Using Web Metrics to Estimate Impact:
II – When Counting Doesn't "Count"

Bob Donahue
WGBH Interactive
August 2006

Abstract

In this article we look at how the use of "counts" for metrics can be greatly improved by providing context tailored to the needs of the project.

1. Introduction

In the last article, we looked at two stats: session length and login frequency to see how their actual distributions provided greater perspective to user populations and behavior, primarily noticing where simple statistics don't produce a picture that is entirely accurate. This time, we'll look at a few stats related to "counts" showing how the addition of contextual information can be helpful to interpreting those results.

At this point I'm also going to veer away from depending on stats that are available to (almost) any web site data miner. This is because TD requires registration[1] and metadata obtained from the registration process provide a substantial boost to the data mining process and the results that can be obtained. In particular, a user's "account" is in two parts: a username chosen by the user but also associated with an "organization" – typically a K–12 school. The primary reason for this setup is to eliminate the problem of "user ID burn" where as time passes all the good names are taken and users who register later than earlier have to come up with extremely cryptic names to satisfy the uniqueness requirement. Instead, for TD, the uniqueness is the combination of username and organization so there can be as many "jsmith" accounts as there are organizations in the database.

Another boost to data mining is that we were able to "seed" the organization table with as many education institutions as we could find. This provided us with a great deal of geographic information and metadata that has become extremely useful for demographic analysis (see below).

2. "How Many…?" Style Questions

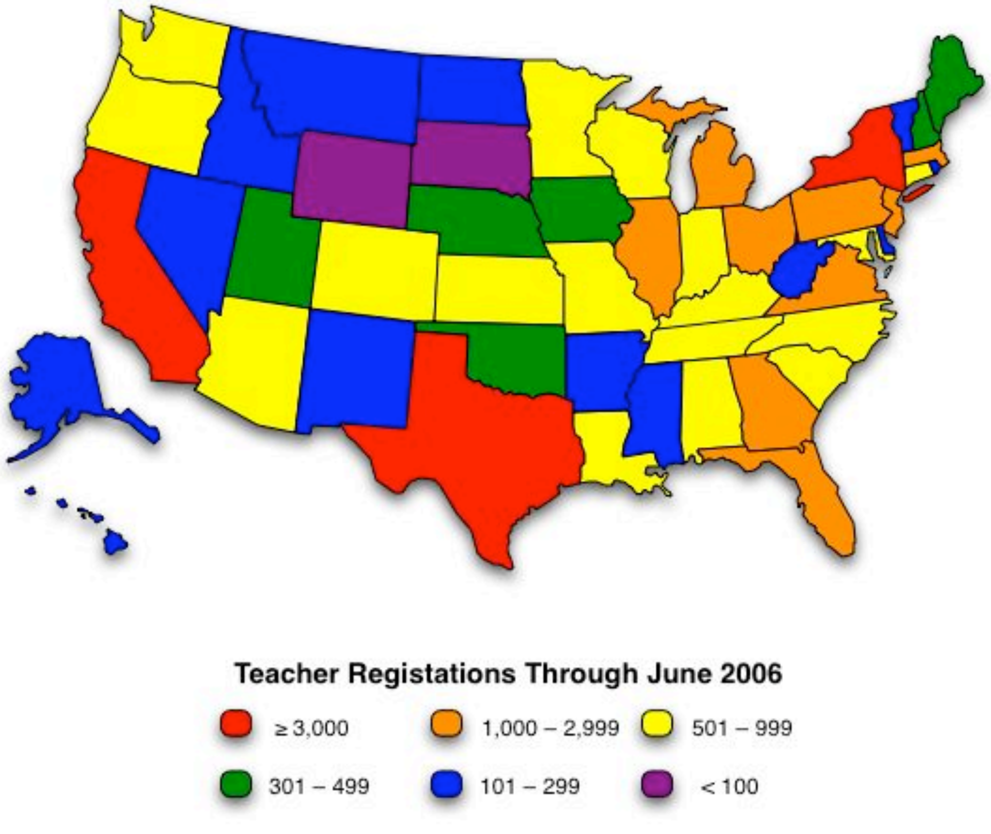
[1] Why? Primarily because the rights clearance for segments of the multimedia files require that the users be from the K–12 community. In order to demonstrate this, registration is required.

