How Google Analytics and Conventional Cookie Tracking Techniques Overestimate Unique Visitors

Max I. Fomitchev
Pennsylvania State University, CS&E
111J IST, University Park, PA 16802
1-814-863-1469
fomitchev@psu.edu

ABSTRACT

We report the results of the analysis of website traffic logs and argue that both unique IP address and cookies vastly overestimate unique visitors, e.g. by factor of 8 in our studies. Google Analytics 'absolute unique visitors' measure is shown to produce similar 6x overestimation. To address the problem we present a new model for relating unique visitors to IP address or cookies.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online information services – *commercial services, data sharing, web-based services.*

General Terms

Algorithms, Measurement, Economics, Experimentation, Human Factors, Theory.

Keywords

Unique visitors, web analytics, web, traffic logs, cookies.

1. THE CASE FOR RECURRING TRAFFIC

Proliferation of online enterprises and web-based commerce has put an unprecedented importance on web traffic analysis [1]. For successful analysis one has to understand that meaningful and useful traffic is usually recurring: i.e. if a user likes the site he or she will periodically return to it as opposed to a user who stumbled upon the site for the first time. To test this hypothesis we have conducted a survey of 100 randomly picked web users (age 20-50, median age 30; 90% male; 80% North America, 15% Europe). The survey revealed that virtually all respondents had a short list of their favorite online resources that they were visiting daily. While these resources included popular portals and search engines they also included their favorite specialty sites. These specialty sites were visited daily by 97% of the respondents; about half of the surveyed users emphasized that they are visiting their favorite sites many times a day, perhaps as frequently as 10/day.

2. TRACKING VISITORS WITH COOKIES

Cookie-tracking is a de-facto standard for unique visitor tracking: when a user visits a site for the first time a new cookie with a unique user ID is generated by the web server and is sent back to the browser for storage. When the same user comes back to the site, the web server retrieves the stored cookie, extracts the unique user ID and thus identifies the user as a returning visitor. On the first glance the cookie tracking mechanism should yield more accurate unique visitor information since in theory the cookies should be far more persistent than the IP addresses. In practice cookies fail to provide the desired degree of reliability because:

1) The same person can access the site from various computers, devices, locations, or browsers thus multiplying cookies;

2) Cookies are deleted by anti-viruses, OS reinstalls, or by user.

To ascertain the accuracy of the cookie tracking method comScore Corporation has recently conducted a study indicating that the cookie-clearing factor alone contributes to 2.5-fold (!) inflation of unique visitor stats [2]. According to comScore 31% of U.S. Internet users cleared their cookies during the month [2]. In our study 43% of users admitted that they clear cookies at least once a week or more often corroborating comScore findings – Figure 1.

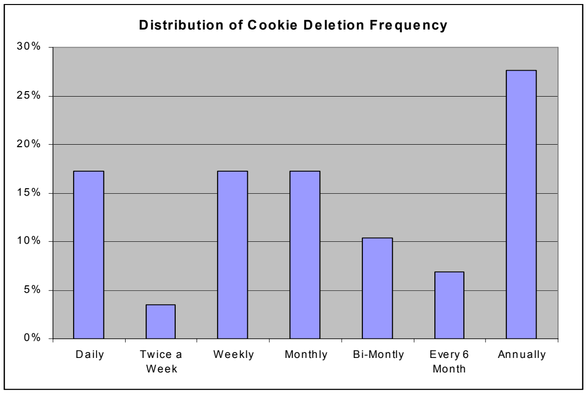

Figure 1. Cookie-clearing frequencies for 100 surveyed users.

3. UNIQUE VISITOR STUDIES

The empirical relationship between unique IPs, cookies and unique visitors was obtained in a series of traffic studies conducted at ultramax-music.com in 2006 and 2008. The results of the two studies were verified against live traffic data and were found to be fully consistent – Figure 2. The 2006 and 2008 studies analyzed traffic obtained from 171 randomly selected registered study participants (age 20-49, median age 30; 90% male; 80% from North America, 15% from Europe). The study period was 28 days and the registered study participants were required to check in daily; their IP address was recorded. During the 2008 study cookie count inflation was also recorded and compared against Google Analytics – Figure 3.

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

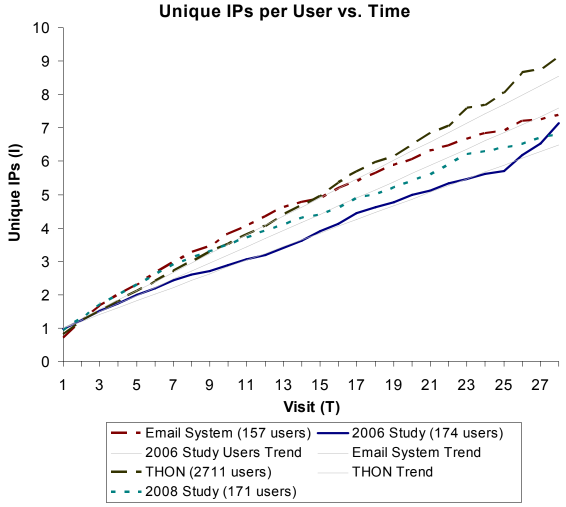

Figure 2. Unique IP inflation with time (2006 & 2008 studies).

