Appendix A Data Construction

Appendix A.1 Internet Archive

The Internet Archive provided server headers for the organization homepages in our sample. The Internet Archive scans these sites regularly to archive in the Wayback Machine. Our sample is at the monthly level. In the event that the Internet Archive scanned one website more than once during a single month, we first use the observation that includes server vendor and vintage information. If there are still multiple observations, we use the server vendor/vintage that appeared most frequently.

Appendix A.2 Server Header Parsing

The raw server headers are parsed into server vendors and vintages using regular expressions. We use the regular expressions included in Wappalyzer, an open source browser extension that detects the presence of various technologies used on a website.

We validate our findings regarding which organizations use open source versus proprietary server software in two ways. First, we find a correlation of almost one between us saying that a website uses Microsoft IIS based on server headers and the presence of other indications that the site is using IIS. In particular, sites that are using ASP.NET, a web framework that runs on IIS, are almost always flagged as running IIS based on the server headers. Second, we examine the correlation between open source usage as documented by the Harte-Hanks technology survey and our server header parsing. We find that organizations with more open source usage in Harte-Hanks are also more likely to be using an open source web server according to our server header parsing.

Appendix A.3 Orbis and Compustat

For each organization homepage, we attached one NAICS code, one location, and one organization legal status. We do this by finding all of the Orbis business records associated with the same URL as the homepage. We then take the record with the highest number of employees and capture the NAICS, location, and legal status from that business record.

In addition, we capture the revenue, employees, total assets, and capital expenditures from the Orbis data. For this data, we find the organizations associated with the same URL as the homepage. We then drop any organization records that are unconsolidated records when a consolidated record exists. From the existing records, we aggregate the financial records by summing across those associated records.

Our Orbis data has more limited coverage of organizations prior to 2010. Therefore, we also merge data from CompuStat for public firms. For each homepage in the dataset, we find the associated CompuStat records for that firm within a year. We sum the financials across those records to construct a single financial record per homepage per year.

When Compustat financial data is available, we use that data instead of the Orbis data. When Compustat is not available, as is the case with all private firms, we use the Orbis financial data.

There is a high degree of correlation between these measures when available in both databases:

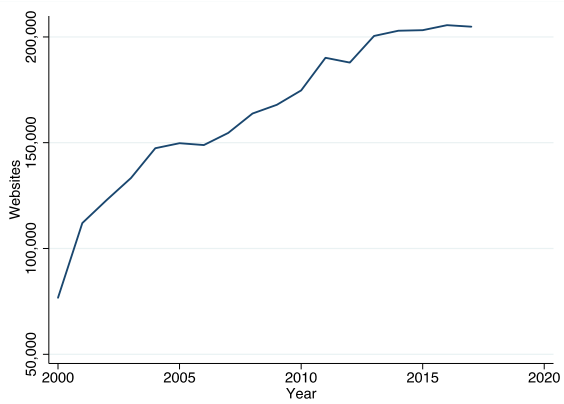
- Total Assets: 0.989
- Employees: 0.973
- Capital Expenditure: 0.908

Appendix B Internet Archive Coverage and Scanning Frequency

The Internet Archive (IA) is not able to scan every website every month. Over time, IA scanned more websites. In Figure B1, we show the number of homepages in the dataset for which IA scanned each year. The number increases every year until recently.

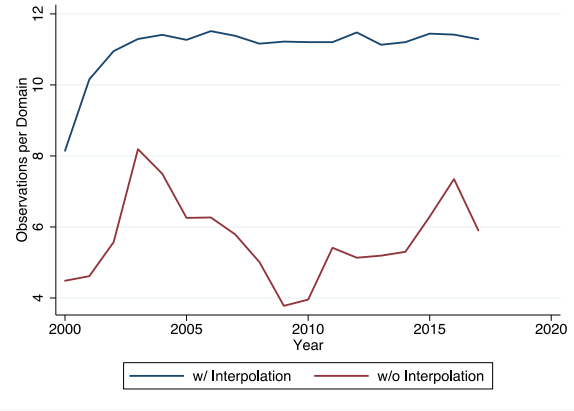
The frequency that homepages were scanned changed over time, however, the impact on our analysis has not. Figure B2 shows the average number of observations per organization in each year. The figure shows this separately for the raw number of sites that IA scanned as well as the number of observations in the data after interpolation. The number of observations per website collected by IA declined between 2004 and 2009 and increased between 2010 and 2017. The number of observations per homepage in the data, however, has remained relatively constant at around 11. The reason for this is because we interpolate observations between observations that have the same server vendor and server vintage number.

Figure 10 Homepages in the Dataset



Note: The above plot shows the number of homepages in the dataset over time.

Figure 11 Observations per Domain in the Dataset



Note: The above plot shows the number of observations per domain in the dataset over time. Lines are shown separately for the full dataset with interpolation and the raw data without interpolation.

