

Advanced Web Metrics Whitepaper

Understanding Web Analytics Accuracy

by Brian Clifton (PhD)
Version 2.0, March 2010

Preface

When it comes to benchmarking the performance of your web site, on-site web analytics measurement is critical. But this information is only accurate if you avoid common errors associated with collecting the data – especially comparing numbers from different sources. This white paper is aimed at web managers, digital marketers and webmasters who want to maximise the accuracy of their data.

Originally published in February 2008, this second edition has been completely revised and updated for 2010.

From The Author





Thank you for downloading this free whitepaper. Documents such as these represent the culmination of a huge effort on my part to research, write and update the contents. My hope is to educate and inform so that you become comfortable with your web visitor data, mitigate error bars, and go on to build your analysis hypothesis on solid foundations.


I would greatly appreciate your feedback - either a tweet, blog comment or rating on this whitepaper's companion blog site would be great.

Brian Clifton



 Add your comments on the blog - [Measuring Success](#)

 Follow my interests and thoughts [@BrianClifton](#)

 Join your peers on the [LinkedIn Group](#)

Copyright Statement: All content © 2010 by Brian Clifton - Copyright holder is licensing this under the Creative Commons License, Attribution-Noncommercial-No Derivative Works 3.0 Unported, <http://creativecommons.org/licenses/by-nc-nd/3.0/>. (This means you can post this document on your site and share it freely with your friends, but not resell it or use as an incentive for action.)

Table of Contents

Introduction	4	Why PPC Vendor Numbers Do Not Match	15
How Web Sites Collect Visitor Data	4	<i>Tracking URLs: Missing Paid Search Click-throughs</i>	15
<i>Page Tags and Logfiles</i>	4	<i>Slow Page Load Times</i>	15
Cookies in Web Analytics	6	<i>Clicks and Visits: Understanding the Difference</i>	16
Understanding Web Analytics Data Accuracy	7	<i>PPC Account Adjustments</i>	16
Issues Affecting Visitor Data Accuracy for Logfiles	7	<i>Keyword Matching: Bid Term versus Search Term</i>	16
<i>Dynamically Assigned IP Addresses</i>	7	<i>Google AdWords Import Delay</i>	16
<i>Client-Side Cached Pages</i>	8	<i>Losing Tracking URLs Through Redirects</i>	16
<i>Counting Robots</i>	8	Data Misinterpretation	17
Issues Affecting Visitor Data From Page Tags	8	Why Counting Uniques Is Meaningless	18
<i>Setup Errors Causing Missed Tags</i>	8	Ten Recommendations For Enhancing Accuracy	18
<i>JavaScript Errors Halt Page Loading</i>	9	Summary	19
<i>Firewalls Block Page Tags</i>	9	Acknowledgements	19
<i>Logfiles “See” Mobile Users</i>	9		
Issues Affecting Visitor Data When Using Cookies	9		
<i>Visitors Rejecting or Deleting Cookies</i>	9		
<i>Users Owning and Sharing Multiple Computers</i>	10		
<i>Latency Leaves Room for Inaccuracy</i>	11		
<i>Offline Visits Skewing Data Collection</i>	11		
Comparing Data From Different Vendors	12		
<i>First-Party Versus Third-Party Cookies</i>	12		
<i>Page tags: Placement Considerations</i>	12		
<i>Did You Tag Everything?</i>	12		
<i>Pageviews: A Visit or a Visitor?</i>	12		
<i>Cookies Timeouts</i>	13		
<i>Page-tag Code Hijacking</i>	13		
<i>Data Sampling</i>	13		
<i>PDF files: A Special Consideration</i>	13		
<i>E-commerce: Negative Transactions</i>	13		
<i>Filters and Settings: Potential Obstacles</i>	13		
<i>Time Differences</i>	14		
<i>Process Frequency: Understanding glitches</i>	14		
<i>Goal Conversions versus Pageviews</i>	14		

Introduction

In the past decade, the Internet has transformed marketing, but anyone expecting to increase their revenue and profitability using the web needs to get their facts straight with respect to web traffic. Of course, the web is a great medium to market and sell products and services. But if you don't understand the behaviour of your web site visitors in sufficient detail, your business is going nowhere.

So it is no great surprise that the business of web analytics has grown in tandem with business use of the Internet. Put simply, web analytics are tools and methodologies used to enable organisations to track the number of people who view their site and then use this to measure the success of their online strategy.

The danger is, too many businesses take web analytics reports at face value and this raises the issue of accuracy. After all, it isn't difficult to get the numbers.

However the harsh truth is web analytics data can never be 100 percent accurate, and even measuring the error bars is difficult.

So what's the point?

First, the good news. Error bars remain pretty constant on a weekly, or even a monthly, basis. Even comparing year-on-year behaviour can be safe as long as there are no dramatic changes in technology or end-user behaviour. As long as you use the same measurement "yard stick", visitor number trends will be accurate.

Here are some examples of accurate metrics:

- 30 percent of my web site traffic came via search
- 50 percent of visitors viewed page X.html
- We increased conversions by 20 percent last week
- Pageviews at our site increased by 10 percent during March

With these types of metrics, marketers and webmasters can determine the direct impact of specific marketing campaigns. The level of detail is critical. For example, you can determine if an increase in pay-per-click advertising spend for a set of keywords on a single search engine – increased the return on investment during that time period. So, as long as you can minimise inaccuracies, web analytics tools are effective for measuring visitor traffic to your online business. The remainder of this document examines, in detail, how inaccuracies arise and how organisations can counter them.

How Web Sites Collect Visitor Data

Page Tags and Logfiles

There are two common techniques for collecting web visitor data – page tags and logfiles.

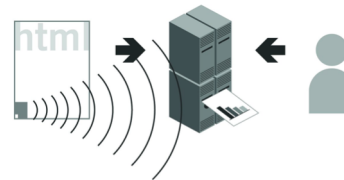


Figure 1 Schematic page tag methodology: Page tags send information to remote data collection servers. The analytics customer views reports from the remote server.

Page tags collect data via the visitor's web browser and send information to remote data-collection servers. The analytics customer views reports from the remote server (see Figure 1). This information is usually captured by Javascript code (known as tags or beacons) placed on each page of your site. Some vendors also add multiple custom tags to collect

additional data. This technique is known as client-side data collection and is used mostly by outsourced, Software as a Service (SaaS) vendor solutions.

