Data Quality in Website Traffic Metrics: A Comparison of 86 Websites Using Two Popular Analytics Services

BERNARD J. JANSEN

Qatar Computing Research Institute, Hamid Bin Khalifa University, Doha, Qatar, jjansen@acm.org

SOON-GYO JUNG

Qatar Computing Research Institute, Hamid Bin Khalifa University, Doha, Qatar, sjung@hbku.edu.qa

JONI SALMINEN

Qatar Computing Research Institute, Hamid Bin Khalifa University, Doha, Qatar, jsalminen@hbku.edu.qa

Estimating website traffic is used in contexts such as business intelligence, competitive analysis, marketing, design, and research. However, the accuracy of popular traffic estimation services is a critical unanswered question. This research compares one year's worth of average monthly analytics metrics data from Google Analytics with those from SimilarWeb for 86 websites of various sizes. The results show statistically significant differences between the two services for total visits, unique visitors, and bounce rates. Using Google Analytics as the baseline, SimilarWeb average values were 19.4% lower for total visits, 38.7% lower for unique visitors, and 25.2% higher for bounce rate. The website rankings between SimilarWeb and Google Analytics for all metrics are significantly correlated, especially for total visits and unique visitors. Finally, the differences between SimilarWeb and Google Analytics measures are systematic, so with Google Analytics metrics from a known site, one can reasonably estimate the Google Analytics metrics for similar sites based on the SimilarWeb values. The implications are that SimilarWeb provides conservative traffic estimates relative to those of Google Analytics and that the tools can be utilized in a complementary fashion in situations where direct site data is not available.

Note: This is a pre-print of a manuscript undergoing peer review. Please cite as: **Jansen, B. J., Jung, S.G., and Salminen, J. (2020). Data Quality in Website Traffic Metrics: A Comparison of 86 Websites Using Two Popular Analytics Services. Tech. Report. http://www.bernardjjansen.com/uploads/2/4/1/8/24188166/traffic_analytics_comparison.pdf**

CCS CONCEPTS • Information systems • World Wide Web • Web applications

Additional Keywords and Phrases: Digital Analytics, Website Traffic, Online Traffic Estimation Services, Analytics Packages

1 INTRODUCTION

Web analytics is the collection, measurement, analysis, and reporting of digital data to enhance insight into the behavior of website visitors [51]. Web analytics is a critical component of business intelligence, competitive analysis, and other domains because countless business decisions are made daily based on website traffic estimates that are obtained from traffic estimation services. Organizations monitor the incoming and outgoing traffic to their sites to identify which pages are popular, determine user interests, and stay abreast of any apparent trends [53]. There are many ways to monitor this traffic, and the gathered data is used for structuring sites, highlighting security problems, indicating bandwidth issues, and assessing organizational key performance indicators (KPIs).

One can group approaches to collecting website analytics data using the focus of data collection efforts, resulting in the emergence of three general methodologies, namely: (a) user-centric, (b) site-centric, and (c) network-centric.

- **User-centric**: Web analytics data is gathered via a panel of users, which is tracked by software installed on users' computers (e.g., a plugin for web browser) [29,36,52,78]. For example, when users install an extension to their

browser, they agree in the license agreement that the data on the websites they visit will be processed and analyzed. The primary advantage here is that the user-centric approach does not rely on cookies or tags; the collected data is focused on actual people. An additional advantage is that one can compare web analytics data across multiple websites. The disadvantage includes recruiting and incentivizing a sufficiently large user panel that is a representative sample of the online population—this is quite challenging, and only a few companies have been able to recruit sizeable user panels (e.g., Alexa). Another disadvantage may be the issue of privacy since many users are not willing to share information on every website that they visit, so users may take efforts to mask their actions.

- **Site-centric**: Web analytics is gathered via software on a specific website (e.g., [3,43,47,54,56,87,120,124]). Most websites use some type of site-centric approach for analytic data gathering, typically employing cookies and/or tagging of pages on the website (e.g., Google Analytics, Adobe Analytics). The primary advantage of this approach lies in the collection of foundational counts (e.g., pages viewed, times accessed, etc.), which is relatively straightforward. Another advantage is that users do not need to install specific software (beyond browsers). However, there are at least three major disadvantages. First, site-centric software is focused on cookies/tags, so these counts may not reflect actual people (i.e., the counts are counts of cookies and tags) or be actual actions of people on the website. Instead, site-centric approaches actually count the number of cookies dropped or tags fired. Second, this approach is susceptible to bots and other forms of analytics inflation tactics. Finally, the site-centric analytics are usually for just one website and only accessible to the owner of that website, making this approach not widely available for business intelligence, marketing, advertising, or other tasks requiring web analytics data from a large number of sites.
- **Network-centric**: Web analytics is gathered via observing and collecting traffic in the network [91,99]. There are various techniques for network-centric web analytics data gathering, with the most common being data purchased or acquired directly from Internet service providers (ISPs). However, there are other data gathering methods, including leveraging search traffic, search engine rankings, paid search, and backlinks [75,113]. The main advantage of the network-centric approach is that one can relatively easily collect analytics on a large number of websites. Also, the setup is relatively easy. Neither users nor websites are required to install any software. The major disadvantage is that there is no information about the on-site actions of the users. A second disadvantage is that major ISPs do not freely share their data, so acquiring this data can be quite expensive. However, other network-centric data can be more reasonably acquired by companies (i.e., SpyFu, SEMRush, Ahrefs), albeit requiring substantial computational, programming, and storage capacity.

Of course, one can use a combination of these methods (e.g., SimilarWeb [103]), but these are three general approaches, with much academic research using one or more of these methods [45,92,101,134]. See Table 1 for a summary of the advantages, disadvantages, and examples of implementations.

Table 1: Comparison of User, Site, and Network-centric Approaches to Web Analytics Data Collection, Showing Advantages, Disadvantages, and Examples of Each Approach at the Time of the Study

Approach	Advantages	Disadvantages	Examples
User-centric	• Focus on people Compare across websites; so can use for business intelligence	• Creating a representative user panel is challenging • User computer software must be installed	• Alexa • ComScore
Site-Centric	• No special user software to install	• Site software must be installed	• Google Analytics • Adobe Analytics

Approach	Advantages	Disadvantages	Examples
	Wide range of analytics for a specific site	• Focus on cookies and tags, not real people • Access limited to website owner; so cannot use for business intelligence among multiple sites	• IBM Analytics
Network-Centric	• Data collection is straightforward • No special software to install for users or sites • Compare across websites; can use for business intelligence	• Data can be challenging to obtain • Limited on-site analytics; generally only between sites data	• Hitwise • SEMRush • SpyFu • AhRefs

While site-centric web analytics tools like Google Analytics can provide analytics results for one's own website, there is often a need to also compare with other websites. Therefore, traffic estimation services, such as SimilarWeb, have become an essential part of web analytics in the business intelligence area [30]. These traffic estimation tools provide web analytics estimate results for one or more websites, a significant feature for many tasks in the competitive intelligence area. These traffic estimation services allow for the benchmarking of web analytics measures and metrics among multiple websites. Traffic estimate tools are essential for a variety of reasons, including competitive analysis, advertising, marketing, domain purchasing, media buying [9,67,74,83,85,96,117], and firm acquisitions [15], along with the use of traffic estimation services in academic research [12]. They are also valuable for accessing the external view of your own website (i.e., what others who do not have access to site-centric analytic data see). These traffic estimation services return a variety of metrics, depending on the platform. However, there are critical questions concerning the accuracy and reliability of these traffic estimation tools that affect billions of dollars in online advertising, firm acquisition, and research. As such, there is a critical need for an assessment of these tools.

In this research, we compare web analytics values from Google Analytics (the industry-standard website analytics platform at the time of the study) and SimilarWeb (the industry-standard traffic estimation platform at the time of the study) using three core web analytics metrics (total visits, unique visitors, and bounce rate) averaged monthly over 12 months for 86 websites. We conduct statistical analysis along several fronts reporting both exploratory and statistical results. We then tease apart the nuanced differences in discussing findings and present both the theoretical and the practical implications of this research.

2 REVIEW OF LITERATURE

Web traffic services have been employed in research and used by researchers for an array of inquiries and topics. These areas include online gaming [132] – which was then correlated with academic performance, social media and multi-channel online marketing [4,79], online community shopping [68], online purchase predictions [60], online research methods [62], social science issues [16], and user-generated content on social media [6,18–20]. These services have also been used in research concerning online interests in specific topics [71,77,133] – including home birth, online branding in social media [66,136], and mobile application usage [61]. They have also been used in studies about website trust and privacy [14,42,108,109,125] – including in the health domain, website design [13,28,65,122], and website popularity and ranking [4,5,17,27,32,68,73,110,111,135] – for a variety of areas, including news and tourism. As shown by this lengthy list of citations, SimilarWeb, and other traffic estimation tools are widely used in peer-reviewed academic research and relied on

for ranking or report metrics from the tools. However, none of these research articles examined the accuracy of these traffic estimation services.

Academic research on this area of traffic estimation accuracy evaluation is limited. Lo and Sedhain [70] evaluate six websites lists, including the ranked list from Alexa (the only one that is still active, as of the date of this study). The researchers examined only the top 100 websites and only compared the rankings among the lists. They concluded that the ranking among the lists differed. This is not surprising given that the methodologies used to create the lists varied in terms of website traffic, number of backlinks, and opinions of human judges. Vaughan and Yang [119] used organizations from the U.S. and China and collected web traffic data for these sites from Alexa Internet, Google Trends for Websites, and Compete (Alexa is the only service still active from this research, as of the date of this study). The researchers report significant correlations between web traffic data and organizational performance measures of academic quality for universities and financial variables for businesses. Napoli, Lavrakas, and Callegaro [81] present some of the challenges and issues with the user-centric analytics approach, namely that the results often do not align with site-centric measures based on results from a ComScore study. The researchers attribute the discrepancies to the sampling of the user panels. Scheitle and fellow researchers [99] examine the rankings of several websites lists, including Alexa but not including SimilarWeb, investigating similarity, stability, representativeness, responsiveness, and benignness in the cybersecurity domain, but not actual web traffic. The researchers report that the ranked lists are not stable and open to manipulation. Pochat and colleagues [91] extend this research by introducing a list that is less susceptible to rank manipulation.

