

Web analytics is growing more sophisticated. We're developing methods for understanding customers, predicting trends and assessing ROI. Every month analytics gurus amaze you with the latest revelations to sharpen your focus and tune your spend. What no one is telling you is that all these systems and numbers are based on inaccurate numbers. The god of web analytics has feet of clay -- 100 percent accuracy is impossible.

Web analysis is based on counting a very limited number of things. People visit websites and read pages. Therefore we can count people, visits and page views. That's all. Financial details are linked to these things, not inherent within them. If I buy PPC from Google, Google is charging me for visits it sends. In other words, it's just counting visits. If all we can measure is people, visits and page views, it's important to understand how accurate we can be about them. The bad news is that we can't assess any of these with perfect accuracy.

We Can't Count Visitors

It's not possible to count people on the web. They don't exist. People don't visit web sites. Their computers do. The exact process is that a browser requests a copy of a page be sent to it from a server. The browser reads that page and uses it to display something on screen. People aren't even reading your site's pages. They're reading what their browser did to copies of those pages. Ask any designer how consistent that process is, then duck.

What few standards there are for web metrics have been laid down by JICWEBS. This is an international body composed of the audit standards bureaus for most countries, including the USA and all European countries. The JICWEBS standard for identifying a unique visitor is that it is the combination of the User Agent and IP address. The User Agent identifies the browser and operating system.

For example, mine is "**Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322).**" MSIE 6.0 tells me it's Internet Explorer 6.0. Windows NT 5.1 tells me I'm running Windows XP. Many people are running IE 6 on XP, so that information alone is not enough to uniquely identify me. I also have an IP address, my internet address. By combining the full User Agent with the IP address, in theory, I am uniquely identified. This is far from accurate.

Every single person inside the Ford corporation has the same IP address. They all go onto the web from the same gateway in Chicago (even the 88,000 in Europe). Corporations hide internal IP addresses for valid security reasons. Most people in Ford have the same browser and operating system (what Ford calls the Global Client). Thus, according to the official standards, more than 320,000 people are the same unique visitor. This will hold true for any corporation with shared internet access and a common standard for their workstations.

On the reverse side, most home users or small businesses will be given a different IP address every time they connect to the web. This means they'll look like a different person every visit. This is OK for unique visitors, but means you're under-counting repeat visitors.

Cookies can improve this, and JICWEBS allows (but does not require) the use of cookies to identify unique visitors. However, people block or remove cookies. This is not just limited to dated, or stored, cookies that last between visits. We've done detailed research into this and estimate between three percent and five percent of all visitors are blocking session-only

cookies as well. The percentage is highest among Unix users, and lowest among Mac users. In other words, the more techie the visitor, the more chance they'll avoid being counted.

What all of this means is that you're probably only getting about 90 percent accuracy with identification of unique visitors.

Monthly stats can be even more misleading on occasion. JICWEBS sets the standard for calculating unique visitors per month when producing audits. The official method for counting unique visitors per month is to calculate how many you got in a single day, and then multiply that by 31. Thus if 100,000 unique visitors came to my news site in a day, I have 3.1 million unique visitors per month. So be aware that "unique visitors per month" may not be counting how many different people actually visited the site that month. Or it may. It depends.

We Can't Count Duration

Think about what happens when someone is reading your site. They ask for a webpage. Then later they ask for another one. The time taken between the two requests is deemed to be spent reading the first page. Add all these durations up and you've got the total time of the visit.

This creates a problem for one-page visits. Since there is no second page, we can't calculate a page duration. Officially a one-page visit is not a visit; it has to be two pages to count as a visit. Some packages won't count the zero duration one-page visits when they determine average visit duration, but you'd be surprised how many do. If you are using one that does, you think people spend half the time on your site they really do.

Now think about what happens when someone reads the last page. There is no duration we can calculate for this page. What this means is that all web analytics packages are under-reporting the time people spend on your site. They have to because they can't tell how long someone spends on the last page -- it never gets calculated.

