Increasing Accuracy for Online Business Growth

A whitepaper by Brian Clifton in conjunction with Omega Digital Media Ltd

Version 0.1, February 2008

Table of Contents

Preface

When it comes to benchmarking the performance of your web site, web analytics 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 marketers and webmasters who want to maximise the accuracy of their data.

About the Author

Brian Clifton (PhD) is an internationally established search engine marketing and web analytics expert who has worked in these fields since 1997. Specialising in web analytics and search marketing, his business was the first UK Partner for Urchin Software Inc., the company that later became Google Analytics.

Brian joined Google in 2005 to define, develop and lead the Web Analytics team for Europe, Middle East and Africa. He is currently working on his first book – Advanced Web Metrics With Google Analytics, to be published by Wiley.

Views expressed in this document are the authors and do not represent Google or any other entity. The names of actual companies and products mentioned herein may be trademarks of their respective owners.

If you have comments about this document, add your views at: www.advanced-web-metrics.com/accuracy-whitepaper.

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 versus 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 a visitor's data through their web browser. This information is usually captured by JavaScript code (known as tags or beacons) placed on each page of your site. The technique is known as client-side data collection and this is used mostly by outsourced, hosted vendor solutions.

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 independent of the visitor's browser. This technique, known as server-side collection, captures all requests made to your web server, including pages, images and PDFs and is most used by 'stand alone' software vendors.

In the past, the easy availability of web server logfiles made this technique the most adopted for understanding the behaviour of visitors to your site. But in recent years, page tags have become more popular. Not only is implementation of page tags easier from a technical point of view, but data management needs are significantly reduced. Why? Because the data is collected and processed by external servers (your vendor), saving web site owners from the expense and maintenance of running software to capture, store and archive information.

It is important to note that both techniques, when considered in isolation, have their limitations. Table 1 summarises 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, the advantages of one counters the disadvantages of the other. This is known as a HYBRID method and some vendors can provide this.

Are there alternatives?

The method you choose depends on your objectives and the technical resources available to you. It is important to keep in mind that, although they're the most commonly used, page tags and logfiles are not the only means available for collecting information about your visitors.

Table 1 – Page Tag versus Logfile Data Collection

Page Tagging	Logfile Analysis
Advantages	**Advantages**
• Breaks through proxy and caching servers - provides more accurate session tracking • Tracks client side events - JavaScript, Flash, Web 2.0 • Captures client-side e-commerce data - server-side access can be problematic • Collects and processes visitor data in near real-time • Allows program updates to be performed by your vendor • Allows data storage and archiving to be performed by your vendor	• 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 mobile visitors by default
Disadvantages	**Disadvantages**
• Setup errors lead to data loss – if you make a mistake with your tags, data is lost and you cannot go back and re-analyse • 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	• Proxy and caching inaccuracies – if a web page is cached, no record is logged on your web server • No event tracking – no JavaScript, Flash, Web 2.0 tracking • Requires program updates to be performed by your own team • Requires storage and archiving to be performed by your own team • Robots multiply visits

Are there alternatives? *(Continued)*

Network data collection devices – sometimes known as 'packet sniffers' – gather web traffic information from routers into 'black box' appliances. The downside of this is that the process can be expensive and complicated, and few vendors offer this method.

Another technique is to use a web server application programming interface (API) or loadable module. These programs extend the capabilities of web servers – enhancing and extending the logged fields – and streaming the captured data to a reporting server in real time.

The humble cookie

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

For web analytics, the main purpose of cookies is to identify users for later use – most often with a visitor ID number. 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. Aside from web analytics, web servers can also use cookie information to present personalised web pages. A returning customer might see a different page from the one a first-time visitor would view, a 'welcome back' message to give them a more individual experience or an auto-login for a returning subscriber.

Cookie facts:

- Cookies are small text files, stored locally, that are associated with visited web site 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 web site domain that a visitor requests directly either by typing in the URL into their browser or 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 web site domain setting the cookie information can retrieve this data. This is a security feature built into all web browsers.
- For third-party cookies, the web site domain setting 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.
- Cookies are no larger than 4 kilobytes.
- A maximum of 50 cookies are allowed per domain for the latest versions of IE7 and Firefox 2. Other browsers may vary (Opera 9 currently has a limit of 30).