Appendix C Representativeness of the Data

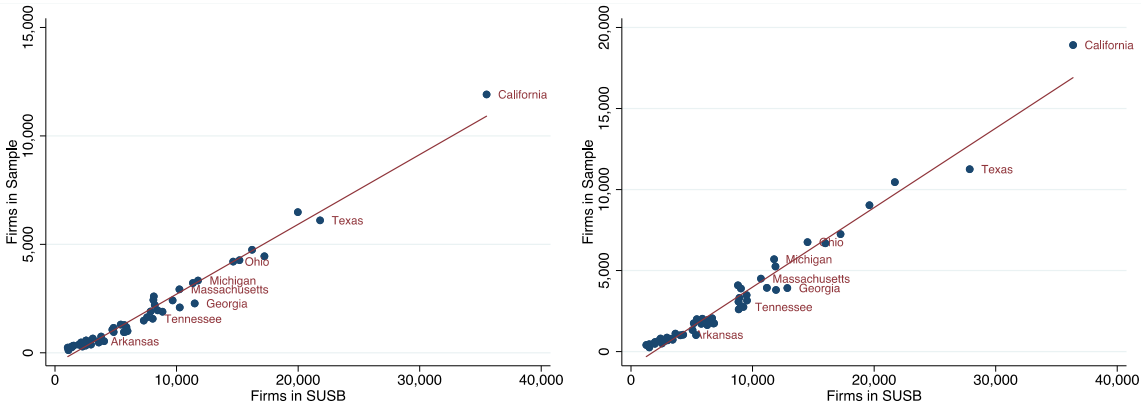
In order to show that the data is representative of US organizations, we compare our sample with data from the Census Bureau's Statistics of US Businesses (SUSB). Our sample includes organizations in the Bureau van Dyke Orbis database that are geographically located in the United States, have a website, and have at least 50 employees at some point between 2000 and 2018. We compare with SUSB's listing of the number of firms with at least 50 employees in the United States.

In Figure 12, we show the number of firms in a state according to the SUSB versus the number of homepages in the study's dataset in 2001. The correlation between these numbers is very high. The number of homepages is a fraction of the number of firms. That could be because multiple organizations have the same homepage (e.g. Walmart and Walmart Vision Centers are sometimes listed as separate firms, but both use the homepage Walmart.com). Another reason could be because some fractions of organizations do not have a website or BvD Orbis does not have the website for that organization.

In Figure 13, we show the number of firms with more than 50 employees by NAICS listed in the SUSB versus the number of homepages in the study's dataset. There is a positive correlation between these numbers, however, there is also more variation. A reason why these

numbers may have more variation is because we associate each website with one organization and one NAICS. All subsidiaries of a corporate group, which share a website, will be grouped under that single NAICS. On the other hand, each subsidiary may actually operate in different NAICS.

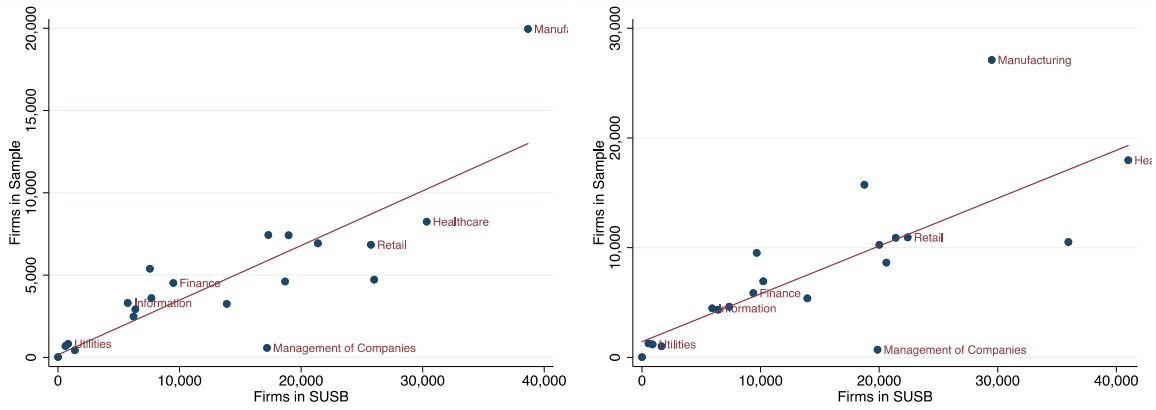
Figure 12 Firms in State versus Number of Homepages in Sample



(a) 2001 (b) 2017

Note: The above plots show the number of firms with greater than 50 employees as listed in the Statistics of US Businesses (SUSB) versus the number of homepages in this study's dataset by state. The plot on the left shows this for the year 2001, while the plot on the right shows it for 2017.

Figure 13 Firms by NAICS versus Number of Homepages in Sample



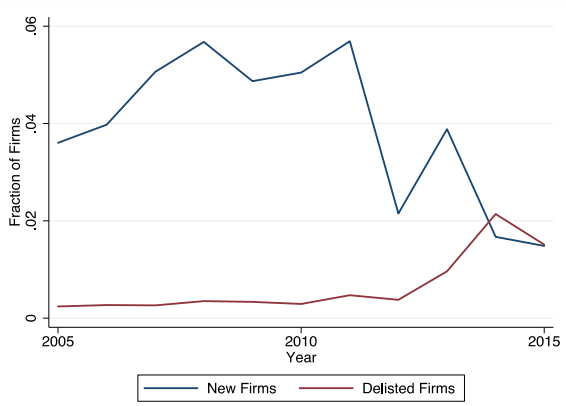
(a) 2001 (b) 2017

Note: The above plots show the number of firms with greater than 50 employees as listed in the Statistics of US Businesses (SUSB) versus the number of homepages in this study's

dataset by NAICS. The plot on the left shows this for the year 2001, while the plot on the right shows it for 2017.

