older, while medium-sized organizations with fewer than 250 employees use server software 8.26% closer to the frontier than their larger counterparts. Productivity analysis is already challenging among startups and small organizations, so this finding suggests those concerns could be extended to IT-intensive medium-sized organizations as well.

There is a substantial heterogeneity across industries in the distance to the technological frontier of the server software used by organizations, as well as a general persistence in ranking. Surprisingly, organizations in the information industry tend to use technology far from the technological frontier. In 2018, the average age of the Apache software an information organization used was 52 months older than the age of the latest Apache release—in other words, the organization was using software that had not been updated for over four years. In the same year, retail merchants used server software 5.1% closer to the frontier, while food and accommodation organizations used software 13.12% closer. Some of the variation in distance to the frontier is due to selection of server software vendor: A portion of frontier organizations in 2018 had switched to Nginx or alternative server software vendors.[26] A complete explanation for the information sector's distance from the technological frontier remains for future research.

What else produces differences in outcomes? One possible explanation is that organizations install updates to the Apache server at different rates and use software of different vintages on average. Moreover, organization churn among vintages also could affect the technology age of organizations' server choices. We observe such behavior in a few different ways in the next figures and tables.

In Figure 7, we see a pattern frequently seen among capital goods, which shows organizations that (later) exit tend to use older vintages of the server software than organizations that (later) continue operating.[27] This difference reflects behavior commonly observed in long-lived capital goods, where organizations cease investing in improving a capital good in advance of retiring it. Again, we view this as one more piece of compelling evidence that organizations treat open source software the same way they treat any other capital good. The only difference between the two is the invisibility of open source.

As with any situation in which users have options among different suppliers of a capital good, switching between suppliers and upgrades into new vintages provides additional insight about the services they receive. To reiterate, this has not been visible until this study. The results

[26] Some organizations utilize alternative servers, such as Lighttpd, Apache Tomcat, and others. In addition, some leading companies, such as Google and Twitter, utilize customized server software.

[27] Exit is defined for an organization as the year in which either Orbis or Compustat say that the organization was delisted or became defunct. If such a year is not listed in those databases, we use the last year in which Internet Archive had collected data for that organization's homepage.

in Figure 8 demonstrate the prevalence of switching among organizations that used Apache and IIS exclusively during 2005. In the left plot, we see that approximately 20% of organizations that used an open source server switched to the proprietary software IIS during the next ten years. This would create new expenses in standard accounting measures. Interestingly, another 20% switched *between* open source servers, from Apache to Nginx, an improvement that would have been unobserved in its entirety.

In contrast, among the organizations that used IIS in 2005, almost 60% switched to using open source servers Apache and Nginx by 2015. For these organizations, software capital assets that had previously appeared on the books would have disappeared when the organization switched to an open source alternative. That is consistent with the increasing prevalence of intangible assets due to veiled inputs.

We use descriptive logistic regressions to study how organizations' decision to switch between server vendors correlate with organization characteristics. The regression coefficient estimates are included in Table 5. We find that for organizations using Apache, they were more likely to switch from Apache to IIS if they are large, in the South or Midwest, and in public administration, utilities, and finance industries. More precisely, large organizations are 21.01% more likely to switch to IIS than smaller organizations.[28] Organizations in the South are 2.98% more likely than Northeastern organizations to make that switch.[29] Finance organizations are 21.98% more likely to switch to IIS than information sector organizations.[30] In contrast, for organizations using IIS, they are more likely to switch from IIS to Apache if they have fewer than 250 employees (18.27%), or if they are in the West versus Northeast (7.44%), or in the accommodation and food industries versus the information sector (15.11%).

In Figure 9, we show how the adoption of Nginx correlates with organization characteristics. Remarkably, we see a general tendency across all organizations to switch to Nginx, with less observable variance correlated with location, size, age, or industry. That does

[28] Organizations with more than 250 employees switch from Apache to IIS at a predicted rate of 43.57%, while smaller organizations switch at a predicted rate of 52.73%.

[29] Apache using organizations in the South switch to IIS at a rate of 46.89%, while those in the Northeast make the switch at 45.33% rate.

[30] Information sector organizations are 43.8% to make the switch from Apache to IIS, while Finance organizations are predicted to make that switch with a probability of 53.47%.

not imply zero variance, however. Organizations in the information industry switched sooner than others, and organizations in finance were slowest of all. During 2015, 13.82% of information industry organizations had adopted Nginx; it would take two more years before Nginx achieve the same adoption rate among finance organizations. That is evidence that this open source software produced useful advances that virtually every organization in the economy could appreciate, and notably, none of it would have been measured. That is evidence of a widespread omission, which, with the benefit of this study's data, can be attributed to the diffusion of one server, which was the "hot product" of the time.

Overall, the foregoing suggests unpriced software and software updates might mislead analyses of the sources of organization productivity. Going back to Figure 1, the patterns reveal that extrapolation based on observable proprietary software usage would lead to proportional mismeasurement between 2001 and 2010. That would lead to overestimates of the contribution of visible software. Since 2010, however, the introduction of new server software makes estimates of software capital based on proprietary software usage negatively correlated with the total actual quality-adjusted software. During more recent years, the correlation of value between that captured by proprietary software usage and that captured by open source usage is negative. If this is not accounted for, the source of productivity gains could be entirely omitted. It is even possible that the gains from visible software could be underestimated.

We bring one final piece of evidence that suggests, again, that a firms' use of web servers, and other unmeasured technology usage proxies for productive assets, which meaningfully explain variations in firms' value-added levels. We do it through a replication of Nagle's (2019) study of the association between open source and firm productivity. Nagle fits firms' value added to a production function with IT capital, non-IT capital, and labor as inputs, following a long line of research into the association of productivity with the use of IT (Hitt and Tambe, 2014). The goal of Nagle's study is to replicate the approach of prior literature, and add evidence of an association between the use of different measures of open source and firm productivity.

Table 6 shows the estimated coefficients of the multifactor productivity analysis for the 1,577 firms that overlap in our sample and Nagle's (2019). We follow Nagle's notation and variable construction for IT capital, non-IT capital, non-IT labor, and unpriced open source

software. Note in particular that, according to this definition, the variable *Non-Pecuniary OSS* includes only firms' use of Linux as open source software. In summary, we replicate Nagle's specification, and then add new exogenous variables, which is the firm's use of server software.

In this exercise, we examine the correlation between input factors and value-added output of firms. If, after conditioning on all other categories of input factors, open source software inputs are significantly correlated with the value-added of a firm then software inputs are correlated with productivity increasing inputs and decisions. This analysis is not aimed at estimating a causal effect of software investments on productivity. Rather, we seek to illustrate that decisions about software inputs, which are often hidden from traditional accounting methods used for studying production, are associated with higher productivity.[31]

Column (1) show the results of a regression of the logarithmically transported value added of a firm on the capital, labor, and open source software (Linux) used by the firm. Column (2) adds interactions between *Non-Pecuniary OSS* and *ITIntensity*, defined as the value of the computer hardware owned by a firm divided by its sales. Column (3) and (4) interact the usage of open source software (Linux) with indicator variables for if the firm operates in industries considered IT producing and finance, insurance, and real estate respectively. Columns (5) through (7) include two additional variables *Server QA Stock* (*OSS*) and *Server QA Stock* (*Prop.*), representing the quality-adjusted server software used by the firm in a year. When a firm used only open source server software in a year, *Server QA Stock* (*Prop.*) is zero, while users of proprietary server software have a *Server QA Stock* (*OSS*) of a non-zero value. Note that the *QA Stock* variable not only captures use of server software, but in the regression could also capture other omitted variables, such as the unobserved complementary technology use that server use proxies for. Columns (8) through (10) include firm fixed effects in order to absorb firm level unobserved heterogeneity.