Understanding Web Analytics Accuracy

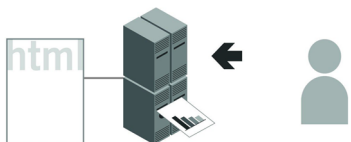


Figure 2 Schematic logfile methodology
The web server logs its activity to a text file that is usually local. The analytics customer views reports from the local server.

Logfiles refer to data collected by your web server independently of a visitor's browser: the web server logs its activity to a text file that is usually local. The analytics customer views reports from the local server, as shown in Figure 2. This technique, known as server-side data collection, captures all requests made to your web server, including pages, images, and

PDFs, and is most frequently used by stand-alone licensed software vendors.

In the past, the easy availability of web server logfiles made this technique the one most frequently adopted for understanding the behaviour of visitors to your site. In fact, most Internet service providers (ISPs) supply a freeware log analyzer with their web-hosting accounts (Analog, Webalizer, and AWstats are some examples). Although this is probably the most common way people first come in contact with web analytics, such freeware tools are too basic when it comes to measuring visitor behaviour and are not considered further in this book.

In recent years, page tags have become more popular as the method for collecting visitor data. Not only is the implementation of page tags easier from a technical point of view, but data-management requirements are significantly reduced because the data is collected and processed by external SaaS servers (your vendor), saving website owners the expense and maintenance of running licensed software to capture, store, and archive information.

Note that both techniques, when considered in isolation, have their limitations. Table 1 summarizes the differences. A common myth is that page tags are technically superior to other methods, but as Table 1 shows, that depends on what you are looking at. By combining both techniques, however, the advantages of one

counter the disadvantages of the other. This is known as a **hybrid** method and some vendors can provide this.

Table 1 – Page Tag versus logfile data collection

Page Tagging	Logfile Analysis
<p>Advantages</p> <ul style="list-style-type: none"> • Breaks through proxy and caching servers—provides more accurate session tracking. • Tracks client-side events—e.g., JavaScript, Flash, Web 2.0 (Ajax). • Captures client-side e-commerce data—server-side access can be problematic. • Collects and processes visitor data in nearly real time. • Allows the vendor to perform program updates for you. • Allows the vendor to perform data storage and archiving for you. 	<p>Advantages</p> <ul style="list-style-type: none"> • Historical data can be reprocessed easily. • No firewall issues to worry about. • Can track bandwidth and completed downloads—and can differentiate between completed and partial downloads. • Tracks search engine spiders and robots by default. • Tracks legacy mobile visitors by default.
<p>Disadvantages</p> <ul style="list-style-type: none"> • Setup errors lead to data loss—if you make a mistake with your tags, data is lost and you cannot go back and reanalyze. • Firewalls can mangle or restrict tags. • Cannot track bandwidth or completed downloads—tags are set when the page or file is requested, not when the download is complete. • Cannot track search engine spiders—robots ignore page tags 	<p>Disadvantages</p> <ul style="list-style-type: none"> • Proxy and caching inaccuracies—if a page is cached, no record is logged on your web server. • No event tracking—e.g., no JavaScript, Flash, Web 2.0 tracking (Ajax). • Requires your own team to perform program updates. • Requires your own team to perform data storage and archiving. • Robots multiply visit counts.

Other Data-Collection Methods

Although logfile analysis and page tagging are by far the most widely used methods for collecting web visitor data, they are not the only methods. Network data-collection devices (packet sniffers) gather web traffic data from routers into black-box appliances. Another technique is to use a web server application programming interface (API) or loadable module (also known as a plug-in, though this is not strictly correct terminology). These are programs that extend the capabilities of the web server—for example, enhancing or extending the fields that are logged. Typically, the collected data is then streamed to a reporting server in real time.

Cookies in Web Analytics

Page tag solutions track visitors by using cookies. Cookies are small text messages that a web server transmits to a web browser so that it can keep track of the user's activity on a specific website. The visitor's browser stores the cookie information on the local hard drive as name-value pairs. Persistent cookies are those that are still available when the browser is closed and later reopened. Conversely, session cookies last only for the duration of a visitor's session (visit) to your site.

For web analytics, the main purpose of cookies is to identify users for later use—most often with an anonymous visitor id. Among many things, cookies can be used to determine how many first-time or repeat visitors a site has received, how many times a visitor returns each period, and how much time passes between visits. Web analytics aside, web servers can also use cookie information to present personalized web pages. A returning customer might see a

different page than the one a first-time visitor would view, such as a “welcome back” message to give them a more individual experience or an auto-login for a returning subscriber.

The following are some cookie facts:

- Cookies are small text files (no larger than 4 Kb), stored locally, that are associated with visited website domains.
- Cookie information can be viewed by users of your computer, using notepad or a text editor application.
- There are two types of cookies: first party and third party.
- A first-party cookie is one created by the website domain. A visitor requests it directly by typing the URL into their browser or by following a link.
- A third-party cookie is one that operates in the background and is usually associated with advertisements or embedded content that is delivered by a third-party domain not directly requested by the visitor.
- For first-party cookies, only the website domain setting the cookie information can retrieve the data. this is a security feature built into all web browsers.
- For third-party cookies, the website domain setting the cookie can also list other domains allowed to view this information. the user is not involved in the transfer of third-party cookie information.
- Cookies are not malicious and can't harm your computer. they can be deleted by the user at any time.
- A maximum of 50 cookies are allowed per domain for the latest versions of IE8 and Firefox 3. Other browsers may vary (opera 9 currently has a limit of 30; Safari and Google Chrome have no limit on the number of cookies per domain).

Understanding Web Analytics Data Accuracy

When it comes to benchmarking the performance of your website, web analytics is critical. However, this information is accurate only if you avoid common errors associated with collecting the data—especially comparing numbers from different sources. Unfortunately, too many businesses take web analytics reports at face value. After all, it isn't difficult to get the numbers. The harsh truth is that web analytics data can never be 100 percent accurate, and even measuring the error bars can be difficult.

So what's the point?