Based on the Orbis database's organization founding date and organization delisted dates, we plot the fraction of organizations in our sample that are newly founded organizations and delisted organizations for each year.

Figure 14 New Organizations and Delisted Organizations in Sample



Note: The above plot shows the fraction of organizations with observations in each year for which that year is the founding year of the organization as well as the fraction for which it is the delisting year of the organization.

Appendix D Hiding Vintage Numbers

Many server software packages allow for hiding the vintage number from the server headers. The Apache server software for example provides six different configurations:[33]

- Full (default): the full server vintage number (e.g. Apache 1.3.6 (Unix) PHP/4.2.2)
- ProductOnly: only the vendor name (e.g. Apache)
- Major: the major vintage number only (e.g. Apache/1)
- Minor: the major and minor vintage numbers only (e.g. Apache/1.3)

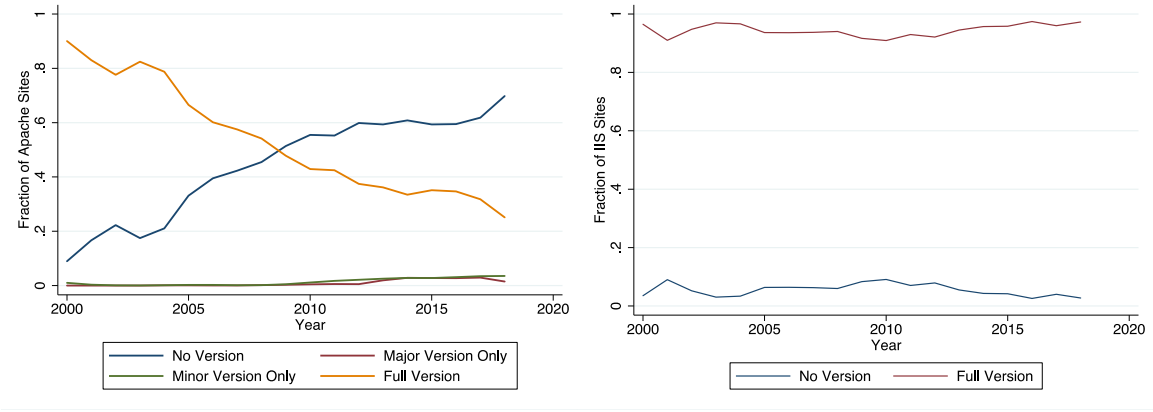
[33] https://httpd.apache.org/docs/2.4/mod/core.html - servertokens

- Minimal: all vintage numbers (e.g. Apache/1.3.6)
- Operating System: all vintage numbers and operating system (e.g. Apache/1.3.6 (Unix))

Our analysis of server vendor switching utilizes any observations where the server vendor name is visible. Our analysis of vintage changes, however, is limited to observations in which all parts of the vintage number are available. We make this restriction because we want to be precise about the distance of a organization to the technological frontier and the technology age of the server software.

In Figure D1, we show the fraction of observations by the information that is visible. Among websites hosted on IIS, the vintage number is almost always shown in the server headers. For sites hosted by Apache, the fraction of sites showing their full server vintage number decreases over time.

Figure 15 Availability of Software Vintage Information



(a) Apache

(b) IIIS

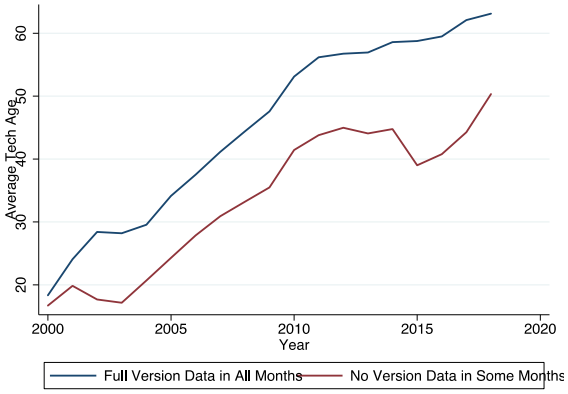
Note: The above plots show the fraction of homepages for which we have complete server vintage information. The left plot shows this for observations using the Apache server, while the plot on the right side shows this for sites using the IIS server.

One could imagine that more security conscious organizations would hide their vintage number in order to make it somewhat more challenging for hackers to exploit known

vulnerabilities with specific software vintages. On the other hand, one could also imagine that organizations that did not update their server software frequently would be more inclined to hide their vintage numbers since their software would be more likely to have vulnerabilities.

We plot the average technology age of the server software used by organizations that have complete server vintage in each month versus organizations that have some months with only the server vendor name but no server vintage data. Figure D2 displays the results of this exercise. The plot shows that organizations hiding their server vintage number on average use newer server software than organizations that leave their server information publicly visible.

Figure 16 Tech Age of Software Used by Organizations Hiding Their Server Vintage Numbers



Note: The above plot shows the average tech age of server software used by organizations that have complete vintage data in all months and organizations with hidden vintage numbers in some months. We use only Apache observations in this plot.