While few academic studies have examined traffic services, even fewer have evaluated the actual traffic numbers, instead focusing on the easily accessible (and usually free) ranked lists. Studies are even rarer still on the performance of SimilarWeb, despite its standing and reputation in the industry. Scheitle and colleagues [99] attribute this absence to SimilarWeb charging for its service, implying that researchers are, perhaps, cheap. Regardless of the reason, the only academic study that we are aware of that explicitly examines traffic numbers, including SimilarWeb, is Prantl and Prantl [92]. This study compares rankings among Alexa, SimilarWeb, and NetMonitor [82] for a set of websites in the Czech Republic, using NetMonitor as the baseline. The research only reports the traffic comparison between SimilarWeb and NetMonitor. The researchers, unfortunately, provide neither detailed exploratory analysis nor statistical analysis of the analytic comparison. Also, NetMonitor uses a combination of site and user-centric measures, so it is unclear how the traffic metrics are calculated. However, the researchers [92] report that SimilarWeb overreports traffic compared to NetMonitor. They also note that SimilarWeb provides traffic estimates with +/- 30% compared to NetMonitor traffic measurements for 49% of the 487 websites.

There are a number of academic studies focusing nearly exclusively on the ranked list. In addition, there have also been several studies by practitioners examining actual analytics from the traffic estimation services, with reported results varying, unfortunately. Some of these studies show traffic estimation services, notably SimilarWeb, reportedly underestimating traffic [37,80,84], as much as 30% to 50%. Other studies show traffic estimation services, notably SimilarWeb, reportedly overestimating traffic [26,46,50,89,93], from 11% for large websites to nearly 90% for small websites [50]. SimilarWeb itself states that its reported values will vary +/- 20% when compared with other services. However, a trend is that SimilarWeb [88,112] consistently ranks as the best or one of the best traffic estimation tools in the industry [46,49,72], as noted by industry practitioners [67,83,85,94,98,115,129]. SimilarWeb consistently outperforms other services [26], and SimilarWeb traffic and other measurements are accurate for many sites in these studies, with reports being as high as 25% [93]. Even when the reported traffic numbers are off, the SimilarWeb results nearly always correlate with the baseline traffic trends, usually compared with Google Analytics data. The correlation is also positive relative to overall accuracy among sites [93].

Although these studies provide insights into the area, there are potential issues with regard to relying on these industry studies, including possible questions on data appropriateness, lack of explicitly defined methods of analysis, and conflicts of interest (as some of these studies are performed by potential competitors of SimilarWeb). Also, many of these studies employed a small number of data points [37,80,89], making statistical analysis challenging. Other studies had a short temporal span [72,84,93], as there can be significant traffic fluctuations for sites depending on the time of year, or mainly high-traffic websites [84], which are easier to estimate. Finally, some studies have imprecise metric reporting [26,50,58,84], raising doubt on the results, or a limited non-impactful set of metrics [37,58,84], which are not central to analytics insights.

Because of these potential issues, there is a critical need for a rigorous academic analysis of traffic estimation services to supplement these industry studies. As some readers may not be aware and as a reminder to others, academic research typically requires detailed descriptions of data collection and justification of methods – often making the raw data available. Also, authors are routinely required to declare any potential conflicts of interest upon submitting a research article. Finally, most research articles undergo a rigorous peer review, where anonymous experts in the area attempt to 'punch holes' in the research data, methods, or findings. These requirements are typically not present in industry studies and highlight the need for academic research in the area.

Given the use of traffic estimation tools in academic research and their widespread use in the practitioner communities, there is a notable lack of research examining the accuracy of these tools. Determining their accuracy is of critical importance, given the widespread reliance in many domains of research and implementation. However, due to the absence of studies in the area, several unanswered questions remain, including: *How accurate are these analytics services? How do they compare with other analytics methods? Are these analytics tools better (or worse) at measuring certain analytics metrics than other methods?* These are some of the motivations for our research. These are essential questions that need addressing for critical evaluation of research findings and business decisions that rely on these services to be made. They are conceptual straightforward; however, they are surprisingly difficult to evaluate in practice.

3 RESEARCH QUESTIONS

Our research objective is to *compare and contrast the reported analytics measures between SimilarWeb and Google Analytics*. To investigate this research objective, we focus on three core web analytics metrics – total visits, unique visitors, and bounce rate – which we define in the methodology section. Although there is a nearly endless list of possible metrics for investigation, these three metrics are central to addressing analytics measurements of frequency, reach, and duration, respectively. As such, they are central to the web analytics analysis of any single website or set of websites. So, in the interest of space and impact of findings, we focus on these three metrics, leaving other metrics for future research.

Given that Google Analytics uses site-centric direct access to website data and SimilarWeb employs a triangulation of datasets and techniques, we would reasonably expect measurements would differ between Google Analytics and SimilarWeb. Furthermore, because Google Analytics is the *de facto* industry standard, we use Google Analytics measures as the baseline for this research. Therefore, our hypotheses (H) are:

- **H01**: SimilarWeb measures of total visits to websites differ from those reported by Google Analytics.
- **H02**: SimilarWeb measures of unique visitors to websites differ from those reported by Google Analytics.
- **H03**: SimilarWeb measures of bounce rates for websites differ from those reported by Google Analytics.

We investigate these hypotheses using the following methodology.

4 METHODOLOGY

Our data collection platforms are Google Analytics and SimilarWeb.

4.1 Google Analytics

Google Analytics is a site-centric web analytics platform and is the most popular site analytics tools in use [123] – that is, the market leader. Google Analytics tracks and reports website traffic for a specific website. It is the accepted baseline for web analytics website data. This tracking by Google Analytics is accomplished via cookies and tags [38]; a tag is a snippet of JavaScript code added to the individual pages. The tags are executed in the JavaScript-enabled browsers of the website visitors. Once executed, the tag sends the visit data to a data server and sets a first-party cookie on cookie-enabled browsers on visitors' computers. The tag must be on a page on the site for Google Analytics to track the web analytics data for that page. Figure 1 shows a Google Analytics dashboard at the time of the study.

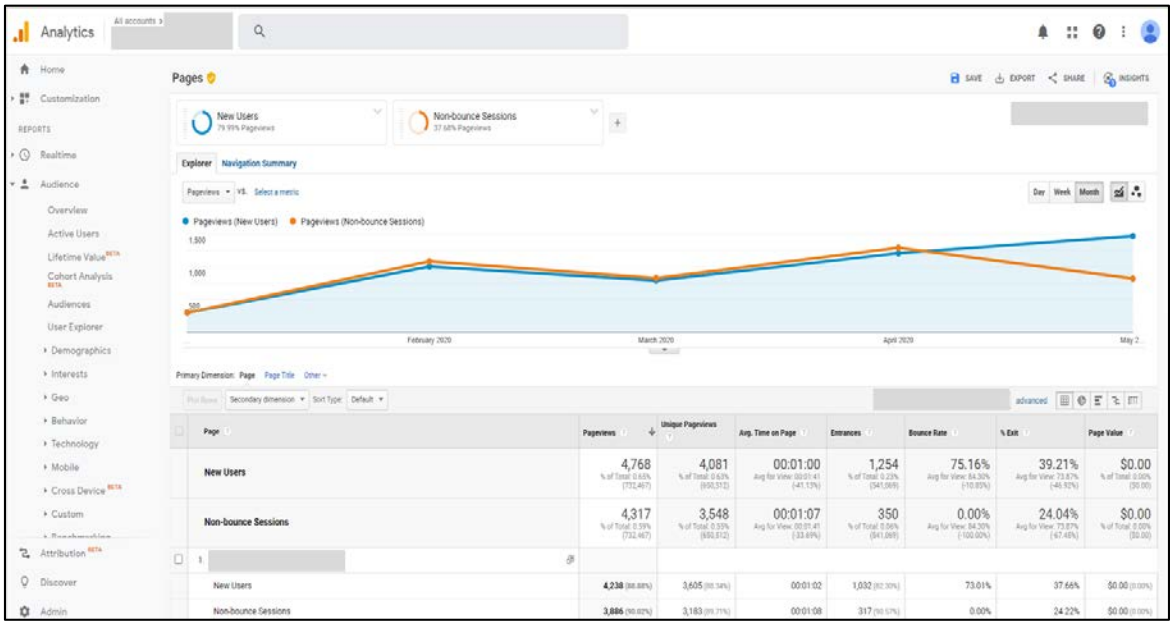


Figure 1: Example of a Google Analytics Dashboard as of the Date of This Study. (note: website-identifying details masked.)

Concerning the data collection, analysis, and reporting algorithms of Google Analytics, they are proprietary. However, enough is known to validate their employment as being the industry standard. The techniques of cookies and the general process of tagging are well-known, although there may be some small nuances in implementation. Google Analytics does employ statistical data sampling techniques, so the reported values are not all direct counts. However, the general overview of the data sampling approach is presented in reasonable detail [29], and the described subsampling is an industry standard methodology [59].

4.2 SimilarWeb

SimilarWeb [88,112] is a traffic estimation service providing web analytics data for one or multiple websites. SimilarWeb uses a mix of user, site, and network-centric data collection approaches for triangulation of data [97,104],

reportedly collecting and analyzing billions of data points per day [103]. SimilarWeb's philosophical approach is that each method has strengths and weaknesses, and the best practice is triangulating multiple algorithms and data sources [104]. Figure 2 shows a SimilarWeb web page dashboard at the time of the study.

