quαntcast

Cookie Corrected Audience Data — White Paper

Leveraging Multiple Data Sources to Calibrate Unique Cookie, Machine, and People Counts in a Direct-Measurement Media Economy

6.24.08

Quantcast Corporation

201 3rd Street, 2nd Floor
San Francisco, CA 94103

Phone: 415.738.4755
Email: contact@quantcast.com

Executive Summary

Digital media is dramatically more fragmented than traditional media. This transition brings enormous opportunity for addressable advertising solutions to deliver more relevant content to consumers, but it has also created a discontinuity in media measurement and planning. Panel-based services (such as comScore and Nielsen//NetRatings) use a sample based methodology to estimate audiences. Unfortunately, panels do not deal well with the fragmentation or piecemeal distribution (think videos, widgets and advertisements) of today's Internet, resulting in inaccurate data, or simply no data at all, for many online media properties.

The Internet enables direct-measurement of every media consumption event, but the atomic unit of this census approach is the cookie which is not necessarily equivalent to the people-based projections from panel (sample) services. People equals reach and is of central importance to media planners.

A new approach to audience measurement leveraging the accuracy and actionability of directly measured data and providing complimentary people based accountability is required for the continued growth and sustainability of the ad-supported media industry.

Addressing this challenge requires innovation in bridging the inherent differences between the directly measureable cookies and estimates of people. This is not unlike challenges that the media industry has previously faced in bridging the "what" (raw traffic) and the "who" (actual people) in Print. The magazine industry provides an analogy to the challenges in the digital space in that circulation is a "census" level data point (it represents the reported number of magazines printed and distributed) which is coupled with panel data for people-based adjustments (readers per copy, demographics, etc). As we shall see, in digital the correction factors are reversed (less people than cookies, rather than more people than circulation) but the hybrid approach leveraging raw circulation and panel data is relevant.

The Quantcast model leverages census level data, directly measured in cooperation with online media publishers, integrated with panel sources in a hybrid statistical model that accounts for the complicating factors of online media measurement, namely cookie deletion and multiple machine usage.

It's complex. But it is not insurmountable. At Quantcast we believe that an audience measurement methodology that starts with the widest array of available data and accounts for marketplace realities is far superior to starting with a limited panel-based view of the world which is used to project outward. As media fragments, and becomes more dynamic, the latter approach breaks down.

This whitepaper provides an overview of Quantcast's approach to translating raw cookie data into actionable, people-based audience intelligence as part of our broader pixel and panel methodology. It explains of how we model cookies, account for deletion and non-acceptance, and incorporate this data within the broader set of audience measurement challenges.

Panels versus Direct-Measurement

Online panel and direct-measurement solutions each bring distinct advantages and disadvantages to the audience measurement arena. More importantly, used on their own, these differing approaches create confusion in the marketplace. As this New York Times article points out, publishers often have a difficult time reconciling their internal census-based web measurement with unique audience estimates from Internet panels (*How Many Site Hits? Depends Who's Counting* – NYT 10/07).

Panels have long been the standard for measuring people-based audiences. It would be incredibly difficult, if not impossible to track every TV viewer or magazine reader, so traditional audience measurement has relied on projections from panels. Online audience research providers like comScore and Nielsen//NetRatings measure individuals on unique machines within their panels of web users (for whom they collect survey data), and extrapolate total audience sizes based on population weightings and enumeration.

Panel-based audience measurement has always had its flaws. In a world of limited media fragmentation (i.e. few media choices), panels were efficient ways to measure what people were consuming – and they were able to adequately account for bias that had material impact to projections. However, in the online world of high-fragmentation (millions of sites, dynamic consumption and distributed experiences), limited sample sizes and unknown/varied biases quickly erode the ability of a panel to provide detailed audience visibility.

While the advertising industry has leveraged syndicated panel data for audience insights, web site operators have increasingly relied on web analytics solutions (Omniture, Webtrends, Google Analytics, etc.) to directly measure all traffic to their own web site for insights on navigation, commerce and promotion activities. Web analytics solutions measure unique cookie counts, sessions, visits, page views and duration metrics via directly measured "log file" data.[1] While this methodology provides high accuracy based on raw traffic (assuming non-human activity is properly filtered), it does not provide any visibility into person-level measures (demographics, duplication, or other criteria).