Appendix E Windows IIS Price Series

When computing the omitted value due to servers that are not priced, we utilize the price of the most popular Windows IIS server in contemporaneous use and take that price as the shadow value of the open source software. The following table shows the most popular Windows IIS vintage by year as well as the price of that proprietary software package:

Year	Most popular Windows IIS Vintage	Price of Windows IIS Standard Edition (2012 $s)	Price of Windows IIS Data Center Edition (2012 $s)
2000	4.0	$1,652.0	
2001	4.0	$1,652.0	
2002	5.0	$1,652.36	
2003	5.0	$1,652.36	
2004	5.0	$1,652.36	
2005	5.0	$1,652.36	
2006	6.0	$1,458.51	
2007	6.0	$1,458.51	
2008	6.0	$1,458.51	
2009	6.0	$1,458.51	
2010	6.0	$1,458.51	
2011	6.0	$1,458.51	
2012	6.0	$1,458.51	
2013	6.0	$1,458.51	

2014	7.5	$1,283.15	$3,209.48
2015	7.5	$1,283.15	$3,209.48
2016	7.5	$1,283.15	$3,209.48
2017	7.5	$1,283.15	$3,209.48

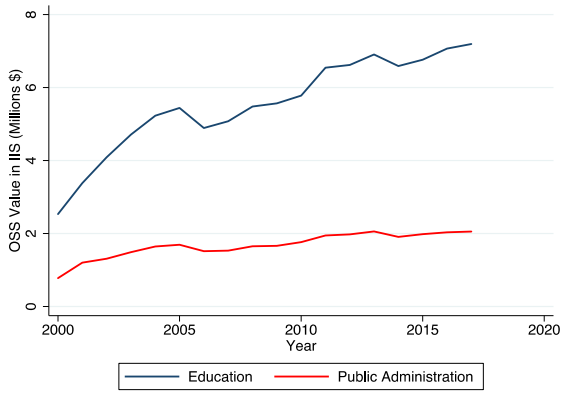
Appendix F Education and Public Administration Organizations

Our estimates of the value of open source software from our sample is scaled to be representative of U.S. firms in Figure 4 and Figure 5. We do this re-weighting the observations in our sample according to the number of firms in the United States of that size according to the Census Bureau's Statistics of U.S. Businesses (SUSB) database.

We also track and estimate the value of web server software used by educational institutions and public administration organizations. Educational institutions include K-12 schools, colleges, and universities. Public administration organizations include state, local, and federal government offices as well as court systems and public works.

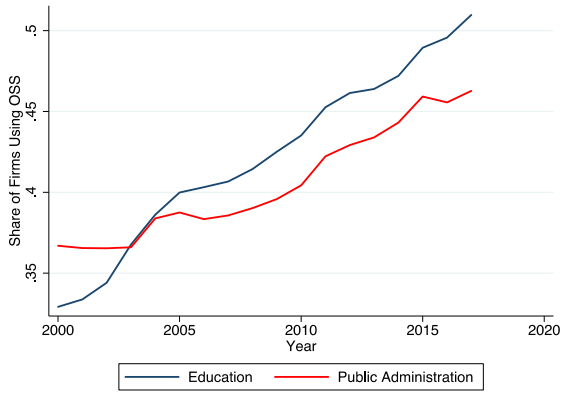
Unfortunately, the SUSB does not provide data on the number of public administration organizations or the number of schools in the U.S. Because these organizations are not precisely comparable to the medium and large businesses that make up the focus of our main analysis, we include figures for the estimated value of software used at these organizations in this section. Note, again, that these figures are estimated on the basis of the organizations in our sample and are not representative of the total value by the full set of such organizations in the U.S.

Figure 17 Omitted Value of Open Source Server for Educational Institutions and Public Administration Organizations



Note: The above plot displays the estimated value of open source servers being used in each year, broken down by organization industry. 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. The plot shows this with organizations binned by industry, defined by two-digit NAICS code. Only NAICS 61 and 92 are used for this figure. 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 18 Open Source Server Usage by Industry

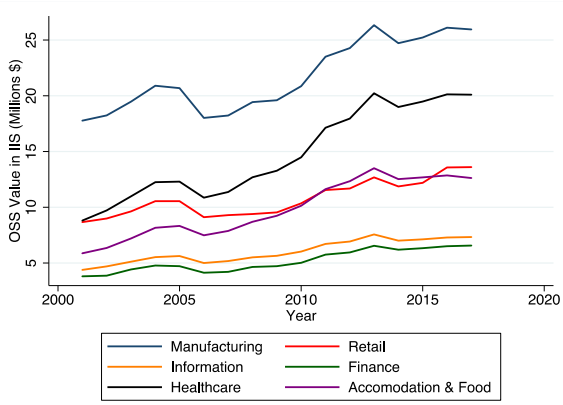


Note: The above plot displays the share of organizations in our sample using an open source server over time broken down by industry. The vertical axis is the share of organizations using OSS within an age bin. The horizontal axis is the year of observation. The organizations are binned by industry, defined by two-digit NAICS code. 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.