The estimated coefficients show that the server stock used by firms is a significant omitted variable from traditional productivity analysis. In our preferred specifications, Columns

[31] Firms' decisions regarding which software inputs to use are endogenous. Firms that are highly productive for unobservable reasons are likely to also select the software that enables them to be most productive. The choice of which software to use may be correlated with many other unobservable input and production decisions by firms that ultimately influence productivity. Proprietary software firms may offer volume discounts for particularly productive firms. Given these potential confounders, the above regression and exercise should only be seen as highlighting that software is a factor input that is correlated with productivity and omitted from previous studies of firm productivity.

(8)-(10) including firm fixed effects, the estimated coefficients on the *Server QA Stock* are small or even slightly negative. The interaction between the open source server stock and *ITIntensity*, however, is large and positive. Similarly, the interaction of server stock and the firm being an IT producer is positive for both open source and proprietary server software. Producers of IT that use open source software are associated with a 13.7% higher value added for each quality-adjusted unit of server capital, while proprietary software IT producers are associated with a 10.7% higher value added on average. This is in contrast with firms operating in the fields of finance, insurance, and real estate, where using more recent server software is associated with no change in value-added levels. We speculate that the large and significant estimates likely not only capture effects due to use of servers but also many complementary technologies that server use proxies for. Consistent with Nagle's conclusion, these estimates suggest measuring use of open source software enables researchers to measure an important mechanism related to productivity that had previously been hidden.

7. Conclusion

Despite the web's essential role in the digital economy, the operations and infrastructure that support the commercial web have remained largely veiled to economic analysis and accounting. In this study, we focus on two problems in standard productivity analysis that are associated with mismeasurement of value—omission and misattribution. Both mismeasurements are created by unpriced software and software updates.

We characterize mismeasurement in economic accounting and productivity estimates by closely examining web server software, an asset that plays a critical role in the digital economic activity. To enable our analysis, we compile the largest dataset ever on business web server use in the United States, with disaggregated information on the usage of web server software and the installation of software updates by over 200,000 medium to large organizations in the United States between 2001 and 2018. Our sample achieves good representation of organizations by geography and industry.

We find that the omission of open source web servers, such as Apache and Nginx, produces a large bias in measuring the economic value of server software, approximately $66 million in 2000 with an increase to between $125 and $315 million by 2018 for our sample. This omission is particularly pronounced among young organizations, smaller organizations, organizations on the West, and organizations in healthcare, lodging and food. We also analyze the dynamic aspects of organizations' choices of server software, such as upgrading to new technologies and switching between different products. We find these dynamic issues further exacerbate mismeasurement.

Mismeasurement is particularly fraught in the most recent decade, because a new server has become widely adopted, and organizations in some industries have substantially converted to open source. Finally, we study how unpriced server software and software updates mislead analyses of the sources of firm productivity. We find that including the data on server software use is a meaningful explanation, both statistically and economically, for variations in firm value-added levels.

Our study contributes to the literature on the measurement of the economic value, as well as to growth attributable to IT, by presenting the most comprehensive analysis of mismeasurement of open source software use, with a sample that is much broader and extends for a longer period of time than previous studies of proprietary versus open source software. We also systematically document a multitude of organization behavior, including choices of open source versus proprietary, upgrades to new technologies, and switching to different products. Moreover, we correlate these behaviors to a variety of organization characteristics and show there is significant heterogeneity in mismeasurement across organization size, age, geography, and industry and how that measurement changes over time.

Throughout the study, we have aspired to analyze open source web servers as if they were one example among many and argue that, as such, web servers illustrate a broad set of omission and misattribution problems that arise from behavior and aspects common to open source software usage and unpriced software updates. Nonetheless, web servers possess one unique feature, namely, their connection to their academic origins. Unlicensed software and shareware are common practices in universities for diffusing new software into new use. Our

results suggest that the type and scale of underestimates found with web servers could arise for similar reasons with other open source software with academic origins.

We were able to demonstrate unobservable IT can mislead productivity analyses, but nothing in the demonstration of productivity mismeasurement was unique to web servers per se. Much of the analysis could be extended to a wide range of additional open source software. Our paper shows unobserved and unmeasured technology such as server software creates an amount of economic value that cannot be ignored. This begs the questions, what would emerge if all open source software were analyzed? The literature on growth attributable to IT should further address issues due to the mismeasurement of the economic value created by open source software use.

This study highlights a number of open questions. We described variance in organization use of servers, but did not fully analyze why organizations made the choices they did, nor why they chose to upgrade when they did. Future research should examine the decision of organizations regarding the timing and direction of investment in their software capital. This includes the choice of when to update software as well as when to switch to a different software vendor. Such a study requires even more data and modeling of switching costs. Our evidence suggests such analysis also will yield insight into veiled value creation that standard productivity analysis mismeasures.

Bibliography

Brynjolfsson, Erik. 1993. "The Productivity Paradox of Information Technology." *Communications of the ACM* 36 (12): 66–77. https://doi.org/10.1145/163298.163309.

Byrne, David, and Carol Corrado. 2019. "Accounting for Innovations in Consumer Digital Services: IT Still Matters." In *Measuring and Accounting for Innovation in the 21st Century*. Vol. 2019. NBER Book Series Studies in Income and Wealth. http://www.federalreserve.gov/econres/feds/files/2019049pap.pdf.

Byrne, David, Stephen D. Oliner, and Daniel E. Sichel. 2013. "Is the Information Technology Revolution Over?" *International Productivity Monitor* 25: 20–36.

Calvino, Flavio, Chiara Criscuolo, Luca Marcolin, and Mariagrazia Squicciarini. 2018. "A Taxonomy of Digital Intensive Sectors," June. https://doi.org/10.1787/f404736a-en.

Corrado, Carol A. 2011. "Communication Capital, Metcalfe's Law, and US Productivity Growth." *Metcalfe's Law, and US Productivity Growth (March 1, 2011)*.

Gordon, Robert J, and Hassan Sayed. 2020. "Transatlantic Technologies: The Role of ICT in the Evolution of U.S. and European Productivity Growth." *National Bureau of Economic Research Working Paper Series* No. 27425. https://doi.org/10.3386/w27425.

Greenstein, Shane, and Frank Nagle. 2014. "Digital Dark Matter and the Economic Contribution of Apache." *Research Policy* 43: 623--631.