"How many users do we have?" "How many page hits did we get last month?" "How many…?" are questions that I frequently receive for the purposes of adding quantitative information to some report or presentation.   Whenever I get these requests, I'm always a little uneasy because it's impossible to know what the requestor's expectations are compared to comparative numbers from other sites.

This is why context is so important.   McDonalds claims to have served "billions and billions" despite the limitation of the actual population of planet Earth (and while there seems to be a franchise on nearly every corner, I rather doubt that Big Macs are much of a staple in Ulaan Bator) and their "stats" *probably* count repeat visitors separately.

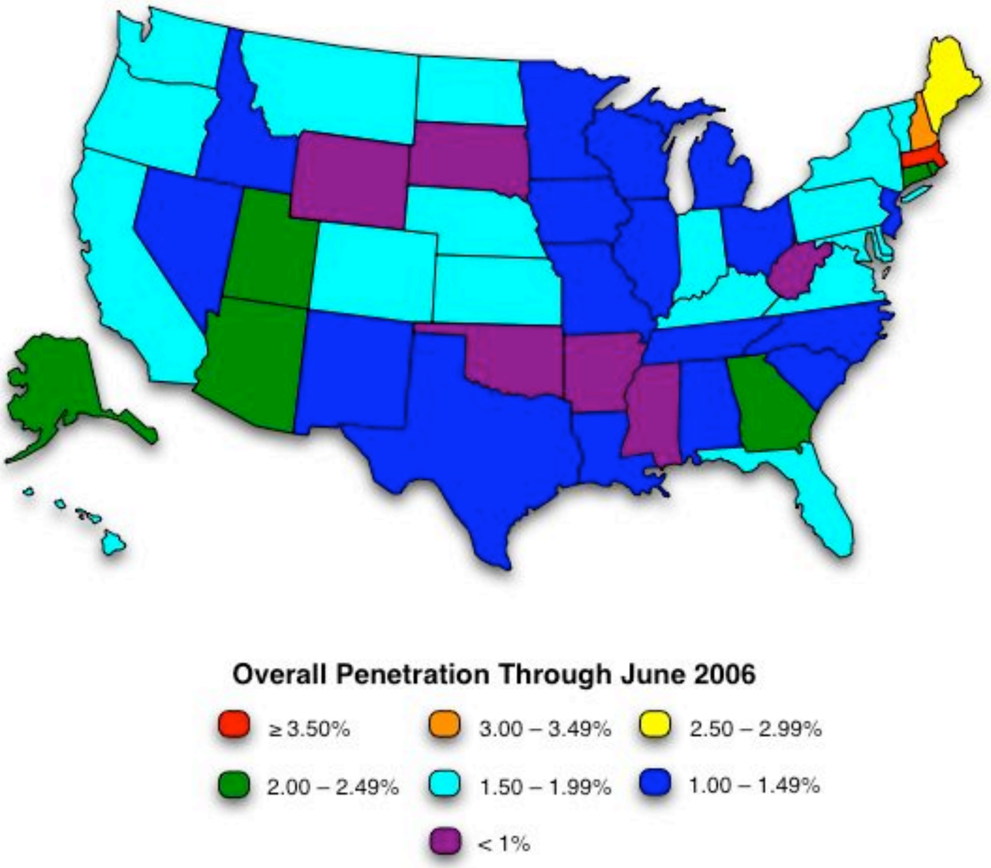
3. Case Study #1: Penetration Versus Counting Within a Market/Population

The first example is a typical one: the number of registered teachers by state.



While there's nice variation in the "tapestry" with a little examination it becomes apparent that there's not much news here because the values are heavily influenced by other factors — primarily the populations of teachers in each state. This makes it hard to ascertain where the "interesting" results might be since you'd expect that the states with higher (or lower) populations should have correspondingly high or low registration rates.