We Can't Count Visits

The JICWEBS definition for a visit is that it is a series of page requests with a gap of no more than 30 minutes between each one. If someone asks for a page 31 minutes after the preceding page, it must be counted as a new visit. You'd be surprised how often this happens with complex products like mortgages, insurance, and other financial products. Generally, the more detailed the page, the more commitment required to buy, the more chance you'll get the occasional page view that exceeds 30 minutes.

On the other hand, what if someone views your site, goes off and compares it with a competitor, then returns after 20 minutes? That still counts as part of the same visit. Technically it constitutes a single visit of two sessions, but no one follows the differentiation of sessions and visits as the standard allows.

What both of these cases illustrate is that our counting of visits is based on an arbitrary selection of 30 minutes as the magic number. For most purposes this is fine, so long as you accept it is our best attempt at a workable number, not an accurate measurement of reality.

Statistical Probabilities, not Counting

Web analysis is statistics, not accounting. Absolute precision is impossible. The problems listed above are an inevitable consequence of the nature of internet technology, not because we don't care or because analytics software is shoddy.

This inaccuracy is okay so long as you don't get too excited about the fine detail. Statistics is fuzzy around the edges, so you shouldn't make decisions based on small differences. Understand that your visitor stats are accurate plus or minus five percent or even 10 percent. Recognize that people are spending a little longer on your site than you can ever know, or maybe a little less. It depends on what you're looking at.

Add a margin of error to financial and ROI calculations. In statistical analysis, there is the concept of "degrees of certainty," what us ordinary folks call "margin of error." It is possible to calculate this with slightly more precision than guesswork. If you want to get into extreme details with your analysis, you need to start incorporating concepts like this into your numbers.

If you design your processes accordingly, the exact numbers shouldn't matter too much. You are where you are today. You want to improve on this. The key to success is to concentrate on trends over time, not individual numbers.

Technical Issues

We should remember that these stats are gathered, processed and delivered by software. The web analytics software industry is new and far from mature. Web analytics software is not perfect, and it introduces inaccuracies of its own into the process. This section discusses the inaccuracies introduced by the software we use.

Log Analysis Issues

Many people use log analysis to get their stats. Log analysis is much less accurate than page-based tracking. Here's why:

Spiders and Robots

Search engine spiders read your site, so does performance monitoring software. Most log analysis software doesn't distinguish between page requests by humans and page requests by software. This inflates the number of page views dramatically. Since search engines go through pages at a rate of one per second, it can also reduce average visit duration and average page read time. This is okay if you know and adjust for it. If you think all the reported activity is coming from people, but your system is not separating out spiders, then you believe that people are making shorter visits than they really are. I believe this is why most designers think the average visit duration to a website is three to four minutes and the average page duration is about 30 seconds. In reality it's about twice that.

SWF Files

SWF files are flash files. Flash is a problem for log analysis. A flash file can be a complete page. It can also be a simple animation inside a page. So when a log analysis product sees

that an SWF file has been viewed, does it count as a page view or not? Most count it as a page view. If you've got flash animations inside your pages, look at your stats again. If you've got flash as both full pages and as page elements, then I doubt you'll be able to get accurate stats from log analysis at all.

Caching and Cache busting

Most browsers keep a copy of each webpage you read. If you hit the back button they serve you that page instead of bothering to ask the server for another copy. Log analysis misses this because the server never saw the second viewing. This accounts for about 30 percent of all page views. Saving pages like this is called "caching." It's a major problem in online advertising because if an advertisement is cached people may not get paid for delivering it. This is why there is so much talk about "cache busting" technology in ad delivery.

It isn't just browsers that cache. Corporate gateways cache commonly requested pages to save time and bandwidth. ISPs may cache for the same reasons. If you're using log analysis for your stats, you're missing about one-third of your activity. Page-based tracking works from the act of reading the page, so it is "cache-busting."

