

STERLING DESIGNS

Helping You Link It All Together

Cellular: (269) 207-6007 E-mail: webmaster@sterlingdesigns.net

Considerations on Website Statistics

(adapted from Stephen Turner's "How the Web Works")

Many people have incorrect ideas about what can and cannot be calculated in relation to website statistics. The incorrect perceptions of these folks are not helped by site statistics programs which claim to calculate things which cannot really be calculated, only estimated. The simple fact is that certain data which we would like to know and which we expect to know are simply not available. And the estimates used by other programs (ones we don't use) are not just a bit off, but can be very, very wrong. For example, if your home page has 10 graphics on, and an AOL user visits it, most programs will count that as 11 different visitors!

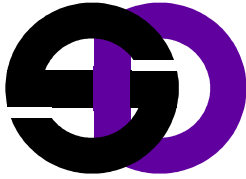
The Basic Model of How the Web Works

When someone visits a website, his or her computer makes one request for the homepage. The owner of the site will know the date and time of the request and which page was requested (the homepage), and the internet address of the person's computer (his or her host). Usually you will also know which page referred the person to your site (for example, a search engine like google.com), and which browser (like Internet Explorer, Netscape, Mozilla, etc.) The person was using. You do not know any usernames or email addresses for the person.

Next, the person's browser looks at the homepage to see if it has any graphics on it. Your site, like most other sites, does have graphics; however some sites do not have graphics. If the person has "image loading" turned on in his or her browser, the computer makes a separate connection to retrieve each of these graphics. The person never logs into your site: the computer just makes a sequence of requests, one for each new file it was necessary to download. The referring page for each of these graphics was your homepage. Let's say there were 10 graphics on your homepage. So far 11 requests have been made to your server (the place where the information necessary to display your site lives).

After all that, the person visits some of your other pages, making a new request for each page and graphic to download. Finally, the person follows a link out of your site. You will never know about that at all. His/her computer, like everyone else's, just connected to the next site without telling you.

Unfortunately, it's not always quite as simple as I just explained above. One major problem is caching. There are two major types of caching. First, a person's browser automatically caches files when they are downloaded. This means that if a person visits your site again, the next day for example, he or she doesn't need to download the whole page and all the graphics again.



STERLING DESIGNS

Helping You Link It All Together

Cellular: (269) 207-6007 E-mail: webmaster@sterlingdesigns.net

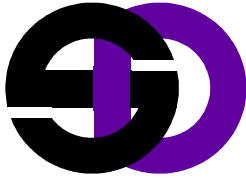
Depending on the browser settings, the person's computer might check with your site that the page hasn't changed: in that case, you will know about it, and the statistics program will count it as a new request for the page. But if the person set his/her browser not to check with your site then the page is read again without you ever knowing about it.

The other sort of cache is on a larger scale. Almost all Internet Service Providers (ISPs; like AT&T, Earthlink, Iserv, etc.) now have their own cache. This means that if someone tries to look at one of your pages and anyone else using the same ISP has looked at that page recently, the cache will have saved it, and will give it out without ever telling you about it. (This applies whatever the browser settings are.) So hundreds of people could read your pages, even though you'd only sent it out once.

Think of caching this way: If I buy a case of pop at Sam's Club, they record that as one purchase. When I drink those cans of pop at home, Sam's doesn't record each one as a sale again. This is like the cache in my browser. Or if I were to sell each can of pop to a friend, Sam's still recorded only one sale. This is like the cache on the ISP's server.

So, your site contains about 12 HTML pages and about 20 graphics. If someone, in visiting your site, viewed all 12 pages and all 20 graphics, this would show up as 32 requests for files. But, you cannot know many things.

- *You can't tell the **identity** of your readers.* Unless you explicitly require users to provide a password, you don't know who connected or what their email addresses are.
 - *You can't tell how many **visitors** you've had.* You can guess by looking at the number of distinct hosts that have requested things from you. Indeed this is what many programs mean when they report "visitors". But this is not always a good estimate for three reasons. First, if users get your pages from a local cache server (as discussed above), you will never know about it. Secondly, sometimes many users appear to connect from the same host: either users from the same company or ISP, or users using the same cache server. Finally, sometimes one user appears to connect from many different hosts. AOL now allocates users a different hostname for every request. So if your home page has 10 graphics on, and an AOL user visits it, most programs will count that as 11 different visitors!
 - *You can't tell how many **visits** you've had.* Many programs, under pressure from advertisers' organizations, define a "visit" or "session" as a sequence of requests from the same host until there is a half-hour gap. This is an unsound method for several reasons. First, it assumes that each host corresponds to a separate person and vice versa. This is simply not true in the real world, as discussed in the previous paragraph. Secondly, it
-



STERLING DESIGNS

Helping You Link It All Together

Cellular: (269) 207-6007 E-mail: webmaster@sterlingdesigns.net

assumes that there is never a half-hour gap in a genuine visit. This is also untrue. I quite often follow a link out of a site, then go “back” in my browser and continue with the first site from where I left off. Should it really matter whether I do this 29 or 31 minutes later? (Incidentally, cookies don't solve these problems. Some sites try to count their visitors by using cookies. This reduces the errors. But it can't solve the problem unless you refuse to let people read your pages who can't or won't take a cookie. And you still have to assume that your visitors will use the same cookie for their next request.)

- *You can't follow a person's **path** through your site.* Even if you assume that each person corresponds one-to-one to a host, you don't know their path through your site. It's very common for people to go back to pages they've downloaded before. You never know about these subsequent visits to that page, because their browser has cached them. So you can't track their path through your site accurately.
- *You often can't tell where they **entered** your site, or where they found out about you from.* If they are using a cache server, they will often be able to retrieve your home page from their cache, but not all of the subsequent pages they want to read. Then the first page you know about them requesting will be one in the middle of their true visit.
- *You can't tell how they **left** your site, or where they went next.* They never tell you about their connection to another site, so there's no way for you to know about it.
- *You can't tell **how long** people spent **reading** each page.* Once again, you can't tell which pages they are reading between successive requests for pages. They might be reading some pages they downloaded earlier. They might have followed a link out of your site, and then come back later. They might have interrupted their reading for a quick game of solitaire. You just don't know.
- *You can't tell **how long** people spent on your site.* Apart from the problems in the previous point, there is one other complete show-stopper. Programs which report the time on the site count the time between the first and the last request. But they don't count the time spent on the final page, and this is often the majority of the whole visit.

It is important to keep these things in mind when reading statistics on a website's performance.

Stephen Turner's original article "How the Web Works" can be found at <http://www.analog.cx/docs/webworks.html>