So, first, let's try to minimize the population effects by normalizing by the *total*[2] number of teachers in each state:

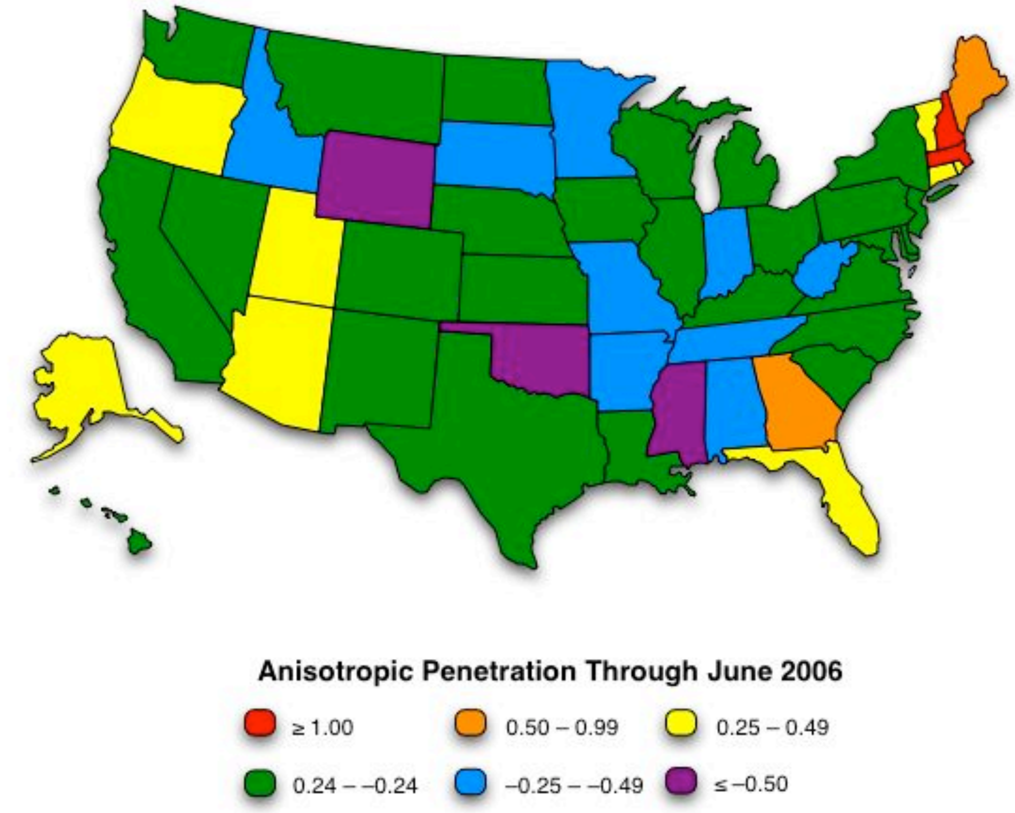


This definitely changes the map with a significant "hot spot" appearing in the New England states reflecting that the initial marketing efforts for TD were concentrated there. But now a few states stand out: Alaska, Utah, Arizona, and Georgia in particular. Note that some states, like Oklahoma and Texas which had what appeared to be a higher than expected number of registrations actually ends up being an under performer in actual penetration.

We can take this a few steps further. First, let's introduce a new metric: anisotropic penetration. This is a fancy term for comparing two normalized maps by dividing one

[2] Note that the normalizing data is not the number of science teachers but the total across all disciplines. This is important because Teachers' Domain like many collections within the NSDL is centered on science. Therefore, the penetration values will always be low because the normalization is taking place with a value that includes teachers who aren't within the targeted user population. In the case of TD, as it expands to include more curriculum disciplines, this normalization will become more appropriate.

by the other. For the map below, it's the percentage of teachers in each state of the US total (so each state's value is their contribution to the whole) normalized by the percentage of TD teachers registered in each state to the US total. The distribution is then re-centered by subtracting 1.0 from each result so that a value of zero indicates the rate of registration is mostly aligned with the fraction of teachers in that state. Thus, positive values indicate a greater than expected registration rate, negative values a shortfall. Here's the final map:



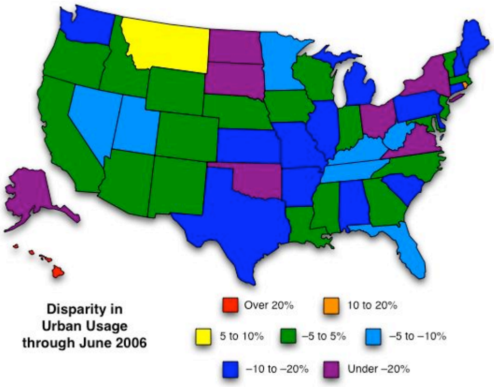
Now we're finally able to clearly see performance/impact in terms of how well we've been able to reach teachers across the US at the state level. Once again it's clear that Alaska, Utah, and Georgia are doing well, but now Florida, Arizona, and Oregon also have relative registration rates that are above the norm. Notice that the three states with the highest number of registered teachers: California, Texas, and New York actually are in the middle of the pack once the data are distilled, showing that their higher relative user counts should've been anticipated based upon their relatively high populations *and* number of teachers there.

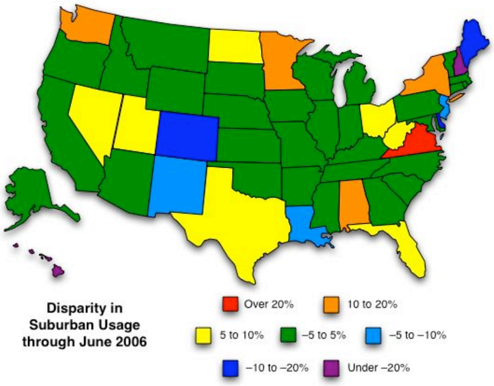
But wait! There's More!

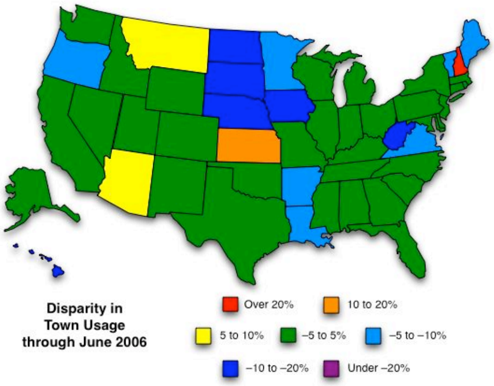
As I mentioned above when we first set up TD to make registration slightly easier we attempted to seed the organization table with as many K–12 schools as we could. The data used to do so also came with quite a bit of metadata that we greedily ingested into our database. Among the data provided was a classification of each school's geographic locale broken down into eight categories:

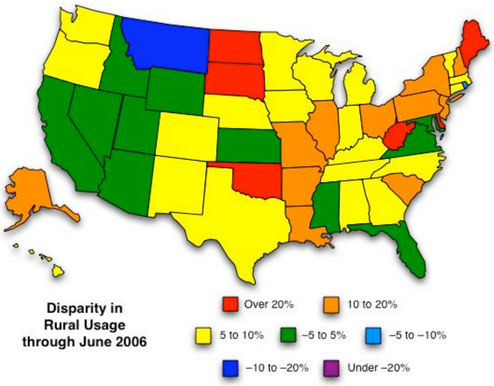
1. Large Central City (pop. ≥ 250,000)
2. Mid-Sized Central City (pop. ≥ 25,000)
3. Urban Fringe of Large Central City
4. Urban Fringe of Mid-Sized Central City
5. Large Town
6. Small Town
7. Rural, Outside Metro
8. Rural Inside Metro

We can apply the same analysis to the teachers' schools to see how our penetration varies by locale type. To simplify things a bit, I'm combining the cities, urban fringes (suburbia), towns, and rural areas so that we have only four locales.