Griliches, Zvi. 2000. *R&D, Education, and Productivity: A Retrospective*. Harvard University Press.

Hall, Robert E., and Dale W. Jorgenson. 1967. "Tax Policy and Investment Behavior." *The American Economic Review* 57 (3): 391–414.

Jorgenson, Dale W. 1963. "Capital Theory and Investment Behavior." *The American Economic Review* 53 (2): 247–59.

Jorgenson, Dale W, Mun S Ho, and Jon D Samuels. 2013. "Economic Growth in the Information Age: A Prototype Industry-Level Production Account for the United States, 1947-2010." *Manuscript, Harvard University*.

Keller, Sallie, Gizem Korkmaz, Carol Robbins, and Stephanie Shipp. 2018. "Opportunities to Observe and Measure Intangible Inputs to Innovation: Definitions, Operationalization, and Examples." *Proceedings of the National Academy of Sciences* 115 (50): 12638–45. https://doi.org/10.1073/pnas.1800467115.

Kim, Do Yoon. 2020. "Product Market Performance and Openness: The Moderating Role of Customer Heterogeneity." Draft.

Korkmaz, Gizem, Claire Kelling, Carol Robbins, and Sallie Keller. 2019. "Modeling the Impact of Python and R Packages Using Dependency and Contributor Networks." *Social Network Analysis and Mining* 10 (1): 7. https://doi.org/10.1007/s13278-019-0619-1.

Lakhani, Karim R, and Eric von Hippel. 2003. “How Open Source Software Works: ‘Free’ User-to-User Assistance.” *Research Policy* 32 (6): 923–43. https://doi.org/10.1016/S0048-7333(02)00095-1.

Lerner, Josh, and Mark Schankerman. 2010. “The Comingled Code: Open Source and Economic Development.” *MIT Press Books*.

Nagle, Frank. 2019. “Open Source Software and Firm Productivity.” *Management Science* 65 (3): 1191–1215. https://doi.org/10.1287/mnsc.2017.2977.

Nordhaus, William D. 2006. “Principles of National Accounting for Nonmarket Accounts.” In *A New Architecture for the US National Accounts*, 143–60. University of Chicago Press.

O’Mahony, Siobhán. 2003. “Guarding the Commons: How Community Managed Software Projects Protect Their Work.” *Research Policy* 32 (7): 1179–98. https://doi.org/10.1016/S0048-7333(03)00048-9.

Robbins, Carol A, Gizem Korkmaz, José Bayoán Santiago Calderón, Daniel Chen, Claire Kelling, Stephanie Shipp, and Sallie Keller. 2018. “Open Source Software as Intangible Capital: Measuring the Cost and Impact of Free Digital Tools.” Draft.

Stiroh, Kevin J. 2002. “Information Technology and the U.S. Productivity Revival: What Do the Industry Data Say?” *American Economic Review* 92 (5): 1559–76. https://doi.org/10.1257/000282802762024638.

Syverson, Chad. 2011. “What Determines Productivity?” *Journal of Economic Literature* 49 (2): 326–65. https://doi.org/10.1257/jel.49.2.326.

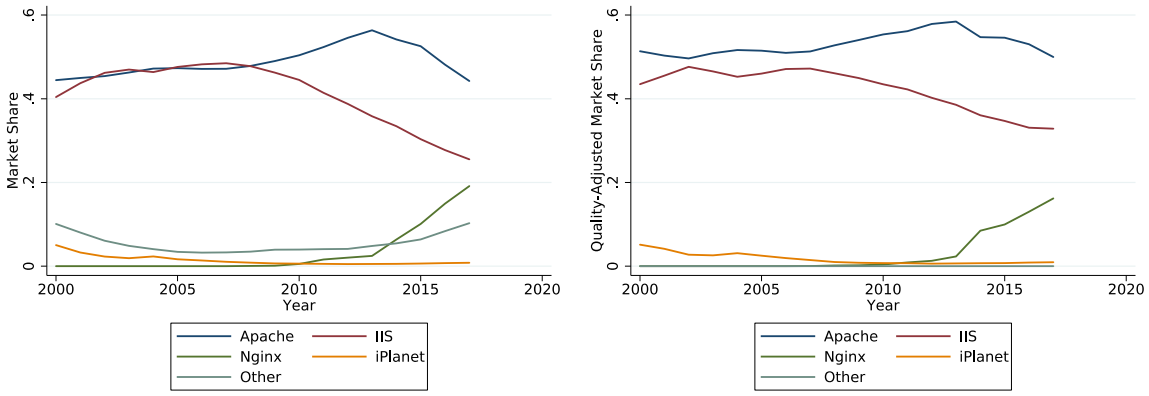
Tambe, Prasanna, and Lorin M Hitt. 2014. “Job Hopping, Information Technology Spillovers, and Productivity Growth.” *Management Science* 60 (2): 338–55. https://doi.org/10.1287/mnsc.2013.1764.

Tambe, Prasanna, Lorin Hitt, Daniel Rock and Erik Brynjolfsson. 2020. "Digital Capital and Superstar Firms." *National Bureau of Economic Research Working Paper Series* No. 28285. https://dio.org/10.3386/w28285.

West, Joel. 2003. “How Open Is Open Enough?: Melding Proprietary and Open Source Platform Strategies.” *Research Policy* 32 (7): 1259–85. https://doi.org/10.1016/S0048-7333(03)00052-0.

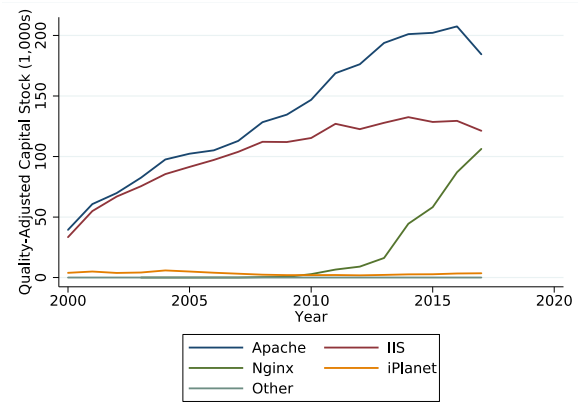
Tables and Figures

Figure 1 Market Shares Over Time



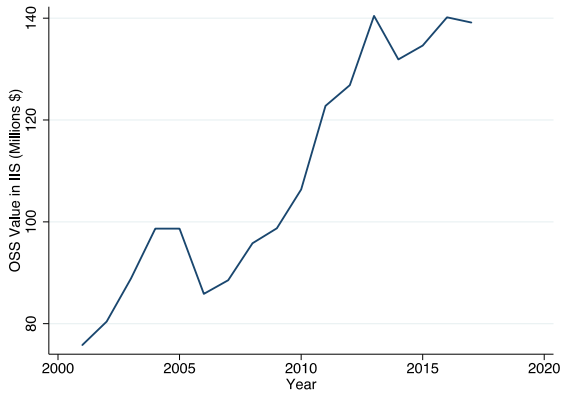
Note: The above plot displays the share of organizations in our sample using a server vendor over time. The vertical axis is the share of organizations. The horizontal axis is the year. As our panel data is at the monthly level, the year observations above are weighted by the number of months within a year that organizations utilized a server vendor.