Wake Turbulence

While page-based tracking may avoid the caching problems of log analysis, it is victim to another back button issue that log analysis avoids. Many people exit a site by repeatedly clicking their back button. Log analysis doesn't pick this up, but page-based tracking does. This means many visits end with a series of one or two-second page views in reverse order from the first half.

There's no official term for this, but I call it "wake turbulence." Most analysis tools don't even recognize this problem, let alone deal with it. It increases the average number of page views per visitor, and reduces the average page duration.

I don't think you can blame software for this one. If you watch people fill in multi-page forms, you'll often see them go back and forth within the sequence, so the system can't automatically eliminate quick views of preceding pages. This would mean you would have to examine the click-streams yourself and decide which were valid reads and which were just wake turbulence. Obviously that's an impossible job. You just have to accept a degree of fuzziness around your stats for visit duration, number of pages read and average page read time.

Time Changes

Think about what happens when daylight savings cuts in. Someone enters your site at 11:45 pm. He or she stays for 30 minutes. Daylight savings starts at midnight and the clocks roll back one hour. The visit finishes at 11:15 pm. At the end of the daylight saving period the clocks go the other way. In this case the 30-minute visit starts at 11:45 and finishes at 1:15.

You'd be surprised how few web analytics packages handle this accurately. In fairness it is a tough one to code for. Some systems can cope with this because they work in GMT, then convert the visit times at the point of reporting. But most use local time. How do they handle this? A surprising number simply discard all the records during the crossover period.

User Resistance

Some of your visitors don't trust you. Some major-name tracking systems are listed as spyware and blocked. Some people block cookies. Some people clean out their cookies regularly. If you are tracking repeat visitor behavior with cookies you have to accept some degree of inaccuracy as people block or remove them.

This is more likely if you have another company gathering and analyzing your cookies. It depends on how they do it. If your site sets the cookie you have less chance of being blocked than if their system sets the cookie. A cookie set by them is called a "third-party cookie." Third-party cookies are much more likely to be blocked than your own. Under some circumstances, in some countries, third-party cookies can even be illegal.

Transversal Losses

Transversal is what you do when you click a hyperlink -- you transverse from one page to the next. Sometimes people click on a link but never arrive at the other end. Browsers crash; they change their mind, and so on.

This is starting to become a source of contention in PPC advertising. Google usually charges me for more visits than I can see arriving in my client's sites. It is usually over by 25 percent or so for my clients. I'm not the only person with this problem. Google believes this is a minor and rare problem, but many users are not so sure.

I have questioned Google on this and they have informed me they use log analysis for calculating click-throughs. I have asked them if they filter out search engines and robots that we know read Google results and follow links. They have replied:

"I would like to reassure you that our system will only count legitimate clicks to your client's ads and will filter out all other traffic. As stated in our Terms and Conditions, we require that all parties using Google AdWords services accept our metrics."

This problem is not unique to Google. It occurs to a greater or lesser degree with all forms of inter-site link activity. This means that your ROI calculations for PPC advertising and affiliate marketing cannot be perfectly accurate, but need to permit a margin of error.

Conclusions

We have to accept that web analytics software is in its infancy. Anyone old enough to remember Pong? Compare that with the latest computer games. Look at your web analytics software and try to imagine the same level of improvement over the next 20 years. Now look at the software you use today again. Understand that this stuff has a long way to go.

We can do great things with web analytics software today compared with five years ago, but we have only just begun. Remember that this is not precision accounting but statistical analysis with flawed software. Work with big margins of error and focus on trends, not detailed numbers. Understand that no matter what you do, a certain degree of guessing is inherent in the process.

Things that throw your stats: necessary uncertainties in Web Analytics
By Brandt Dainow (bd@thinkmetrics.com)

Life's full of uncertainties and web analytics is no different. Somehow we all manage to get by.