Despite the pitfalls, error bars remain relatively constant on a weekly, or even a monthly, basis. Even comparing year-by-year behaviour can be safe as long as there are no dramatic changes in technology or end-user behaviour. As long as you use the same yardstick, visitor number trends will be accurate. For example, web analytics data may reveal patterns like the following:

- Thirty percent of site traffic came from search engines.
- Fifteen percent of site revenue was generated by product page x.html.
- We increased subscription conversions from our email campaigns by 20 percent last week.
- Bounce rate decreased 10 percent for our category pages during March.

With these types of metrics, marketers and webmasters can determine the direct impact of specific marketing campaigns. The level of detail is critical. For example, you can determine if an increase in pay-per-click advertising spending—for a set of keywords on a single search engine—increased the return on investment during that time period. As long as you can minimize

inaccuracies, web analytics tools are effective for measuring visitor traffic to your online business.

Conflicting Data Points Are Common

A UK survey of 800 organizations revealed that almost two-thirds (63 percent) of respondents say they experience conflicting information from different sources of online measurement data ("Online Measurement and Strategy Report 2009," Econsultancy.com, June 2009).

Next, I'll discuss in detail why such inaccuracies arise, so you can put this information into perspective. The aim is for you to arrive at an acceptable level of accuracy with respect to your analytics data. Recall from Table 1 that there are two main methods for collecting web visitor data—logfiles and page tags—and both have limitations.

Issues Affecting Visitor Data Accuracy for Logfiles

Logfile tracking is usually set up by default on web servers. Perhaps because of this, system administrators rarely consider any further implications when it comes to tracking.

Dynamically Assigned IP Addresses

Generally, a logfile solution tracks visitor sessions by attributing all hits from the same IP address and web browser signature to one person. This becomes a problem when ISPs assign different IP addresses throughout the session. A U.S.-based comScore study (http://www.comscore.com/Press_Events/Presentations_Whitepaper

[s/2007/Cookie Deletion Whitepaper](#)) showed that a typical home PC averages 10.5 different IP addresses per month. Those visits will be counted as 10 unique visitors by a logfile analyzer. This issue is becoming more severe, because most web users have identical web browser signatures (currently internet explorer). As a result, visitor numbers are often vastly over counted. This limitation can be overcome with the use of cookies.

Client-Side Cached Pages

Client-side caching means a previously visited page is stored on a visitor's computer. In this case, visiting the same page again results in that page being served locally from the visitor's computer, and therefore the visit is not recorded at the web server.

Server-side caching can come from any web accelerator technology that caches a copy of a website and serves it from their servers to speed up delivery. This means that all subsequent site requests come from the cache and not from the site itself, leading to a loss in tracking. Today, most of the Web is in some way cached to improve performance. For example, see Wikipedia's cache description at <http://en.wikipedia.org/wiki/Cache>.

Counting Robots

Robots, also known as spiders or web crawlers, are most often used by search engines to fetch and index pages. However, other robots exist that check server performance—uptime, download speed, and so on—as well as those used for page scraping, including price comparison, e-mail harvesters, competitive research, and so on. These affect web analytics because a logfile solution will also show all data for robot activity on your website, even though robots are not real visitors.

When counting visitor numbers, robots can make up a significant proportion of your pageview traffic. Unfortunately, these are difficult to filter out completely because thousands of home grown and

unnamed robots exist. For this reason, a logfile analyzer solution is likely to over count visitor numbers, and in most cases this can be dramatic.

Issues Affecting Visitor Data From Page Tags

Deploying a page tag on every single page is a process that can be automated in many cases. However, for larger sites 100 percent correct deployment is rarely achieved. Perhaps it is because the page tag is hidden to the human eye or there is so much other data available that those errors often go unnoticed for long periods. Having a full deployment is crucial to the accuracy and validity of data collected by this method.

Setup Errors Causing Missed Tags

The most frequent error by far observed for page tagging solutions comes from its setup. Unlike web servers, which are configured to log everything delivered by default, a page tag solution requires the webmaster to add the tracking code to each page. Even with an automated content management system, pages can and do get missed.

In fact, evidence from analysts at Maxamine (<http://www.maxamine.com>)—now part of Accenture Marketing Sciences—who used their automatic page auditing tool has shown that some sites claiming that all pages are tagged can actually have as many as 20 percent of pages missing the page tag—something the webmaster was completely unaware of. In one case, a corporate business-to-business site was found to have 70 percent of its pages missing tags. Missing tags equals no data for those pageviews.

JavaScript Errors Halt Page Loading

Page tags work well, provided that Javascript is enabled on the visitor's browser. Fortunately, only about 1 to 3 percent of Internet users have disabled Javascript on their browsers, as shown in Figure 3. However, the inconsistent use of Javascript code on web pages can cause a bigger problem: Any errors in other Javascript on the page will immediately halt the browser scripting engine at that point, so a page tag placed below it will not execute.

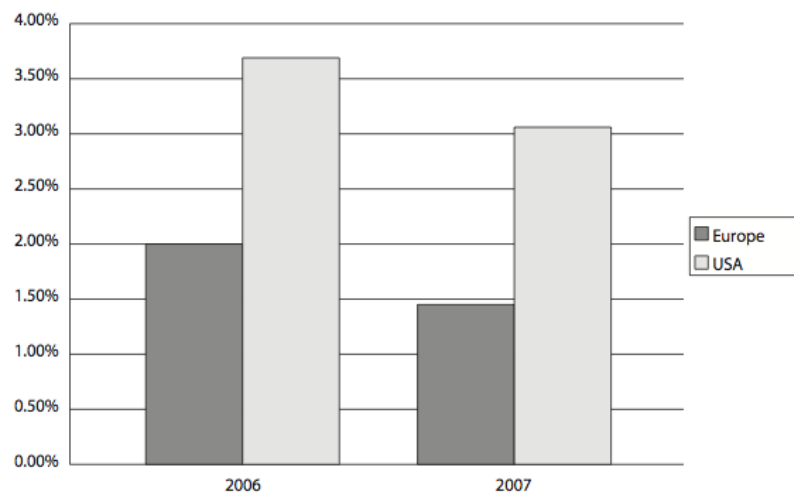


Figure 3 Percentage of Internet users with Javascript-disabled browsers. Source: 1,000,000,000 visits across multiple industry web properties using IndexTools (www.visualrevenue.com/blog—Dennis R. Mortensen)

Firewalls Block Page Tags

Corporate and personal firewalls can prevent page tag solutions from sending data to collecting servers. In addition, firewalls can also be set up to reject or delete cookies automatically. Once again, the effect on visitor data can be significant. Some web analytics

vendors can revert to using the visitor's IP address for tracking in these instances, but mixing methods is not recommended. As discussed previously in "issues affecting visitor data accuracy for logfiles" (comScore report), using visitor IP addresses is far less accurate than simply not counting such visitors. It is therefore better to be consistent with the processing of data.