Figure 2 Quality Adjusted Capital Stock



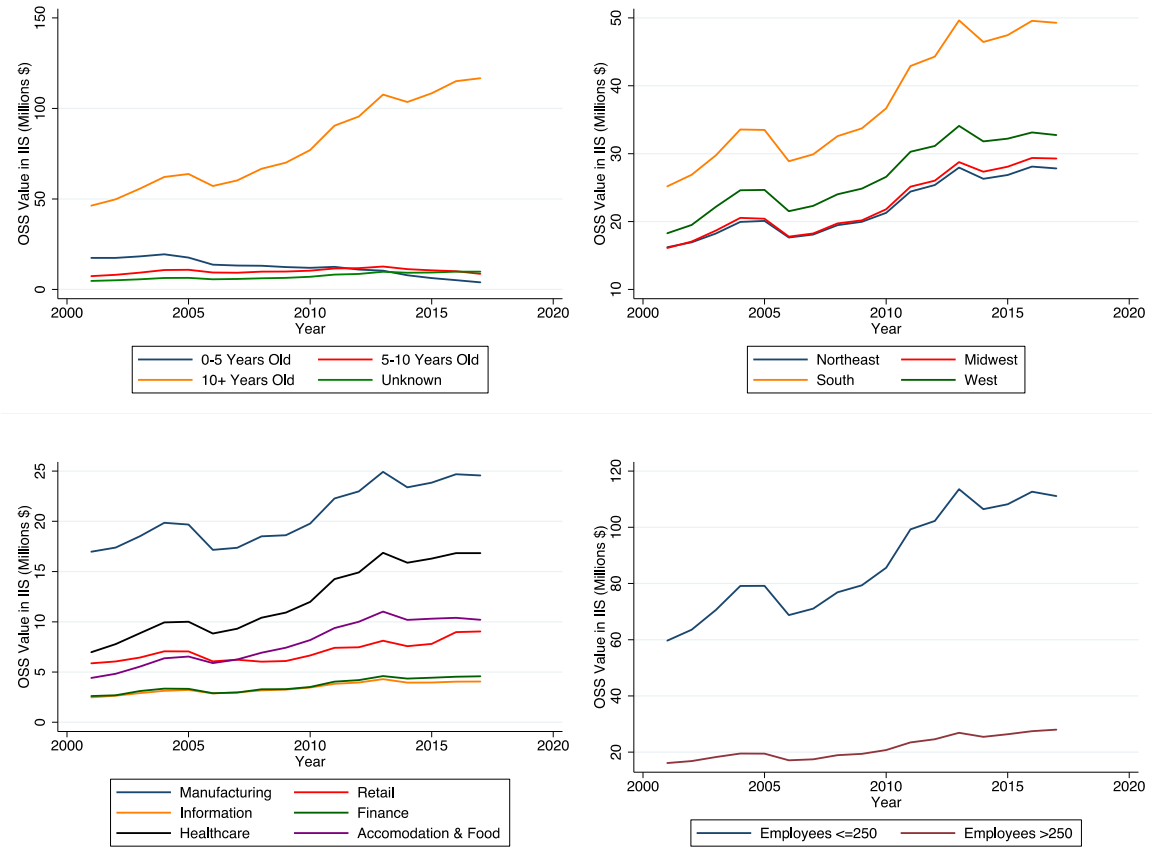
Note: The above plot displays the total quality adjusted capital stock ($QACAP_t$) of each server for our sample. The vertical axis is the sum of units of quality based on the inverse of the CPI for the server vintages and scaled by the number of servers using that vintage in that year. The horizontal axis is the year. As our panel data is at the monthly level, the year observations above are weighted by the number of months within a year that organizations utilized a server vendor.

Figure 3 Omitted Value of Open Source Servers



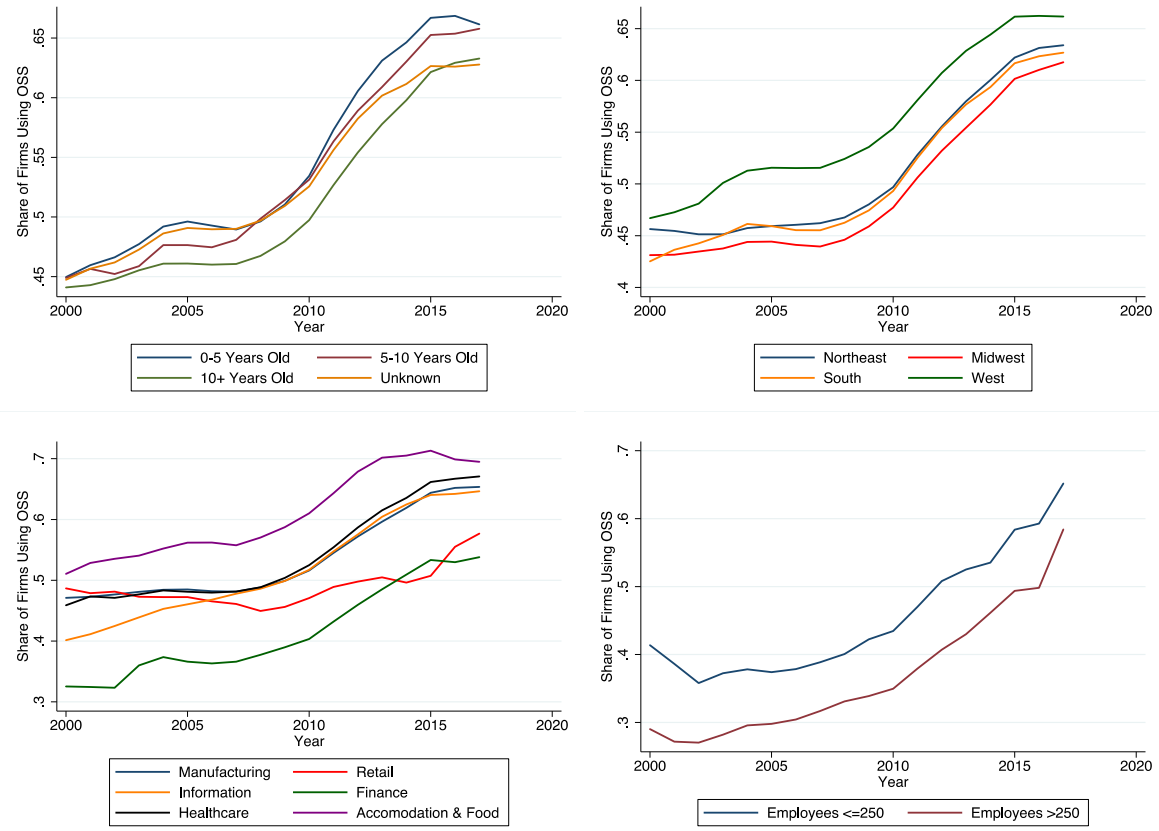
Note: The above plot displays the estimated value of open source servers being used in each year for our sample. For each observation in the dataset in which a organization utilizes the Apache or Nginx web server, we find the price of the Windows server 10-user package containing the most widely used vintage of the proprietary Microsoft IIS server in that year. The vertical axis displays the total of these shadow values multiplied by ten for the ten-user license. Prices are deflated to 2012 dollars. Observations are weighted for representativeness by state and NAICS with data from the Census SUSB.

