Using Web Metrics to Estimate Impact:
III – Growing Pains?

Bob Donahue (bob_donahue@wgbh.org)
WGBH Interactive
November 2006

Abstract

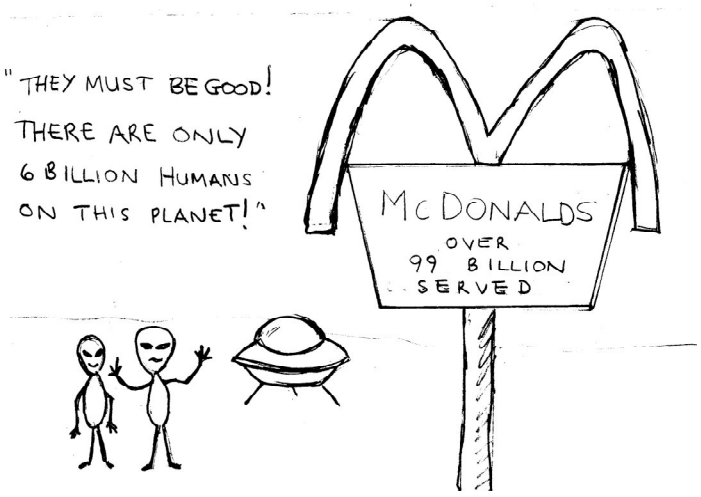
What do you do if there isn't a straightforward way to normalize your web statistics? Inside, we'll look at a few tricks to get around that problem.

1. Introduction

In the previous reports I've placed a great amount of emphasis on shying away from using "open-ended" statistics such as a raw "# of visitors" versus normalized statistics in which the added perspective greatly increases their worth. However, this isn't always practical, producible, or even possible depending upon the nature of the statistics and the availability of the information needed to provide the necessary normalization.

So sometimes, what's needed is essentially a way to compare the data set to itself. For data in a time series (i.e., where the *x* axis is some range of time), you can flip things around so that what you're looking at is a measure of growth instead of counts.

2. Yet Another Diatribe Against Raw Counts



Outside of many McDonald's, they tell you the numbers they've served is in excess of 99 billion. If that refers to customers, it must include repeat visits, unless they've used a time machine to reach most of their clientele who are primarily space aliens. And while the number sounds impressive, it doesn't really say much in terms of impact. Or, it might refer to burgers, but then it doesn't say who the recipients are (or if they're all human…). Or it might refer to the number of times a transaction occurred, but then that might or

might not include orders that don't include burgers… In short, I don't know what it means, but it certainly sounds like a huge number and I suspect that that's truly the only message, left entirely to the interpretation of the viewer of the sign with the expectation that there are very few things with that many zeroes in it to use for comparison.

3. "What Will I Ever Use This For?" — Well, Now You Know!

So back to "real world" metrics: here's where we get to have some fun with math. Dust off your old algebra textbook and turn to the chapter on logarithms. Believe it or not, the answer is hidden in there.

One of the properties of logarithms is expressed in the *folding time* of a process. One example that you've heard of is radioactive decay, which is the basis for (among other things) carbon dating using the half-life of isotopic ratios[1]. The formula looks like this:

$$N \propto ke^{-t/\tau}$$

where N is the count of something, t is the time index, τ is the folding time, and k is a scaling factor. OK, so this is great if you "think" in base e instead of base 10. However, the way to understand this equation is this: τ represents the length of time it takes for N to change by a factor of e (or whatever base you're using). Put another way:

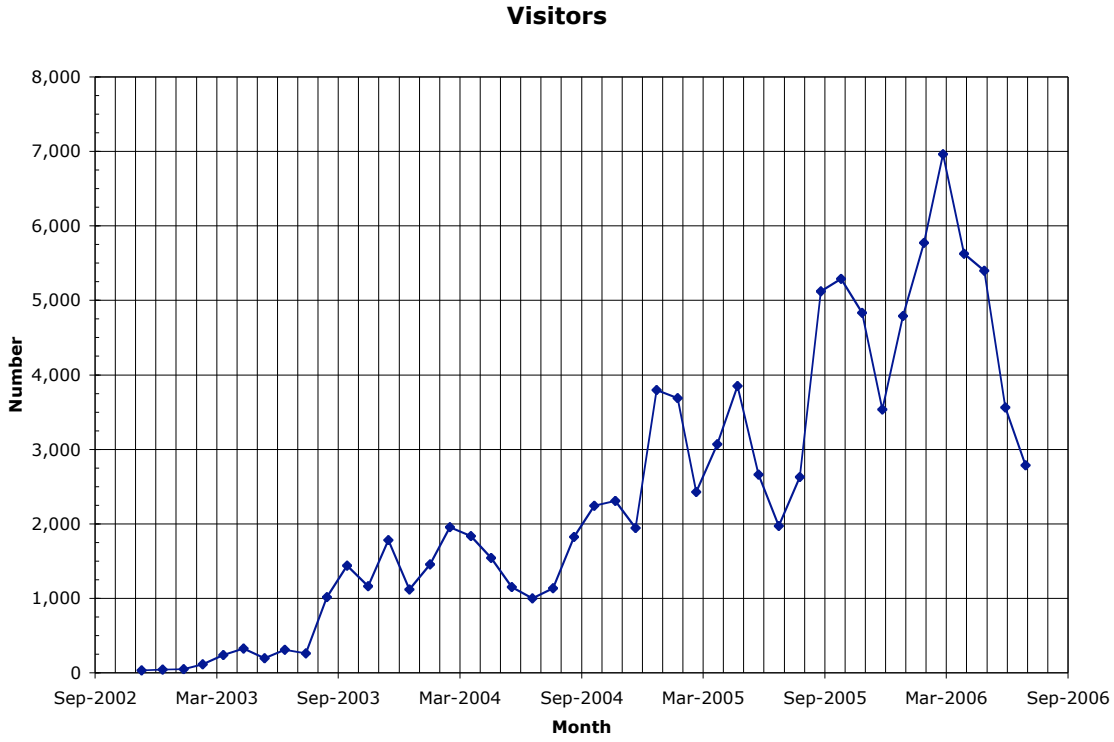
$$\ln N = \ln k - t/\tau$$

which now resembles a linear equation with intercept $\ln a$ and slope $-t/\tau$.