Logfiles "See" Mobile Users

A mobile web audience study by comScore back in January 2007 (www.comscore.com/press/release.asp?press=1432) showed that in the United States, 30 million (or 19%) of the 159 million U.S. Internet users accessed the Internet from a mobile device. At that time, the vast majority of mobile phones did not understand Javascript or cookies, and hence only logfile tools were able to track visitors who browsed using their mobile phones.

However, thanks mainly to the phenomenal success of the iPhone, mobile visitors on your website can now be tracked with page tag web analytics, because the browser software is very similar to that found on regular laptops and PCs, that is, where both Javascript and cookies are used.

Issues Affecting Visitor Data When Using Cookies

Cookies are a very simple, well-established way of tracking visitors. However, their simplicity and transparency (any user can remove them) presents issues in themselves. The debate of using cookies or not remains a hot topic of conversation in web analytics circles.

Visitors Rejecting or Deleting Cookies

Cookie information is vital for web analytics because it identifies visitors, their referring source, and subsequent pageview data. The

current best practice is for vendors to process first-party cookies only. This is because visitors often view third-party cookies as infringing on their privacy, opaquely transferring their information to third parties without explicit consent. Therefore, many anti-spyware programs and firewalls exist to block third-party cookies automatically. It is also easy to do this within the browser itself. By contrast, anecdotal evidence shows that first-party cookies are accepted by more than 95 percent of visitors.

Visitors are also becoming savvier and often delete cookies. independent surveys conducted by Belden Associates (2004), Jupiterresearch (2005), Nielsen//Netratings (2005) and comScore (2007) concluded that cookies are deleted by at least 30 percent of internet users in a month.

Users Owning and Sharing Multiple Computers

User behaviour has a dramatic effect on the accuracy of information gathered through cookies. Consider the following scenarios:

Same user, multiple computers

- Today, people access the Internet in any number of ways – from work, home, or public places such as Internet cafes. One person working from three different machines results in three cookie settings, and all current web analytics solutions will count each of these anonymous user sessions as unique.

Different users, same computer

- People share their computers all the time, particularly with their families, and, as a result, cookies are shared too (unless you log off or switch off you computer each time it is used by a different person). In some instances, cookies are deleted deliberately. For example, Internet cafes are set up to do this automatically at the end of each session. So even if a visitor uses that cafe regularly and works from the same

machine, a web analytics solution will 'see' them as a different and new visitor every time.

Correcting Data for Cookie Deletion and Rejection

Calculating a correction factor to account for your visitors either deleting or rejecting your web analytics cookies is quite straightforward. All you need is a website that requires a user login. That way you can count the number of unique login IDs and divide it by the number of unique users your web analytics tool reports. The result is a correction factor that can be applied to subsequent data (number of unique visitors, number of new visitors, or number of returning visitors).

Having a website that requires a user login is, thankfully in my view, quite rare, because people wish to access information freely and as easily as possible. So, although the correction-factor calculation is straightforward, you most probably don't have any login data to process. Fortunately, a small number of websites can calculate a correction factor to shed light on this issue. These include online banks and popular brands such as Amazon, FedEx, and social network sites, where there is a real user benefit to both having an account and (most importantly) using it when visiting the site.

A specific example is Sun Microsystems Forums (<http://forums.sun.com>), a global community of developers with nearly 1 million contributors. A 2009 study by Paul Strupp and Garrett Clark, published at <http://blogs.sun.com/pstrupp/>, reveals some interesting data.

When using third-party cookies:

- 78% is the correction factor for monthly unique users.
- 20% of users delete (more correctly defined as lose) their measurement cookie at least once per month.

- 5% of users block the third-party measurement cookie.

When using first-party cookies:

- The correction factor improves to 83%.
- Percentage of users who delete their measurement cookie at least once per month decreases to 14%.
- Percentage of users who block the first-party measurement cookie drops to less than 1%.

Note that this is a tech-savvy audience—those who can delete/block an individual cookie without a second thought.

An interesting observation from the study that Paul himself highlights, is the relatively small value of the correction factor. That is, when using a first-party cookie, a more precise unique visitor count is 0.83 multiplied by the reported value. Putting this into context, as part of the analysis, 30% of users who used more than one computer in a month to visit the forum were removed from the data prior to analysis. This indicates that multiple-device access happens more frequently than cookie deletion.

It is tempting to think that this data can be used to correct your own unique visitor counts. However, the correction factor is a complicated function of cookie deletion, multiple computer use, and visitor return frequency. These factors will almost certainly be different for your specific website. Nonetheless, it is a useful rule-of-thumb guide.

Latency Leaves Room for Inaccuracy

The time it takes for a visitor to be converted into a customer (latency) can have a significant effect on accuracy. For example,

most low-value items are either instant purchases or are purchased within seven days of the initial website visit. With such a short time period between visitor arrival and purchase, your web analytics solution has the best possible chance of capturing all the visitor pageview and behaviour information and therefore reporting more accurate results.

Higher-value items usually mean a longer consideration time before the visitor commits to becoming a customer. For example, in the travel and finance industries, the consideration time between the initial visit and the purchase can be as long as 90 days. During this time, there's an increased risk of the user deleting cookies, reinstalling the browser, upgrading the operating system, buying a new computer, or dealing with a system crash. Any of these occurrences will result in users being seen as new visitors when they finally make their purchase. Offsite factors such as seasonality, adverse publicity, offline promotions, or published blog articles or comments can also affect latency.