Data collection issues affecting logfiles

One IP address registers as one person

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 Internet service providers (ISPs) assign different IP addresses throughout the session.

A recent US based comScore study (www.comscore.com/request/cookie_deletion_white_paper.pdf) showed that a typical home PC averages 10.5 different IP addresses per month. In which case those visitors will be counted as 10 unique visitors by a logfile analyser. This issue is becoming more severe as 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 by the use of cookies.

Cached pages are counted once

Client-side caching is where a visitor's computer stores a web page they've visited. The next time they look at that page, it will be served locally from their computer. This means that the site visit will not be recorded at the web server. Server-side caching is made possible by 'web accelerator' technology. This caches a copy of a web site to speed up delivery. It means that all subsequent requests a visitor makes to view that page are also served from the cache and not the site itself, again affecting visitor tracking. Today, most of the web is cached to improve performance. For example see Google's use of cache at www.google.com/intl/en/help/features.html#cached .

Robots multiply figures

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, etc) as well as those used for page scraping (price comparison, email harvesters, competitive research, etc). These affect web analytics because a logfile solution will also show all data for robot activity on your web site even though they 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 ones exist. For this reason, a logfile analyser solution is likely to over-count visitor numbers and in most cases this can be dramatic.

Logfiles see mobile users

All is not lost for logfile analysers. A mobile web audience study by comScore for January 2007 (www.comscore.com/press/release.asp?press=1432) showed that in the U.S., 30 million (or 19 percent) of the 159 million U.S. Internet users accessed the Internet from a mobile device.

For the vast majority of commercial of websites, the number of pageviews from mobile phones is currently very small in comparison with normal computer access. However, this number will continue to grow in the coming years. In fact, Japan and many parts of Asia are currently experiencing an explosive growth in mobile Internet access.

As most mobile phones do not yet understand JavaScript or cookies, logfile tools are able to track visitors who browse using their phones - something page tag solutions cannot do. The next generation of mobile phones is already increasing mobile pageview volume. Some can be tracked by JavaScript and cookies, such as the iPhone. However, maybe a superior tracking method will evolve for tracking mobile visitors.

Data collection issues affecting page tags

Setup errors cause missed tags

The setup of page tags causes a number of issues when trying to track visitors to a site. Where web servers automatically log everything, a page tag solution relies on the webmaster to add hidden tag codes to each page. Pages can get missed, even with automated page tagging or content management systems.

In fact, evidence from analysts at MAXAMINE who used their automatic page auditing tool (www.maxamine.com) 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. You can imagine the effect that might have on your visitor-tracking statistics.

JavaScript errors halt page loading

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

Firewalls block page tags

Another issue stems from corporate and personal firewalls that 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 "One IP address registers as one person", the comScore report shows that 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.

Data collection issues when using cookies

Visitors can reject or delete cookies

Cookie information is vital for web analytics because it uniquely identifies the visitor, their referring source and subsequent pageview data to them. The current best practice is for vendors to process first-party cookies only. The reason is 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 for the visitor to do this within the browser itself. By contrast, anecdotal evidence shows that first-party cookies are accepted by 95+ percent of visitors.

Visitors are also becoming savvier and often delete cookies. Independent studies 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 own and share 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.

Latency leaves room for inaccuracies

Web analytics accuracy can be affected by the time it takes for a visitor to become a customer – also known as 'latency'. For example, most low-value items are either instant purchases or made within seven days of the customer's initial visit to the web site. This short timeframe leaves little room for changes to a user's Internet setup, so your web analytics solution has the best possible chance of capturing all visitor pageview and behaviour information and reporting more accurate results.

With higher-value items, it is usually 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 their browser, upgrading their operating system, buying a new computer, or dealing with a system crash. Any of these occurrences can result in the user being 'seen' as a new visitor when they finally make their purchase. Off-site factors such as seasonality, adverse publicity, offline promotions or published blog articles/comments can also affect latency.