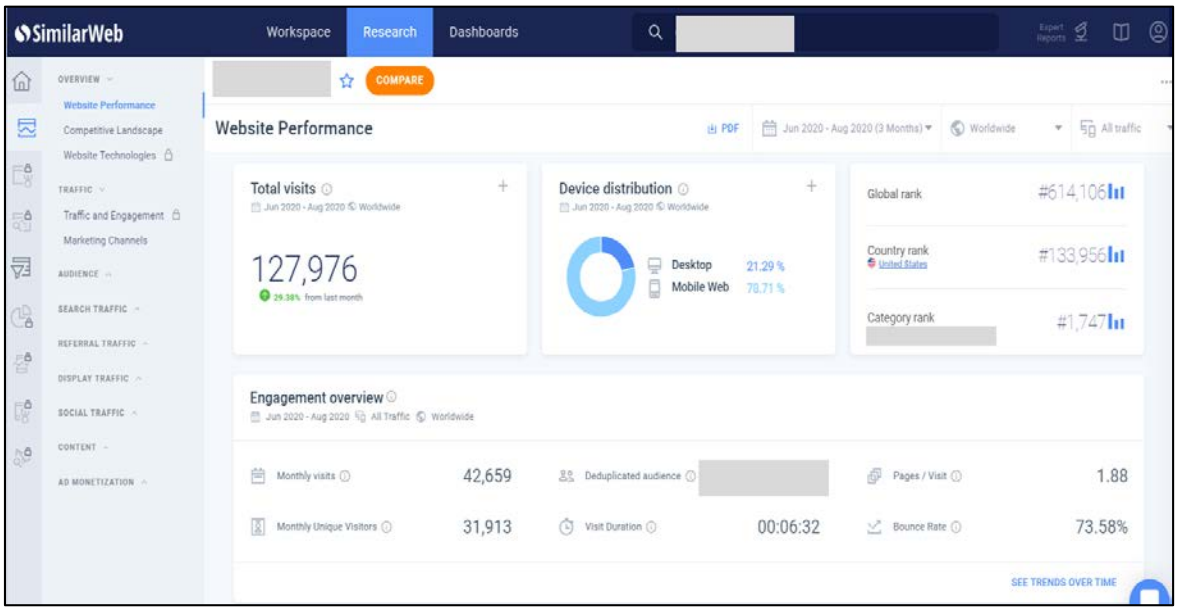


Figure 2: Example of a SimilarWeb Webpage Dashboard as of the Date of This Study (note – website-identifying details masked.)

Regarding the data collection, analysis, and reporting algorithms of SimilarWeb, they are proprietary, but again, enough is known to validate the general implementation. The SimilarWeb foundational principle of triangulating user, site, and network-centric data collection data [97,104] is academically sound, with triangulating data and methods widely used in and advocated by academia [29,57]. SimilarWeb data collection, analysis, and reporting methodology are outlined in fair detail [103], although, like Google Analytics, the proprietary specifics are not provided. However, from the documentation that is available [102–105,112,130], the general approach is to collect data from three primary sources, which are: (a) a reportedly 400 million worldwide user panel [102], (b) specific website analytics [104], and (c) ISP and other traffic data [104]. These are supplemented with publicly available data sources (e.g., population figures). Each of these datasets will have some overlap (i.e., the web analytics data from one collection method will also appear in one or both of the other collection methods). With the collected data, augmented with publicly available data [104], SimilarWeb uses statistical techniques and ensemble machine learning approaches to generate web analytics estimates. These estimates can then be compared to the overlapped data to make algorithmic adjustments to the predictions. This is a more complex approach relative to Google Analytics; however, SimilarWeb's task is much more complicated. In sum, the general techniques employed by SimilarWeb are standard methodologies [25,69,97,131], academically sound, and industry standard.

4.3 Data Collection Procedure

For our analysis, we identify a set of websites with traffic estimation by SimilarWeb and having Google Analytics linked to SimilarWeb [105], thereby making their Google Analytics values available. If a website has a Google Analytics link, SimilarWeb, using the paid version, offers the option of reporting either the SimilarWeb or the Google Analytics numbers

for these websites. This premium feature allows us to compare the SimilarWeb estimates and the Google Analytics numbers for our identified web analytics metrics of total visits, unique visitors, and bounce rates.

To identify our pool of possible websites, we employ the Majestic Million [3]. The Majestic Million list of websites, first released in October 2012, is creative commons licensed and derives from Majestic's web crawler. The Majestic Million list ranks sites by the number of /24 IPv4-subnets linking to that site [9], which is used as a proxy for website popularity.

Using this large, open-licensed, and readily available list as the seed listing, we started at the top, submitted the link to the SimilarWeb application program interface (API), and checked whether SimilarWeb provided traffic estimates or if the website linked its Google Analytics to the SimilarWeb service. If the website had both SimilarWeb estimates and Google Analytics metrics, we included it as a candidate website for our research. If not, the website was excluded. We then proceeded to the next website on the list and repeated the submission and verification process. Our goal was to identify at least 30 websites, which is the 'rule of thumb' needed for statistical analysis means and standard deviations.

However, with considerable effort, we continued these steps until we identified 91 websites. There were five websites where the values from Google Analytics and SimilarWeb values differed by orders of magnitude. As there seemed to be no discernible patterns among these five websites upon our examination, we excluded them as outliers and reserved them as candidates for future study. This left us with 86 websites for analysis. We concluded that this was more than adequate for our research, as it was well more than what was required for statistical analysis [21].

For the privacy of the companies' websites and given that these web analytics comparisons are a paid business product of SimilarWeb, we have determined not to make the specific links publicly available. However, we outline our methodology in detail so that those interested can recreate our research.

4.4 Data Analysis

We employ the paired t-test for our analysis. The paired t-test compares two means that are from the same units. The purpose of the test is to determine whether there is statistical evidence that the mean difference between paired observations on a particular outcome is significantly different from zero. We transformed our data to parametric via the Box-Cox transformation [24] by using the log-transformation function: log(variable + 0.5). The data is successfully normalized, though a bit of skewness exists, as the data is weighted toward lower traffic numbers using the log transformation. Despite the existing skewness, previous work shows that a method such as the paired t-test is robust in these cases [23,48]. The use of transformation ensured that our statistical approach is valid for the attributes of the data set. We then executed the paired t-test among three groups to test the differences between the means of total visits, unique visitors, and bounce rates. We conducted the statistical testing on the transformed values; however, for clarity, we report the non-transformed values for means, standard deviation, maximum, minimum, and median.

Using the SimilarWeb API, we collect the reported values for total visits, unique visitors, and bounce rate for each month over 12 months (September 1, 2019, through August 31, 2020, inclusive) for each of the 86 websites on our list. We then average the monthly values for each metric for each platform to get the values that we use in our analysis. We use the monthly average to mitigate any specific monthly fluctuation. For example, some websites have seasonal fluctuations in traffic. Some websites may experience outages during specific months or denial of service attacks. Using the monthly average over a 12-month period helps mitigate the possible short-term variations.

Our three measures, total visits, unique visitors, and bounce rate, are considered core metrics in the domain of web analytics [11,51,63]. A metric is typically a number, such as a count or a percentage. However, the method for how each of these metrics is specifically counted or calculated may vary by platform or service; therefore, it is crucial to understand

these differences. Plus, the conceptual understanding may differ from the specific ability of a method for tracking in implementation. Table 2 presents an overview of these metrics.

Table 2 Comparison of Definitions of Total Visits, Unique Visitors, and Bounce Rate Conceptually and for the Two Traffic Platforms: Google Analytics and SimilarWeb

Definition of:	Total Visits	Unique Visitors	Bounced Rate
Conceptually	Sum of times that all people go to a website during a measurement period. A measure of frequency.	Sum of actual people who have visited a website at least once during a measurement period. A measure of reach.	Ratio of bounced visits divided by all visits to a website during a given period. A bounced visit is the act of a person immediately leaving a website before any interaction can reasonably occur A measure of duration.
Practically	Sum of times at least one page of a website has been loaded into a browser during a measurement period.	Sum of distinct tracking measures requesting pages from a website during a given period determined by a method such as cookie, tag, or plugin.	Ratio of single-page visits divided by all visits to a website during a given period (i.e., single page visits divided by all visits)
Google Analytics	Sum of single visits to a website consisting of one or more pageviews during a measurement period. The default visit timeout is 30 minutes, meaning that if there is not activity for this visit on the website for more than 30 minutes, then a new visit will be reported if another interaction occurs.	Sum of unique Google Analytics tracking code and browser cookies that visit a website at least once during a measurement period.	Ratio of single-page visits divided by all visits to a website during a measurement period (i.e., the percentage of all visits on a website in which a single page is viewed and triggered only a single request to the Analytics server). Single-page sessions have a session duration of 0 seconds since there are no subsequent server hits after the first one that would let Analytics calculate the length of the session.
SimilarWeb	Sum of times at least one page of a website has been loaded into a browser during a measurement period Subsequent page views are included in the same visit until the user is inactive for more than 30 minutes. If a user becomes active again after 30 minutes, that counts as a new visit. A new session will also start at midnight.	Sum of computing devices visiting a website within a geographical area and during a measurement period.	Ratio of single page visits by all visits for a website within a geographical area and during a measurement period.

5 RESULTS

5.1 Exploratory Results

Our 86 websites represent companies based in 26 countries, as shown in Table 3. We used the country classifications provided by SimilarWeb, and we verified the classifications based on our own assessment of the websites and links.

Table 3 Host Country of Organization for 86 Websites in Study

Country	No.	%
United States	43	50.0%
India	6	7.0%
Russian Federation	6	7.0%
Japan	4	4.7%
United Kingdom	4	4.7%
France	3	3.5%
Israel	3	3.5%
Spain	2	2.3%
One each (Belarus, Belgium, Canada, Chile, China, Cuba, Ecuador, Germany, Madagascar, Malaysia, Nigeria, Taiwan, Turkey, Ukraine, United Arab Emirates)	15	17.4%
	86	100.0%

The 86 organizational websites are from the following 19 industry verticals, as shown in Table 4. We used the industry classifications provided by SimilarWeb [106,107], and we verified the classifications based on our own assessment of the websites and company background material provided.

Table 4 Industry Vertical of Organization for 86 Websites in Study

Website Category	No.	%
News and Media	36	41.9%
Computers Electronics and Technology	10	11.6%
Arts and Entertainment	9	10.5%
Science and Education	5	5.8%
Community and Society	4	4.7%
Finance	4	4.7%
Business and Consumer Services	2	2.3%
E-commerce and Shopping	2	2.3%
Gambling	2	2.3%
Travel and Tourism/	2	2.3%
Vehicles	2	2.3%
Health	1	1.2%
Hobbies and Leisure	1	1.2%
Home and Garden	1	1.2%
Jobs and Career	1	1.2%
Law and Government	1	1.2%
Lifestyle/Beauty and Cosmetics	1	1.2%
Lifestyle/Fashion and Apparel	1	1.2%
Sports	1	1.2%
	86	100.0%

The 86 organizational websites' types are shown in Table 5. We used the site type classifications provided by SimilarWeb, and we verified the classification based on our own assessment of the website content and features. Content sites are websites that, well, provide content as their primary function. Transactional websites are those sites that are primarily selling a product. 'Other' refers to those websites that do not neatly fit into the other two categories.