Offline Visits Skewing Data Collection

It is important to factor in problems that are unrelated to the method used to measure visitor behaviour but that still pose a threat to data accuracy. High-value purchases such as cars, loans, and mortgages are often first researched online and then purchased offline. Connecting offline purchases with online visitor behaviour is a long-standing enigma for web analytics tools. Currently, the best-practice way to overcome this limitation is to use online voucher schemes that visitors can print and take with them to claim a free gift, upgrade, or discount at your store. If you would prefer to receive your orders online, consider providing similar incentives, such as web-only pricing, free delivery if ordered online, and the like.

Another issue to consider is how your offline marketing is tracked. Without taking this into account, visitors who result from your offline campaign efforts will be incorrectly assigned or grouped with other referral sources and therefore skew your data.

Comparing Data From Different Vendors

As shown earlier, it is virtually impossible to compare the results of one data-collection method with another. The association simply isn't valid. However, given two comparable data-collection methods—both page tags—can you achieve consistency? Unfortunately, even comparing vendors that employ page tags has its difficulties. Factors that lead to differing vendor metrics are described in the following sections.

First-Party Versus Third-Party Cookies

There is little correlation between the two because of the higher blocking rates of third-party cookies by users, firewalls, and anti-spyware software. For example, the latest versions of Microsoft Internet Explorer block third-party cookies by default if a site doesn't have a compact privacy policy (see www.w3.org/P3P).

Page tags: Placement Considerations

Page-tag vendors often recommend that their page tags be placed just above the `</body>` tag of your HTML page to ensure that the page elements, such as text and images, load first. This means that any delays from the vendor's servers will not interfere with your page loading. The potential problem here is that repeat visitors, those more familiar with your website navigation, may navigate quickly, clicking onto another page before the page tag has loaded to collect data. Clearly, the longer the delay, the greater the discrepancy will be.

Tag placement was investigated in a 2009 whitepaper by tagMan.com. Their study of latency effects revealed that approximately 10 percent of reported traffic is lost for every extra second a page takes to load. In addition, moving the Google Analytics page tag from the bottom of a page to the top increased the reported traffic by 20%.

Stone Temple Consulting conducted a similar study in 2007. Their results showed that the difference between a tracking tag placed at the top or bottom of a page accounted for a 4.3% difference in unique visitor traffic. This was attributed to the 1.4 second difference in executing the page tag.

In addition, non-related Javascript placed at the top of the page can interfere with Javascript page tags that have been placed lower down. Most vendor page tags work independently of other Javascript and can sit comfortably alongside other vendor page tags—as shown in the Stone Temple Consulting report in which pages were tagged for five different vendors. However, Javascript errors on the same page will cause the browser scripting engine to stop at that point and prevent any Javascript below it, including your page tag, from executing.

Did You Tag Everything?

Many analytics tools require links to files such as PDFs, Word documents, or executable downloads or outbound links to other websites to be modified in order to be tracked. This may be a manual process whereby the link to the file needs to be modified. The modification represents an event or action when it is clicked, which sometimes is referred to as a virtual pageview. Comparing different vendors requires this action to be carried out several times with their specific codes (usually with Javascript). Take into consideration that whenever pages have to be coded, syntax errors are a possibility. If page updates occur frequently, consider regular website audits to validate your page tags.

Pageviews: A Visit or a Visitor?

Pageviews are quick and easy to track; and because they require only a call from the page to the tracking server, they are very similar among vendors. The challenge is differentiating a visit from a visitor; and because every vendor uses a different algorithm, no single algorithm results in the same value.

Cookies Timeouts

The allowed duration of timeouts—how long a web page is left inactive by a visitor—varies among vendors. Most page-tag vendors use a visitor-session cookie timeout of 30 minutes. This means that continuing to browse the same website after 30 minutes of inactivity is considered to be a new repeat visit. However, some vendors offer the option to change this setting. Doing so will alter any data alignment and therefore affect the analysis of reported visitors. Other cookies, such as the ones that store referrer details, will have different timeout values. For example, Google Analytics referrer cookies last six months. Differences in these timeouts between different web analytics vendors will obviously be reflected in the reported visitor numbers.

Page-tag Code Hijacking

Depending on your vendor, your page tag code could be hijacked, copied, and executed on a different or unrelated website. This contamination results in a false pageview within your reports. By using filters, you can ensure that only data from your domains are reported.

Data Sampling

This is the practice of selecting a subset of data from your website traffic. Sampling is widely used in statistical analysis because analyzing a subset of data gives very similar results to analyzing all of the data, yet can provide significant speed benefits when processing large volumes of information. Different vendors may use different sampling techniques and criteria, resulting in data misalignment.

PDF files: A Special Consideration

For page tag solutions, it is not the completed PDF download that is reported, but the fact that a visitor has clicked on a PDF file link.

This is an important distinction as information on whether or not the visitor completes the download – for example a 50-page PDF file – is not available. Therefore, a click on a PDF link is reported as a single event or pageview.

Note: The situation is different for logfile solutions. When you view a PDF file within your web browser, Adobe Reader can download the file one page at a time, as opposed to a full download. This results in a slightly different entry in your web server logfile, showing an HTTP status code 206 (partial file download). Logfile solutions can treat each of the 206 status code entries as individual pageviews. When all the pages of a PDF file are downloaded, a completed download is registered in your logfile with a final HTTP status code of 200 (download completed). Therefore, a logfile solution can report a completed 50-page PDF file as 1 download and 50 pageviews.

E-commerce: Negative Transactions

All e-commerce organizations have to deal with product returns at some point, whether because of damaged or faulty goods, order mistakes, or other reasons. Accounting for these returns is often forgotten within web analytics reports. For some vendors, it requires the manual entry of an equivalent negative purchase transaction. Others require the reprocessing of e-commerce data files. Whichever method is required, aligning web visitor data with internal systems is never bulletproof. For example, the removal or crediting of a transaction usually takes place well after the original purchase and therefore in a different reporting period.

Filters and Settings: Potential Obstacles

Data can vary when a filter is set up in one vendor's solution but not in another. Some tools can't set up the exact same filter as another

tool, or they apply filters in a different way or at a different point during data processing.

Consider, for example, a page-level filter to exclude all error pages from your reports. Visit metrics such as time on site and page depth may or may not be adjusted for the filter depending on the vendor. This is because some vendors treat page-level metrics separately from visitor-level metrics.

