P E R S P E C T I V E S

PHOENIX CENTER FOR ADVANCED LEGAL & ECONOMIC PUBLIC POLICY STUDIES

Approximating the Distribution of Broadband Usage from Publicly-Available Data

George S. Ford, PhD*

May 31, 2012

Introduction

If a broadband service provider imposes a monthly usage cap of 250 gigabytes ("GB"), how many of its customers would hit the limit? What if it were 200GB? 100GB? 50GB? Oddly, while we can look up on a broadband map what type of broadband is available at every address in the United States,[1] we have no off-the-shelf answers to many basic questions about Internet usage levels (of which I am aware).[2]

In this PERSPECTIVE, I attempt to provide some rough guidance on how Internet usage varies across users, and do so using publicly-available information. In fact, my calculations are based on only two data points that, when combined with the pattern Internet usage is known to follow (that is, the statistical distribution of usage), permit the full pattern of usage levels across connections to be approximated. A check on the accuracy of this approximation is conducted using other publicly-available data. Finally, an example of how to use this information, drawing from claims made by Comcast about usage levels and caps, is provided.

My approach to approximating usage patterns may be useful for variety of policy issues. For example, when addressing universal service for broadband, the level of service that qualifies as "broadband" will have to be parameterized. Knowledge of the usage distribution may aid in establishing these service level definitions that can be described as "reasonably comparable to those services provided in urban areas."[3] Usage patterns may also be relevant for modeling the construction costs of broadband networks, particularly for parties that do not have access to high-quality internal data on usage patterns. The analysis may also be useful in assessing claims made in the current debate over usage-based pricing and usage caps.

... we can look up on a broadband map what type of broadband is available at every address in the U.S., but we have no off-the-shelf answers to many basic questions about Internet usage levels.

Data and Distributions

Most of the available data on broadband usage presents scant detail on the underlying pattern of usage levels across broadband connections. This pattern reflects the statistical distribution of usage, which permits the calculation of the probability that that usage levels take on a particular value in the population. Publicly-available data often provides information such as averages and medians, perhaps along with a few statistics related to the quantiles of the distribution (e.g., how much of the total traffic do the top 1% of users consume). While

informative, this information does not permit interested parties to compute other statistics of interest. If the full statistical distribution of usage were known, however, then many statistics of interest could be calculated with ease.

The most recent study, of which I am aware, is an October 2010 study of Internet traffic by Cisco.[4] This study of international broadband usage reports that the average broadband connection consumed 14.9 GB per month. A year prior, the consumption was only 11.4 GB per month, so usage is growing by a whopping 31% per year. While the average usage levels are interesting, the average usage level is not much to work with if you are trying to approximate the underlying distribution of broadband usage. However, this study also reports that the top 1% of users accounted for 20% of Internet traffic, and the top 10% of users accounted for 60% of traffic. Now we are talking—these data points are far more useful.

But how do you convert a few data points into a full distribution of Internet usage? It turns out to be easier than you may think. Research reveals that nearly all communications traffic, including Internet traffic, can be approximated with high accuracy by the log-normal distribution.[5] Thus, approximating the broadband usage distribution involves finding a lognormal distribution that accurately matches the few data points from the Cisco report.

The general shape of the log-normal distribution is illustrated in Figure 1. As the figure shows, across many parameter values, the distribution is skewed right with a long tail.[6] Most of the observations have lower values, and the long tail implies that a disproportionate share of the total activity level is accounted for by the larger users. Given the skew, the mean will be larger than the median. The shape of the distribution matches some data reported in the Federal Communications Commission's OBI TECHNICAL PAPER NO. 4.[7]

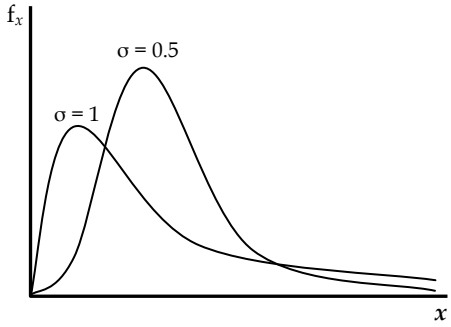

Figure 1. Log-Normal Distribution (μ = 0)