Table 5: Website Type for the 86 Websites in Study

Site Type	No.	%
Content	50	58.1%
Other	34	39.5%

Site Type	No.	%
Transactional	2	2.3%
	86	100.0%

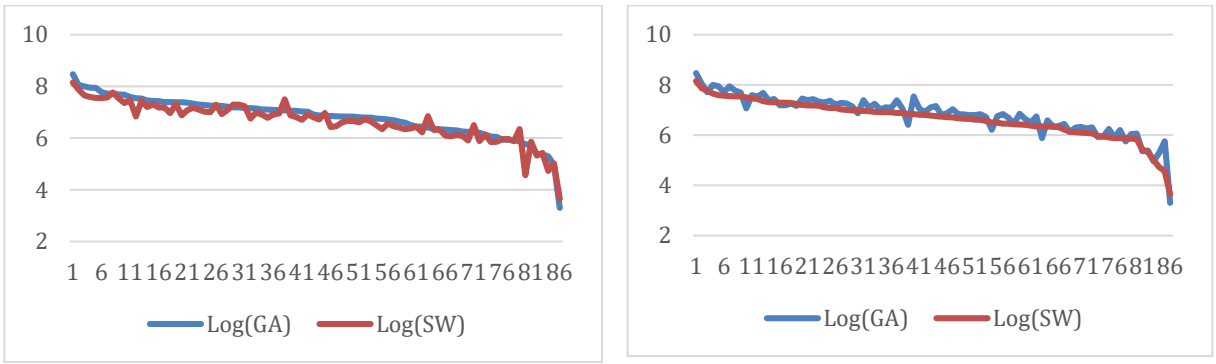
5.2 H01: SimilarWeb's measurements of total visits to websites differ from those reported by Google Analytics.

A paired t-test was conducted to compare the number of total visits reported by Google Analytics and SimilarWeb. There was a significant difference in the reported number of total visits for Google Analytics (M = 19.8 million, SD = 37 million) and SimilarWeb (M = 12.3 million, SD = 19.6 million); $t(85) = 6.43$, $p < 0.01$. These results indicate that there is a difference in the number of total visits between the two approaches. Specifically, our results show that SimilarWeb's reported number of total visits is statistically lower than the values reported by Google Analytics. **Therefore, H01 is fully supported: SimilarWeb's measurements of total visits to websites differ from those reported by Google Analytics.**

The number of total visits for all 86 websites was 1,703.5 million (max = 292.5 million; min = 1,998, med = 7.8 million), as reported by Google Analytics, and 1,060.1 million (max = 140.8 million; min = 4,443; med = 5.9 million), as reported by SimilarWeb. Using the total aggregate visits for all 86 websites by using Google Analytics as the baseline, SimilarWeb underestimated by 7.5 million (19.4%) total visits. Using Google Analytics numbers as the baseline for total visits, SimilarWeb overestimated 15 (17.4%) sites and underestimated 66 (76.7%) sites. The two platforms were nearly similar (+/- 5%) for 5 (5.8%) sites.

Ranking the websites by total visits based on Google Analytics and SimilarWeb, we then conducted a Spearman's rank correlation. Results indicated that there is a significant, strong positive association between the rankings of Google Analytics and SimilarWeb, ($r_s(85) = 0.954$, $p < 0.001$).

Graphically, we compare the reported total visits between Google Analytics and SimilarWeb in Figure 6, with Figure 6a ranked by Google Analytics and Figure 6b ranked by SimilarWeb. As shown in Figure 6, generally, the differences in the number of total visits between Google Analytics and SimilarWeb are relatively stable except for the low traffic websites.



(a) (b)

Figure 6: Ranked Listing of Total Visits Reported by Google Analytics (a) and SimilarWeb (b). Differences Remain Fairly Stable Except for Low Traffic Websites (i.e., the differences increase for low traffic websites)

This finding implies that, although the reported total visits values differ between the two platforms, the trend of the values for the set of websites is generally consistent. So, if one is interested in a ranking (e.g., "Where does website X rank within this set of websites based on total visits?"), then SimilarWeb values will generally align with those of Google Analytics for those websites. However, if one is specifically interested in numbers (e.g., "What is the number of total visits

to each of N websites?), then the SimilarWeb total visit numbers will be ~20% below those reported by Google Analytics, on average.

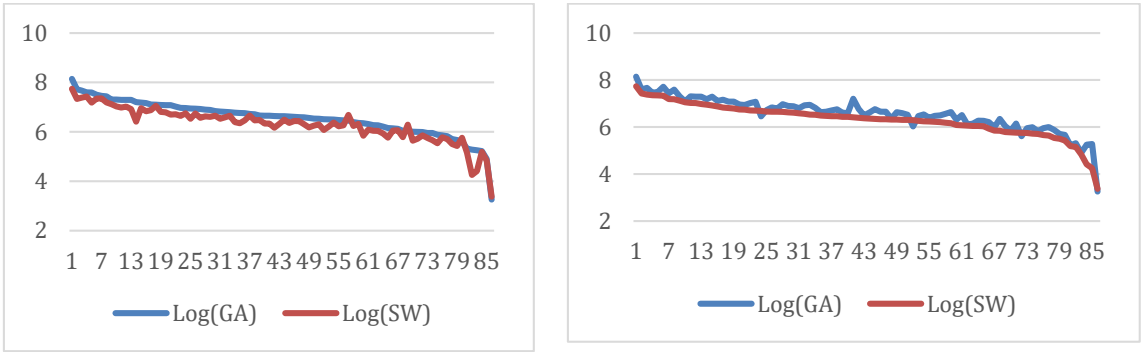
5.3 H02: SimilarWeb measures of unique visitors to websites differ from those reported by Google Analytics.

A paired t-test was conducted to compare the number of unique visitors reported by Google Analytics and SimilarWeb. There was a significant difference in unique visitors for the Google Analytics (M= 9.7 million, SD= 17.6 million) and the SimilarWeb (M= 5.1 million, SD= 7.9 million) conditions; $t(85)= 12.60$, $p < 0.01$. These results indicate that there is a difference in the number of unique visitors between the two approaches. Specifically, our results show that the reported number of unique visitors by SimilarWeb is statistically lower than the values reported by Google Analytics. **Therefore, H02 is fully supported: SimilarWeb measures of unique visitors to websites differ from those reported by Google Analytics.**

The total number of unique visitors for all 86 websites was 834.7 million (max= 138.1 million; min= 1,799; med= 4.3 million) reported by Google Analytics and 439.0 million (max= 54.6 million; min= 2,361; med= 2.3 million) reported by SimilarWeb. Using the aggregate unique visitors for all 86 websites, using Google Analytics as the baseline, SimilarWeb underestimated by 395.6 million (38.7%) unique visitors. Using Google Analytics numbers as the baseline, SimilarWeb overestimated four (4.7%) sites and underestimated 82 (95.3%) sites.

Ranking the websites by unique visitors based on Google Analytics and SimilarWeb, we then conducted a Spearman's rank correlation. Results of the Spearman correlation indicated that there is a significant strong positive association between the ranking of Google Analytics and SimilarWeb: ($r_s(85) = 0.967$, $p < .001$).

Graphically, we compared the reported unique visitors between Google Analytics and SimilarWeb in Figure 7, with Figure 7a ranked by Google Analytics and Figure 7b ranked by SimilarWeb. As shown in Figure 7, generally, the differences in the number of unique visitors between Google Analytics and SimilarWeb are stable except for the low-traffic websites.



(a) (b)

Figure 7: Ranked Listing of Unique Visitors Reported by Google Analytics (a) and SimilarWeb (b). Differences Remain Fairly Stable Except for Low Traffic Websites (i.e., the differences increase for low-traffic websites)

This finding indicates that, while the reported values for unique visitors differ between the two platforms, the trend of the values for the set of websites is mostly consistent. So, if one is interested in a ranking (e.g., "Where does website X rank within this set of websites based on unique visitors?"), then SimilarWeb values will generally align with those of Google Analytics for those websites. However, if one is specifically interested in numbers (e.g., "What is the number of

unique visitors to each of N websites?), then the SimilarWeb unique visitors numbers will be ~40% below those reported by Google Analytics, on average.

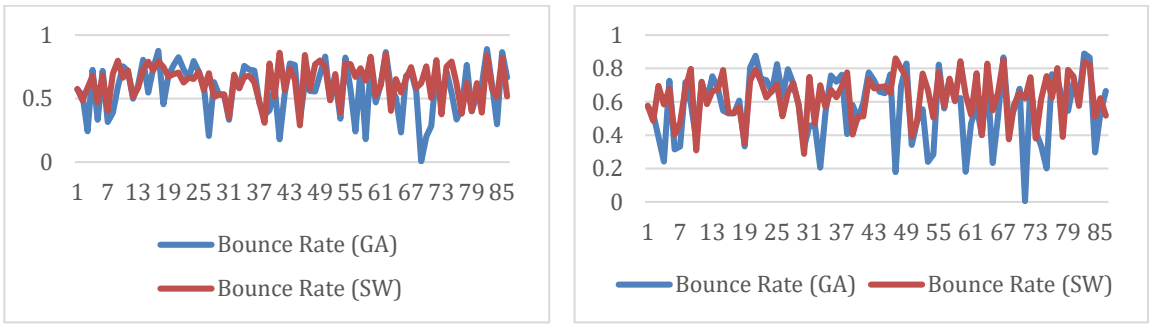
5.4 H03: SimilarWeb measures of bounce rates for websites differ from those reported by Google Analytics.

A paired t-test was conducted to compare bounce rates reported by Google Analytics and SimilarWeb. There was a significant difference in the bounce rates between the Google Analytics (M= 56.2%, SD= 20.4%) and the SimilarWeb (M= 63.0%, SD= 13.8 %) conditions; $t(85)=-2,96$, $p < 0.01$. Specifically, our results showed that the reported bounce rates by SimilarWeb were significantly higher than those reported by Google Analytics, fully **supporting H03: SimilarWeb measures of bounce rates for websites differ from those reported by Google Analytics.**

The average of bounce rates for all 86 websites was 56.2% (SS= 20.4%, max= 88.9%; min= 20.4%; med= 59.2%) reported by Google Analytics and 63.0% (SS= 13.8%, max= 86.0%; min= 28.8%; med= 65.3%) as reported SimilarWeb. Using Google Analytics as the baseline, SimilarWeb estimated 6.8% more than the average bounce rate. Additionally, SimilarWeb overestimated 35 (40.7%) sites and underestimated 31 (36.0%) sites. The two platforms were nearly similar (+/- 5) for 20 (23.3%) sites.