Time Differences

A predicament for any vendor when it comes to calculating the time on site or time on page for a visitor's session involves how to calculate for the last page viewed. For example, time spent on pageA is calculated by taking the difference between the visitor's timestamp for pageA and the subsequent timestamp for pageB, and so on. But what if there is no pageC; How can the time on page be calculated for pageB if there is no following timestamp?

Different vendors handle this in different ways. Some ignore the final pageview in the calculation; others use an onUnload event to add a timestamp should the visitor close their browser or go to a different website. Both are valid methods, although not every vendor uses the onUnload method. The reason some vendors prefer to ignore the last page is that it is considered the most inaccurate from a time point of view—perhaps the visitor was interrupted to run an errand or left their browser in its current state while working on something else. Many users behave in this way; that is, they complete their browsing task and simply leave their browser open on the last page while working in another application. A small number of pageviews of this type will disproportionately skew the time-on-site and time-on-page calculations; hence, most vendors avoid this issue.

Note: Google Analytics ignores the last pageview of a visitor's session when calculating the time-on-site and time-on-page metrics.

Process Frequency: Understanding glitches

the frequency of processing is best illustrated by example: google Analytics does its number crunching to produce reports hourly. however, because it takes time to col- late all the logfiles from all of the data-collecting servers around the world, reports are three to four hours behind the current time. in most cases, it is usually a smooth pro- cess, but sometimes things go wrong. For example, if a logfile transfer is interrupted, then only a partial logfile is processed. because of this, google collects and reprocesses all data for a 24-hour period at the day's end. other vendors may do the same, so it is important not to focus on discrepancies that arise on the current day.

Goal Conversions versus Pageviews

Using Figure 4 as an example, assume that five pages are part of your defined funnel (click-stream path), with the last step (page 5) being the goal conversion (purchase). During checkout, a visitor goes back up a page to check a delivery charge (step A) and then continues through to complete payment. The visitor is so happy with the simplicity of the entire process that she then purchases a second item using exactly the same path during the same visitor session (step B).

Depending on the vendor you use, this process can be counted in various ways, as follows:

- Twelve funnel page views, two conversions, two transactions
- Ten funnel page views (ignoring step A), two conversions, two transactions
- Five funnel page views, two conversions, two transactions
- Five funnel page views, one conversion (ignoring step B), two transactions

Most vendors, but not all, apply the last rationale to their reports. That is, the visitor has become a purchaser (one conversion); and this can happen only once in the session, so additional conversions

Understanding Web Analytics Accuracy

(assuming the same goal) are ignored. For this to be valid, the same rationale must be applied to the funnel pages. In this way, the data becomes more visitor-centric.

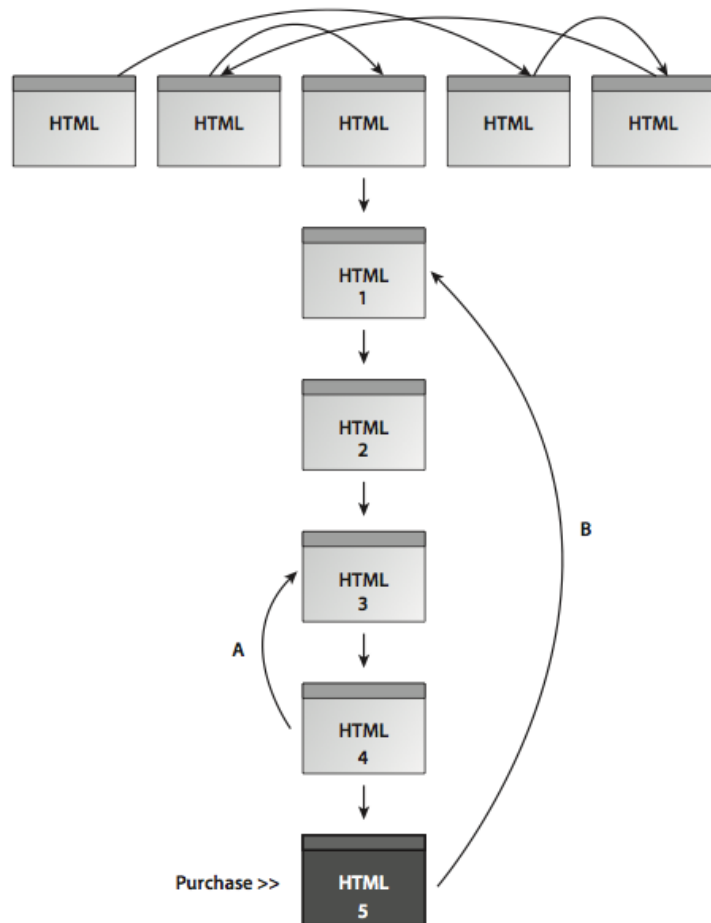


Figure 4 A visitor traversing a website, entering a five-page funnel, and making two transactions

Note: in the above example, the total number of pageviews is 12 and should be reported as such in all pageview reports. It is the funnel and goal conversion reports that will be different.

Why PPC Vendor Numbers Do Not Match

If you are using pay-per-click (PPC) networks, you will typically have access to the click-through reports provided by each network. Quite often, these numbers don't exactly align with those reported in your web analytics reports. This can happen for the reasons described in the following sections.

Tracking URLs: Missing Paid Search Click-throughs

Tracking URLs are required in your PPC account setup in order to differentiate between a non-paid search engine visitor click-through and a paid click-through from the same referring domain – Google.com or Yahoo.com, for example. Tracking URLs are simple modifications to your landing page URLs within your PPC account and are of the form `www.mysite.com?source=adwords`. Tracking URLs forgotten during setup, or sometimes simply assigned incorrectly can lead to such visits incorrectly assigned.

Slow Page Load Times

As previously discussed, the best practice location for web analytics data-collection tags is at the bottom of your pages—just above the `</body>` HTML tag. If your PPC landing pages are slow to download for whatever reason (server delays, page bloat, and so on), it is likely that visitors will click away, navigating to another page on your site or even to a different website, before the data-collection tag has had chance to load. The chance of this happening increases the longer the page load time is. The general rule of thumb for what constitutes a

Understanding Web Analytics Accuracy

long page load is only two seconds (see www.akamai.com/html/about/press/releases/2009/press_091409.html).