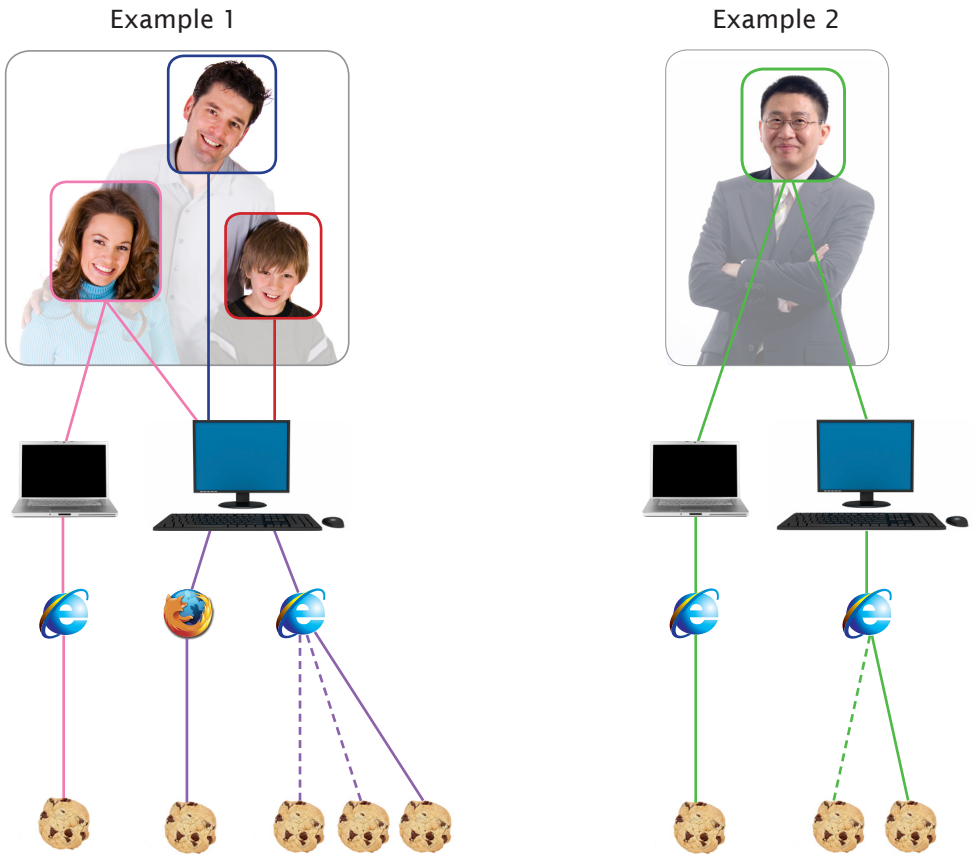
Direct measurement utilizes cookies, small text files recording a visit of an internet browser to a particular web site, as the atomic measurement unit. Cookies are blind to the person who is doing the internet browsing and in isolation do not necessarily accurately reflect the precise number of people visiting a given web destination. The complicating issues include cookie deletion, non acceptance of cookies by a browser, multiple machine/device use and multiple people using the same machine.

For example, when web users delete Internet cookies, the same browser can generate multiple cookies for a given site. As a result the number of cookies is greater than the number of people. Additionally, the many-to-many relationship between people and internet accessible machines adds complexity to media measurement: multiple people in a household may use one machine (and access the same sites), and individual users may use multiple machines to access a given site. The varied way in which all these factors may combine results in unique differences between cookies and people for every media property.

Figure 1 (next page) provides a representation of how cookie deletion and machine use impact the relationship between people and cookies. In the first example, three family members visit ABC.com. The mother accesses the site from Internet Explorer on her work laptop and generates a single cookie (recording her 20 visits, site preferences, page views and duration), she also uses Internet Explorer on the PC at home – she does not delete her cookies. From their home PC, the father access ABC.com through Internet Explorer and deletes his cookies twice in March (generating three cookies in total, the same cookies as the mother for the home PC[2]). Their son also accesses ABC.com on the home PC using Firefox, generating an additional cookie. Directly measured server-side counts total five cookies in March for the three people in this example. In Example 2, a single male accesses ABC.com from both his home laptop and work PC using Internet Explorer. He does not delete cookies on his home machine, but his work PC automatically deletes cookies monthly (generating two cookies in March). Directly measured server-side counts total three cookies for this single person in March.

1 There are differences between server-side and client-initiated log analysis but they are not important to our discussion here.
2 Here assuming the same login on the machine.

