Using Web Metrics to Estimate Impact:
I – The Lawlessness of Averages

Bob Donahue
WGBH Interactive
July 2006

Abstract

This is the first in a series of articles to help NSDL partners find ways to uncover measurements of education impact from available web metrics. To start out, we'll look at a couple of "standard" metrics to see how reality compares to expectations in user populations and behavior.

1. Introduction

An often-overlooked step in the process of building an education web site – once the site is launched and everyone involved in the production breathes a sigh of relief – is actually stepping back and waiting to see if it was truly worth the effort. While some groups have the luxury of taking an "If you build it, they will come…" approach, for most of us it's the detection and measurement of post-launch impact based upon usage that becomes the next priority. Yet, for the most part, the statistics used to base a determination of "impact" are generally misunderstood, and like the "telephone" game become even more muddled as they're passed from party to party. Then, they almost always become condensed into a sentence fragment buried within a report, or as a bullet point in a presentation.

There's a danger there, both because it invites misunderstanding, but mostly because there's almost always a great deal of information to uncover if one takes the time to look for it, In addition, it's important to make the effort to understand why the results obtained came to pass. Over the course of this series of articles I'll make my own attempt to uncover measurements of impact relevant to the K–12 community using statistics derived from the Teachers' Domain project[1].

2. Tools of the Trade: Logs, Stats, and Demographics

It basically comes down to this: the more data you track, the more you can learn from it. Having access to your complete web logs is fundamentally important because it provides you with a way to check results, dig deeper, etc. Services like Omniture are invaluable for digesting the huge amounts of data that pile up, (esp. because they also allow you to customize your own stats – this will be one of the topics in an upcoming report).

[1] `http://www.teachersdomain.org/`

The two case studies below only strictly require web server logs to distill the necessary information. The customized setup for the TD site makes this a little easier because we require registration and link users to an organization (usually a K–12 school) from which we can derive additional information – in particular, geospatial data). In both cases we came into the study with a set of expectations, assumptions, and even preliminary results…

… in both cases we were in for a few surprises.

3. Case Study 1: Beware the Means of March! (or April, or May…)

As part of my data mining duties, I track a few dozen stats monthly. As you'd expect, this also means that I receive a few e-mail requests for "results" every month, usually desired in the "10 words or less" category. One of these has been "How long do users stay on the site?" Well – there's the long answer, and the short answer…

The (right – for all the wrong reasons) short answer: based on both Omniture stats and my own accounting – about 10 minutes on average. This particular value made many people relieved because the general concept for a TD session was having a teacher bring the site up to use one of the resources in the classroom – say a video – and the process of showing the clip, answering questions, etc., was expected to be about 10 minutes, based upon the known length of the videos and from observing focus groups when the site was first developed.

Yay. Success! Well, if you believe the "10 words or less" version of the results.

But is it truly a *representative* picture? Hmmm. Well, this is a good time to revisit what we mean when we say "average" and more importantly, what assumptions are going into that concept.

We all know how to calculate the mean: take the sum of all of the samples and divide by the number of samples. But this assumes:

1. The sample is representative of the true population. In this case, our sample (probably taken over a calendar month) needs to show the range of behaviors of the users in proportion to the frequency of those behaviors. Why wouldn't that be the case? Well, one occasional situation is having something on your site listed on someone else's site as being particularly timely, "neat," or some other criteria, causing a spike in hits but from a group of people who aren't your target audience with a pattern of behavior that might not mimic that of your target audience;

2. The sample has only one distribution, and that the distribution can be described by formulae and statistical concepts that are generally referenced/used.

If you were to try to picture a "perfect" distribution, you'd probably think of something like a "bell curve" or Gaussian with the mean sitting right atop the distribution such that:

$$mean \approx median \approx mode$$

where the median is the value such that 50% of the distribution lies below its value and 50% is above, and the mode is the point in the distribution with the highest frequency. So, in a perfect Gaussian-like distribution, the three are equal.

However, many common situations do not follow this kind of distribution at all! For example, if you roll a die and keep track of the number of 1's, 2's, etc., with a large enough sample (and assuming no one has rigged the die) the "average" is 3.5 (which doesn't correspond to any of the actual measurements, the median is also 3.5, and there's no true mode since all values have the same frequency. So, is it really beneficial to rely on stats that presume a single set of behaviors?

