The RedEye Report

The inaccuracy of IP and Cookie-based
online Management Information
and what IS professionals can do about it

"Online tracking can be very complex. The 'RedEye Report' helped me see through the smoke and mirrors and to manage our business better."

Pat Gildea, npower, e-delivery Manager

Red Eye International Ltd
26 Grosvenor Gardens
London
SW1W 0GT

0845 094 1118

Executive summary

The vast majority of Web sites still use IP-Based server logs to analyse web usage. It is commonly recognised that IP-based tracking systems are the least accurate and so some companies who require more robust MI have turned to cookie based systems.

RedEye has provided Cookie-based web tracking solutions to UK corporations since 1997. Our clients require robust data to enable them to create more profitable customer relationships by getting more customers spending more money, more often.

We have monitored the ability of Cookies to build a picture of individuals' long-term usage of a Web site. However the consumer backlash against Cookies means a significant number of Internet users regularly delete them reducing the accuracy of even these systems to unacceptable levels for data-driven direct marketing. To counter this, RedEye has developed new software that uses log-in details as the primary method for identifying individuals.

Although opinions abound about the accuracy of both technologies only recently (Jupiter Report 2005) have there been any other empirical evidence. This has led to a great deal of confusion that this report dispels.

RedEye has developed technology to overcome this problem which also allows us to measure the accuracy of IP and cookie systems by comparing them to a known set of data. We publish the results from those comparisons here for the first time in 2003. Since the study, there is continual work with RedEye's clients to keep track of these errors to monitor and react to consumer activity

The study is based on analysis of two of the UK's largest ecommerce web sites, www.asda.com and www.williamhill.co.uk. The results are staggering:

> **The IP-based approach overestimated unique visitors by up to 7.6 times in a single month while a Cookie-based approach overestimated unique users by up to 2.3 times in the same period.**

The report concludes that log-on IDs combined with session cookies should be used for data-driven marketing (reactivating lapsed customers for example). For sites where access is not restricted to registered users, a Cookie-based approach combined with appropriate weightings is the only way to ensure that online management information is sufficiently accurate for strategic decision making.

The online management information used by the majority of companies today is so fundamentally flawed that decisions based on it are likely to be dangerously inappropriate.

The value of this study to managers

In the face of growing hostility to cookies and as online technology develops a debate rages about the relative merits and accuracies of IP and cookie-based web analytics. This is not a trivial issue.

As more commerce is transacted online, managers need information to allow them to:

- Value online businesses
- To invest advertising budget
- To understand customer loyalty
- To discriminate between profitable and loss making activity
- To calculate if conversion rates warrant expensive site redesign or not.

Suppliers of tracking technology offer a range of conflicting opinions about the accuracy of the information on which managers make these decisions. However in keeping with our long-term commitment to Clients and our leading position in the industry, RedEye has conducted an empiric study of information accuracy using data from two of our Clients (and two of the largest ecommerce Web sites in the UK) www.asda.com and www.williamhill.co.uk.

We believe this study helps online professionals better understand the data they have now, highlights what they can do to improve it and assists them with their IT investment decisions.

What we did

A detailed description of our methodology can be found in the appendices. To summarise here RedEye ran a full test comparing the two main technologies of IP-based and Cookie-based tracking.

The key to the study was a robust set of raw data about which accurate information on consumer behaviour was known. This known set can be created by RedEye because our advanced technology can use log-in details as the basis for the unique identifier giving completely accurate information at a customer account level.

Our 'control' results were verified using an independent third-party web analytics tool and were found to be more than 99% accurate.