Appendix G Replication of Findings Using Secondary NAICS Codes

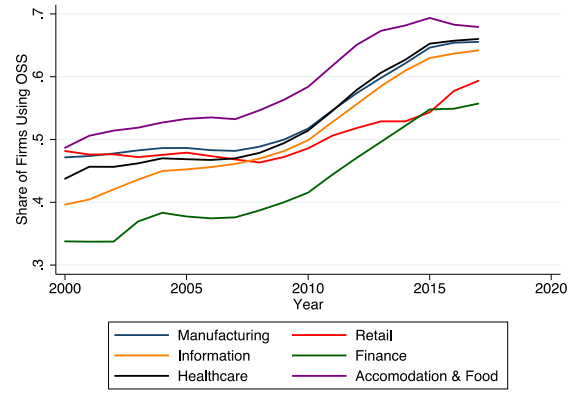
In the main text of the paper, we associate each website with an organization and each organization with a single industry based on the "core" NAICS code of its main line of business. In the below figures, we replicate the same exercises as above but instead treating each organization in the data as being in each of the distinct two-digit NAICS associated with either the organization's primary or secondary industries. Because most organizations operate within a single 2-digit NAICS, these figures are quite similar to the analogs in the main text which used just the primary NAICS for classification.

Figure 19 Omitted Value of Open Source Servers, Breakdowns by Organization Characteristics



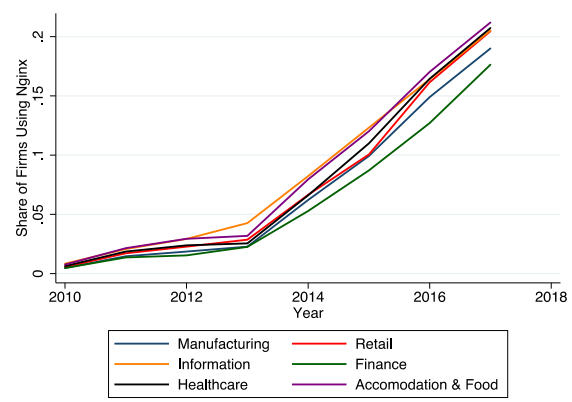
Note: The above plot displays the estimated value of open source servers being used in each year, broken down by organization industry. For each observation in the dataset in which a organization utilizes the Apache or Nginx web server, we find the price of the Windows server 10-user package containing 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 vertical axis is the OSS value for organizations within an age bin. The horizontal axis is the year of observation. The organizations are binned by industry, defined by two-digit NAICS code. Only six NAICS categories are shown in this figure. 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 20 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 industry. 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 organizations binned by industry, defined by two-digit NAICS code. 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 21 Adoption of Nginx



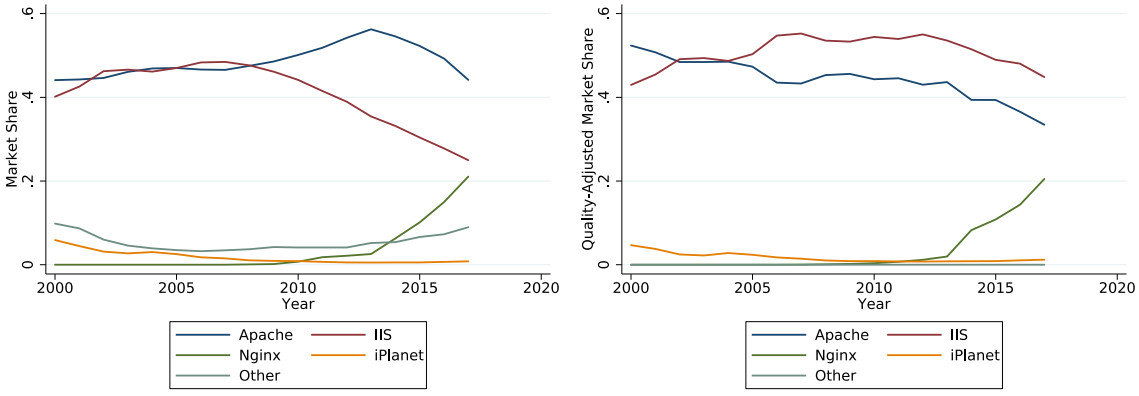
Note: The above plot shows the share of organizations using the Nginx server, broken down by organization industry. The vertical axis is the share of organizations using Nginx within an age bin. The horizontal axis is the year of observation. The organizations are binned by industry, defined by two-digit NAICS code.

Appendix H Replication of Findings at the Subsidiary Level

A unit of observation in our main sample is a website domain in a month. We use that level of analysis because many organizations make decisions regarding web server infrastructure at the level of the parent organization.