The lognormal distribution may be defined by two parameters. On the log scale, the parameter μ is called the *location* parameter and σ the *spread* parameter. A useful property of the lognormal distribution is that the quantiles are preserved under monotonic transformations. Thus, it is the *spread* parameter than can be adjusted in order to match the quantiles from the Cisco report, because the quantiles are not sensitive to the *location* of the distribution. Once the spread is determined, the location can be calculated using

$$\mu = \ln(\bar{x}) - \frac{1}{2}\sigma^2 \qquad (1)$$

where $\bar{x}$ is the assumed average GB per month (e.g., 14.9 in the Cisco report).

... based on what we know about Internet usage, the distributions fit the data well.

The Cisco study reports that the top 1% of users accounted *for more than* 20% of Internet traffic. However, in the graphic, the number is 20%, suggesting the usage level was close to but slightly higher than 20%. According to the report, the top 10% of users accounted for 60% of traffic, and this statement was without qualification. Undoubtedly, the share was not

exactly 60%, but I will assume it is very close to that level.

Armed with these two data points, my analysis of the lognormal distribution indicated that a spread parameter of 1.535 produced and underlying distribution consistent with the Cisco data.[8] With this parameterization, the top 10% of users account for 60% of the traffic, and the top 1% of users account for 21.5% of the traffic. These results are highly comparable to distribution described by Cisco. Note also that the top 20% of users account for about 76% of the total traffic. Once more, the nearly ubiquitous 80:20 rule is found to apply.

Using the data from the Cisco report, my analysis indicated that σ is 1.535, and from Equation (2), I can calculate μ to be 1.52. The underlying distribution matching the reported statistics in the Cisco report is approximated by the distribution log-$\mathcal{N}$(1.52, 1.535).

Much can be done with this information using a spreadsheet. For example, say you wanted to know what the threshold GBs per month for the top 1% of users is? Given the mean of 14.9 GB per month, the threshold level of usage could be computed in Microsoft Excel with the formula

$$= \text{LOGINV}(0.99,1.52,1.535) \approx 163.$$

So, about 1% of users consumed 163 GB or more per month, according to the Cisco data (dated October 2010). Or, you could ask how many users would be affected by a GB cap of 163 GB? The Excel formula is

$$= 1 - \text{LOGNORMDIST}(163,1.52,1.535) \approx 0.01,$$

which matches the result from the previous example.

Suppose, alternately, you wanted to know how many customers would be affected by a 50 GB cap. The answer is

$$= 1 - \text{LOGNORMDIST}(50,1.52,1.535) \approx 0.06,$$

so only about 6% of users would hit the 50 GB limit in a given month. Or, say you wanted to know the GB threshold for a cap affecting 10% of the customers. The answer is

$$= \text{LOGINV}(0.90,1.52,1.535) \approx 33,$$

so a cap of about 33 GB would affect 10% of the customers.

Importantly, all of these calculations are based on very specific assumptions about the *location* and *spread* of the distribution. If other assumptions are made about either parameter, then modification to the inputs in the Excel formulas is necessary (or whatever statistical package is being used). To adjust the underlying distribution, the first step is to determine the spread (σ) and then compute μ using Equation (1).

In some reports on broadband usage, the median rather than the average may be reported. If so, you can compute μ using,

$$\mu = \ln(Median), \qquad (2)$$

and the same formula can be used, along with Equation (1), to convert the mean to a median if desired. For example, if you knew the mean but wanted to know the median, then the median can be computed by using,

$$Median = \exp(\ln(\bar{x}) - 0.5\sigma^2), \qquad (3)$$

where Equation (1) is substituted into Equation (2) to produce Equation (3). An example is provided below.