Against this known set of data, IP and cookie methods were applied and the results compared.

www.asda.com and www.williamhill.co.uk were chosen because:

- There is a high propensity for customers to make multiple repeat purchases in a 28 day period
- Logged in users account for more than half of all page impressions on both sites
- The sites appeal to very different types of consumers

Standard Web Metrics – the figures used by the majority of businesses

Table 1 shows the web data inaccuracies over a single day

Data collection method	Pages viewed	Total Visitors	Repeat Visits	Total Visits
True activity[1]	100	100	100	100
IP Address	100[2]	261	1010[3]	361
Cookie data	100	113	90	99

Using the IP address, the average number of distinct visits reported each day was overstated by up to 3.6 times

Using the cookie, the average number of repeat visitors reported each day was understated by up to 0.9 times – this is due to a visitor logging back on their account before the 30 minute time out of the session had elapsed

Table 2 shows the web data inaccuracies over seven days

Data collection method	Pages viewed	Total Visitors	Repeat Visits	Total Visits
True activity	100	100	100	100
IP Address	100	502	305	361
Cookie data	100	158	117	98

Using the IP address, the average number of repeat visitors reported each week was overstated by up to 3.1 times

Using the cookie, the average number of repeat visitors reported each week was overstated by up to 0.2 times (caused by some regular customers deleting their cookie between visits or accessing the site from a different computer)

Table 2 shows the web data inaccuracies over 28 days

Data collection method	Pages viewed	Total Visitors	Repeat Visits	Total Visits
True activity	100	100	100	100
IP Address	100	760	276	361
Cookie data	100	228	155	99

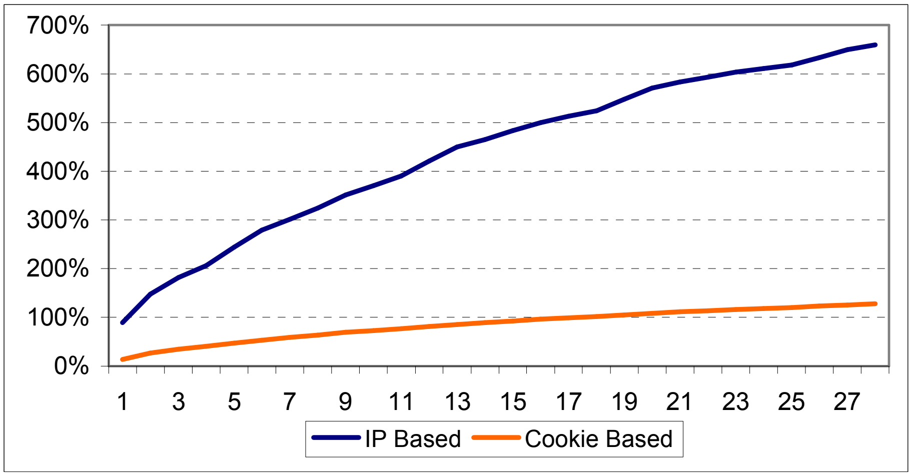
Using the IP address, the total number of visitors reported each month was overstated by up to 7.6 times

The total number of visitors reported using a Cookie-based approach was overstated by up to 2.3 times for a 28-day period

[1] True activity is accounted for by each visitor logging into their online account, and this data has been indexed to 100 to protect client confidentiality

[2] This measurement excludes the issue of page caching at the ISP and browser level

[3] Figure may be skewed due to a relatively small sample size across one day



Graph 1 illustrates how the inaccuracies of identifying returning visitors mount up. As the graph suggests, the level of error increases over time where a line representing log-in data on the graph would be a flat line along the bottom (i.e. at 0%).

Whilst the core study was done over a 28-day period, it was extended to 90 days for one site to see how the errors changed over time. The error for cookie-based analysis more than doubled from 28 days to 90 days.

Beyond Top Line Metrics

In addition to high-level management information, many companies now need to mine their data to customer level. Whilst relatively few companies today use their data to drive personalised marketing campaigns, this is something that increasing numbers would like to be able to do. This section looks at how accurate the data that drives such activities will be with an IP-based or Cookie-based approach.

To give an indication of how suitable each technology is for more advanced purposes we have measured the percentage of visits that were successfully identified as having started at the same time, contained the same pages in the same order, and finished at the same time compared to our control.

