Economics Letters 128 (2015) 62–68

Contents lists available at ScienceDirect

Economics Letters

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

The size distribution of websites

Steven Schmeiser

Department of Economics, Mount Holyoke College, 50 College Street, South Hadley, MA 01075, United States

HIGHLIGHTS

- The upper tail of the website size distribution follows Zipf's law.
- The result holds for the world, Germany, and the United States.
- Web traffic from China does not follow this pattern.

ARTICLE INFO

Article history:
Received 12 November 2014
Received in revised form
30 December 2014
Accepted 16 January 2015
Available online 22 January 2015

JEL classification:
D3
L1

Keywords:
Internet
Size distribution
Power law
Zipf's law

ABSTRACT

The upper tail of the size distribution of websites follows a power law with slope close to one (Zipf's law). This finding is robust to measuring website size by unique visitors and page views, and holds for the United States, Germany, and the world. Web traffic in China has less support for a power law.

1. Introduction

Many size distributions of interest in economics take on the form of a power law with a slope coefficient ζ near one (Zipf's law). Well studied examples of this phenomenon include city sizes[1] and the size distribution of firms.[2] Here, I use OLS regressions and MLE estimates of the Pareto distribution to show that the upper tail of websites also follows a power-law distribution with ζ close to one. The observation that the upper tail of websites follow Zipf's law provides a strong constraint for models that either predict or exogenously specify a website size distribution, and contributes to the growing list of empirical power laws in economics.

I analyze data on the top 30,000 sites visited worldwide, as well as the top 30,000 sites viewed by users in China, Germany, and the United States. Size is measured two ways: by number of unique visitors, and by number of page views. Size distributions for Germany, the US, and the world all follow a power law, while the evidence for China is weaker.

To my knowledge this is the first paper to document the size distribution of websites in the economics literature. A few notable examples from computer science include Adamic and Huberman (2000), who obtain data on sites accessed by America Online users in 1997 and find evidence of a power law distribution, but do not report the slope coefficient. Adamic and Huberman (2002) revisit the AOL data and plot the data against a Pareto distribution with slope 1. Breslau et al. (1999) conduct six traces on academic, ISP, and corporate networks between 1996 and 1998 and find that web requests from a fixed group of users follow a power law distribution with slope ranging from 0.64 to 0.83 depending on the network. Last, Albert et al. (1999) investigate the topological properties of the World-Wide Web and find that the tail distributions of incoming and outgoing links on web pages follow a power law with slopes 2.1 and 2.45, respectively. Here, I use recent data collected from an Internet data company and find a slope coefficient close to one for many combinations of geography and tail size.

E-mail address: steven@schmeiser.org.

[1] Gabaix (1999), Gabaix and Ioannides (2004) and Holmes and Lee (2010).

[2] Luttmer (2007) and Axtell (2001).

http://dx.doi.org/10.1016/j.econlet.2015.01.019
0165-1765/

Table 1
Summary statistics for reach and page views. Reach and page views are both measured on a per-million basis within the specified geographic market.

	World		China		Germany		US	
	Reach	Pageviews	Reach	Pageviews	Reach	Pageviews	Reach	Pageviews
Site 1	496,000	113,250	779,700	106,510	778,000	97,480	834,200	194,900
Site 10	64,930	5675	185,400	15,433	87,100	4497	99,500	6562
Site 100	10,430	648	8,500	876	12,600	591	13,490	556
Site 1000	1,380	61	700	66	1,838	74	1,640	57
Site 10,000	163	6.38	170	9.4	190	8.5	177	6.5
Site 30,000	29	0.8	12	0.41	53	1.1	40	0.95

Table 2
OLS estimates of the slope coefficient ζ, with the null hypothesis that the data follows Zipf's law. Standard errors are constructed according to Gabaix and Ibragimov (2011). Estimates in **bold** are those for which the null hypothesis is not rejected.