A Lookup Table Approach

For the less analytically inclined, it is also possible to create a table that can be used to calculate the threshold levels of the distribution with nothing more than a mean usage level. This approach is made possible by the stability of the quantiles to monotonic transformations.

Let $\bar{x}$ be the average usage, and let x_p be the threshold usage level that cutoffs the largest p percent of users. I can define a threshold markup factor (on the mean) for any given mean, assuming spread constant, as

$$\lambda_p = x_p / \bar{x}. \quad (4)$$

Thus, if the average usage ($\bar{x}$) is 20 GB per month, and the markup factor is 2 for the top 10% of users ($\lambda_{0.10} = 2$), then the threshold level of usage that defines the top 10% of users is 40 GB per month. Using this approach, I can avoid setting the mean for the distribution, and look up the markup factors from a table.

Table 1. Tails of the Usage Distribution*

Top Users	Share of Traffic	λ_p	Cutoff (Avg. 14.9 GB)
1%	21%	10.98	163.6
2%	30%	7.22	107.6
3%	36%	5.54	82.5
4%	41%	4.51	67.2
5%	46%	3.84	57.2
10%	60%	2.20	32.8
15%	69%	1.51	22.5
20%	76%	1.12	16.7

* $\sigma = 1.535$

Table 1 summarizes the λ markup factors under the assumption that σ is 1.535. If you wish to know the cutoff value of the top 1% of users, then from Table 1 observe that $\lambda_{0.01} = 10.98$. For a mean traffic use of 14.9 GB, the cutoff is 163 GB, the same as before. The 10% cutoff value, alternately, is $\lambda_{0.1} = 2.2$, so at an average usage of 14.9 GB, the threshold usage level for the top 5% of users is 33 GB per month (also as before). Appendix A contains more detail on the λ values.

Comparison to Other Data

Another report that contains data on the quantiles of broadband usage is a 2009 study by equipment manufacturer Sandvine.[9] This report, *2009 Global Broadband Phenomena*, states that the top 1% of users consume 25% of the traffic. The top 20% of users account for 80% of the traffic. A spread parameter of 1.65 generates an underlying distribution that matches the data points from this report, a value that is very similar to the 1.535 for the Cisco data. It appears that an assumed value on σ in the ballpark of 1.6 provides a good approximation of the underlying distribution of broadband usage from these two studies based on international usage data.[10] More evidence on this issue is presented next.

An Example: Comcast's 250 GB Cap

Until recently, Comcast imposed a 250 GB per month usage cap on its residential broadband customers.[11] The company claimed that for about 99% of its residential customers, the cap presented no issue whatsoever.[12] This anecdote suggests that only 1% of Comcast's residential users consume 250 GB per month or more. Comcast also indicated that its median customer consumes about 8 GB to 10 GB per month. Was Comcast's claim legitimate?

First, since Comcast reports the median usage level, I need to convert the median to the mean. To be conservative, assume the median in 8 GB per month. Using Equation (2), I compute $\mu = 2.08$. Assuming a lognormal distribution with this μ and $\sigma = 1.535$, I can compute the share of customers affected by the cap using the Excel formula,

= 1 - LOGNORMDIST(250,2.08,1.535),

which renders a value of 0.012, indicating that about 1.2% of users are affected by the cap. Or, I can use the Excel formula,

= LOGINV(0.90,2.08,1.535),

to discover that the top 1% of customers use at least 285 GB per month. Thus, Comcast's claim is certainly in the ballpark of what public data

tell us, particularly given the likely accuracy of my approximations.

... Comcast's claim is certainly in the ballpark of what public data tell us, particularly given the likely accuracy of my approximations.