Ranking websites by bounce rate based on Google Analytics and SimilarWeb, we then conducted a Spearman's rank correlation. Results of the Spearman correlation indicated a somewhat strong significant positive association between the ranking of Google Analytics and SimilarWeb, ($rs(85) = 0.734$, $p < .001$).

Graphically, this is illustrated in Figure 8, where we compare bounce rates between Google Analytics and SimilarWeb with Figure 8a ranked by Google Analytics and Figure 8b ranked by SimilarWeb. Figure 8 shows some correlation as well as substantial variations site to site.



(a) (b)

Figure 8: Ranked Listing of Bounce Rates Reported by Google Analytics (a) and SimilarWeb (b). Differences Are Correlated but with Substantial Variations Among Websites

This finding indicates that, although SimilarWeb and Google Analytics reported similar bounce rates for more than 20% of the sites, the difference between the values for the other 80% for the two platforms was quite high. We address the possible reasons for this high discrepancy later.

6 DISCUSSION

Table 6 presents a summary of our findings for the 86 websites using average monthly total visits, unique visitors, and bounce rates during a 12-month period.

Table 6 Summary of Results Comparing Google Analytics and SimilarWeb for Total Visits, Unique Visitors, and Bounce Rate. Difference uses Google Analytics as the Baseline. Results based on Paired t-Test for Hypotheses Supported or Not Supported

Metric / Service	Google Analytics	SimilarWeb	Difference	Hypotheses
Total Visits	1,703,584,207	1,060,137,189	(643,447,018) (19.4%)	Fully Supported – The reported values differ
Unique Visitors	834,656,530	439,016,436	(395,640,094) (38.7%)	Fully Supported – The reported values differ
Bounce Rate	56.2%	63.0%	6.8%	Fully Supported – The reported values differ
Metric / Service	Google Analytics	SimilarWeb	Difference	Hypotheses

	Number of Sites (Relative to Google Analytics Values) Where SimilarWeb Numbers Were:			
	Higher	Lower	Similar (+/- 5%)	
Total Visits	15 (17.4%)	66 (76.7%)	5 (5.8%)	SimilarWeb values will generally be lower than Google Analytic
Unique Visitors	4 (4.7%)	82 (95.3%)	0 (0.0%)	SimilarWeb values will generally always be lower than Google Analytic
Bounce Rate	35 (40.7%)	31 (36.0%)	20 (23.3%)	SimilarWeb values will generally vary compared to Google Analytic

As shown above, statistical testing of all three hypotheses is statistically significant, so **all three hypotheses are supported**. The reported values for the metrics of total visits, unique visitors, and bounce rates for Google Analytics and SimilarWeb differ significantly. The website rankings by each service are significantly correlated, so it seems that these ranked lists can be used for research and other purposes, with the caveat highlighted in [91,99]. These analyses compare the two services' precision (i.e., how close measured values are to each other).

However, the underlying question motivating our research remains this: *How accurate are the reported metrics from traffic estimation services (i.e.,* how close are the measure values to the "true" values)*?* Regardless of the statistical testing results, this motivational question is more challenging to address. In reality, there is one "true" number of visits, visitors, and bounces. However, is it realistic to expect any web analytics service ever to match perfectly with reality? Moreover, what is the reality in terms of web analytics? It is a misconception to view web analytics data collecting as "counting." In most cases, web analytics is not counting; instead, it is "measuring." It is well known that for any measure, there will be an error rate (+/- n%) [22]. No measure or measurement tool is perfect, and web data can be particularly messy [114].

Although one might lean toward considering metrics reported by Google Analytics as the "gold standard" for website analytics (and justifiably so in many cases), it is also well known within the industry that Google Analytics has tracking issues, and many Google Analytics accounts are incorrectly set up [7,8,64,118]. There are also cases where other analytics methods might be more appropriate. Google Analytics relies on one data collection approach: basically a cookie and tagging technique. There are certainly cases (e.g., cookies, incognito browsing, etc.) when this method is not accurate (e.g., unique visitors). Furthermore, Google Analytics might not be installed correctly or the same on all websites. Therefore, these issues result in problems with Google Analytics being seen as the "gold standard."

To investigate our motivation research question concerning the accuracy of SimilarWeb as a traffic estimation service, we conduct a deductive analysis using a likelihood of error [126]. For each of our metrics, we analyze what makes theoretical sense for which web analytics approach would result in the most accurate measurement, Google Analytics or SimilarWeb? We discuss our analysis of each metric below.

Bounce rate: A high bounce rate is undesirable. The bounce rate means that someone comes to a site and immediately leaves without taking any relevant action. For this metric, **both** Google Analytics [40] and SimilarWeb are conceptually incorrect due to the practical issues of measuring a bounce visit [31]. For a meaningful visit measurement, there has to be an entry point (where the person came to the site) and an exit point (where the person left the site). If there is no endpoint, both Google Analytics and SimilarWeb count, it as a single page visit and, therefore, a bounce because there is no exit.

Obviously, there are many situations where there is a relevant action taken on a site, but there is no exit point [34]. For example, there can be an e-commerce site where a potential consumer arrives on a product page, reads the content, and takes no other action at that time. Another case is a newspaper site where an audience member comes to the site, scans the headlines, reads the article snippets, but takes no other action, such as clicking [55]. In each of these cases, the visit could last several minutes or longer. However, since there is no exit page (i.e., no second page), both Google Analytics and SimilarWeb would count these example visits as bounces.

So, we can reasonably assume both Google Analytics and SimilarWeb are overcounting bounces. This may be why the values vary substantially between the two services. However, since bounce is a site-centric specific measure, we would expect Google Analytics to be at least more precise (if not more accurate) then SimilarWeb when measuring bounce rate on a single given site. However, SimilarWeb's panel data may help correct this somewhat for a set of sites, which Google Analytics does not measure. So, if one needs to examine the bounce rate of several websites, Google Analytics cannot be used since website owners usually do not make their web analytics data available to the public.

In the end, conceptually, both Google Analytics and SimilarWeb are surely overcounting the number of bounced sessions, thereby increasing the bounce rate. In terms of mechanical metrics, one would expect Google Analytics to be better for an individual site. SimilarWeb might be expected to give a reasonable bounce rate estimate for some sites due to their user-centric panel data, and bounce rates are generally high, especially for sites that are more highly trafficked. This reasonableness in results from both Google Analytics and SimilarWeb is borne out in our statistical analysis above, where the two services were more in agreement for the larger traffic sites for bounce rates.

Total Visits: This seems like a straightforward site-centric metric for which Google Analytics should excel. Although there is the room for some noise in the visits, such as housekeeping visits (i.e., visits from internal company personnel for site maintenance), bot-generated visits [121], purchased traffic, or hacking attacks that might not conceptually meet the definition of a visit, it is difficult to imagine how a traffic estimation service could be better than a site-centric service in this regard. The use of the site and network-centric data collection data employed by traffic estimation services like SimilarWeb would not mitigate the noise mentioned above; however, the user-centric panel data might compensate for some of the noise issues for at least a high traffic website. However, in general, one would expect Google Analytics to be more accurate in measuring visits than SimilarWeb. However, Google Analytics data is generally not available for multiple websites, so relying on Google Analytics for multiple sites is not a practical option. For these cases, one would have to employ a traffic estimation service, such as SimilarWeb. Based on our analysis above, values for total visits would be under Google Analytic measurements by ~20%.

Unique Visitors: Finally, we consider unique visitors. In this case, perhaps surprisingly, one would expect the greater likelihood of error to be with the site-centric measurements, resulting in SimilarWeb measures being more accurate.

Site centric services, such as Google Analytics, typically rely on a combination of cookies and tags to measure unique visitors. This would generally result in an overcount of unique visitors by the services. For example, the expected life cycle of a computer is three to five years [2,33], meaning a person changing computers would be registered as a new visitor. The market share of browsers changes considerably over the years [10,100], meaning when someone has changed browsers, he/she would be registered as a new visitor. Studies also show that 40% of Internet users clear cookies on a daily, weekly,

or monthly basis [1,127], and about 3.7% of users disable all cookies [86,128]. (i.e., each of these actions would trigger a unique visitor count when visiting a website). However, some studies point to a much higher rate, with more than 30% of users deleting cookies in a given month [1]. Many people also use the incognito mode on their browsers [41,44], which triggers a new visitor count in Google Analytics [35,95]. Also, many people have multiple devices (i.e., personal computer, work computer, smartphone, tablet, etc.), with about 50% of Americans using four Internet-enabled devices [90,116], so each device would be counted as a unique visitor even if the same person is using multiple devices.

For these reasons, the unique visitor number measured using the cookie approach would surely lead to an overcount using site-centric metrics. *How much of an overcount?* Based on the issues just outlined, it seems that, for Google Analytics, a 20% overestimate in monthly unique visitors to 30% overestimate for more extended periods seems conservative and reasonable. However, more precise measures would require an in-depth study and is a task for future research.

For unique visitors, it seems that panel data, such as those that similar web and other network-centric services use, might be more accurate. However, this might only hold for larger websites. For lower-traffic websites, it is not clear that panel data would be accurate as there is not enough traffic to these sites to generate reasonable statistical analysis. Generally, for unique visitors, it seems that Google Analytics would most likely overestimate the number of unique visitors to the website. SimilarWeb might be more accurate, due to its user panel data approach, for the larger traffic websites but have questionable accuracy (either over- or underestimating) for the smaller traffic websites. Again, this conclusion is borne out by our analysis above, where the difference between Google Analytics and SimilarWeb increased for the smaller websites (see Figure 7).

6.1 Theoretical Implications

We highlight three theoretical implications of this research, which are:

- **Triangulation of Data, Methods, and Services**: There seems, at present, to be no single data collection approach (user, site, or network-centric) or web analytics service (including Google Analytics and SimilarWeb) that would be effective for all metrics, context, or business needs. Therefore, a triangulation of services, depending on the data, method of analysis, or need, seems to be the most appropriate approach.
- **Discrepancies with Implementation**: Regarding precision, we have established a difference between the two services, and we know the general methodologies and metrics calculations. However, the nuances of implementation (for both services) have, as of the date of this study, not been independently audited. So, we cannot say, in practice, which is the best implementation for a given metrics. Again, this points to the need for a triangulation of services.
- **Discrepancies with Reality**: Precision does not mean accuracy for either Google Analytics or SimilarWeb. We have already outlined potential issues with all three of the metrics examined (total visits, unique visitors, bounce rates), where the mechanics of application are not aligned with conceptual definitions of what these metrics are supposedly measuring. This situation calls for both continued research into improved measures and a realization that the reported values (from both Google Analytics and SimilarWeb) should not be viewed necessarily as "truth." Rather, the values are reported measures with some error rates (+/-).