Tables 4-6 show the lowest percent of correct records found in the data for both sites across different time frames using an IP-address and cookie approach

Table 4 shows web data inaccuracies across a single day

Data collection method	Paths	Browser History	Tracked Completely
True activity	100%	100%	100%
IP Address	14%	16%	43%
Cookie data	93%	72%	81%

Table 5 shows web data inaccuracies across seven days

Data collection method	Paths	Browser History	Tracked Completely
True activity	100%	100%	100%
IP Address	14%	14%	27%
Cookie data	93%	40%	63%

Table 6 shows web data inaccuracies across 28 days

Data collection method	Paths	Browser History	Tracked Completely
True activity	100%	100%	100%
IP Address	14%	8%	22%
Cookie data	93%	22%	50%

Across all time frames, there is only a 14% chance that a path reported using the IP-address approach actually happened.

This 14% chance of path accuracy will also lead to similar errors in entry pages, exit pages and dwell times.

Across a whole month, only 22% of visitors can be tracked completely if you use the IP-Address approach.

Accurate path analysis can be carried out using a cookie based approach.

Less than a quarter of cookie based profiles were correct after 28 days rendering cookies useless for long term personalisation

Half of all visitors to one of the sites deleted their cookie, or accessed the site from another PC over the 28 day period

What causes these errors?

IP-based Figures

IP analysis causes the greatest inaccuracies for a range of reasons including:

- The IP address that is recorded is hardly ever the IP address of the visitor's computer. It is more usually the IP address of a computer at the Internet Service Provider (called a proxy server).
- Originally ISPs allocated a dedicated proxy server (and IP address) for each user for as long as they remained connected to the Internet. This is a very inefficient way to use the available hardware and the new generation of ISPs load balance proxy servers on a page-by-page basis. This means that each user's IP address can change each time

they request a new page. One person looking at three pages may be interpreted to be three people looking at one page each. Impossible data about entry and exit pages are one symptom of this problem.

- Company users accessing the Internet from one external-facing server can share the same IP address. This means that if two people in an office visit the same web site they will appear to be the one person visiting twice.
- Numbers of users and web domains is continually going up but there are a fixed number of IP addresses available, so different users frequently share the same IP address. To save costs ISPs also tend to maintain only a small number of IP addresses, which they share amongst their entire user base.

All these factors mean that IP based tracking systems can over and under count the number of visitors to a site. IP addresses are simply incapable of accurately tracking individuals. This report proves that the argument that the various errors cancel themselves out on an aggregated basis is not true.