If I assume that Comcast is being very precise in its statements, then I could better describe the underlying distribution as having a spread (σ) of 1.48, whereby 250 GB per month cap would affect exactly 1% of the customer base. The reduction in σ implies that more customers are heavier users (see Figure 1), and this difference has a few possible explanations. First, it may be that the cable operator customer base consumes more bandwidth on average than do others customers. There are reasons to believe this may be so. Second, it may be that U.S. users are heavy users relative to the rest of the world. The Cisco and Sandvine study look at international data. Third, it may be that the spread is falling over time. The Cisco and Sandvine reports are dated 2010 and 2009, respectively. Newer data from Cisco or Sandvine may help decipher which of these explanations is most valid (or, at least, which isn't). Notably, the reduction in σ does not materially affect the distribution; the top 1% of users still account for 20% of the traffic, and the top 10% of users account for 58% of the traffic. Also, for Comcast, I estimate the average (not median) residential user consumes 33 GB per month.

If I assume that Comcast's median usage is 10 GB, then about 1.8% of customers are affected by the 250 GB cap (σ = 1.535). By reducing σ to 1.39, I can reduce the number of affected customers to 1%, as reported by Comcast. Since 1.8% would likely be reported as 2%, then I think it is reasonable to conclude that the σ for a U.S. based cable operator's customer base is currently about 1.4. The reduction in σ to 1.39 implies that the top 1% of users account for 17.6% of the traffic, and the top 10% account for 54% of the traffic. So, while the change is σ does impact the shares, the change is not large. In sum, this analysis suggests that an assumption of a σ in the 1.4 to 1.6 range is generally supported by the available data.[13]

Conclusion

In this PERSPECTIVE, I have attempted to shed some light on the underlying usage distribution of broadband Internet service. To date, the public has had little insight into the underlying distribution, and has been forced to rely on a smattering of descriptive statistics. Using very little data and assuming an underlying theoretical distribution which has been shown to be a good approximation for communications and Internet traffic, I approximate the underlying distribution of usage. Given the rising incidence of usage caps on broadband connections, the approximated distributions provided here may be interest to policymakers and other interested parties.

Notably, my analysis passes no judgment on usage caps. It is a statistical analysis. Nor does my analysis differentiate between fixed and mobile broadband connections, except to the extent the underlying data does so. Obviously, usage levels for home connections are likely to be substantially greater than for mobile connections.

Finally, while I find this analysis somewhat compelling, the truth is that I am approximating a distribution without the ability to perform a thorough check on its quality across a range of values. Nevertheless, based on what we know about Internet usage, the distributions fit the data well. I suspect more can be done than is accomplished here, perhaps using data that is not publicly available. It would be interesting to see if my findings hold up to scrutiny.

Appendix A. Threshold Factors (λ_p)

	σ = 1.535		σ = 1.65		σ = 1.4	
Top Users	% Traffic	λ_p	% Traffic	λ_p	% Traffic	λ_p
1%	21%	10.96	25%	11.93	18%	9.76
2%	30%	7.19	34%	7.60	26%	6.66
3%	36%	5.51	41%	5.69	31%	5.25
4%	41%	4.51	46%	4.62	36%	4.37
5%	46%	3.85	50%	3.87	40%	3.76
6%	49%	3.34	54%	3.33	44%	3.32
7%	52%	2.97	57%	2.93	47%	2.97
8%	55%	2.67	60%	2.61	50%	2.69
9%	58%	2.42	62%	2.34	52%	2.46
10%	60%	2.21	64%	2.12	55%	2.26
11%	62%	2.03	66%	1.94	57%	2.10
12%	64%	1.87	68%	1.78	59%	1.95
13%	66%	1.74	70%	1.64	61%	1.82
14%	68%	1.62	72%	1.52	63%	1.71
15%	69%	1.51	73%	1.42	64%	1.60
16%	71%	1.42	74%	1.32	66%	1.51
17%	72%	1.33	76%	1.24	67%	1.43
18%	73%	1.25	77%	1.16	69%	1.35
19%	74%	1.18	78%	1.09	70%	1.28
20%	76%	1.12	79%	1.03	71%	1.22
21%	77%	1.06	80%	0.97	72%	1.16
22%	78%	1.01	81%	0.91	73%	1.11
23%	79%	0.96	82%	0.87	75%	1.06
24%	80%	0.91	83%	0.82	76%	1.01
25%	81%	0.87	84%	0.78	77%	0.97
26%	81%	0.83	84%	0.74	78%	0.92
27%	82%	0.79	85%	0.70	78%	0.89
28%	83%	0.75	86%	0.67	79%	0.85
29%	84%	0.72	86%	0.64	80%	0.82
30%	84%	0.69	87%	0.61	81%	0.78