Figure 4 Omitted Value of Open Source Servers, Breakdowns by Organization Characteristics



Note: The above plot displays the estimated value of open source servers being used in each year, broken down by organization age, geography, industry and size. For each observation in the dataset in which a organization utilizes the Apache or Nginx web server, we find the price of the Windows server 10-user package containing the most widely used vintage of the proprietary Microsoft IIS server in that year. The vertical axis displays the total of these shadow values multiplied by ten for the ten-user license. Prices are deflated to 2012 dollars. Observations are weighted for representativeness by state and NAICS with data from the Census SUSB. The top left plot shows this with organizations binned by the age of the organization. Organization ages are computed as the difference between the year the observation is made and the year of incorporation of the organization. The vertical axis is the OSS value for organizations within an age bin. The horizontal axis is the year of observation. The top right plot shows this with organizations binned by their geographic region as defined by Census regions. The bottom left shows this with organizations binned by industry, defined by two-digit NAICS code. Only six NAICS categories are shown in this figure. The bottom right shows this with organizations binned by size. As our panel data is at the monthly level, the year observations above are weighted by the number of months within a year that organizations utilized a server vendor.

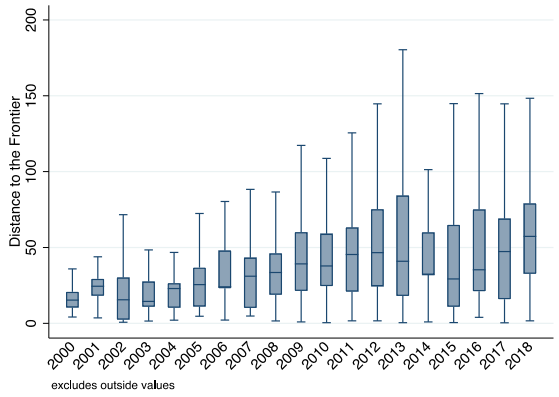
Figure 5 Open Source Server Usage, Breakdowns by Organization Characteristics



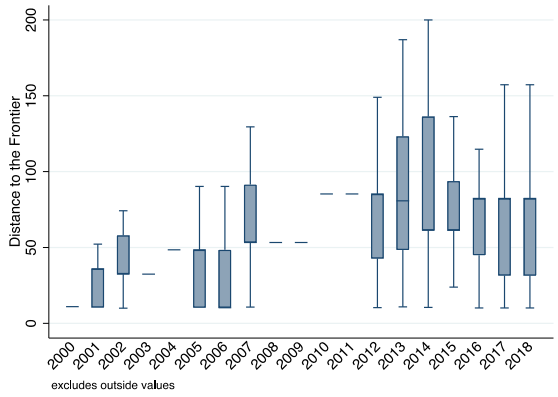
Note: The above plot displays the share of organizations in our sample using an open source server over time, broken down by organization age, geography, industry and size. The top left plot shows this with organizations binned by the age of the organization. Organization ages are computed as the difference between the year the observation is made and the year of incorporation of the organization. The vertical axis is the share of organizations using OSS within an age bin. The horizontal axis is the year of observation. The top right plot shows this with organizations binned by their geographic region as defined by Census regions. The bottom left shows this with organizations binned by industry, defined by two-digit NAICS code. The bottom right shows this with organizations binned by size. As our panel data is at the monthly level, the year observations above are weighted by the number of months within a year that organizations utilized a server vendor.

Figure 6 Distance to the Tech Frontier by Server Vendor

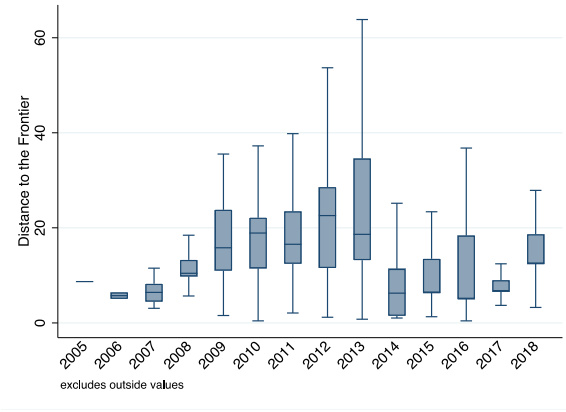
(a) Apache



(b) IIS

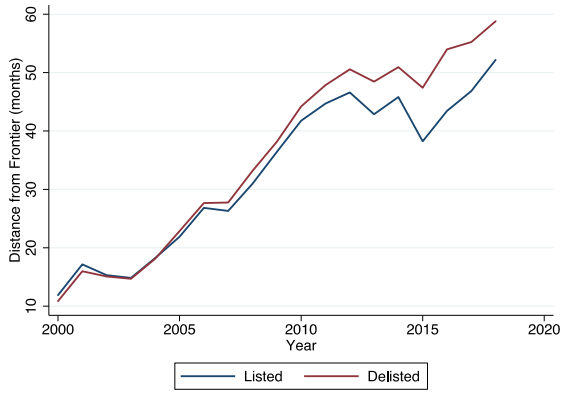


(c) Nginx



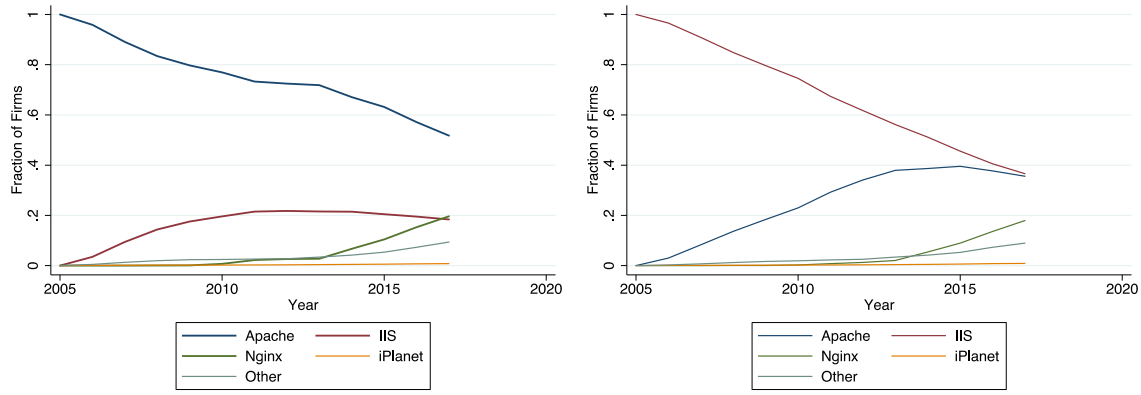
Note: The above plot shows the interquartile range of the distance of servers to the technological frontier (DTF_t). The definition of DTF_t is the number of months since the server vintage used by a organization was released minus the number of months since the latest vintage of that server vendor's software was released. We drop situations in which a organization used a beta server vintage prior to the general release of that vintage. The year observations are weighted by the number of months of observations within that year. The top left are Apache servers, the top right are IIS servers, and the bottom are Nginx servers.

