Almost with inexorable momentum, the Internet hurls itself into new territory. Some time ago, more than two billion humans had adopted at least one Internet-enabled device in some form, and nobody doubts that another two billion will accrue soon. New webpages increasingly find ways to inform readers, as more information in a variety of formats continues to be layered on the basic system of data internetworking.

That growth has been measured in a variety of dimensions. Today I would like to report on some research to measure one aspect of the Web’s growth, which I did with Frank Nagle, a doctoral student at Harvard Business School (see http://ssrn.com/abstract=2336372). We sought to figure out how much Apache served web surfers in the United States. That is not a misprint. Apache is the name for the most popular webserver in the world. It is believed to be the second most popular open source project after Linux.

Why do this? Measuring Apache is a key step in understanding the underlying economics. Because it’s free, Apache’s value is easy to mismeasure, and that makes its economics easy to misunderstand.

What is Apache?

Apache descended from software invented at the National Center for Supercomputing Applications (NCSA) at the University of Illinois, which also was the home of the Mosaic browser. Apache arose from server software that worked with Mosaic. It was called the NCSA HTTPd server. This was the most widely used HTTP server software in the research-oriented “early days” of the Internet.

While the University of Illinois successfully licensed the Mosaic browser for millions of dollars, the server software first became available for use as shareware, with the underlying code available to anyone, without restriction. Many webmasters took advantage of the shareware by adding improvements as needed or by communicating with the lead programmer, Robert McCool. McCool, however, left the university (along with others) to work at Netscape in the middle of 1994, and thereafter, webmasters and web participants lost their coordinator.

By early 1995, there were eight distinct versions of the server in widespread use, each with some improvements that the others did not include. These eight teams sought to coordinate further improvements. They combined their efforts, making it easier to share resources and improvements and build further improvements on top of the (unified) software. The combination of eight versions was called Apache, ostensibly because one of the founders respected the assertive reputation of one tribe of native North Americans—and also because others called the software “a patchy webserver.”

Informally at first and more formally over time, the group adopted the practices of open source. Skipping a long history, Apache became an essential component in the customer-facing commercial transactions of many firms, as well as in the procurement activities supported by electronic commerce. Furthermore, Apache is used as the base for many other commercial products, such as the IBM HTTP Server, which comes bundled with the IBM WebSphere Application Server.

It’s well known from publicly available statistics that Apache is disproportionately used to host websites that receive large amounts of traffic. Apache hosts 57 percent of the million busiest websites (see http://news.netcraft.com/archives/2013/08/09/august-2013-webserver-survey.html).

Apache’s basic economics derive from the lack of prices for the software. The absence of pecuniary transactions first arose at the beginning of Apache’s existence. It continued as Apache adopted open source practices. As with other open source software, Apache eschews standard marketing and 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.

In sum, Apache plays an important role in operating the Internet, but it never...
**continued from p. 80**

goes on sale or directly generates revenue. It does not produce the typical markers of economic value, and, thus, despite its ubiquity, it is easy to overlook.

**How much Apache?**

Although data on the number of websites hosted via Apache HTTP Server is readily available from public sources, data on the number of actual Apache HTTP Servers used is not. Additionally, existing public data does not clearly identify the location or country for these servers. However, because web servers are primarily used to host public webpages, and are directly reachable via the Internet, Frank Nagle and I were able to count them.

Take it with a grain of salt. Apache HTTP Servers can be used internally by organizations, so our calculation can be considered a lower bound on the number of actual Apache HTTP Servers in use. Furthermore, a number of different network architectures—including load balancing, elastic and cloud computing, and so on—allow multiple web servers to run on one IP address, which would also lead our collection method to underestimate the true capacity of Apache HTTP Servers.

We first identified the full list of IPv4 addresses registered to US organizations, available from the American Registry for Internet Numbers. As of 15 October 2011, when we undertook this experiment, there were 1.54 billion IPv4 addresses allocated in the US.

It was too costly to scan every IPv4 address, so we took a random sampling of 15,865,522 addresses, which is just over 1 percent. For each IPv4 address in our sample, we checked to see if the system was running a web server. If it was, we determined whether the server ran Apache, Microsoft IIS, or anything else, including unidentified servers. This method gave us what we sought—a census of server use and its characteristics, which otherwise is not available.

The details are straightforward for someone technically skilled in web programming and administration, although they’re tedious to report in this context.

This method identified “outward” facing servers.

This approach has one other principal drawback. One server may support a large or small number of pages. This method will be proportional to Apache’s actual importance in the economy when the size of use is uncorrelated with our measurement strategy (that is, no selection bias), and our sample size is large. In principle, this feature makes the small sample sizes potentially problematic. We did not find any symptoms of the problems, but small samples should be used with caution, as would occur with narrow geographies or industries.

Here is the answer: Of the 15,865,522 addresses in our sample, we found that 195,885 (1.23 percent) were running a webserver. The other 98.77 percent of the IPs scanned were either inactive or were devices that were not web servers on standard TCP ports. Of these 195,885 web servers, 44,211 (22.57 percent) were running Apache and 24,222 (12.37 percent) were running Microsoft IIS. Apache and IIS account for 34.94 percent of all web servers in our sample. The remaining web servers were either unidentifiable or were running a different web server, such as nginx or a proprietary web server. For example, Google has developed its own internal web server that it uses in place of a publicly available web server.

If we extrapolate these numbers to the full US IPv4 space, we estimate that there are 18,981,268 outward-facing webservers in the United States, 4,284,049 of which are running Apache. If the rest of the world looks like the US (a leap, to be sure), then—continuing this extrapolation to the entire range of IP addresses in the world, of which there are 3.706 billion—there would be 10,288,264 Apache servers in the world.

If publicly released data on worldwide websites (see http://news.netcraft.com/archives/2013/08/09/august-2013-webserver-survey.html) is accurate, then our estimates suggest there are 33 websites per Apache server. This is plausible because there must be a very skewed number of web pages per web server.

While some Apache servers serve only a single website, many are used by hosting facilities that host hundreds of websites.

Is that a lot? That equates to between 1.3 percent and 8.7 percent of the stock of prepackaged software in private fixed investment in the United States. That looks like a lot to me, especially for one piece of software.

Ponder that for a moment. The micro-mechanisms that create measurement issues for economic accounting of open source software are not unique to Apache. They are common to several Internet inventions that diffused into commercial use without formal market transactions and licenses, and where open source institutions supported deployment and use. Other prominent examples from this time period are Linux, software built around TCP/IP, and software built on top of the World Wide Web. That is a lot of stuff.

Furthermore, while Linux and Apache are two of the most recognized open source software projects, there are many others that play an important role in the digital economy but aren’t accounted for in any productivity measures, such as Perl, PHP, or Firefox, as well as a creative common license in a not-for-profit setting, such as in Wikipedia. Frank and I showed that the missing value is large in one specific instance, which suggests, perhaps, a big missing value in general.

Shane Greenstein is the Kellogg Chair in Information Technology at the Kellogg School of Management, Northwestern University. Contact him at greenstein@kellogg.northwestern.edu.