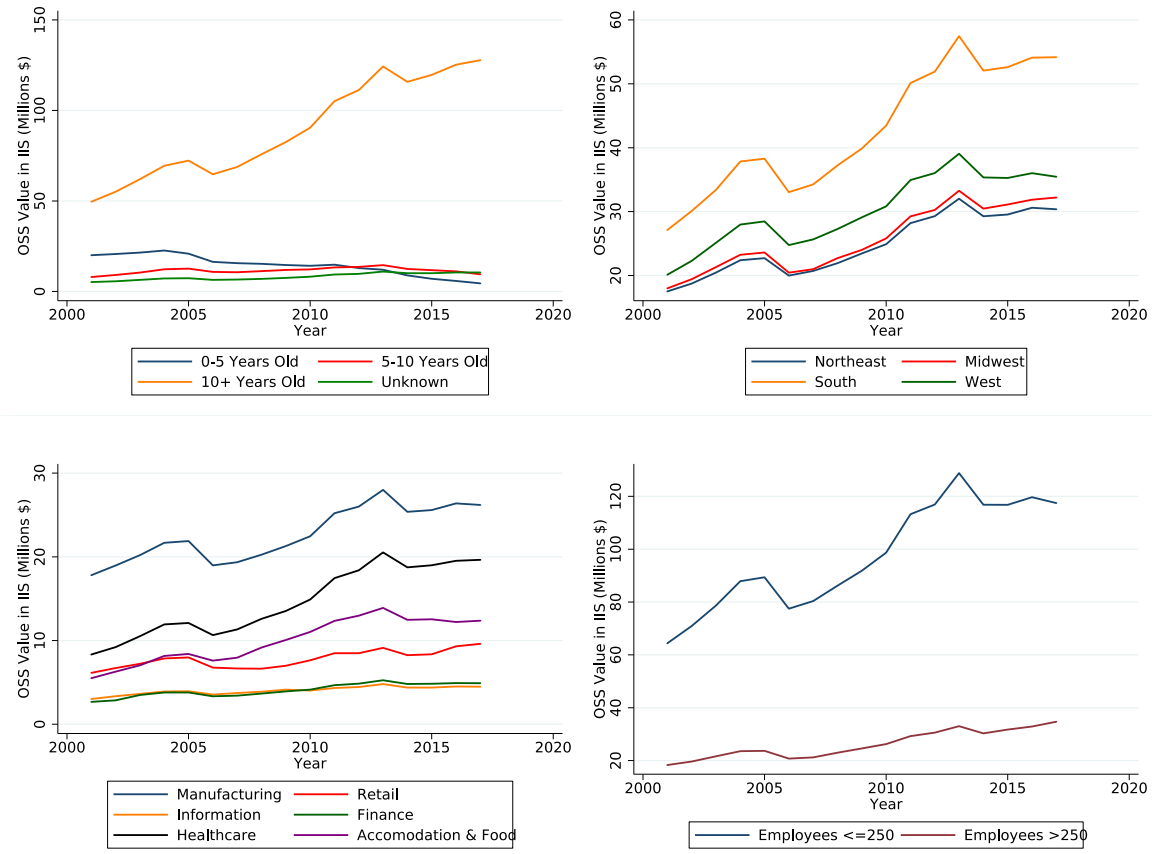
In this section, we replicate our results with the level of analysis being a subsidiary of an organization, when a subsidiary exists, and otherwise the parent organization. Because the website of the subsidiary may still be the same as the parent organization, this has the effect of essentially giving more weight to the observations of organizations with larger numbers of subsidiaries.

Figure 22 Market Shares Over Time



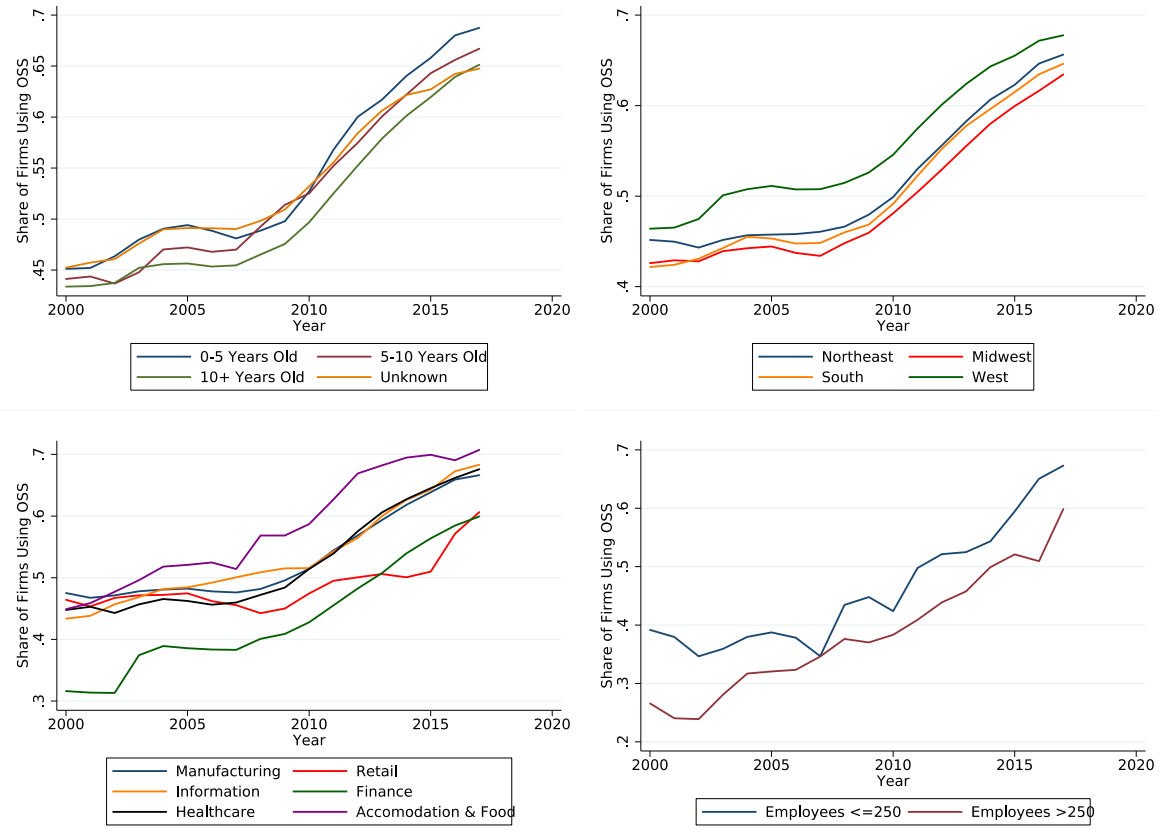
Note: The above plot displays the share of organizations in our sample using a server vendor over time. The vertical axis is the share of organizations. The horizontal axis is the year. As our panel data is at the monthly level, the year observations above are weighted by the number of months within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