NOTES:

* **Dr. George S. Ford** *is the Chief Economist of the Phoenix Center for Advanced Legal and Economic Public Policy Studies.* *The views expressed in this* PERSPECTIVE *do not represent the views of the Phoenix Center or its staff. He is grateful to Randy Beard for helpful suggestions.*

1 www.broadbandmap.gov.

2 This issue is of increasing policy relevance. *See, e.g.*, Letter from Melissa Newman, CenturyLink, to Marlene H. Dortch, *Ex Parte*, WC Docket No. 10-90 (March 30, 2012) (available at: http://apps.fcc.gov/ecfs/document/view?id=7021905144).

3 47 USC § 254(b)(3).

4 *Cisco Visual Networking Index: Usage*, Cisco (October 25, 2010) (available at: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/Cisco_VNI_Usage_WP.pdf), at p. 7.

5 The approximation is primarily of the tails, which is of most interest. *See, e.g.,* J. Gao, *Modeling Bursty Traffic in Computer Communications Networks*, Unpublished Manuscript (2004) (available at: http://wireless.ece.ufl.edu/seminar/gao.pdf); I. Antoniou, V. Ivanov, Valery Ivanov & P.V Zrelov, *On the Log-Normal Distribution of Network Traffic*, PHYSICA D: NONLINEAR PHENOMENA, Volume 167, Issues 1–2, 1 (July 2002) at 72–85; A.B. Downey, *Lognormal and Pareto Distributions in the Internet*, 28 Computer Communications 790-801 (2005); R. Andrade, A. Lisser, N. Maculan & G. Plateau, *Telecommunication Network Capacity Design for Uncertain Demand*, 29 COMPUTATIONAL OPTIMIZATION AND APPLICATIONS 127-146 (2004); V. A. Bolotin, *Modeling Call Holding Time Distributions for CCS Network Design and Performance Analysis*, 12 IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS 433-438 (1994).

6 A discussion of the lognormal distribution is provided at: http://www.itl.nist.gov/div898/handbook/apr/section1/apr164.htm.

7 *Broadband Performance,* OBI TECHNICAL PAPER NO. 4, Federal Communications Commission (2010) (available at: http://transition.fcc.gov/Daily_Releases/Daily_Business/2010/db0813/DOC-300902A1.pdf), at Exhibit 18.

8 This task was accomplished by generating log normal distributions across a range of values for σ (with $\mu = 0$) until the data matched the properties reported in the Cisco report. All calculations were in Eviews 7.2 (available at: www.eviews.com).

9 *2009 Global Broadband Phenomena*, Sandvine (2009) (available at: http://www.sandvine.com/downloads/documents/2009%20Global%20Broadband%20Phenomena%20-%20Executive%20Summary.pdf).

10 For this spread, the top 1% of users account for 23% of the traffic, the top 10% account for 62% of the traffic, and the top 20% account for 77.5% of the traffic.

11 *See, e.g.,* M. Reardon, *Comcast Ditches 250GB Data Cap, Tests Tiered Pricing* (May 17, 2012) (available at: http://news.cnet.com/8301-1023_3-57436489-93/comcast-ditches-250gb-data-cap-tests-tiered-pricing).

12 http://customer.comcast.com/help-and-support/internet/common-questions-excessive-use.

13 AT&T reports that its top 2% of users account for 20% of traffic (http://www.att.com/esupport/internet/usage.jsp#fbid=P80_h5RKxOE). This distribution is not as skewed as the others, suggesting a σ of about 1.2.