Figure 7 Distance to the Tech Frontier for Active and Delisted Organizations



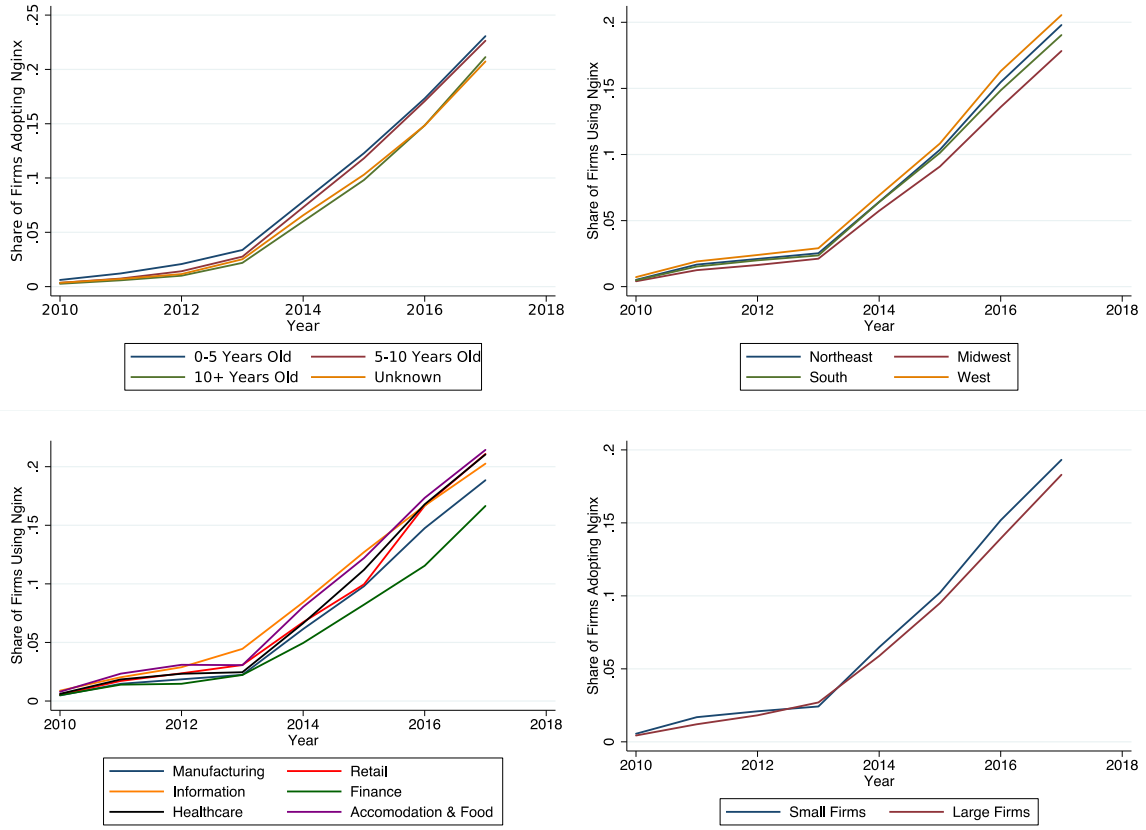
Note: The above plot displays the average distance to the tech frontier (DTF_t) among active and delisted organizations using Apache server software. The definition of DTF_t is the number of months since the server vintage used by a organization was released minus the number of months since the latest vintage of that server vendor's software was released. We define an organization as delisted in a year if either Orbis or Compustat say that the organization was delisted or became defunct in that year. If such a year is not listed in those databases, we use the last year in which IA had collected data for that organization's homepage.

Figure 8 Switching Server Vendors



Note: The above plots show the fraction of organizations using server software from each vendor over time. On the left we restrict to the subsample of organizations that used Apache exclusively during 2005. On the right we restrict to the subsample of organizations that used IIS exclusively during 2005.

Figure 9 Adoption of Nginx



Note: The above plot shows the share of organizations using the Nginx server, broken down by organization age, geography, industry and size. The top left plot shows this with organizations binned by the age of the organization. Organization ages are computed as the difference between the year the observation is made and the year of incorporation of the organization. The vertical axis is the share of organizations using Nginx within an age bin. The horizontal axis is the year of observation. The top right plot shows this with organizations binned by their geographic region as defined by Census regions. The bottom left shows this with organizations binned by industry, defined by two-digit NAICS code. The bottom right shows this with organizations binned by size. As our panel data is at the monthly level, the year observations above are weighted by the number of months within a year that organizations utilized a server vendor.

Table 1 Summary of Sources of Data

Data	Description	Frequency	Source
Organization characteristics	230,611 organizations in the US with at least 50 employees at some point between 2000 and 2018; organizations are identified by their homepages; variables include location, NAICS code, number of employees and year of incorporation	Yearly panel from 2000 to 2018	For public firms, CompuStat database; when CompuStat data is not available, Bureau van Dyke's Orbis database
Organizations' server software usage	Vendor and vintage of server software of 213,956 homepages of the above US organizations	Monthly panel from 2000 to 2018	The Internet Archive[32]
CPI deflator	Consumer Price Index for "Computer Software and Accessories," on a December 1997=100 base	Monthly series from 2000 to 2018	Bureau of Labor Statistics, USA
Windows IIS price series	Price of the Windows server 10-user package containing the most widely used vintage of the proprietary Microsoft IIS server in that year	Yearly series from 2000 to 2018	Authors' compilation (see Appendix E)
State and industry weights	Weights used for weighting observations by state and NAICS for representativeness	Yearly from 2000 to 2018	Statistics of U.S. Businesses (SUSB)

Note: The sources indicate the sources of raw data. We have extensively cleaned and reshaped the data. Please see Section 4 Data, Section 5 Measurement, and the Appendices for details of data construction.