Figure 23 Omitted Value of Open Source Servers, Breakdowns by Organization Characteristics



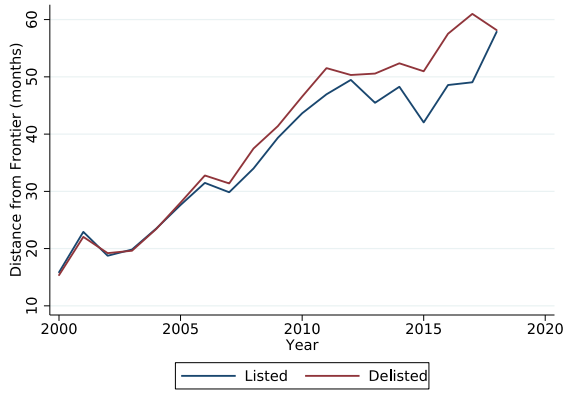
Note: The above plot displays the estimated value of open source servers being used in each year, broken down by organization age, geography, industry and size. The above plot includes subsidiaries. For each observation in the dataset in which a organization utilizes the Apache or Nginx web server, we 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 24 Open Source Server Usage, Breakdowns by Organization Characteristics



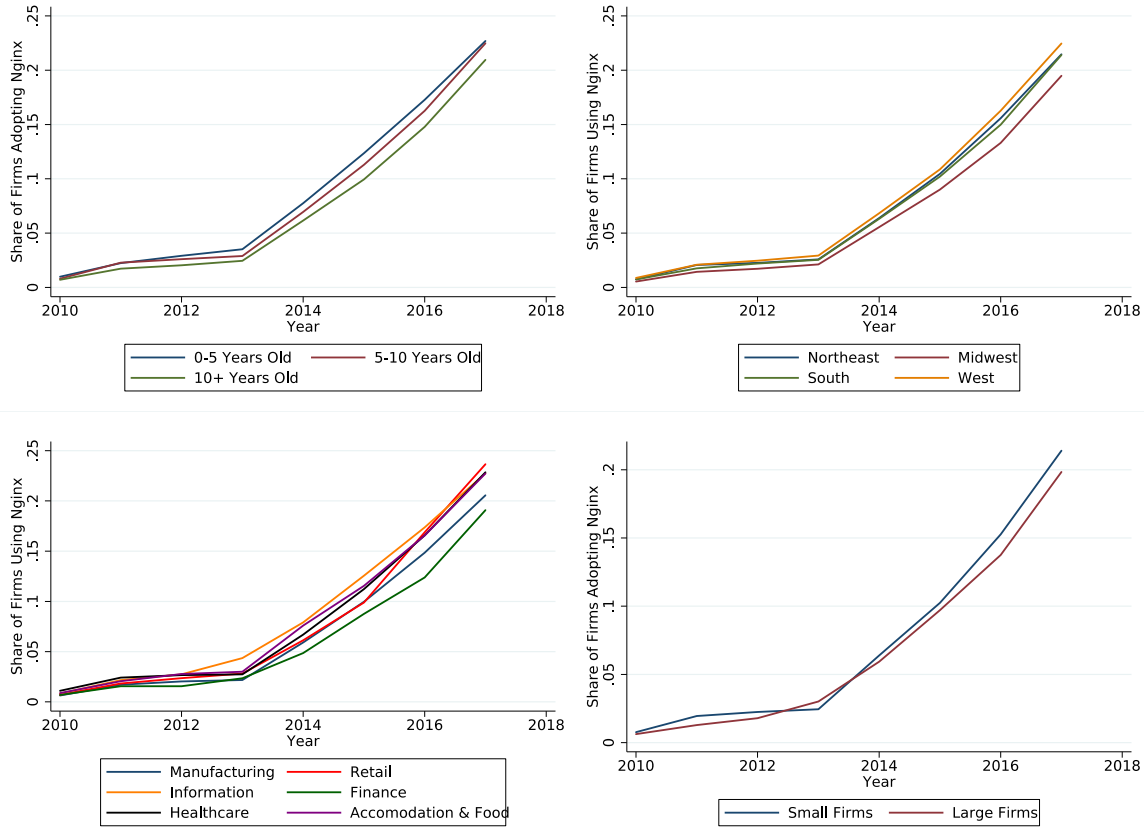
Note: The above plot displays the share of organizations, including subsidiaries, in our sample using an open source server over time, broken down by organization age, geography, industry and size. The top left plot shows this with organizations binned by the age of the organization. Organization ages 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 25 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, including subsidiaries, 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 26 Adoption of Nginx



Note: The above plot shows the share of organizations, including subsidiaries, 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.