Disparity in Urban Usage through June 2006


Disparity in Suburban Usage through June 2006


Disparity in Town Usage through June 2006


Disparity in Rural Usage through June 2006

By comparing these four maps, we find some very interesting trends:

1. First, it's clear that our penetration in more rural areas is generally better than in urban areas (although Montana is the sole exception in both the urban and rural

maps!);

2. Things are flattest at the town level, although hot spots still exist (New Hampshire and Kansas);

3. States whose overall penetration disparities are extremely high or low often show concentrations in a particular locale. For example, Oklahoma, which has a low overall penetration, still has one of the highest relative penetrations in rural areas.

4. Only five states have little disparity in any locale: California, Idaho, Maryland, Mississippi, and Wyoming[3].

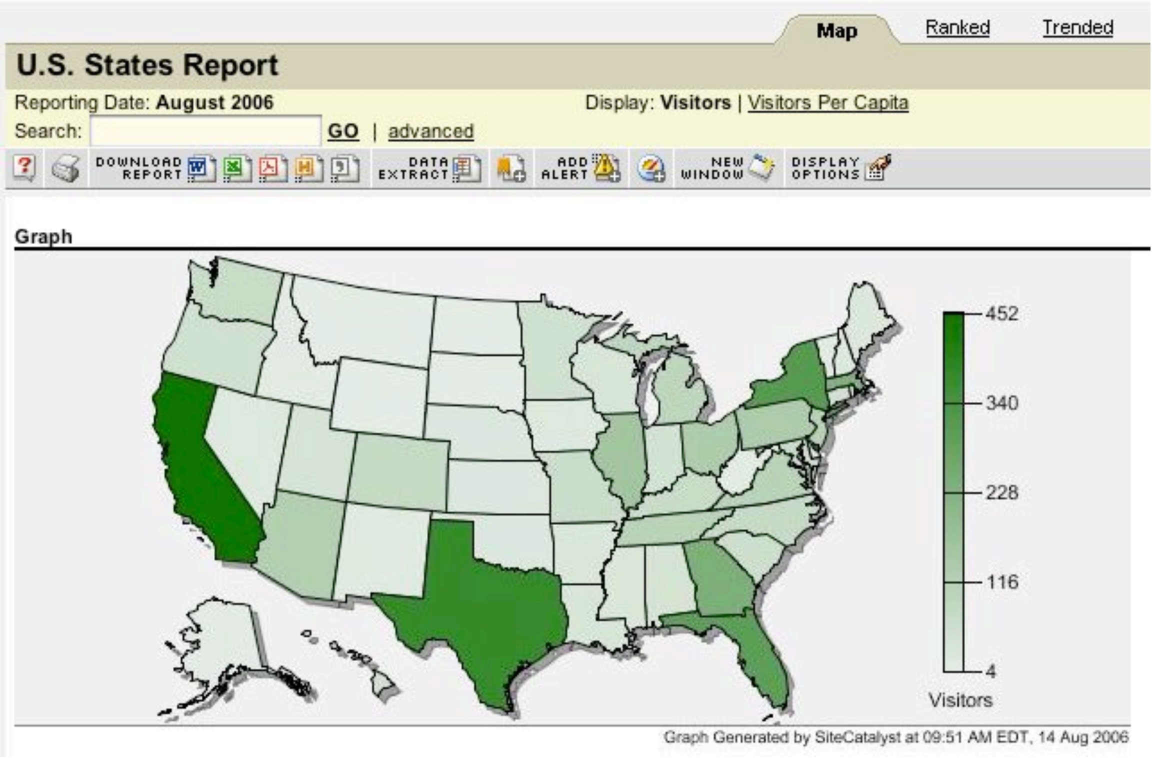
How does this help? Well, it's a qualitative measure of impact and shows trends in populations that identify whom you've had more success reaching, and who is lagging behind. An important thing to realize is that these maps show that for many states, the "state as a whole" information provides a really fuzzy picture because going one level deeper reveals distinct differences.