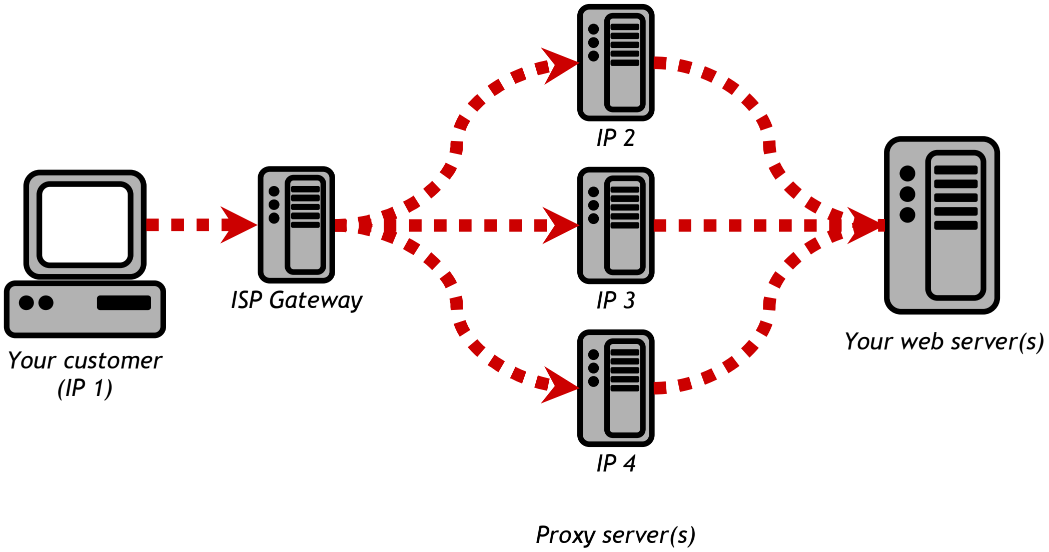


Figure 1 – How the customer's true IP address is hidden behind their ISP. The same customer can appear as any of the IP2, IP3 or IP4 within a short period of time but never as IP1.

Cookie-based figures

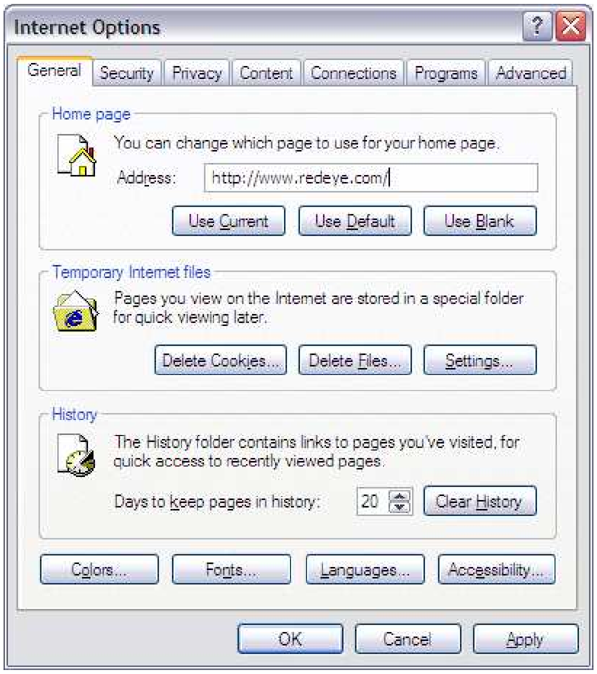
Whilst Cookie visitor figures were much more accurate than IP-based, alone they were still too inaccurate to provide a clear basis on which to make certain decisions.

Under a Cookie-based approach the visitor is required to accept a unique reference that is used every time they request a page from the site. The visitor can choose whether or not to accept the Cookie via the settings in their Web browser. Early versions of Internet Explorer automatically accept all Cookies by default.

Internet Explorer 6 gives the user more control over Cookies, enabling them to accept Cookies based on who they are from, how long they are retained for, and for what purpose any data associated with them will be used.

As well as making it easier for people to control what types of Cookies are accepted, Internet Explorer 6 also makes it easier for people to delete Cookies by pressing a button on the first page of the Internet Options.

RedEye believes that having the button next to the ‘Delete Files’ button encourages people to delete Cookies as part of good housekeeping and to save space. This is almost always unnecessary given that Cookies usually take up less than 100 bytes.



Home page

You can change which page to use for your home page.