6.2 Practical Implications

We highlight three practical implications of the findings, which are:

- **Use of Google Analytics and SimilarWeb**: Findings of our research show that, in general, SimilarWeb estimates for total visits and number of unique visitors will generally be lower than those reported by Google Analytics, but the correlation between the two platforms is high for these two metrics. So, if one is solely interested in a ranking of a set

of websites for which one does not have the Google Analytics data, the SimilarWeb metrics are a workable proxy. If one is interested in the actual Google Analytics traffic for a set of websites, one can use the SimilarWeb estimates and increase by about 20% for total visits and about 40% for unique visitors. As a caveat, the Google Analytics unique visitor's numbers are probably an overcount, so the SimilarWeb values may be more in line with reality. As an easier 'rule of thumb', we suggest using a 20% adjustment (i.e., increase SimilarWeb estimates) for both metrics based on the analysis findings above.

- **Verification of Analytics for a Single Website**: In general, Google Analytics is a site-centric web analytics platform, so it would be a reasonable service to use for a single website that one owns and has access to. However, comparing traffic values from Google Analytics to those of SimilarWeb (or other traffic estimation services) may be worthwhile. Given the reported number of sites with Google Analytics improperly setup [7,8], large discrepancies between values from Google Analytics and SimilarWeb might indicate a problem in the website's analytics setup.
- **Estimating Google Analytics Traffic for Multiple Websites**: As shown above, the differences between Google Analytics and SimilarWeb metrics for total visits and unique visitors are systematic (i.e., the differences stay relatively constant). This means, if you have Google Analytics values for one site, you adjust and use a similar difference for the other websites to get reasonable traffic numbers to those from Google Analytics. This technique is valuable in competitive analysis situations where you are comparing multiple sites against a known website and want the Google Analytics values for all sites.

6.3 Limitations, Future Research, and Strengths

Limitations and Future Research: The first limitation concerns data quality. In the absence of ground truth, we primarily measure precision and not the accuracy of the two web analytics services. As noted, there are inconsistencies between the two platforms. So, the analytics data that decision-makers may perceive as accurate, objective, and correct may not have these qualities due to the several potential sources for errors outlined above. Future research should be undertaken by the major web analytics services to provide metric values with confidence intervals to depict them as ranges rather than exact values. Another limitation is that the source codes and specific implementation for these platforms are not available, so the specifics and nuances of the implementation cannot be verified. Future research could focus on using open-source analytics platforms, such as Matomo [76], to tease apart some of these metric implementations. Other future research involves replication studies with different sets of websites, other traffic estimation services, other metrics, and analysis of specific website segments based on country, type, size, or industry vertical.

Strengths: There are several strengths of this research. First, we use two popular web analytics services. Second, we employ many websites, 86, with various attributes, ensuring that the sample size was robust. Third, we collect data over an extended period of 12 months to account for short periods of fluctuation with the website traffic measures. Fourth, we report and statistically evaluate three essential and core web analytic metrics – total visits, unique visitors, and bounce rates. Fifth, we discuss and offer both theoretical and practical implications of our research. To our knowledge, this is one of the first and one of the most extensive academic examinations and analyses of these popular web analytics services.

6.4 Conclusion

For this research, we compared three monthly analytics metrics from Google Analytics to those from SimilarWeb based on 12 months of data for 86 worldwide websites. Findings show statistically significant differences between the two services for total visits, unique visitors, and bounce rates. Compared to Google Analytics, SimilarWeb values were ~20% lower for total visits, ~40% lower for unique visitors, and ~25% higher for average bounce rate. The rankings of all three

metrics are significantly correlated between SimilarWeb and Google Analytics. The measurement differences are systematic between the two analytics services. The implications are that SimilarWeb provides conservative traffic estimates relative to Google Analytics. These web analytics tools can be complementarily utilized in various contexts, especially when needing analytics data for multiple websites.

REFERENCES

[1] Dr Magid Abraham, Cameron Meierhoefer, and Andrew Lipsman. 2007. The Impact of Cookie Deletion on the Accuracy of Site-Server and Ad-Server Metrics: An Empirical Comscore Study. (2007), 19.

[2] Jason Acimovic, Francisco Erize, Kejia Hu, Douglas J. Thomas, and Jan A. Van Mieghem. 2018. Product Life Cycle Data Set: Raw and Cleaned Data of Weekly Orders for Personal Computers. *M&SOM* 21, 1 (May 2018), 171–176. DOI:https://doi.org/10.1287/msom.2017.0692

[3] Haris Ahmed, Dr Tahseen, Waleej Haider, Muhammad Asad, Shardha Nand, and Saher Kamran. 2017. Establishing Standard Rules for Choosing Best KPIs for an E-Commerce Business based on Google Analytics and Machine Learning Technique. *International Journal of Advanced Computer Science and Applications* 8, (January 2017). DOI:https://doi.org/10.14569/IJACSA.2017.080570

[4] Kusum L. Ailawadi and Paul W. Farris. 2017. Managing Multi- and Omni-Channel Distribution: Metrics and Research Directions. *Journal of Retailing* 93, 1 (March 2017), 120–135. DOI:https://doi.org/10.1016/j.jretai.2016.12.003

[5] Hüseyin Akcan, Torsten Suel, and Hervé Brönnimann. 2008. Geographic web usage estimation by monitoring DNS caches. In *Proceedings of the first international workshop on Location and the web* (LOCWEB '08), Association for Computing Machinery, New York, NY, USA, 85–92. DOI:https://doi.org/10.1145/1367798.1367813

[6] Alan B. Albarran. 2013. *The Social Media Industries*. Routledge.

[7] Alex Ramadan. 2019. Common Google Analytics Setup Errors and Omissions. *UpBuild*. Retrieved October 9, 2020 from https://www.upbuild.io/blog/common-google-analytics-setup-errors/

[8] Amanda Gant. 2020. Inaccurate Google Analytics - Why Google Analytics is Wrong and How to Fix It. *Orbit Media Studios*. Retrieved October 5, 2020 from https://www.orbitmedia.com/blog/inaccurate-google-analytics-traffic-sources/

[9] Amin Shawki. 2013. 6 Free Analytics Tools to Help You Understand Your Competitor's Web Traffic. *InfoTrust*. Retrieved October 2, 2020 from https://infotrust.com/articles/6-free-analytics-tools-to-help-you-understand-your-competitor-s-web-traffic/

[10] Jari Arkko. 2020. The influence of internet architecture on centralised versus distributed internet services. *Journal of Cyber Policy* 5, 1 (January 2020), 30–45. DOI:https://doi.org/10.1080/23738871.2020.1740753

[11] Avinash Kaushik. 2007. Web Analytics Standards: 26 New Metrics Definitions. *Occam's Razor by Avinash Kaushik*. Retrieved October 6, 2020 from https://www.kaushik.net/avinash/web-analytics-standards-26-new-metrics-definitions/

[12] M. Bakaev, V. Khvorostov, S. Heil, and M. Gaedke. 2017. Web Intelligence Linked Open Data for Website Design Reuse. In *ICWE*. DOI:https://doi.org/10.1007/978-3-319-60131-1_22

[13] Maxim Bakaev, Vladimir Khvorostov, Sebastian Heil, and Martin Gaedke. 2017. Web Intelligence Linked Open Data for Website Design Reuse. In *Web Engineering* (Lecture Notes in Computer Science), Springer International Publishing, Cham, 370–377. DOI:https://doi.org/10.1007/978-3-319-60131-1_22

[14] Himani Bansal and Shruti Kohli. 2019. Trust evaluation of websites: a comprehensive study. *International Journal of Advanced Intelligence Paradigms* 13, 1–2 (January 2019), 101–112. DOI:https://doi.org/10.1504/IJAIP.2019.099946

[15] Omri Barzilay. 2017. How SimilarWeb Helps Investors Make Decisions About Their Portfolio. *Forbes*. Retrieved October 2, 2020 from https://www.forbes.com/sites/omribarzilay/2017/11/09/meet-similarweb-one-of-wall-streets-secret-weapons/

[16] Ivana Batarelo Kokić and Sani Kunac. 2019. Media Coverage of School Behaviour Issues: A Content Analysis of Digital Media Messages. (2019). Retrieved October 4, 2020 from https://www.bib.irb.hr/1007563

[17] Samantha Bates, John Bowers, Shane Greenstein, Jordi Weinstock, Yunhan Xu, and Jonathan Zittrain. 2018. *Evidence of Decreasing Internet Entropy: The Lack of Redundancy in DNS Resolution by Major Websites and Services*. National Bureau of Economic Research. DOI:https://doi.org/10.3386/w24317

[18] Bruce G. Vanden Bergh, Mira Lee, Elizabeth T. Quilliam, and Thomas Hove. 2011. The multidimensional nature and brand impact of user-generated ad parodies in social media. *International Journal of Advertising* 30, 1 (January 2011), 103–131. DOI:https://doi.org/10.2501/IJA-30-1-103-131

[19] Andreas Blombach, Natalie Dykes, Stefan Evert, Philipp Heinrich, Besim Kabashi, and Thomas Proisl. 2019. A New German Reddit Corpus. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, 278–279.

[20] Andreas Blombach, Natalie Dykes, Philipp Heinrich, Besim Kabashi, and Thomas Proisl. 2020. A Corpus of German Reddit Exchanges (GeRedE). In *Proceedings of the 12th Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, 6310–6316. Retrieved October 4, 2020 from https://www.aclweb.org/anthology/2020.lrec-1.774

[21] Dennis D. Boos and Jacqueline M. Hughes-Oliver. 2000. How Large Does n Have to be for Z and t Intervals? *The American Statistician* 54, 2 (2000), 121–128. DOI:https://doi.org/10.2307/2686030

[22] Marit L. Bovbjerg. 2020. Random Error. In *Foundations of Epidemiology*. Oregon State University. Retrieved October 9, 2020 from https://open.oregonstate.education/epidemiology/chapter/random-error/

[23] G. E. P. Box and S. L. Andersen. 1955. Permutation Theory in the Derivation of Robust Criteria and the Study of Departures from Assumption. *Journal of the Royal Statistical Society. Series B (Methodological)* 17, 1 (1955), 1–34.

