

Research: Papers

Copyright (c) 1994 by Donna L. Hoffman and Thomas P. Novak

August 18, 1994

This paper appears in print as: Hoffman, D.L. & Novak, T.P. (1994), "How Big is the Internet," *HotWired*, Aug. 18.

How Big Is the Internet?

Peter Lewis of the *New York Times* caused a stir on the Internet with his August 10, front-page, article "Doubts are Raised on Actual Number of Internet's Users." Lewis cast doubt upon the commonly cited number of 20 to 30 million Internet users, quoting John Quarterman as saying "Suppose there were really only two million or three million." A deflation of market size by a magnitude of ten is certainly cause for alarm. But are there grounds to sound the alarm?

Quarterman's lower estimate is explained in a [June 1994 FAQ](#) in which he makes the following points:

- the "best figures there are" are from *his* company's January 1994 Internet Demographic Survey, rather than competitor [Mark Lottor's Internet Domain Survey](#),
- reachable hosts as determined from a survey should be used as the baseline count of Internet hosts, rather than Lottor's estimate of reachable hosts calculated from a sample of hosts in the Domain Name System,
- the "real factor for users per Internet host" is about 3.5, rather than 7.5 or even 10 users/host as is assumed by other researchers.

Let's look at each point in turn.

1. The best figures there are.

Quarterman's survey was sent to postmasters of nearly 5000 Internet domains. Now, no offense to our local postmaster, but since he doesn't respond to our emails, we can't imagine him taking the time to respond to Quarterman's survey! Thus, we suspect that Vanderbilt University may not be represented in the Internet Demographic Survey. Indeed, Quarterman's FAQ notes that only 13% of received responses were useable. This is not very encouraging, indicating that our postmaster would be in good company if he did not respond. For such a high involvement product category, this response rate is *way* too low, and introduces bias of unknown magnitude and direction. The results are simply not projectable.

But, assuming for the moment that our postmaster *did* muster up the effort to respond, we are concerned how he would have reacted to the survey. Unfortunately, Quarterman's survey violates just about every rule of survey design! For example, a basic rule of survey research is "don't ask people questions they cannot answer." The Vanderbilt postmaster is a terrific guy (even though he doesn't answer our emails), but we think he would have a tough time with:

- total people in your organization: _____
- network users who send mail outside your domain: _____
- computers reachable with ICMP ECHO (ping) from the Internet: _____

- o percentages of your users in the following age categories: (list of eight age categories)

There is an expression for the requests above -- GIGO or "Garbage In, Garbage Out." Frankly, we can't imagine postmasters, let alone anyone else, answering these questions with anything better than wild guesses (unless of course they've done their own surveys -- which, as far as we know, they haven't).

Face it, these are *tough* questions that require serious legwork to answer. We really must question the quality of the data received, and we are certainly not convinced that in comparison to Lottor's estimates, Quarterman's are the "best figures there are."

2. Reachable hosts.

Quarterman insists that Lottor's raw host numbers are too high because "a lot of hosts on networks...are deliberately firewalled so you can't get there from the Internet proper." Thus, only reachable (i.e., "pingable") hosts should be used. Sound reasonable? Let's think about it.

A colleague across the hall has a Mac on the Internet, but he is not pingable. A co-author of ours at the University of Pittsburgh who connects to the Internet and uses Mosaic from his 486 machine is also not pingable. Vanderbilt University has 100 Apple Remote Access users who are not pingable, although they are using Mosaic and other Internet services from home. The Owen Graduate School of Management has 400 full-time MBA students, who are not pingable when using the Mac and Pentium machines in our computer lab to access the Internet. Our guess is that there are a whole lot of machines on the Internet which are not pingable - but which are also not behind firewalls or serving as routers. If this is the case, using the reachable hosts methodology will grossly underestimate the number of people on the Internet.

The problem is that a focus on "reachable hosts" is biased toward server applications rather than client applications. Surveys of Internet usage need to focus on the end user. Unfortunately, neither Lottor's nor Quarterman's survey focuses upon the end user *customer*.

3. Users per host.

Quarterman's FAQ claims the real factor for users per host is about 3.5. This is apparently based upon the numbers from his Internet Demographic Survey. As you might guess, we have some problem believing the numbers from that survey, since it includes open-ended, excruciatingly-detailed questions addressed to overburdened postmasters. Lottor, in reporting the results of his January 1993 Internet Domain Survey, says that some people have suggested 10 per host. Quarterman throws around other suggestions of 5 and 7.5. What's the "real number?" Face it - no one knows!

At this point, the amazing thing about the size of the Internet in our minds is that *no one* really has a very good idea how large it is! Approaches such as those taken by Lottor and Quarterman attempt to derive the number of users by making two assumptions: 1) the number of hosts ("reachable" or not) and 2) the number of users per host.

Improvement in measurement methodology is needed to nail down both of these numbers. For the number of hosts, we need a better definition of a valid host than "pingable." For the number of users per host, we really need to obtain *distributions* of users per host for various host segments (like .edu and .com, for starters). Quite likely, the mean will be a very poor measure of central tendency when hosts such as aol.com -- with a million users having limited Internet access -- are lumped together with our single-user workstations.

Current approaches to estimating the size of the Internet are akin to estimating the number of people in the United States by sampling the number of buildings, without regard to their function or contents. There is another way to go measure the usage of the Internet. A way that is market-driven and customer-oriented. Rather than inferring the number of users by counting and sampling machines, sample the *users*.

This opens up the question, "what is a user?" Our anecdotal evidence suggests that users go through a progression of adoption stages, starting with email, moving on to Usenet news groups and other text-based Internet services, and graduating to hypermedia applications such as Mosaic. All of these types of usage need to be tracked.

The Internet has evolved dramatically in size and economic importance. It is high time for the first Internet Users Sample Survey. This survey should include the larger group of individuals with any kind of network access. Note that we're not talking about a proprietary survey where information is sold to those firms willing and able to pay, but a large-scale global

sample survey of the current market size of individuals with network access. Such a survey should be conducted on a regular (at a minimum, annual) basis. This information is critical for the development of electronic commerce. It is foolhardy to base strategic business decisions upon the numbers currently available.

Thus, Lewis' article is, indeed, cause for alarm. Not because there are "only" two or three million users of the Internet, but because it is clear that we don't really have a clue *how many* users there really are.

Donna L. Hoffman and Thomas P. Novak are Associate Professors of Management at the Owen Graduate School of Management at Vanderbilt University, where they research the marketing implications of commercializing the Internet.