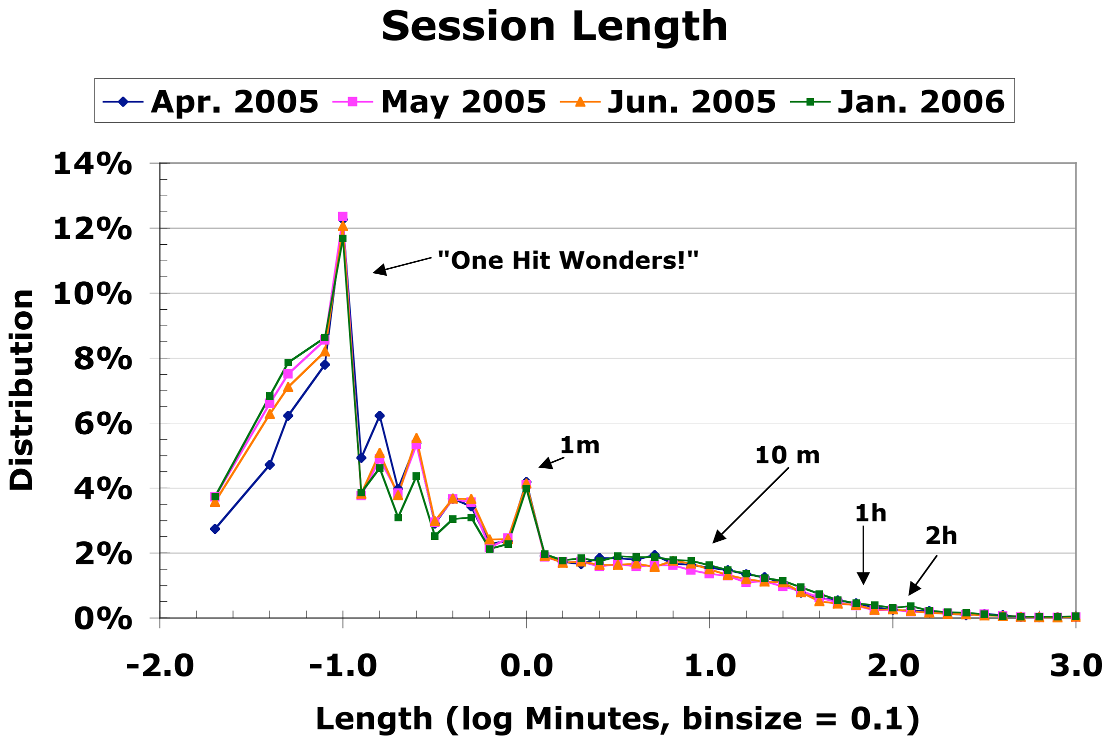
As a check I went back to the data and immediately realized that something was horribly amiss! While the mean of the data was about 10 minutes, the *median* was far far lower indicating that my distribution was not much of a nicely shaped Gaussian. One possibility that could produce such a disparity would involve outliers on both ends skewing the distribution such that the mean just happened to lie where our expectations had placed it but with a distribution nothing like what we had assumed.

What is the actual distribution?

Every login to the TD site is logged in our database with a unique ID that carries through the session. By looking at the time of login and the time of the last record in the Apache server log I could construct a value for session duration. However, this value comes with an inherent flaw: it's always an *underestimate* of the true duration because there's no way to track the length of time the user spent on the last page they visited.

Here's what I found sampling from four different months[2]:

[2] Why logarithmic units? Partially to compress the time scale and binning, and partly because we tend to characterize things like session length terms of time scales rather than the actual values themselves, e.g.: seconds (a "short" visit) vs. minutes (a "regular" visit) vs. hours (a "long" visit).