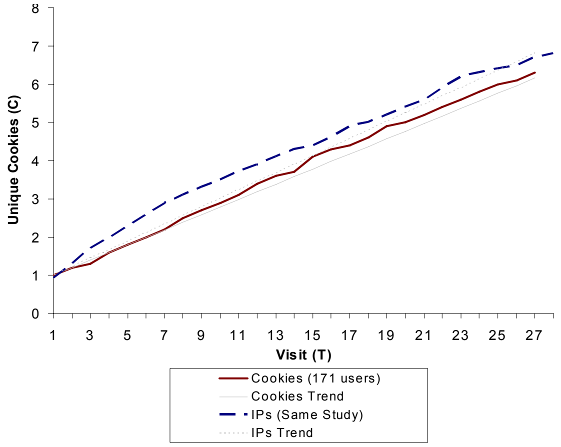

Figure 3. Cookie count inflation with time (2008 study).

The unique IP (as well as the unique cookie) to unique visitor ratio – I_0 – depicted on Figures 2 and 3 can be approximated as:

$$I_0 = 1 + X\,T_N \qquad (1)$$

where X is the average inflation factor. For the relationships on Figure 4 the inflation factor X is 0.28 for THON, 0.24, and 0.22 for the 2006 and 2008 studies. T_N is the time expressed as a count of visits. The count of visits T_N can be linked to time t using the visitation period T: $T_N = t / T - 1$ (2)

The equation (1) allows writing the following empirical formula relating unique visitors U to unique IPs (or cookies) I:

$$I \equiv I_0 U = U(1 + X\,T_N) \qquad (3)$$

where X is the inflation factor (0.2-0.3 from our data) and in general depends on the traffic nature and volume.

During the 2008 study a total of 991 cookies were recorded for 171 study participant thus giving 5.8x unique visitor inflation. Google Analytics reported a similar figure of 949 cookies or the 5.6x unique visitor inflation factor. *In other words Google's 'absolute unique visitors' are not at all unique: the inflation depends on the visitation frequency and grows linearly with time.*

4. ADDITIONAL DATA, LONGER TIME

To validate our results we have obtained traffic data for an Internet email system provided by SureHosting.com. The plot reflecting the ratio of unique IPs to unique visitors is shown on Figure 2, thin dashed and dotted line and on Figure 4, solid line. The plot reveals 35x overestimation resulting 210 visits by 157 users with average inflation factor $X = 0.34$ for the first 30 visits. To verify the observed trend on a larger dataset yet another sample of web logs was obtained, this time for 2,246 users accessing the THON website at Penn State – Figure 2; thick dashed and dotted line; Figure 4, dashed line. This larger dataset spans 50 visits and produces a consistent match with the other datasets for the first 28 visits. The trends observed on Figure 2-4 indicate ~2x unique visitor audience overestimation in a week, ~6-7x overestimation in a month, ~14x overestimation in two months and ~80x overestimation in a year.

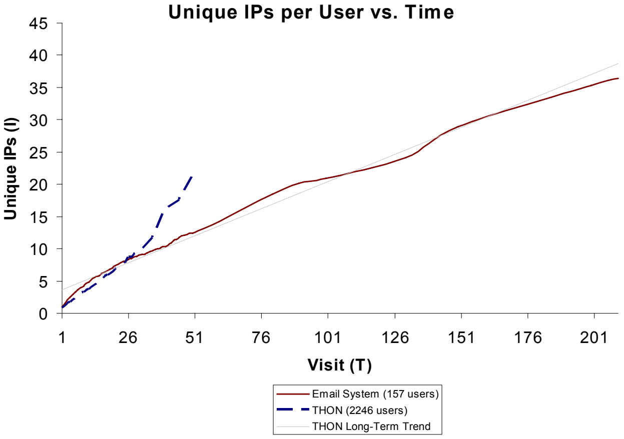

Figure 4. Unique IP count inflation from THON data.

5. CONCLUSION

Unique IP counts as a measure of unique visitors is dubious at best and in the worst case may overestimate the true size of the unique visitor base by an order of magnitude or more depending on the sampling interval size and visitation frequency. Cookies, unfortunately and surprisingly, are almost just as bad in estimating the unique visitors as the unique IP addresses and grossly overestimate the size of the unique visitor base, too. The large noted discrepancy between unique cookies and unique visitors raises doubts in the accuracy of published unique visitor stats used to solicit advertising money. Google Analytics – the industry leader in the user tracking is too fooled by periodic cookie clearing and the multitude of Internet access locations/devices and in our case overestimated the number of unique visitors by a factor of six during the month-long study.

6. REFERENCES

[1] M. Arlitt. Characterizing Web User Sessions. ACM SIGMETRICS Performance Evaluation Review 28 Issue 2 (2000), 50-63.

[2] A. Lipsman. Cookie-Based Counting Overstates Size of Web Site Audiences, comScore, Press Release (2007), http://www.comscore.com/press/release.asp?id=1389