Offline visitor considerations

It is important to factor in problems unrelated to the method used to measure visitor behaviour but which 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 a visitor can print and take with them to claim a free gift, upgrade or discount at your store. If you would prefer to receive online orders, provide similar incentives, such as web-only pricing, free delivery if ordered online etc.

Another issue to consider is how your offline marketing is tracked. Without taking this into account, visitors that result from your offline campaign efforts will be incorrectly assigned or grouped with other referral sources and therefore skew your data. Using vanity URLs with redirection techniques are currently the way to do this.

Comparing data from different vendors

As has been shown, it is virtually impossible to compare the results of one data collection method with another. The association simply isn't valid. But given two comparable data collection methods – page tags – can you achieve consistency? Unfortunately even comparing vendors that employ page tags has its difficulties.

Factors that lead to differing vendor metrics include:

Cookies: First party versus third party

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 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 web site navigation, may navigate quickly, clicking on to another page before the page tag has loaded to collect data.

This was investigated in a recent study by Stone Temple Consulting (www.stonetemple.com/articles/analytics-report-august-2007-part2.shtml). They showed the difference between placing a tracking tag at the top of a page and one placed at the bottom, accounted for a 4.3 percent difference in unique visitor traffic for the same vendor's tool. Their hypothesis for the cause was the 1.4 second delay between loading the top of the page and the bottom page tag. Clearly the longer the delay the greater the discrepancy will be.

Also don't forget that 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 from other JavaScript and can sit comfortably alongside other vendor page tags – as shown in the Stone Temple Consulting report where 5 tools where compared on the same web pages. 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.

Tagging: Covering your bases

If you've tagged all of your web pages, what about tracking files that can't be page tagged, such as PDF, DOC, XLS and EXE? This may be a manual process, where the link to the file needs to be modified. This modification represents an event/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 web site audits to validate your page tags.

Pageviews: A visit or visitor?

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

How do different vendors compare?

The Stone Temple Consulting report referred to earlier (www.stonetemple.com/articles/analytics-report-august-2007.shtml), compared 5 different web analytics vendors with best practice implementations, simultaneously on 7 different websites. The results revealed that despite the very different technologies used, pageview counts varied only by +/-10 percent in most cases.

Cookies: Taking time out

The duration of timeouts – when 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 web site after 30 minutes inactivity is considered to be a repeat visit. However, some vendors offer the option to change this setting. Doing this can put numbers out significantly and 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 codes: Ensuring security

Depending on your vendor, your page tag code could be hijacked, copied and executed on a different or unrelated web site. This contamination results in a false pageview within your reports. Ensure hostname include filters are set up to record data from your web site domains only.

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 viewing a PDF file within your web browser, Acrobat 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 a 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 status code of 200 (download completed). So, a logfile solution can report a completed 50-page PDF file as one download and 50 pageviews.

E-commerce: Negative transactions

All e-commerce organisations have to deal with product returns at some point, whether it's because of damaged or faulty goods, order mistakes or other reasons. Accounting for these within web analytics reports is often forgotten about. 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 bullet-proof. For example, the removal/credit of a transaction usually takes place well after the original purchase and, therefore, in a different reporting time period.

Filters and settings: Potential obstacles

Data can vary if 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 different point in time during data processing.

Goal conversions versus pageviews: Establishing consistency

Consider a visitor traversing through your checkout process – as illustrated in Figure 3.

Five of these pages are part of your defined funnel – or 'click stream path' – with the last step (page 5) being the goal conversion or transaction. During checkout, a visitor goes back up a page to check a delivery charge (label A) and then continues through to complete payment. The visitor is so happy with the simplicity of the whole process, they then go and purchase a second item using the same path during the same visitor session (label B).

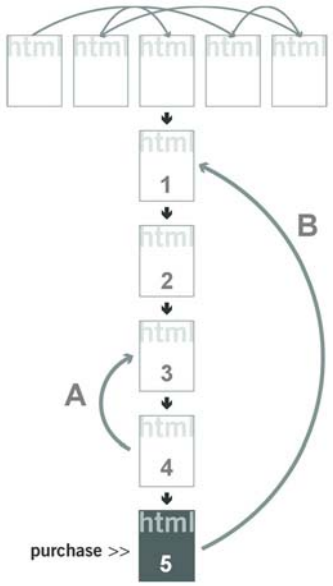

Figure 3 - A visitor traversing through a web site, entering a 5 page funnel, and making two transactions.