[24] G. E. P. Box and D. R. Cox. 1964. An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)* 26, 2 (1964), 211–252.

[25] Jason Brownlee. 2019. A Tour of Machine Learning Algorithms. *Machine Learning Mastery*. Retrieved October 6, 2020 from https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

[26] Bruce Hogan. 2020. How to Analyze Competitor Website Traffic. *SoftwarePundit*. Retrieved October 2, 2020 from https://www.softwarepundit.com/seo/competitor-taffic-analysis

[27] Pablo de Carlos, Noelia Araújo, and José Antonio Fraiz. 2016. The new intermediaries of tourist distribution: Analysis of online accommodation booking sites. *The International Journal of Management Science and Information Technology (IJMSIT)* 19 (2016), 39–58.

[28] Kalpita Chakrabortty and Ebbin Jose. 2018. Relationship Analysis between Website Traffic, Domain Age and Google Indexed Pages of E-commerce Websites. *IIM Kozhikode Society & Management Review* 7, 2 (July 2018), 171–177. DOI:https://doi.org/10.1177/2277975218770028

[29] Chun Choo, Brian Detlor, and Don Turnbull. 2000. Information Seeking on the Web: An Integrated Model of Browsing and Searching. *First Monday* 5, (March 2000). DOI:https://doi.org/10.5210/fm.v5i2.729

[30] Alistair Croll and Sean Power. 2009. *Complete Web Monitoring: Watching your visitors, performance, communities, and competitors*. O'Reilly Media, Inc.

[31] Daniel Schneider and Ruth M. Trucks. 2020. Bounce Rate: What you need to know and how to improve. *SimilarWeb*. Retrieved October 4, 2020 from https://www.similarweb.com/corp/blog/bounce-rate/

[32] Dr Biswajit Das and Jyoti Shankar Sahoo. Social Networking Sites – A Critical Analysis of Its Impact on Personal and Social Life. *International Journal of Business and Social Science* 2, 14 , 222–228.

[33] David Sarokin. 2020. What Is the Life Span of the Average PC? | Small Business - Chron.com. Retrieved October 6, 2020 from https://smallbusiness.chron.com/life-span-average-pc-69823.html

[34] Sukru Eraslan, Yeliz Yesilada, and Simon Harper. 2020. "The Best of Both Worlds!";: Integration of Web Page and Eye Tracking Data Driven Approaches for Automatic AOI Detection. *ACM Trans. Web* 14, 1 (January 2020), 1:1–1:31. DOI:https://doi.org/10.1145/3372497

[35] Eric Fettman. 2015. A Sweet Treat, But Users Delete: Cookies and Cookie Deletion in Google Analytics. *Cardinal Path*. Retrieved October 6, 2020 from https://www.cardinalpath.com/blog/cookies-and-cookie-deletion-in-google-analytics

[36] Vagner Figueredo de Santana and Felipe Eduardo Ferreira Silva. 2019. User Test Logger: An Open Source Browser Plugin for Logging and Reporting Local User Studies. In *Universal Access in Human-Computer Interaction. Theory, Methods and Tools* (Lecture Notes in Computer Science), Springer International Publishing, Cham, 229–243. DOI:https://doi.org/10.1007/978-3-030-23560-4_17

[37] Frank Olivo. Is SEMRush Accurate? A Comparison with My Site's Analytics. *Sagapixel*. Retrieved October 2, 2020 from https://sagapixel.com/seo/is-semrush-accurate/

[38] Google. 2020. Google Analytics Set up the Analytics global site tag - Analytics Help. Retrieved October 5, 2020 from https://support.google.com/analytics/answer/1008080?hl=en

[39] Google. 2020. Google Analytics About data sampling - Analytics Help. Retrieved October 6, 2020 from https://support.google.com/analytics/answer/2637192?hl=en
[40] Google. 2020. Google Analytics Bounce rate - Analytics Help. Retrieved October 4, 2020 from https://support.google.com/analytics/answer/1009409?hl=en
[41] Google. 2020. Google Browse in private - Computer - Google Chrome Help. Retrieved October 6, 2020 from https://support.google.com/chrome/answer/95464?co=GENIE.Platform%3DDesktop&hl=en
[42] Aloysius Bernanda Gunawan. 2020. Socialization of Terms of Use and Privacy Policy on Indonesian e-commerce Websites. *Journal of Sosial Science* 1, 3 (July 2020), 41–45. DOI:https://doi.org/10.46799/jsss.v1i3.35
[43] Ulrich Gunter and Irem Önder. 2016. Forecasting city arrivals with Google Analytics. *Annals of Tourism Research* 61, (November 2016), 199–212. DOI:https://doi.org/10.1016/j.annals.2016.10.007
[44] Hana Habib, Jessica Colnago, Vidya Gopalakrishnan, Sarah Pearman, Jeremy Thomas, Alessandro Acquisti, Nicolas Christin, and Lorrie Faith Cranor. 2018. Away From Prying Eyes: Analyzing Usage and Understanding of Private Browsing. In *Fourteenth Symposium on Usable Privacy and Security.*, Baltimore, MD, USA, 18.
[45] Huy Hang, Adnan Bashir, Michalis Faloutsos, Christos Faloutsos, and Tudor Dumitras. 2016. “Infect-me-not”: A user-centric and site-centric study of web-based malware. In *2016 IFIP Networking Conference (IFIP Networking) and Workshops*, 234–242. DOI:https://doi.org/10.1109/IFIPNetworking.2016.7497222
[46] Joshua Hardwick. 2018. Find Out How Much Traffic a Website Gets: 3 Ways Compared. *SEO Blog by Ahrefs*. Retrieved October 4, 2020 from https://ahrefs.com/blog/website-traffic/
[47] Daqing He, Ayşe Göker, and David J Harper. 2002. Combining evidence for automatic Web session identification. *Information Processing & Management* 38, 5 (September 2002), 727–742. DOI:https://doi.org/10.1016/S0306-4573(01)00060-7
[48] David A. Hull. 1993. Using Statistical Testing in the Evaluation of Retrieval Experiments. Retrieved October 7, 2020 from https://openreview.net/forum?id=Hy4WPU-dWS
[49] Inside Market Report. 2020. Competitor Analysis Tools Market: Size, Production, Prospects, Consumption, Cost Structure, Competitive Landscape 2020-2025 – The Daily Chronicle. Retrieved October 2, 2020 from https://thedailychronicle.in/news/1892984/competitor-analysis-tools-market-size-production-prospects-consumption-cost-structure-competitive-landscape-2020-2025/
[50] Ioana Pupec. 2017. We analyzed 1787 eCommerce websites with SimilarWeb and Google Analytics and that’s what we learned - Omniconvert Blog. *ECOMMERCE GROWTH Blog*. Retrieved October 2, 2020 from https://www.omniconvert.com/blog/we-analyzed-1787-ecommerce-websites-similarweb-google-analytics-thats-we-learned.html
[51] Bernard J. (Jim) Jansen. 2009. Understanding User-Web Interactions via Web Analytics. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 1, 1 (January 2009), 1–102. DOI:https://doi.org/10.2200/S00191ED1V01Y200904ICR006
[52] Bernard J. Jansen and Michael D. McNeese. 2005. Evaluating the effectiveness of and patterns of interactions with automated searching assistance. *Journal of the American Society for Information Science and Technology* 56, 14 (2005), 1480–1503. DOI:https://doi.org/10.1002/asi.20242
[53] Jiepu Jiang, Daqing He, and James Allan. 2014. Searching, browsing, and clicking in a search session: changes in user behavior by task and over time. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval* (SIGIR ’14), Association for Computing Machinery, New York, NY, USA, 607–616. DOI:https://doi.org/10.1145/2600428.2609633
[54] Tingting Jiang, Yu Chi, and Huiqin Gao. 2017. A clickstream data analysis of Chinese academic library OPAC users’ information behavior. *Library & Information Science Research* 39, 3 (July 2017), 213–223. DOI:https://doi.org/10.1016/j.lisr.2017.07.004
[55] Tingting Jiang, Qian Guo, Shunchang Chen, and Jiaqi Yang. 2019. What prompts users to click on news headlines? Evidence from unobtrusive data analysis. *Aslib Journal of Information Management* 72, 1 (January 2019), 49–66. DOI:https://doi.org/10.1108/AJIM-04-2019-0097
[56] Tingting Jiang, Jiaqi Yang, Cong Yu, and Yunxin Sang. 2018. A Clickstream Data Analysis of the Differences between Visiting Behaviors of Desktop and Mobile Users. *Data and Information Management* 2, 3 (December 2018), 130–140. DOI:https://doi.org/10.2478/dim-2018-0012
[57] Todd D. Jick. 1979. Mixing Qualitative and Quantitative Methods: Triangulation in Action. *Administrative Science Quarterly* 24, 4 (1979), 602–611. DOI:https://doi.org/10.2307/2392366

[58] João Aguiar. 2020. SimilarWeb vs SEMrush - Comparing Website Traffic (2020 Case Study). *Mobidea Academy*. Retrieved October 2, 2020 from https://www.mobidea.com/academy/similarweb-vs-semrush-website-traffic/

[59] Jonathan Weber. 2016. How Accurate Is Sampling In Google Analytics? | Bounteous. Retrieved October 6, 2020 from https://www.bounteous.com/insights/2016/03/03/how-accurate-sampling-google-analytics/

[60] Saar Kagan and Ron Bekkerman. 2018. Predicting Purchase Behavior of Website Audiences. *International Journal of Electronic Commerce* 22, 4 (October 2018), 510–539. DOI:https://doi.org/10.1080/10864415.2018.1485084

[61] Seok Kang. 2014. Factors influencing intention of mobile application use. *International Journal of Mobile Communications* 12, 4 (January 2014), 360–379. DOI:https://doi.org/10.1504/IJMC.2014.063653

[62] David Karpf. 2012. Social Science Research Methods in Internet Time. *Information, Communication & Society* 15, 5 (June 2012), 639–661. DOI:https://doi.org/10.1080/1369118X.2012.665468

[63] Avinash Kaushik. 2007. *Web Analytics: An Hour a Day* (1st Edition ed.). Sybex, Indianapolis, Ind.