Clicks and Visits: Understanding the Difference

Remember that PPC vendors, such as Google AdWords, measure clicks. Most web analytics tools measure visitors who can accept a cookie. Those are not always going to be the same thing when you consider the effects on your web analytics data of cookie blocking, Javascript errors, and visitors who simply navigate away from your landing page quickly—before the page tag collects its data. Because of this, web analytics tools tend to slightly underreport visits from PPC networks.

PPC Account Adjustments

Google AdWords and other PPC vendors automatically monitor invalid and fraudulent clicks and adjust PPC metrics retroactively. For example, a visitor may click your ad several times (inadvertently or on purpose) within a short space of time. Google AdWords automatically investigates this influx and removes the additional click-throughs and charges from your account. However, web analytics tools have no access to these systems and so record all PPC visitors. For further information on how Google treats invalid clicks, see:

<http://adwords.google.com/support/bin/topic.py?topic=35>

Keyword Matching: Bid Term versus Search Term

The bid terms you select within your PPC account and the search terms used by visitors that result in your PPC ad being displayed can often be different: think 'broad match'. For example, you may have set up an ad group that targets the word 'shoes' and solely relies on broad match to match all search terms that contain the word 'shoes'. This is your bid term. A visitor uses the search term

'blue shoes' and clicks on your ad. Web analytics vendors may report the search term, the bid term or both.

Google AdWords Import Delay

Within your AdWords account, you'll see that data is updated hourly. This is because advertisers need this information to control budgets. Google Analytics imports AdWords cost data once a day. This is for the data range minus 48 to 24 hours from 23:59 the previous day (so AdWords cost data is always at least 24 hours old).

Why the delay? because it allows time for the AdWords invalid-click and fraud- protection algorithms to complete their work and finalize click-through numbers for your account. therefore, from a reporting point of view, the recommendation is to not compare AdWords visitor numbers for the current day. this recommendation holds true for all web analytics solutions and all PPC advertising networks.

Note: Although most of the AdWords invalid click updates take place within hours, final adjustments may take longer. For this reason, even if all other factors are eliminated, AdWords numbers and web analytics reports may never match exactly.

Losing Tracking URLs Through Redirects

Using third-party ad-tracking systems—such as Adform, Atlas Search, Blue Streak, DoubleClick, Efficient Frontier, and SEM Director—to track click-throughs to your website means your visitors are passed through redirection URLs. This results in the initial click being registered by your ad company, which then automatically redirects the visitor to your actual landing page. The purpose of this two-step hop is to allow the ad-tracking network to collect visitor statistics independently of your organization, typically for billing purposes. Because this process involves a short delay, it may prevent some visitors from landing on your page. The result can be a small loss of data and therefore failure to align data.

Understanding Web Analytics Accuracy

More important, and more common, redirection URLs may break the tracking parameters that are added onto the landing pages for your own web analytics solution. For example, your landing page URL may look like this:

```
http://www.mysite.com/?source=google&medium=ppc&campaign=Jan10
```

When added to a third-party tracking system for redirection, it could look like this:

```
http://www.redirect.com?http://www.mywebsite.com?source=google&medium=ppc&campaign=Jan10
```

The problem occurs with the second question mark in the second link, because you can't have more than one in any valid URL. Some third-party ad-tracking systems will detect this error and remove the second question mark and the following tracking parameters, leading to a loss of campaign data.

Some third-party ad-tracking systems allow you to replace the second ? with a # so the URL can be processed correctly. If you are unsure of what to do, you can avoid the problem completely by using encoded landing-page URLs within your third-party ad-tracking system, as described at the following site: www.w3schools.com/tags/ref_urlencode.asp.

Note: From experience, the most common reasons for discrepancies between PPC vendor reports and web analytics tools arise from:

- Tracking URLs failing to distinguish paying and nonpaying visitors
- Slow page downloading
- Losing data via third-party ad-tracking redirects

Data Misinterpretation

The following are not accuracy issues. However, they point out that data is not always so straightforward to interpret. Take the following two examples:

- New visitors plus repeat visitors does not equal total visitors.

A common misconception is that the sum of the new plus repeat visitors should equal the total number of visitors. Why isn't this the case? Consider a visitor making his first visit on a given day and then returning on the same day. They are both a new and a repeat visitor for that day. Therefore, looking at a report for the given day, two visitor types will be shown, though the total number of visitors is one. It is therefore better to think of visitor types in terms of "visit" type - that is, the number of first-time visits plus the number of repeat visits equals the total number of visits.

- Summing the number of unique visitors per day for a week does not equal the total number of unique visitors for that week.

Consider the scenario in which you have 1,000 unique visitors to your website blog on a Monday. These are in fact the only unique visitors you receive for the entire week, so on Tuesday the same 1,000 visitors return to consume your next blog post. This pattern continues for Wednesday through Sunday.

If you were to look at the number of unique visitors for each day of the week in your reports, you would observe 1,000 unique visitors. However you cannot say that you received 7,000 unique visitors for the entire week. For this example, the number of unique visitors for the week remains at 1,000.

Why Counting Uniques Is Meaningless

The term uniques is often used in web analytics as an abbreviation for unique web visitors, that is, how many unique people visited your site. The problem is that counting unique visitors is fraught with problems that are so fundamental, it renders the term uniques meaningless.

As discussed earlier, cookies get lost, blocked, and deleted—nearly one-third of tracking cookies can be missing after a period of four weeks. The longer the time period, the greater the chance of this happening, which makes comparing year-on-year uniques invalid, for example. In addition, browsers make it very easy these days for cookies to be removed—see the new “incognito” features of the latest Firefox, Chrome, and Internet Explorer browsers.

However, the biggest issue for counting uniques is how many devices people use to access the Web. For example, consider the following scenario:

- You and your spouse are considering your next vacation. Your spouse first checks out possible locations on your joint PC at home and saves a list of website links.
- The next evening you use the same PC to review these links. Unable to decide that night, you email the list to your office, and the next day you continue your vacation checks during your lunch hour at work and also review these again on your mobile while commuting home on the train.
- Day 3 of your search resumes at your friend’s house, where you seek a second opinion. Finally, you go home and book online using your shared PC.