Clearly this is not a simple Gaussian-shaped curve. The jaggedness on the left side of the distribution (1 minute and shorter) is due to the binning of the data. Longwards of 1 minute the distribution evens out and does have that Gaussian-like appearance that one might expect from a single population. Note that the peak of the long "hump" is just short of 10 minutes (but remember all values in this distribution are underestimated in time). These are most likely the users that we had in mind when we developed the site i.e., the "window shoppers": users who come to the site and peruse the resource offerings. Another thing to notice is that the curves change only slightly from month to month[3] – a reassurance that what I'm seeing is representative of the actual user behavior given the incomplete sample.

What can we conclude about this distribution? Well, first, there is more than one population of TD user – something that hadn't been considered when we planned the site, and this newly-identified group is significant in size. But who are they, and are their needs being met?

[3] The April 2005 distribution is slightly different at short session lengths. This was the first month that our Test Drive system was in place. Comparing the behavior of test drivers and non-test drivers is a series of tests that have not yet been completed.

Their short session length gives a clue: the peak is at around 6 seconds for the session (minus the time spent on the last page!). Checking a handful of these sessions manually I found that they tended to come from external searches (e.g., Google, Yahoo!, etc.) with very specific search terms (e.g., "carbon cycle diagram") with a search result that guided them directly to the TD resource on that topic. Thus, it's probable that as opposed to the "window shoppers" these are the "bread, eggs, and milk" users: they have a very short shopping list, aren't interested in "window shopping" a site, and want to locate something, quickly evaluate it, and then bookmark, copy, or use it and move on to the next thing on their "to do" list.

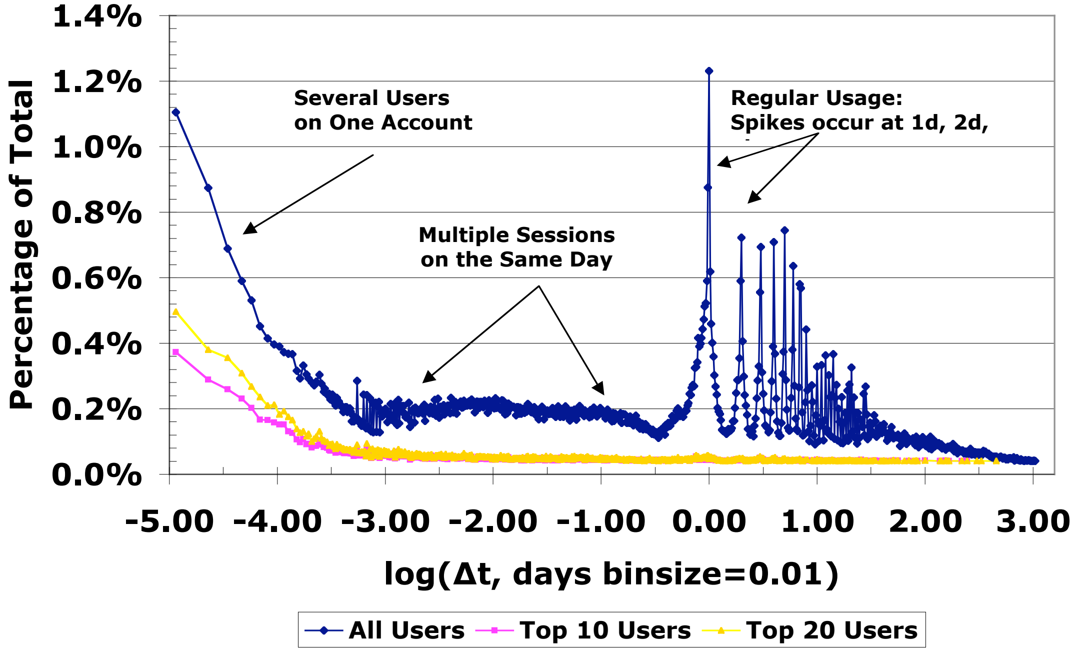
What I learned: there can be more than one population of users. But beware: their definitions of "success" regarding the site might not be the same as the model used in its construction – something to consider when contemplating subsequent future development in order to incorporate potentially different (and possibly disparate) needs.

4. Case Study 2: Why Won't Users Behave and Make My Life Easier?

Another statistic I generate is login frequency. This one is a little more robust than session length because: 1) it only counts registered users and in particular those who visit the site multiple times, which can taken as indicative of interest; and 2) we know definitively the time and date of the beginning of the session even if we don't know how long the user spends on the last page visited. Here, the time between each pair of successive logins by the same user is counted individually across all users (excluding Test Drivers).