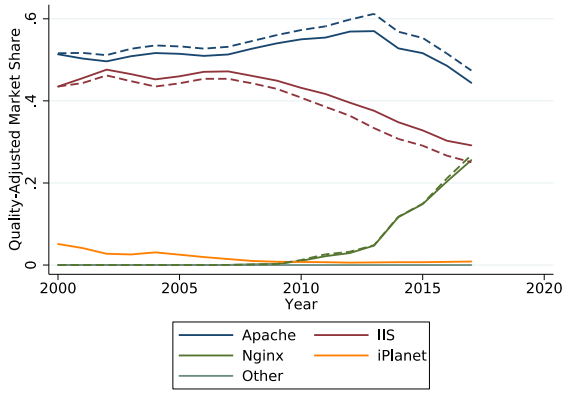
Appendix I Upper and Lower Bounds on the Quality Adjusted Capital Stock of Server Software

Our measure of Quality-Adjusted Capital Stock (QACAP) relies on information about the vintage of server software being used. There are two measurement issues that could cause a bias in this measure. First, Apache and Nginx users are more likely to hide their version numbers in recent years (see Figure 15). Therefore, computing the QACAP using only the observations in our sample in which software vintages are visible would undervalue the Apache and Nginx servers. Second, organizations that hide their server versions are typically closer to the technological frontier than organizations that leave their version numbers visible (see Figure 16).

Therefore, when discussing QACAP, we provide both a lower and an upper bound on the QACAP. For our lower bound, we interpolate the last visible server version for observations after an organization made their server version number not visible. This provides the lower bound because organizations are likely to update their server software over time. For our upper bound, we interpolate the most recent version of server software by server vendor used.

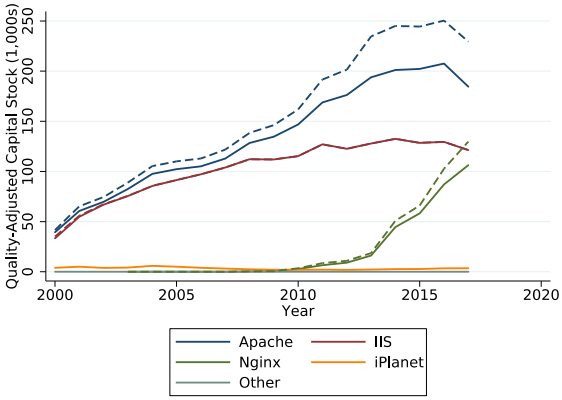
In the following two figures, we show the quality adjusted market share, QACAP, and omitted value of all open source servers. The solid line shows the version when using the last seen version and the dashed line shows the version when using the most recent version of that server software. Therefore, the solid line is the lower bound and the dashed line is the upper bound. The bounds are relatively consistent with the long-term trends showing the robustness of our findings.

Figure 27 Quality Adjusted Market Share Bounds 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. The solid line interpolates the hidden server versions with the last seen version, while the dashed line interpolates with the latest server version.

Figure 28 Quality Adjusted Capital Stock

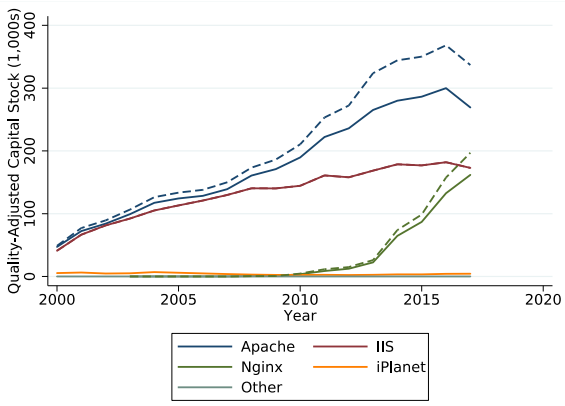


Note: The above plot displays the total quality adjusted capital stock ($QACAP_t$) of each server for our sample. The vertical axis is the sum of units of quality based on the inverse of the CPI for the server vintages and scaled by the number of servers using that vintage in that year. The horizontal 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. The solid line interpolates the hidden server versions with the last seen version, while the dashed line interpolates with the latest server version.

Appendix J Alternative Measure of the Quality Adjusted Capital Stock of Server Software

In the main text of this paper, we compute the quality adjusted capital stock of server software using the inverse CPI as quality weights. In this section, we replicate the results using the quality weights developed by Bryne & Corrado (2019). The overall pattern is the same as the one shown in Figure 2, however, the total QACAP is estimated to be somewhat higher using these weights.

Figure 29 Quality Adjusted Capital Stock based on Bryne & Corrado (2019)



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 quality weights of Bryne & Corrado (2019) 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. The solid line interpolates the hidden server versions with the last seen version, while the dashed line interpolates with the latest server version.