4. That's All Well and Good, But I Don't Have an Organization Table!

You're in luck! There's a new feature in Omniture's Site Catalyst that mimics this for visits. Under "Traffic" then "GeoSegmentation" then "U.S. States" you can get a map showing the number of visitors from a particular state[4].

[3] This begs the question: Why? Could it be that there are programs in those states that provide resources across the entire state (technical infrastructure, etc.)?

[4] How does it know where the users are? It's an educated guess based upon the IP number of the user's computer. AOL users are counted differently, and I imagine that other nation-wide ISPs probably have "fuzzier" data than very local ISPs.



U.S. States Report

Reporting Date: **August 2006**

Display: **Visitors** | Visitors Per Capita

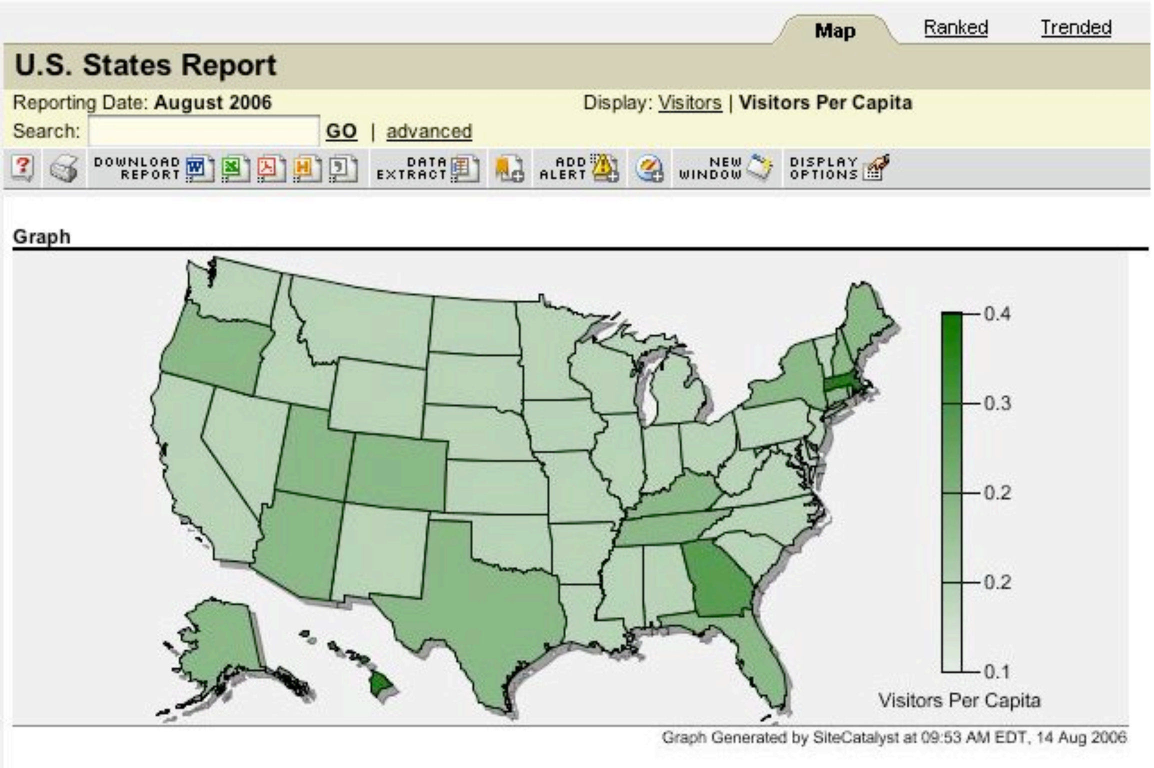
Search:

Graph

Graph Generated by SiteCatalyst at 09:51 AM EDT, 14 Aug 2006

However, this suffers from the same problem as before: the absolute count isn’t helpful because each state has a different population. So you’d expect that TX, CA, etc. would out-weigh WV, HI, ID, etc.