[32] https://archive.org/

Table 2 Summary of Results

Question	Methods	Results	Figures and Tables
How large is the omission of economic value created by open source web server software?	Descriptive plots contrasting usage of open source vs. proprietary server software over time; Estimating the omitted value, using the prices of Windows IIS vintages as the proxy for the shadow value of open source servers	Comparable usage of open source vs. proprietary server software between 2000 and 2010, increasing usage after 2010, indicating increasing omitted value of open source over time; Total value of omission in 2018 between $4.5 billion and $11.3 billion	Figures 1, 2, 3
Does the omitted value vary by organizations' age, geography, industry and size?	Descriptive plots showing heterogeneity of open source usage and omitted value by organization characteristics; Logistic regression predicting open source usage using organization characteristics	Younger organizations, West Coast organizations, organizations in healthcare and accommodations, smaller organizations use open source server software significantly more; Omission of economic value is heterogeneous across organizations characteristics	Figures 4, 5; Table 3
How do dynamic factors affect mismeasurement?	Descriptive plots showing heterogeneity in organizations' behavior 1) updating server vintage, 2) switching server vendors, 3) adopting the new server product Nginx; Logistic regression predicting decisions to update and switch using organization characteristics	Large dispersion in technology age of organizations' choices of servers and this dispersion increases over time, indicating heterogeneity in upgrading behavior; Significant switches between server vendors, software capital assets could appear or disappear on the books; heterogeneity in switching behavior;	Figures 6, 7, 8, 9; Tables 4, 5

		Extensive adoption of Nginx across all organization characteristics, growing mismeasurement due to the entry of the new product	
Does unpriced server software mislead analyses of sources of firm productivity?	Estimating multifactor production function using server software stock as an additional variable (specifications replicate those in Nagle (2019))	Firms' use of server software, and the unmeasured technology usage that proxies for, meaningfully explain variations in firms' value added level	Table 6

Table 3 Logit Predicting OSS Usage

	1(OSS)
Employees, >250	-0.082***
	(0.002)
Young	0.021***
	(0.002)
Geographic Region	
-- Midwest	-0.023***
	(0.003)
-- South	-0.011***
	(0.002)
-- West	0.036***
	(0.003)
Industry	
-- Mining	-0.027*
	(0.015)
-- Utilities	-0.134***
	(0.015)
-- Construction	0.012
	(0.010)
-- Manufacturing	-0.031***
	(0.010)
-- Wholesale	-0.065***
	(0.011)
-- Retail	-0.083***
	(0.011)
-- Transportation	-0.049***
	(0.011)
-- Information	-0.035***
	(0.011)
-- Finance	-0.144***
	(0.011)
-- Real Estate	-0.070***
	(0.011)

-- *Profession*	-0.069***
	(0.011)
-- *Management*	-0.139***
	(0.016)
-- *Admin Support*	-0.028***
	(0.011)
-- *Education*	-0.117***
	(0.011)
-- *Healthcare*	-0.015
	(0.010)
-- *Arts and Entertainment*	-0.090***
	(0.011)
-- *Accommodation and Food*	0.047***
	(0.011)
-- *Other*	-0.047***
	(0.011)
-- *Public Admin*	-0.143***
	(0.012)
-- *Unavailable*	-0.032***
	(0.011)
Year Fixed Effect?	Y
Number of Organization-Year Observations	2,947,067
Psuedo-R^2	.027

Note: The above table shows the results of a logistic regression. The dependent variable is whether or not an organization used open source software as its primary server software in a year. Marginal effects are shown above. Agriculture is the omitted industry and Northeast is the omitted region. Standard errors are clustered at the organization level.

Table 4 Distance to Frontier of Software Used by Organization

	Distance to Frontier		
	Apache	IIS	Nginx
Employees, >250	0.666***	-0.883***	9.458***
	(0.177)	(0.110)	(0.577)
Young	-1.963***	-1.588***	-1.910
	(0.142)	(0.100)	(0.601)
Geographic Region			
-- *Midwest*	-0.543	-1.341***	0.227
	(0.231)	(0.148)	(0.723)
-- *South*	-0.560***	-1.132***	-1.612**
	(0.210)	(0.134)	(0.640)
-- *West*	0.269	-1.319***	-1.664**
	(0.225)	(0.153)	(0.684)
Industry			
-- *Mining*	-2.881**	0.013	0.545
	(1.352)	(0.906)	(3.668)
-- *Utilities*	-0.657	-0.334	12.489***
	(1.200)	(0.743)	(4.204)
-- *Construction*	-0.702	1.448**	-3.002
	(0.915)	(0.610)	(2.448)
-- *Manufacturing*	0.394	1.328**	5.454*
	(0.894)	(0.595)	(2.419)
-- *Wholesale*	0.237	0.995	5.199 **
	(0.925)	(0.614)	(2.517)
-- *Retail*	0.685	-2.676***	11.429***
	(0.919)	(0.612)	(2.489)
-- *Transportation*	-0.186	1.598***	-0.323
	(0.964)	(0.640)	(2.604)
-- *Information*	0.534	-0.393	6.838***
	(0.953)	(0.656)	(2.610)
-- *Finance*	3.179***	-0.214	14.772***
	(0.964)	(0.619)	(2.652)
-- *Real Estate*	-1.171	-0.261	10.158***
	(0.983)	(0.651)	(2.687)

-- *Profession*	-1.143	0.012	8.479***
	(0.919)	(0.610)	(2.501)
-- *Management*	2.150	1.253	16.218***
	(1.412)	(0.838)	(4.669)
-- *Admin Support*	-0.225	0.659	1.008
	(0.939)	(0.627)	(2.533)
-- *Education*	-2.982***	-2.246***	9.097***
	(0.920)	(0.609)	(2.555)
-- *Healthcare*	-1.973**	0.349	1.757
	(0.907)	(0.603)	(2.436)
-- *Arts and Entertainment*	-2.672***	-4.540***	8.440***
	(0.961)	(0.628)	(2.709)
-- *Accommodation and Food*	-2.014**	-2.457***	1.595
	(0.918)	(0.618)	(2.479)
-- *Other*	-1.335	-1.681***	7.359***
	(0.959)	(0.636)	(2.646)
-- *Public Admin*	-0.868	-0.802	3.502
	(1.053)	(0.667)	(3.061)
-- *Unavailable*	-1.107	-0.673	1.901
	(0.932)	(0.617)	(2.540)
Year Fixed Effect?	Y	Y	Y
Number of Organization-Year Observations	1,218,514	1,224,484	109,849
R^2	0.252	0.424	0.0525

Note: The above table displays regression estimates. The dependent variable is the distance to the tech frontier (DTF_t). The definition of DTF_t is the number of months since the server vintage used by an organization was released minus the number of months since the latest vintage of that server vendor's software was released. The left column shows this for Apache users. The middle column shows this for IIS users. The right column shows this for Nginx users. The covariates include the geographic region of the organization, the NAICS of the organization, and fixed effects for the year of the observation.

Table 5 Logit Predicting Switching Vendor

	1(Switched Vendor)	1(Switched Apache to IIS)	1(Switched IIS to Apache)
Employees, >250	-0.003***	0.095***	-0.092***
	(0.001)	(0.003)	(0.003)
Young	0.009***	-0.022***	0.024***
	(0.001)	(0.002)	(0.002)
Geographic Region			
-- Midwest	0.000	0.029***	-0.027***
	(0.001)	(0.003)	(0.003)
-- South	0.006***	0.014***	-0.012***
	(0.001)	(0.003)	(0.003)
-- West	0.004***	-0.043***	0.043***
	(0.001)	(0.003)	(0.003)
Industry			
-- Mining	0.004	0.036*	-0.032**
	(0.004)	(0.018)	(0.018)
-- Utilities	-0.013***	0.150***	-0.149***
	(0.003)	(0.017)	(0.017)
-- Construction	0.007***	-0.018	0.019
	(0.003)	(0.012)	(0.012)
-- Manufacturing	0.002	0.042***	-0.042***
	(0.003)	(0.012)	(0.012)
-- Wholesale	0.003	0.081***	-0.081***
	(0.004)	(0.013)	(0.012)
-- Retail	0.026***	0.069***	-0.075***
	(0.003)	(0.013)	(0.012)
-- Transportation	0.000	0.053***	-0.052***
	(0.004)	(0.013)	(0.013)
-- Information	0.003	0.035**	-0.042***
	(0.003)	(0.014)	(0.013)
-- Finance	0.005*	0.135***	-0.142***
	(0.003)	(0.013)	(0.013)
-- Real Estate	0.005	0.089***	-0.087***
	(0.003)	(0.014)	(0.012)