Address: http://www.redeye.com/

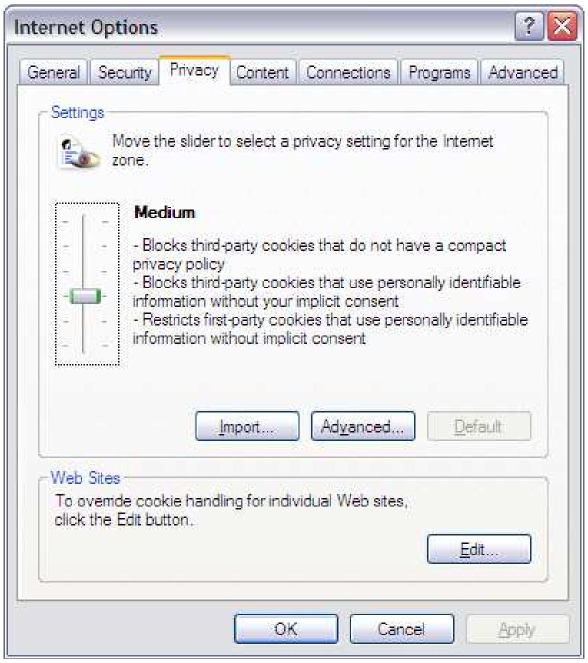
Temporary Internet files

Pages you view on the Internet are stored in a special folder for quick viewing later.

History

The History folder contains links to pages you've visited, for quick access to recently viewed pages.

Days to keep pages in history: 20



Settings

Move the slider to select a privacy setting for the Internet zone.

Medium

- Blocks third-party cookies that do not have a compact privacy policy
- Blocks third-party cookies that use personally identifiable information without your implicit consent
- Restricts first-party cookies that use personally identifiable information without implicit consent

Web Sites

To override cookie handling for individual Web sites, click the Edit button.

The other reason why Cookie-based data could never be perfect is because people can access the Internet from different locations. RedEye commissioned independent research to shed some light on whether it was access from multiple machines or people deleting Cookies (or both) that led to such a large error in Cookie-based data.

Conducted by NOP, the study of 1,000 British Internet Users (definition: spend at least one hour online in average week, representing 80% of the GB Internet base) suggests that even larger errors may exist on some sites:

- 50% of respondents said they had used more than one computer in the last three months
- 70% of respondents said their computer was used by more than one person
- 20% said that they only accepted session-Cookies

- 71% of respondents were aware of Cookies and accepted them. Of these, only 18% did not know how to delete Cookies. 55% were deleting them on a monthly basis
- 89% of respondents who knew what Cookies were and how to delete them said that they had deleted them at least once in the last three months

RedEye believes that as more consumers become aware of Cookies in the future, more will delete Cookies and Cookie-based tracking will be less accurate as a result. This means that Cookie-based systems will not be able to meet the future demands of marketers looking for accurate data about individual customer behaviour to support a customer-centric marketing approach.

What is your data good for?

Based on our analysis we have created a best practice guideline showing what you can do with your data based on the method of collection you use.

If you use an IP-based tracking solution the following metrics can be obtained to a reasonable standard (unless you web servers are affected by page cashing at the ISP or browser level):

- Pages viewed
  - Pages Impressions by URL
  - Page Impressions by Content Area
  - Banner Impressions
  - Referrers

Using a Cookie-based tracking solution the following metrics can be obtained to a reasonable standard:

- Visits
  - Frequency of Visit
  - Pages per Visit
  - Time per Visit
  - Visit Conversion Rate
  - Revenue Per Visit

- Correct Paths
  - Page Time
  - Path Analysis
  - Entry Pages
  - Exit Pages
  - Session Based Marketing ROI Analysis

If you weight the results from a cookie-based tracking system the following metrics can be obtained to a reasonable standard over a longer period. Without weightings cookie tracking should only be used to analyse a single days activity.

- Visitors

- Conversion Ratio
- Revenue per Visitor
- New ÷ Repeat Visitors
- Banner Reach
- Banner Frequency

The following metrics and activities can **only** be achieved by a solution that re-identifies visitors **by log-in** or similar:

- Complete Browsing History
- Lifetime Marketing ROI Analysis
- Registered Visits ÷ Visitors
- Customer Visits ÷ Visitors
- % of new visitors who become repeat visitors
- eCRM emails based on past browsing or purchasing

Conclusion

The adage that once a problem gets measured it gets managed holds true.

Whilst people have long suspected the inadequacies of IP/Browser-based systems, few could have realised the limitations of the technique that was used to value 'dot com' businesses during the 90s. Clearly, as a marketing and business aid, IP-based data is of little use and companies using bargain-basement systems to support business critical decisions do so at their peril.