Fortunately, there’s a “Display: Visitors per Capita” option which mitigates the issue:


U.S. States Report

Reporting Date: August 2006

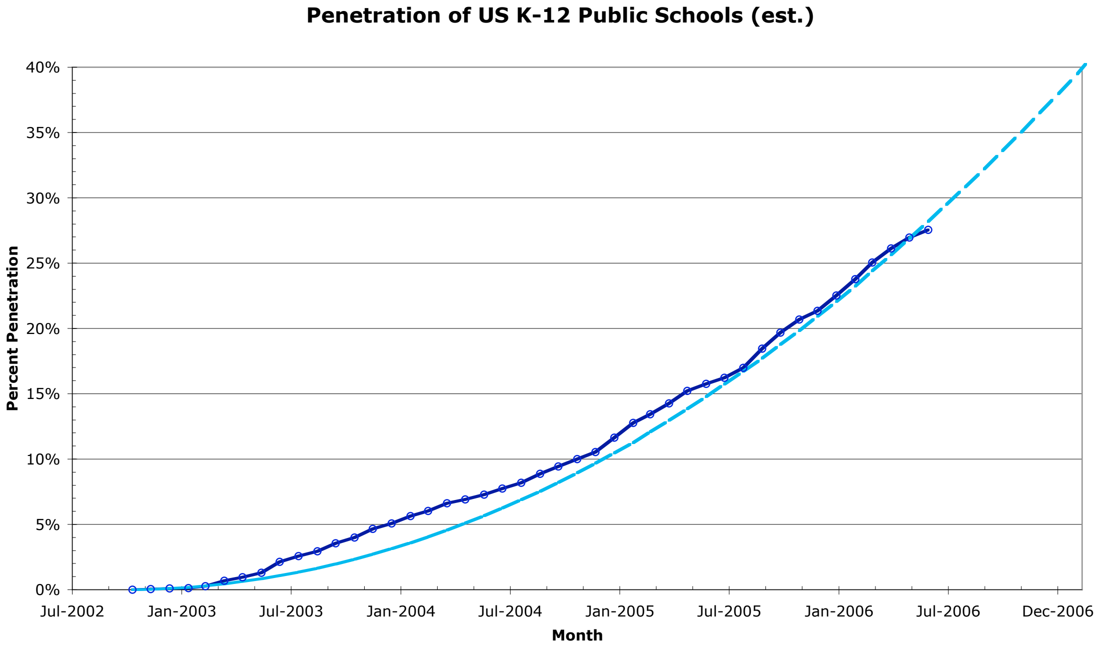
Display: Visitors | Visitors Per Capita

Graph Generated by SiteCatalyst at 09:53 AM EDT, 14 Aug 2006

As before, this “flattens” things out, and reveals different hot spots. In this case, over the first half of August, Georgia seems to have higher-than-average traffic, as does Hawaii!

5. Case Study #2: Counting Against a (Mostly) Known Population Size

Another great side effect of the seeded organization table is that it gives me a good idea of the number of K–12 schools in the U.S. Since all the registered users have to associate themselves with one of them, I can track what percentage of U.S. public schools have TD users.



This is the time series of U.S. K–12 public school penetration since the initial launch of Teachers' Domain by month. The dashed line isn't a fit to the data – it's a model of growth that's similar in characteristics to the observed growth as a comparison. In any case, as of June 2006 we can account for at least[5] 25,000 schools. That's just over 27 1/2% (11 out of every 40) of the entire set of U.S. public schools. This normalization provides extremely helpful context in reporting since without knowledge of how many schools there are in the sample, the number 25,000 could seem high or low whereas the percentage lets you know exactly where you stand. Furthermore, the change over time shows the degree to which growth has been sustained and if it correlates with changes to the site.

This is a good example of why you simply cannot "tell the story" with a single number as a bullet point in a Powerpoint™ presentation. The plot not only shows the data in an easy-to-understand context, but you can also follow the trend with time and correlate changes with project milestones.

6. Coming Up

In the next installment, we'll start thinking about user behaviors and how to identify what that means for impact and assessment.

[5] It's "at least" because schools that were added after the seeding do not have the necessary metadata to distinguish whether they are private schools which aren't included in this particular metric.