	World			China			Germany			US		
$N =$	1000	10,000	30,000	1000	10,000	30,000	1000	10,000	30,000	1000	10,000	30,000
Reach												
$\hat{\zeta}$	1.137	1.098	1.050	0.873	1.260	0.814	1.176	1.064	1.045	1.096	1.045	**0.998**
	(0.051)	(0.016)	(0.009)	(0.040)	(0.018)	(0.007)	(0.053)	(0.015)	(0.009)	(0.050)	(0.015)	(**0.008**)
R^2	0.9988	0.9997	0.9959	0.9942	0.9491	0.9074	0.9982	0.9982	0.9968	0.9954	0.9992	**0.9960**
Pageviews												
$\hat{\zeta}$	**0.970**	**1.010**	0.892	0.908	1.039	0.665	1.090	1.091	0.872	**0.994**	1.044	0.903
	(**0.043**)	(**0.014**)	(0.007)	(0.041)	(0.015)	(0.005)	(0.049)	(0.015)	(0.007)	(**0.044**)	(0.015)	(0.007)
R^2	**0.9979**	**0.9995**	0.9810	0.9982	0.9900	0.9029	0.9913	0.9978	0.9694	**0.9955**	0.9991	0.9778

2. Data

The data is gathered from Alexa Top Sites.[3] Alexa ranks websites according to their Alexa Traffic Rank, which is a combination of reach and page views (defined below). The rankings are calculated based on a three month period ending in September 2014 and are aggregated to the domain level. For instance,

www.example.com/index.html
www.example.com/subdir/anotherpage.html and
www.subdomain.example.com/index.html

are all included under www.example.com.[4] The data is collected from users of Alexa's browser toolbar and other web data sources.[5]

The data includes two measures: reach and page views. Reach is defined as the number of unique visitors per million users that a website receives on a given day. Reach divided by 1,000,000 gives the fraction of the web-using population that visits the site. Page views are defined as the number of URL requests for a site, per million page requests. Page views divided by 1,000,000 gives the fraction of page requests that were directed at the given site. Multiple requests from the same user for the same URL on the same day are only counted once. This could result in under-reporting of sites that dynamically update content that users access at the same URL, such as. Both reach and page views are reasonable measures of website size. Reach emphasizes the size of a website's audience, while page views proxy for both the amount of activity a site receives and opportunities to display advertisements (a primary source of website revenue).

I collect the top 30,000 sites for the world and three countries: China, Germany, and the United States. Country is determined by the location of the visitor, not the location of the website. The reach and page view measures described above are for a given geography. For example, in the Germany data, a site with reach of 100,000 would indicate that ten percent of German web users visit that site.

Verisign estimates that of the 2.5 billion Internet users worldwide, China has the most (618 million) and the US has the second most (254 million).[6] The top 30,000 sites account for about 76% of world page views, 95% in China, 80% in Germany, and 82% in the United States. However, 30,000 sites represent only a small fraction of the 271 million web domains.[7] Summary statistics for the data are presented in Table 1. Within a country, the reach (per-million) of the top site is greater than the reach for the top site worldwide, as the top site in a country is more targeted to the unit of measurement. As can be seen in the table, an order of magnitude increase in rank roughly corresponds to an order of magnitude decrease in size for many of the combinations of geography and size measure.

3. Analysis

Let $i \in \{1, 2, 3, \ldots\}$ denote the rank of a website, with $i = 1$ being the largest, and let $S_{(i)}$ denote the size of the i'th largest site. Regressing the log of rank on log of size is a common method to measure the fit of data to a power law. I run the regression

$$\ln i = \text{const} - \hat{\zeta} \ln S_{(i)} + \epsilon_i \tag{1}$$

with the null hypothesis that $\zeta = 1$ to find the slope coefficient on each combination of geographic area, size measure, and tail sizes (1000, 10,000, and 30,000). As outlined in Gabaix and Ioannides (2004), standard errors reported by statistical software packages are incorrect due to correlation in the residuals generated by the ranking procedure. I therefore compute standard errors of $\hat{\zeta}(N/2)^{-1/2}$ as in Gabaix and Ibragimov (2011). Results are reported in Table 2. The null hypothesis is rejected for most combinations of tail size and geography, however it is not rejected for the US with tail size 30,000 when measured by reach, and it is not rejected for the world or US for small tail sizes when size is measured by page views.

With large sample sizes, it is perhaps not surprising that the null hypothesis is rejected in many cases. Gabaix (2009) cautions against testing whether or not data follows Zipf's law by statistical

[3] http://aws.amazon.com/alexa-top-sites/. Alexa is an Amazon company.
[4] The company does note that it separates out personal home pages and blogs when possible.
[5] http://aws.amazon.com/awis/faqs/.
[6] http://www.verisigninc.com/assets/domain-name-report-april2014.pdf.
[7] *ibid.*