Figure 1: Examples of the Relationship Between Unique People and Unique Cookies



Example 1:

1 PC with 1 unique user
1 PC with 3 unique users
3 unique browsers generating 5 cookies in a given month

Example 2:

2 PCs with 1 unique user each
1 PC with 1 unique user
2 browsers generating
3 cookies in a given month

Cookie deletion, multiple machine use and more than one person per machine are critical factors to address as part of a census-based audience measurement approach.

Cookie Deletion in Context

What are cookies? A cookie is a small text file that your browser uses to identify you to individual sites. For example, a cookie might be used to automatically log you in to your web-based email, or may store your personal preferences on your favorite news site. Cookies are also used to determine aggregate statistics regarding the level of activity that a particular web site might experience.

Cookies improve the overall web experience for consumers and they also offer marketers an opportunity to improve ad delivery and targeting relevance – a benefit that is just starting to be leveraged at scale.

Why (and how) do cookies get deleted? Web users may consciously or unconsciously delete their cookies for a variety of reasons. Security software like “Panicware” and “Windowwasher” automatically deletes cookies on a regular basis. Some browsers such as IE6 have an option allowing users to delete cookies after every browser session. Users can also manually delete their cookie files by taking simple actions within their browser. As web services and platforms evolve, additional situations resulting in added complexity to the cookie issue are expected.

Prior Cookie Deletion Estimates: The online advertising community has spent significant time focused on cookie deletion. Various industry studies (including ComScore, Jupiter Research, Atlas Institute) give a glimpse into the cookie deletion phenomenon that Quantcast has studied in detail. They have focused on quantifying the percentage of the online audience that deletes their cookies on a regular basis (depending on the study, 39%

to 55%). ComScore's study concluded that unique cookie counts could be overestimating people by up to 150% depending on the frequency of site visitation. The example used by comScore to illustrate this overstatement is Yahoo!, which serves as a home page or mail application for many users (a fact that drives unusually high frequency of visitation). However, for the majority of properties on the web, the impact of cookie deletion on unique audience estimates is much lower.

The factors that impact cookie deletion are complex and dynamic, and the impact they have on audience measurement is varied by individual property. While direct measurement solutions must address cookie deletion as they attempt to account for people, they cannot do so based on aggregated industry averages. As illustrated later in this paper, the impact of cookie deletion on unique audience estimates varies substantially based on the type of website measured. At Quantcast, we have a dynamic model to adjust for cookie deletion based on the attributes of a given web site.

Additionally, the overall magnitude of cookie deletion could dramatically change at any point if, for example, a browser enhancement launches that impacts cookie deletion in material ways, relative to current standards. For this reason, any approach employed by measurement services also must be flexible and accommodate marketplace change.

The Quantcast Approach

Quantcast has a hybrid approach to web site audience measurement that leverages access to multiple complementary data sources. The Quantcast Quantified Publisher program captures over 75 billion media consumption events every month, generated by more than 1.4 billion cookies (data as of June, 2008). What's more, many of our Quantified Publisher partners share anonymous identifiers with us that are independent of cookies. Our model also includes several panels which provide for people-based reference points and calibration which are free of cookie deletion. We triangulate across this mass of data with different collection processes, biases and issues. Our models take into account visit frequency, time periods, the likelihood of multiple computer usage and even the impact of multiple people using the same computer to deliver people based estimates.

Our model for translating unique cookies to people has been validated using hold-out samples and independent data sets. Further, our model is dynamic and recalibrated on an ongoing basis to reflect the evolving nature of Internet traffic patterns.

Translating Unique Cookie Counts to People

By studying the rate of cookie deletion across tens of thousands of sites (via both directly measured and panel based data) Quantcast is able to isolate factors that impact relationships between unique cookies, machines and people. Quantcast statisticians are dedicated to understanding the dynamic, property-driven factors that impact cookie per machine, number of unique users per machine, as well as the number of machines that are used by people.

The primary factors impacting translation from cookies to people are:

- *Time period under analysis* – During any short period of time, for example one day, there is a limited opportunity for a user to visit a website, delete their cookie and then visit the same website again (and be re-cookied). When a longer time period is the basis for evaluation, there is a much greater chance that a user could visit multiple times, and that cookie deletion would result in duplicative counts (of people).

- *Typical visit frequency* – Coupled with the time period factor is the issue of typical visit frequency. Consider the example of a user who deletes cookies every day. If a particular site has a group of very addicted users who return frequently, even if a small number delete cookies daily, the result will significantly inflated numbers of cookies relative to the number of actual people who visited the site. In such cases it's not unusual to see an average of two (or more) cookies for every user over the course of a month - of course, most users only result in one cookie, but a small number generate many cookies. Consider the same cookie deleting example, but this time for a site that tends to have mostly passers-by, i.e. visitors that only go to the site once over a given time period. In this extreme, cookie deleting users

have no impact on the total number of cookies for the site, because they don't return to the site after deleting the cookie.