Depending on the vendor you use, this can be counted differently as follows:

- 12 funnel pageviews, 2 conversions,2 transactions
- 10 funnel pageviews (ignoring step A), 2 conversions, 2 transactions
- 5 funnel pageviews, 2 conversions, 2 transactions
- 5 funnel pageviews, 1 conversion (ignoring step B), 2 transactions

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

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.

Process frequency: Understanding glitches

This is best illustrated by example: Google Analytics does its number crunching to produce reports hourly. However, because it takes time to collate 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 process, 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.

Why paid search numbers often don't match

If you are using paid networks, i.e. pay-per-click (PPC), you will typically have access to the click-through reports provided by each network. Quite often, these numbers don't align with those reported in your web analytics reports.

This happens for the following reasons:

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.

Clicks and visits: Understanding the difference

It is important to remember that PPC vendors, such as Google AdWords, measure clicks. Most web analytics measure visitors that 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 under report visits from PPC networks.

Paid search: Important 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: A careful execution

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 is this behind? Because it allows time for the AdWords invalid click and fraud protection processes to complete their work and finalise click through numbers for your account. So, from a reporting point of view, it is important not to compare AdWords' visitor numbers for the current day. This is the same for all web analytics solutions and all PPC advertising networks.

Also bear in mind that, 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.

Third-party ad tracking redirects: Weighing in the factors

Using third-party ad tracking systems – such as Atlas Search, Blue Streak, DoubleClick, Efficient Frontier and SEM Director, for example – to track click-throughs to your web site 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 2-step hop is to allow the ad tracking network to collect visitor statistics independently of your organisation, typically for billing purposes. As this process involves a short delay, it may prevent some visitors from waiting. The result can be a small failure to align data.

In addition, 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=08

If added to a third-party tracking system for redirection, it could look like this:
www.redirect.com/?http://www.mywebsite.com?source=google&medium=ppc&campaign=08

The problem occurs with the second question mark – '?' in the second link – because you can't have more than one in a URL. Some third-party ad tracking systems will detect this error and remove the second '?' and the proceeding 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 process 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:
www.w3schools.com/tags/ref_urlencode.asp

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:

1. 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.

2. 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.

Summary and recommendations

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 showed a 10.5 percent reduction, for example, this figure will be accurate, regardless of the web analytics tool that was 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.

Here are 10 recommendations for web analytics accuracy:

1. Select the data collection methodology based on what best suits your business needs and resources.
2. Be sure to select a tool that uses first-party cookies for data collection.
3. 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.
4. Remove or report separately all non-human activity from your data reports, such as robots and server performance monitors.
5. Track everything. Don't limit tracking to landing pages. Track your entire web site's activity, including file downloads, internal search terms and outbound links.
6. Audit your web site for page tag completeness regularly. Sometimes, site content changes result in tags being corrupted, deleted or simply forgotten.
7. Display a clear and easy-to-read privacy policy (required by law in the European Union). This establishes confidence with your visitors because they better understand how they're being tracked and are less likely to delete cookies.
8. Avoid making judgements on data that is less than -hours old because it's often the most inaccurate.
9. Test redirection URLs to guarantee they maintain tracking parameters.
10. Ensure that all paid online campaigns use tracking URLs to differentiate from non-paid sources.

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 allow you to benchmark the performance of your web site. Achieving this means you're in a better position to then drive the performance of your online business.

Insight makes all the difference. Because there is so much room for error, web analytics is not 100 percent accurate, and taking web analytics reports at face value can be very misleading, even damaging. But measuring trends gives you more insight and knowledge of what's to come, as trends paint a clearer picture of what was. This knowledge will maximise the accuracy of your data and is a critical approach for success.

Acknowledgements

With thanks to the following people for their generous feedback in compiling this whitepaper: Sara Andersson, Alan Boydell, Jean-Baptiste Creusat, Tim Lee, Andrew Miles, Nick Mihailovski, Nicola Rae, Alex Ortiz-Rasado, Tomas Remotigue, Daniel Silander.