-- *Profession*	0.003	0.086***	-0.087***
	(0.004)	(0.013)	(0.012)
-- *Management*	0.001	0.118***	-0.130***
	(0.004)	(0.019)	(0.019)
-- *Admin Support*	0.005*	0.034***	-0.034***
	(0.003)	(0.013)	(0.013)
-- *Education*	-0.004	0.138***	-0.137***
	(0.003)	(0.013)	(0.013)
-- *Healthcare*	0.007**	0.020*	-0.021
	(0.003)	(0.012)	(0.012)
-- *Arts and Entertainment*	0.009**	0.103***	-0.108***
	(0.003)	(0.014)	(0.013)
-- *Accommodation and Food*	0.013***	-0.046***	0.047***
	(0.003)	(0.013)	(0.012)
-- *Other*	0.006**	0.067***	-0.069***
	(0.003)	(0.014)	(0.013)
-- *Public Admin*	-0.013***	0.177***	-0.176***
	(0.003)	(0.015)	(0.014)
-- *Unavailable*	0.021	0.050	-0.051
	(0.023)	(0.084)	(0.078)
Year Fixed Effect?	Y	Y	Y
Number of Organization-Year Observations	1,980,940	1,784,798	1,787,544
Psuedo-R^2	0.016	0.030	0.025

Note: Logistic regressions. Marginal effects are shown above. Agriculture is the omitted industry and Northeast is the omitted region. Standard errors are clustered at the organization level.

Table 6 Misattribution and Productivity

	(1) VA	(2) VA	(3) VA	(4) VA	(5) VA	(6) VA	(7) VA	(8) VA	(9) VA	(10) VA
IT Capital (IT_{it})	0.054***	0.059***	0.053***	0.054***	0.071***	0.060***	0.060***	0.031***	0.031***	0.029***
	0.008	0.008	0.008	0.008	0.009	0.009	0.009	0.007	0.007	0.007
Non-IT Capital (K_{it})	0.260***	0.258***	0.259***	0.260***	0.258***	0.256***	0.263***	0.059	0.055	0.06
	0.014	0.014	0.014	0.014	0.015	0.015	0.015	0.06	0.058	0.062
Non-IT Labor (L_{it})	0.725***	0.720***	0.724***	0.725***	0.709***	0.716***	0.718***	0.810***	0.804***	0.817***
	0.017	0.017	0.017	0.017	0.019	0.019	0.019	0.047	0.046	0.048
Non-Pecuniary OSS	-0.004***	-0.004***	-0.006***	-0.004***	-0.004***	-0.002**	-0.004***	-0.001	0.001	-0.001
	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Non-Pecuniary OSS x $ITIntensity_{it}$		0.343***			0.086			0.161		
		0.106			0.17			0.135		
Non-Pecuniary OSS x *IT Producer*			0.014***			-0.006*			-0.006**	
			0.002			0.003			0.003	
Non-Pecuniary OSS x *FIRE*				0.004			0.002			0.003
				0.003			0.003			0.003
Servers QA Stock (*OSS*)					-0.066	-0.031	-0.053	-0.047	-0.157***	-0.032
					0.05	0.051	0.05	0.038	0.038	0.038
Servers QA Stock (*Prop.*)					-0.029	-0.014	-0.04	-0.038	-0.185***	-0.044
					0.059	0.058	0.06	0.044	0.043	0.045
Servers QA Stock (*OSS*) x $ITIntensity_{it}$					-0.474			3.037***		
					1.53			1.172		
Servers QA Stock (*Prop.*) x $ITIntensity_{it}$					-9.606***			-3.056		
					3.054			2.162		
Servers QA Stock (OSS) x IT Producer						-0.282***			1.111***	
						0.05			0.148	
Servers QA Stock (*Prop.*) x *IT Producer*						-0.315***			1.240***	
						0.043			0.205	
Servers QA Stock (*OSS*) x *FIRE*							-0.236			0.066
							0.161			0.144

Servers QA Stock (*Prop.*)							-0.298			0.081
x *FIRE*							0.204			0.183
Year Fixed Effect?	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Industry Fixed Effect	NAICS-2	NAICS-2	NAICS-2	NAICS-2	NAICS-2	NAICS-2	NAICS-2	N	N	N
Firm FE	N	N	N	N	N	N	N	Y	Y	Y
Number of firm-year obs.	10,355	10,354	10,355	10,355	8,371	8,371	8,371	8,281	8,281	8,281
Number of firms	1,566	1,566	1,566	1,566	1,380	1,380	1,380	1,290	1,290	1,290
R^2	0.93	0.931	0.931	0.931	0.928	0.933	0.928	0.975	0.976	0.975

Note: We restrict the sample of this exercise to the 1,577 firms that overlap in our study's data and Nagle (2019)'s data. We follow Nagle (2019)'s notation and variable construction for value added (VA_{it}), IT capital (IT_{it}), non-IT capital (K_{it}), non-IT labor (L_{it}), unpriced open source software ($Non\ Pecuniary\ OSS_{it}$), IT intensity ($ITIntensity_{it}$) and IT producer ($ITProducer_{it}$). According to this definition, the variable IT_{it} includes both the value of IT hardware at the firm and three times the value of IT labor at the firm. IT hardware is computed by multiplying the firm's stock of PCs and physical servers by the average price of a PC or server that year. The variable $Non\ Pecuniary\ OSS_{it}$ includes only firms' use of Linux as unpriced open source software. $ITIntensity_{it}$ is constructed by dividing the deflated value of the IT hardware at the firm in a given year by deflated sales in that year. $ITProducer_{it}$ is an indicator variable equal to 1 if the firm operates in an industry considered to be IT-producing. We introduce new variables $FIRE_{it}$, $Server\ QA\ Stock\ (OSS)$, and $Server\ QA\ Stock\ (Prop.)$. $FIRE_{it}$ is an indicator variable equal to 1 if the firm is in Finance, Insurance or Real Estate. $Server\ QA\ Stock\ (OSS)$ and $Server\ QA\ Stock\ (Prop.)$ represent the quality adjusted server software used by the firm in a year. When a firm used open source server software in a year, Server QA Stock (Prop.) is zero, while users of proprietary server software have a Server QA Stock (OSS) of zero. Note that the QA Stock variable not only captures use of server software, but also captures other unobserved complementary technology use that server use is a good proxy for. Standard errors are clustered at the firm level. All regressions include controls listed in Nagle (2019), Table 7, except for columns (8)-(10) which include firm fixed effects rather than industry fixed effects.