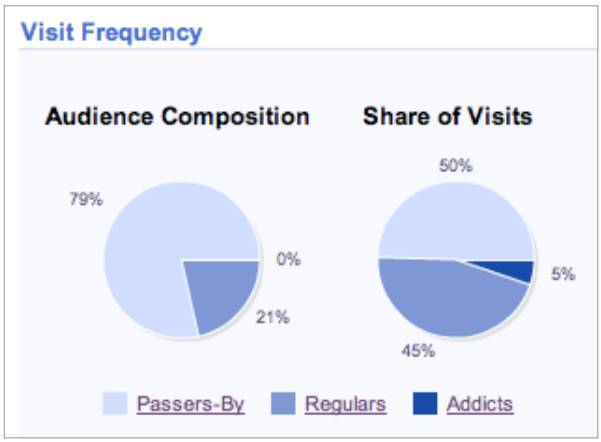
- *Percentage of usage from work or public machines* – sports, mail, and weather sites are all examples of sites that get accessed both at home and at work.

- *Type of site* – portals for example, tend to have multiple people per machine accessing them.

To best illustrate how Quantcast's dynamic cookie to people translation model addresses the complexities of today's online environment, here are two very different property data sets.

Example 1 – "Reference" Site

These sites have a relatively low frequency of visit profile. As the following Quantcast analysis confirms, the majority of this site's visitors only frequent the site 1X per month (Passers-By = 1X, Regulars = 2-29X, and Addicts = 30X+ per month):

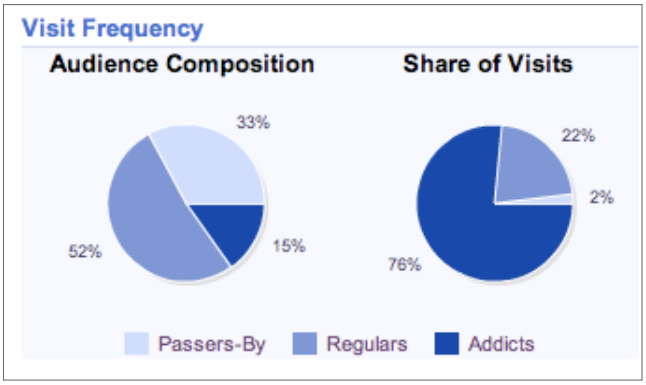


It does, however reach a large number of internet users every month, is consumed at both home and work, and generates most of its traffic from the US. Based on Quantcast's property-centric model, there is a fairly limited deflation in cookie to people counts (a drop from 25.8 million cookies in the US, to 22.2 million people – or a 14% decrease.

Monthly Traffic Summary	US	Global
Page Views per Month	75,138,688	134,618,952
Average Page Views per Visit	2.11	2.21
Visits per Month	35,624,309	60,962,099
Cookies per Month	25,759,835	43,448,640
Home Traffic	64 %	67 %
Work/Education Traffic	36 %	33 %
People per Month	22,215,781	38,141,063
Monthly Page Views per Person	3.38	3.53
Monthly Visits per Person	1.60	1.60

Example 2 – High-Profile Blog (with frequent daily updates)

Unlike sites represented by Example 1, these types of properties have a relatively high frequency of visit profile. A small number of "addicts" generate the majority of site visits.



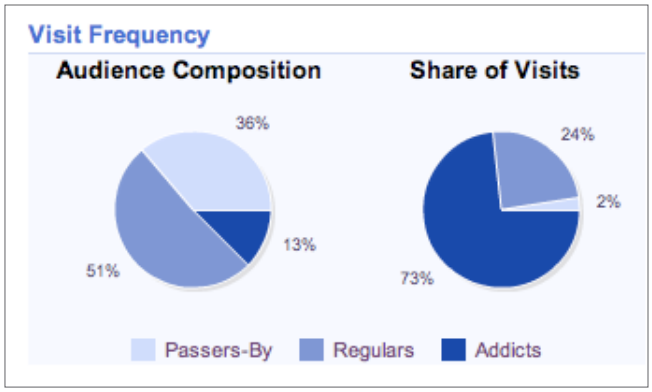
The site in this example, however, reaches a smaller number of internet users every month and is also consumed at both home and work. In this particular case, based on Quantcast's property-centric model, there is a significant translation in cookie to people counts (a drop from 6.4 million cookies in the US, to 3.1 million people – or a 52% decrease).