This scenario is actually very common—particularly if the value of the purchase is significant, which implies a longer consideration

period and the seeking of a second opinion from a spouse, friends, or work colleagues (the Sun Microsystems study discussed earlier estimated the percentage of users using more than one computer in a month to visit the same website as 30 percent).

Simply put, there is not a web analytics solution in the world that can accurately track this scenario, that is, to tie the data together from multiple devices and where multiple people have been involved, nor is there likely to be one in the near future.

Combining these limitations leads to large error bars when it comes to tracking uniques. In fact, these errors are so large that the metric becomes meaningless and should be avoided where possible in favor of more accurate “visit” data. That said, if you must use unique visitors as a key metric, ensure the emphasis is on the trend, not the absolute number.

Ten Recommendations For Enhancing Accuracy

1. Be sure to select a tool that uses first-party cookies for data collection.
2. Don’t confuse visitor identifiers. For example, if first-party cookies are deleted, do not resort to using IP address information. It is better simply to ignore that visitor.
3. Remove or report separately all non-human activity from your data reports, such as robots and server-performance monitors.
4. Track everything. Don’t limit tracking to landing pages. Track your entire website’s activity, including file downloads, internal search terms, and outbound links.
5. Regularly audit your website for page tag completeness (at least monthly for large websites). Sometimes site content changes result in tags being corrupted, deleted, or simply forgotten.
6. Display a clear and easy-to-read privacy policy (required by law in the European union). This establishes trust with your visitors

Understanding Web Analytics Accuracy

because they better understand how they're being tracked and are less likely to delete cookies.

7. Avoid making judgments on data that is less than 24 hours old, because it's often the most inaccurate.
8. Test redirection URLs to guarantee that they maintain tracking parameters.
9. Ensure that all paid online campaigns use tracking URLs to differentiate from non-paid sources.
10. Use visit metrics in preference to unique visitor metrics because the latter are highly inaccurate.

These suggestions will help you appreciate the errors often made when collecting web analytics data. Understanding what these errors are, how they happen, and how to avoid them will enable you to benchmark the performance of your website. Achieving this means you're in a better position to then drive the performance of your online business.

Summary

So, web analytics is not 100 percent accurate and the number of possible inaccuracies can at first appear overwhelming. However, get comfortable with your implementation and focus on measuring trends rather than precise numbers. For example, web analytics can help you answer the following questions:

- Are visitor numbers increasing?
- By what rate are they increasing (or decreasing)?
- Have conversion rates gone up since beginning PPC advertising?
- How has the cart abandon rate changed since the site redesign?

If the trend shows a 10.5% reduction, for example, this figure should be accurate, regardless of the web analytics tool that was used. These examples are all high-level metrics, though the same accuracy can also be maintained as you drill down and look at, for example, which specific referrals (search engines, affiliates, social networks),

campaigns (paid search, email, banners), keywords, geographies, or devices (PC, Mac, mobile) are used.

When all the possibilities of inaccuracy that affect web analytics solutions are considered, it is apparent that it is ineffective to focus on absolute values or to merge numbers from different sources. If all web visitors were to have a login account in order to view your website, this issue could be overcome. In the real world, however, the vast majority of internet users wish to remain anonymous, so this is not a viable solution.

As long as you use the same measurement for comparing data ranges, your results will be accurate. This is the universal truth of all web analytics.

Acknowledgements

With thanks to the following people for their generous feedback in compiling this whitepaper: Sara Andersson, Nick Mihailovski, Alex Ortiz-Rasado, Tomas Remotigue.


retweet
this article


About The Author


Brian Clifton (PhD), is an independent author, consultant and trainer who specialises in performance optimisation using Google Analytics and related tools. Recognised internationally as a Google Analytics expert, his latest book, the second edition of *Advanced Web Metrics with Google Analytics* is used by students and professionals world-wide.

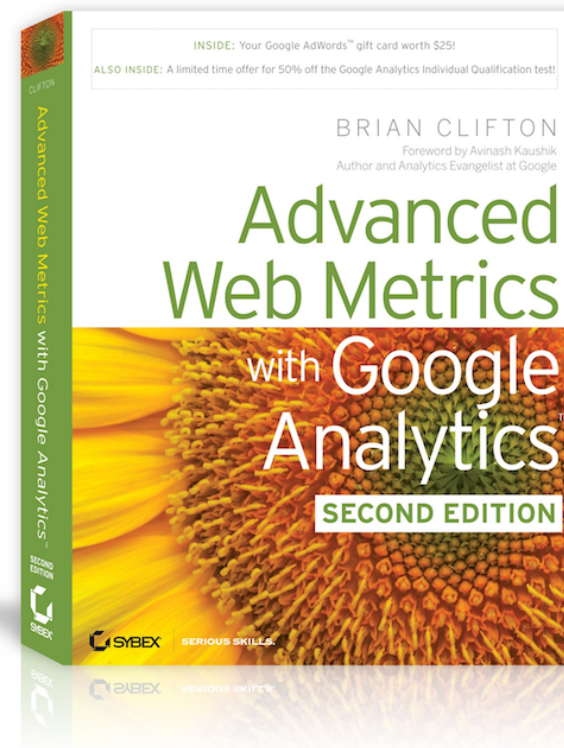
Brian has been involved in web design and SEO since as far back as 1997, when he built his first website and started defining best practise to advise clients. From 2005-8 he was Head of Web Analytics for Google EMEA, defining the adoption strategy and building a team of pan-European product specialists from scratch. A legacy of his work is the online learning centre for the Google Analytics Individual Qualification (GAIQ).

Brian is the Founder, CEO and Senior Strategist for GA-Experts.com – a company specialising in performance optimisation using Google Analytics and related products for global clients.

 Add your comments on the blog - [Measuring Success](#)

 Follow my interests and thoughts [@BrianClifton](#)

 Join your peers on the [LinkedIn Group](#)



Advanced Web Metrics with Google Analytics is available from:
[Amazon](#) (including Kindle), [Barnes & Noble](#) and directly from [Wiley](#).

A [PDF ebook](#) is also available