This report will come as mixed news to marketers. It is likely that they have far fewer people visiting their site than they previously thought but that both the repeat visit rate and conversion ratio will be significantly better that their management information is currently showing them.

The report provides mixed news for online advertisers too. Their campaigns may not be reaching as many people as they have previously believed but the problems with using Cookies to track responses over a 28-day period means their return on investment may be as much as double what they think it is.

Companies looking to personalise their online customer communications using Cookie-based data should pay particular attention to the results. The ability to track consumers using only Cookies degrades significantly over time. Using this technique for data collection means that marketers cannot be sure that they are sending the right message to the right person at the right time.

The question that comes to mind when reviewing these results is how much we can believe those old clichés of the online world; "customer conversion is very poor, on average only 3.5%" and "75% of all baskets are abandoned during a session". With this new evidence it seems likely that the average web site conversion rate could be in the region of 10 to 15%.

RedEye's approach is to identify visitors via log-in wherever possible and we will soon release a solution that automatically weights Cookie-based data when visitors cannot be positively identified.

In our work with our Clients we have discovered that the correct weighting is different for sites depending on the profile (IT sophistication, home vs. office use and sensitivity to privacy etc) of the particular customer base.

If you would like to examine if your data is able to support your business objectives and how to improve the accuracy of your management information then please call Bertie Stevenson on +44 (0) 20 7953 0268.

Copyright Statement

Appendices

Methodology

The advanced technology that RedEye has developed allows log-in details to be combined with Cookies to give completely accurate information at a customer account level. Extracting all visits to a web site, where a Cookie was accepted and the customer logged-in, provided RedEye with a complete set of results to benchmark against.

The study was conducted using data provided by two of the largest ecommerce Web sites in the UK, www.asda.com and www.williamhill.co.uk. The study was conducted over two different 28-day periods chosen at random between March and October 2003.

It was decided not to filter any of the data to remove non-human traffic, as non-human traffic is highly unlikely to log-in.

To verify the results RedEye created pseudo IP addresses from the Cookie and customer reference number of each user. These were processed using a third-party software tool in "IP-only mode". The visitor and visits metrics agreed with those obtained using RedEye's own technology by more than 99%.

Visit Definition

For the purposes of the test, visits were defined as:

- Continuous activity from a visitor.

- If no activity was detected from a visitor for 30 minutes then the visit was deemed to have ended.
- All visits were terminated at midnight to ensure visit figures could be summed over different date ranges and would still reconcile.

IP-Based User Identification Method (IP Combined with User Agent)

The full IP address was combined with the full "HTTP Environment String" to provide the primary identifier for individual users. The "Environment String" includes information about the operating system and browser being used and often includes other information such as the ISP that provided the browser, for example – "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; CDsource=v2.15e.03)" is one generated by the UK's largest ISP Freeserve.

The main limitation with IP/Browser based analysis is that the IP Address used is often one at the ISP level and not the one for the machine the customer is actually using. This does not only mean that multiple people can have the same IP address with a short period of time but also that a visitor's IP can change continuously during a session.

The main limitation with IP/Browser based analysis is that the IP Address used is often one given to the ISP and the machine the customer is using. This not only means that multiple people can have the same IP address but also that a visitor's IP can change continuously during a session.

Cookie-Based Method (Persistent Cookies Only)

RedEye's tracking software requests users store persistent Cookies, which contain a unique 36-digit number. If Cookies are deleted or only session Cookies are accepted, then a new unique value will be given to them the next time they visit the site.

Log-In Based Method (Log in combined with Session Cookie)

This method combines Cookies, with a unique identifier that is passed whenever the user logs in to a site. All activity is then tied back to the specific unique login reference, regardless of the Cookie. Therefore if a person has three different Cookies during the four-week period, the data from these three Cookies will be combined under one using the person's unique identifier, thus reporting the customers' activity, as opposed to the Cookies activity.