Note how the visits and page views per person increases proportionately when compared to the corresponding numbers on a per cookie basis.

Monthly Traffic Summary	US	Global
Page Views per Month	145,734,456	200,078,287
Average Page Views per Visit	2.67	2.73
Visits per Month	54,491,203	73,320,180
Cookies per Month	6,441,716	9,121,243
Home Traffic	69 %	72 %
Work/Education Traffic	31 %	28 %
People per Month	3,149,961	4,631,924
Monthly Page Views per Person	46.27	43.20
Monthly Visits per Person	17.30	15.83

Example #3 – A Social Network

Lastly, here is an example of a social network with significant international usage. Like Example #2, this site enjoys heavy addict consumption, as illustrated by the following Quantcast visit frequency analysis.



In the US, this social network is relatively small, and unlike our prior two examples, it has limited workplace consumption (users of this site tend to be much younger, and not in the workforce). Based on Quantcast's property-centric model, there is a significant deflation in cookie to people counts in the US data (a drop from 5.7 million cookies to 3.5 million people – or a 40% decrease).

Notice, however, that internationally, this site is much larger. When looking at global data, the deflation is even more pronounced (a drop from 96.5 million cookies to 52.7 million people – or a 45% drop.

Monthly Traffic Summary	US	Global
Page Views per Month	1,021,610,553	21,351,935,602
Average Page Views per Visit	31.78	31.03
Visits per Month	32,142,088	688,201,186
Cookies per Month	5,722,728	96,492,745
Home Traffic	81 %	86 %
Work/Education Traffic	19 %	14 %
People per Month	3,461,156	52,652,590
Monthly Page Views per Person	295.16	405.52
Monthly Visits per Person	9.29	13.07

Direct Measurement and Panel Calibration: Providing the Key for Cookie to People Translation

While the internet (and digital distribution) have made the consumer media experience much more relevant and interactive, audience measurement has become a more complicated science. The variables that impact audience projection are wide-ranging and dynamic by property. It is clear that as fragmentation, device proliferation, and consumer control expand, these challenges will only grow.

Building scalable data models which provide capabilities for volume processing and dynamic analysis are critical requirements for the media industry. This white paper provides an initial view into the benefits that a changed approach can yield.

As an industry, we have a variety of options:

- *The Status Quo* – Publishers can continue to rely on directly-measured cookie-based log-file data, and marketers can continue use of people-based panel services, preferring consistency to accuracy

- *A Blanket Adjustment* – We can assume (and apply) a standard industry conversion factor to all directly measured site data based on the estimated percentage of people who delete cookies. While this would facilitate an "easy" approach, it would not provide sellers or buyers of advertising with the intelligence needed to make effective investment decisions.

- *Hybrid Modeling & Calibration* – By using multiple data sources as input into a property-centric model, the industry could incorporate best-of-breed data sources (actual number of cookies as well as machine usage data and 3rd party panel inputs). This approach takes multiple factors into account, and provides a highly dynamic model to accommodate the fragmented landscape.

Quantcast's groundbreaking cookie to people translation model demonstrates that direct measurement is a viable way to estimate audience size. We expect models to continually be enhanced, and to become even more sophisticated as more of the media landscape moves to digital distribution with a "two way" path.

We will continue to innovate.

About Quantcast

Quantcast (www.quantcast.com), was founded by a team of engineers and mathematicians committed to addressing the digital media market's need for transparent and actionable data. Through its services, Quantcast provides advertisers and publishers unmatched abilities to measure, organize, and transact based on directly-measured audience data.

The company's scientific advisory board is led by two of the world's most prominent applied statisticians – Prof. Trevor Hastie and Prof. Jerome Friedman, both of Stanford University.

Quantcast's Quantified Publisher program is available to web properties of any size and provides free direct-measurement for web sites, blogs, videos, games and widgets. The service has experienced rapid adoption by top media companies including CBS, Hachette-Filipacchi, FOX Entertainment, Digg, Wordpress, BizJournals, Bloomberg, HuffingtonPost, National Geographic, Belo, eHarmony, McClatchy, Babycenter.com, Scientific American and Time Inc.'s SI.com, CNNMoney.com and People.com.

Quantcast Corporation

201 3rd Street, 2nd Floor
San Francisco, CA 94103

Phone: 415.738.4755
Email: contact@quantcast.com