The situation for our purposes has nothing whatsoever to do with decay but it works for growth too: all that really changes is that the minus sign in the exponent turns to a plus sign. But it still has that pesky "e" in it. What they didn't probably stress in those algebra classes is that the same equation "works" no matter which base (e.g., 10) is chosen (although it will affect the value of time scale, τ). So, what's a good number to use? Strangely enough, 2 is a great choice for our purposes, but it never gets as much attention as 10 or e, despite making things a little easier to comprehend (and more importantly, easy to explain to your audiences).

Let's take one of our regular "open" stats: visits per month. The time series might look like this:

[1] The term "half life" sort of gives it all away…



While there is obvious long-term growth, there's also plenty of variation (mostly owing to school vacation schedules). While resisting the temptation to fixate on the scale of the *y* axis, how can we process the data to produce a useful, meaningful, and easily understandable message? Let's go back to that formula: if we take the logarithm of both sides it ends up looking like this:

$$\ln N = \ln k + t / \tau_e \,, \text{ or in base 2, } \log_2 N = \log_2 k + t / \tau_2 \,,$$

because of the handy identity $\log_x N = \log N / \log x$, and remembering that the τ's will be different depending on which base is used. OK what does this get you? Well, now you've got a simple ratio on the right hand side of the equation such that a difference of 1.0 on the left indicates the *doubling time* for *N*. Put another way, solving for τ tells you how fast *N* is growing over time. So, let's convert the above figure such that we show the $\log_2$ number of visits over time:

You can still recognize the same fluctuations from the previous diagram (particularly the summer vacation dips), but now a few things are clearer. First, there are definitely *two* eras of different growth: a "ramp up" period from late 2002 (when the site was first launched) through the summer of 2003, after which there's a sustained and persistent growth to the present. What's that growth? Well, the equation above has the same form for a straight line where the slope is $1/\tau$, or that the doubling time, τ, is the inverse slope. After a little linear correlation homework (and some statistics) you end up with these results:

Inverse slope = "doubling time" = 15 ± 2 months.

Plainly, "the number of unique visitors per month doubles over the space of about 15 months." For completeness, here are the formulae you need to get the value for the error bars for the slope σ_a and intercept σ_b in the linear equation $y = ax + b$, where a is the slope and b the intercept:

$\sigma_a^2 = N\,\sigma^2 / \Delta$;
$\sigma_b^2 = \sigma^2\,\Sigma x_i^2 / \Delta$, where N is the number of points in the sample. Δ is:
$\Delta = N\,\Sigma x_i^2 - (\Sigma x_i)^2$, and σ^2 is:
$\sigma^2 = \Sigma\,(y_i - a - bx_i)^2 / (N - 2)$.

But that's the error bar for the *slope* and not the inverse slope. To get that, there's a little additional footwork dealing with propagation of errors[2]. Here, the inverse slope is $t = 1/a$ and the corresponding uncertainty in t is (without showing its derivation)[3]:

$$\sigma_t = t\,\sigma_a / a.$$

3. Not Just a One-Trick Pony

This process also works for cumulative data. How about total visits, including repeat visits? In this case, repeat visits by the same users in the same time period are counted separately.

Again, we have a ramp-up period and then a sustained growth of total visits. Performing the same analysis (from 9/03 to the present), we find a doubling time of 5.8 ± 0.1 months. The "story" here is "How long does it take to log twice as many visits as I've had *up to* the present time?" whereas the previous calculation asks "How quickly will be getting

---

[2] My favorite reference for this sort of "stuff" is Bevington and Robinson's "Data Reduction and Error Analysis for the Physical Sciences" (ISBN 0-07-911243-9).

[3] In general, there are partial derivatives involved for every variable in the equation. I should point out that this formula *only* works for this specific equation; other ratios with different numerators have slightly more complicated reduced equations for the uncertainty.

twice as many visitors as I did last month?" (assuming nothing changes on the site to affect growth).

Note that this puts different sites within the same perspective. A "new" site (or part of a site) might not yet have the number of visitors that a well-established site does, so the overall numbers aren't easily comparable. Here, you can test different sites (or parts of sites) against each other for comparative growth (or stagnation) regardless of the actual numbers.

4. Comparing Growth

So, once I jumped on the "growth bandwagon" I started finding many places to apply it in my metrics. Here's a final example showing the comparative doubling times for teacher registrations to the Teachers' Domain site (http://www.teachersdomain.org) throughout the US:

Even though our "top" five states in terms of total registration are CA, NY, TX, MA, and FL[4], registration is proceeding most quickly in other states: AK, IA, and KY. This metric, along with penetration can alert our marketing gurus to see where past efforts have had the most traction (if they were targeted to specific areas) as well as indicate opportunities for future targeted marketing efforts.

[4] That's not surprising since they also have the highest populations (or electoral college votes since that's also population driven) by state.