If you thought the last distribution was strange, look at this one:

Login Frequency Histogram



There’s quite a bit of fine structure but things break down into three regions:

1. Several “spikes” starting at 1 day, at one day intervals. If you blow this diagram up, you can follow this pattern out to nearly a month. The location of these spikes suggest that when many users access TD, they do so around the same time of day (although not every day);

2. A broad “hump” between a few minutes and ~12-18 hours. These are the situations where people log in two or more times a day and the lack of finer structure suggests that these inter-access intervals are more random in nature;

3. Shortward of a few minutes, there’s a huge “tail” down to the 1-second level!

While I was surprised at the character of the distribution, it’s the short-frequency tail that blew me away. It meant that many users were re-logging on every few seconds. Could this be some weird bug that was causing session IDs to be blown away[4]? Since each user’s session was added to the distribution individually, it finally occurred to me to see what the pattern is for the heaviest users of TD (determined solely by their total number of logins).

[4] A “dead end” thought was that perhaps this could be explained by users having cookies turned off in their browser, but if that were the case, then the session ID wouldn’t be set and wouldn’t appear in the login table used to generate the diagram.

What I found in the login record is that the same user's account would begin a session ~20-30 times within the same 1–2 minutes on the same day. The distribution from just the top 20 users (representing a mere 0.028% of the entire user sample) accounts for almost 50% of the observed short frequency tail!

What behavior explains this? One likely candidate: a teacher gives out her TD password to her students who all log on to TD in the computer lab at the same time to do independent work on the site. Note that this also can account for some of the short-length sessions above if the instructions given to the students had them go to a specific URL to view a single resource on the TD site.

What I learned: users don't behave the way our model expected them to. More importantly, this discovery reveals implications which potentially affects the analysis and interpretation several other stats: e.g., 1) student users are underestimated; 2) data concerning our "Top" users (based upon number of logins) is heavily corrupted by these other usage patterns; 3) site protections/features based on user "type" (e.g., teachers vs. students) may not be as effective as we want if teachers are letting students use their accounts, and so on.

5. Getting Back to Impact

How does this relate to educational impact? In order to make a statement about impact there are two things that ought to be known:

1. Who are my users? Can the important populations be identified?

2. What is their range of behaviors? Are their needs being accommodated within the site design (user interface, etc.)?

3. How does unanticipated behavior affect other metrics, especially those most important to measuring impact?

However, what these two case studies clearly show is that it's not always straightforward to understand exactly what the user populations are at any given time, and that their perception of how to "best" use the site might not align with the design expectations.

Therefore, how can the situation be improved?

1. Learn as much about user populations as possible. In the case of TD, since we require registration we can collect user metadata as part of the registration process. This helps to further identify trends and to correlate with external source of demographic data (interim results on this topic will be the focus of a later report);

2. Learn as much about user behavior. This is more difficult because even a small web site generates huge amounts of server log data and it is extremely difficult to

isolate particular patterns – something I've only just begun attacking. However, in terms of getting started, this is another place where Omniture is particularly helpful as they have an entire section of reports related to path analysis. Here are a few reports I've found useful:

- Full Paths – This report shows the top paths by count or percent. One thing to notice is that you can filter which paths are shown by several criteria: date range, minimum path length, specific entry page, or paths that hit a particular page! (It's these non-default options that contain a lot of power, and tend to be overlooked.)

- Path Length – Although the temptation is to concentrate on hitting a new "record" of path length, the helpful information here is the rate of fall-off of path length and matching the short paths with particular behaviors (e.g., search engine entries).

- Page Flow/PathFinder – These two pages allow you to construct fairly elaborate tests if you're interested in determining how users get from page "A" to page "B" or where they're being led off the expected path. This in particular helped me identify a navigation and UI weakness on the TD site where several users' paths had an odd sequence to it. Upcoming site redesigns will rectify that situation.

In the next report, we'll start to look at what you can do with geospatial information and census demographics, more pitfalls, and why concepts like "billions and billions served" may not be a particularly helpful way to represent usage statistics.