[64] Kevin Bloom. 2020. Why Your Google Analytics Data is Wrong and How To Fix It. *Hinge Marketing*. Retrieved October 9, 2020 from https://hingemarketing.com/blog/story/why-your-google-analytics-data-is-wrong-and-how-to-fix-it

[65] Karol Król and Jozef Halva. 2017. Measuring Efficiency of Websites of Agrotouristic Farms from Poland and Slovakia. *Economic and Regional Studies / Studia Ekonomiczne i Regionalne* 10, 2 (June 2017), 50–59. DOI:https://doi.org/10.2478/ers-2017-0015

[66] Charlie Albert Lasuin, Azizah Omar, and T. Ramayah. 2015. Social media and brand engagement in the age of the customer. Kuching, Sarawak, MALAYSIA, 138–144. Retrieved October 4, 2020 from http://icoec.my/index.php/proceedings/5-icoec-2015-proceedings/103-social-media-and-brand-engagement-in-the-age-of-the-customer

[67] Laura Joint. 2016. Metrics we measure • PR Resolution — by CoverageBook. *PR Resolution — by CoverageBook*. Retrieved October 2, 2020 from https://resolution.coveragebook.com/metrics-we-measure/

[68] P. Leitner and T. Grechenig. 2009. Scalable Social Software Services: Towards a Shopping Community Model Based on Analyses of Established Web Service Components and Functions. In *2009 42nd Hawaii International Conference on System Sciences*, 1–10. DOI:https://doi.org/10.1109/HICSS.2009.377

[69] Jimmy Lin and Alek Kolcz. 2012. Large-scale machine learning at twitter. In *Proceedings of the 2012 international conference on Management of Data - SIGMOD '12*, ACM Press, Scottsdale, Arizona, USA, 793. DOI:https://doi.org/10.1145/2213836.2213958

[70] Bruce W N Lo and Rosy Sharma Sedhain. 2006. How Reliable Are Website Rankings? Implications for E-Business Advertising and Internet Search. 2 (2006), 233–238.

[71] Arwid Lund and Mariano Zukerfeld. 2020. Profiting from Open Audiovisual Content. In *Corporate Capitalism's Use of Openness: Profit for Free?*, Arwid Lund and Mariano Zukerfeld (eds.). Springer International Publishing, Cham, 199–239. DOI:https://doi.org/10.1007/978-3-030-28219-6_5

[72] Margaret Kashuba. 2020. SimilarWeb vs. SEMrush: Which Offers More Accurate Data? | CustomerThink. Retrieved October 2, 2020 from https://customerthink.com/similarweb-vs-semrush-which-offers-more-accurate-data/

[73] Estela Marine-Roig. 2014. A Webometric Analysis of Travel Blogs and Review Hosting: The Case of Catalonia. *Journal of Travel & Tourism Marketing* 31, 3 (April 2014), 381–396. DOI:https://doi.org/10.1080/10548408.2013.877413

[74] Mark Macanas. SimilarWeb vs Google Analytics Traffic Data Mismatch, Explained. *TechPinas : Philippines' Technology News, Tips and Reviews Blog*. Retrieved October 2, 2020 from https://www.techpinas.com/2018/06/SimilarWeb-vs-Google-Analytics.html

[75] Alberto Martín-Martín, Enrique Orduna-Malea, Mike Thelwall, and Emilio Delgado López-Cózar. 2018. Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics* 12, 4 (November 2018), 1160–1177. DOI:https://doi.org/10.1016/j.joi.2018.09.002

[76] Matomo. 2020. Matomo - The Google Analytics alternative that protects your data. *Analytics Platform - Matomo*. Retrieved October 9, 2020 from https://matomo.org/

[77] Christopher C. Miller and Karin A. Fox. 2017. 834: Americans view widely varied blog advice about home birth. *American Journal of Obstetrics and Gynecology* 216, 1 (January 2017), S478–S479. DOI:https://doi.org/10.1016/j.ajog.2016.11.743

[78] Ben Miroglio, David Zeber, Jofish Kaye, and Rebecca Weiss. 2018. The Effect of Ad Blocking on User Engagement with the Web. In *Proceedings of the 2018 World Wide Web Conference* (WWW '18), International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 813–821. DOI:https://doi.org/10.1145/3178876.3186162

[79] S. A. Movsisyan. 2016. Social media marketing strategy of Yerevan brandy company. *Annals of Agrarian Science* 14, 3 (September 2016), 243–248. DOI:https://doi.org/10.1016/j.aasci.2016.08.010

[80] Steve Mulder. 2014. How to Compare Your Site Metrics to Your Local Competitors. Retrieved October 2, 2020 from http://digitalservices.npr.org/post/how-compare-your-site-metrics-your-local-competitors

[81] Philip M. Napoli, Paul J. Lavrakas, and Mario Callegaro. 2014. Internet and mobile ratings panels. In *Online Panel Research: A Data Quality Perspective*. 387–407. Retrieved October 4, 2020 from http://www.wiley.com/WileyCDA/WileyTitle/productCd-1119941776.html

[82] Nielsen. 2015. NetMonitor. *Nielsen Admosphere*. Retrieved October 3, 2020 from https://www.nielsen-admosphere.eu/products-and-services/internet-measurement/netmonitor/

[83] Kenny Novak. 2019. SimilarWeb vs Alexa: Which Traffic Estimator is More Precise? *Growtraffic Blog*. Retrieved October 2, 2020 from https://growtraffic.com/blog/2019/03/similarweb-alexa-which-precise

[84] Olha Diachuk, Pavel Loba, and Olga Mirgorodskaya. 2020. Comparing accuracy: SEMrush vs SimilarWeb. *owox*. Retrieved October 2, 2020 from https://www.owox.com/blog/articles/semrush-vs-similarweb/

[85] Olly Finley. 2020. SEMrush vs SimilarWeb: What is the Best Tool for Media Buyers? *Blog lemonads*. Retrieved October 2, 2020 from https://www.lemonads.com/blog/semrush-vs-similarweb-what-is-the-best-tool-for-media-buyers/

[86] Opentracker. Third-Party Cookies vs First-Party Cookies. *Opentracker*. Retrieved October 6, 2020 from https://www.opentracker.net/article/third-party-cookies-vs-first-party-cookies-2/

[87] A. Ortiz-Cordova and B. J. Jansen. 2012. Classifying Web Search Queries in Order to Identify High Revenue Generating Customers. *Journal of the American Society for Information Sciences and Technology* 63, 7 (2012), 1426–1441.

[88] Osman Husain. 2015. Forget the garage – this multi-million dollar company started in a jewelry store. *Tech in Asia*. Retrieved October 2, 2020 from https://www.techinasia.com/the-story-of-similarweb

[89] Patrick Langridge. 2016. How Accurate Are Website Traffic Estimators? Retrieved October 2, 2020 from https://www.screamingfrog.co.uk/how-accurate-are-website-traffic-estimators/

[90] Pew Research Center. 2019. Demographics of Mobile Device Ownership and Adoption in the United States. *Pew Research Center: Internet, Science & Tech*. Retrieved October 6, 2020 from https://www.pewresearch.org/internet/fact-sheet/mobile/

[91] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. 2019. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. *Proceedings 2019 Network and Distributed System Security Symposium* (2019). DOI:https://doi.org/10.14722/ndss.2019.23386

[92] David Prantl and Martin Prantl. 2018. Website traffic measurement and rankings: competitive intelligence tools examination. *International Journal of Web Information Systems* 14, 4 (January 2018), 423–437. DOI:https://doi.org/10.1108/IJWIS-01-2018-0001

[93] Rand Fishkin. 2015. The Traffic Prediction Accuracy of 12 Metrics from Compete, Alexa, SimilarWeb, & More. *SparkToro*. Retrieved October 2, 2020 from https://sparktoro.com/blog/traffic-prediction-accuracy-12-metrics-compete-alexa-similarweb/

[94] Richard D. Pace. 2013. SimilarWeb: Fuzzier and Warmer than Alexa - Everything PR. *Everything PR News*. Retrieved October 2, 2020 from https://everything-pr.com/similarweb-alexa/

[95] Silver Ringvee. 2019. How to Detect and Track Incognito Users with Google Analytics and Google Tag Manager. *Reflective Data*. Retrieved October 6, 2020 from https://reflectivedata.com/how-to-detect-track-incognito-users-with-google-analytics-and-google-tag-manager/

[96] Roy Hinkis. Traffic and Engagement Metrics and Their Correlation to Google Rankings. *Moz*. Retrieved October 4, 2020 from https://moz.com/blog/traffic-engagement-metrics-their-correlation-to-google-rankings

[97] Neil Salkind. 2010. Triangulation. In *Encyclopedia of Research Design*. SAGE Publications, Inc., 2455 Teller Road, Thousand Oaks California 91320 United States. DOI:https://doi.org/10.4135/9781412961288.n469

[98] Sam Romain. 2018. Website Traffic Estimators: The Best Tools for the Job | Romain Berg. Retrieved October 2, 2020 from https://www.romainberg.com/website-traffic-estimators-the-best-tools-for-the-job/, https://www.romainberg.com/website-traffic-estimators-the-best-tools-for-the-job/

[99] Quirin Scheitle, Oliver Hohlfeld, Julien Gamba, Jonas Jelten, Torsten Zimmermann, Stephen D. Strowes, and Narseo Vallina-Rodriguez. 2018. A Long Way to the Top: Significance, Structure, and Stability of Internet Top Lists. *Proceedings of the Internet Measurement Conference 2018* (2018), 478–493. DOI:https://doi.org/10.1145/3278532.3278574

[100] Shanhong Liu. Desktop internet browser market share 2015-2020. *Statista*. Retrieved October 6, 2020 from https://www.statista.com/statistics/544400/market-share-of-internet-browsers-desktop/

[101] Abhinav Shukla, Shruti Shriya Gullapuram, Harish Katti, Karthik Yadati, Mohan Kankanhalli, and Ramanathan Subramanian. 2017. Evaluating content-centric vs. user-centric ad affect recognition. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (ICMI '17), Association for Computing Machinery, New York, NY, USA, 402–410. DOI:https://doi.org/10.1145/3136755.3136796

[102] SimilarWeb. 2018. SimilarWeb Marketing Solution: The Most Reliable and Comprehensive Data on Competitor and Market Strategies.

[103] Similarweb. 2020. SimilarWeb Data Methodology. *SimilarWeb Data Methodology*. Retrieved October 4, 2020 from http://support.similarweb.com/hc/en-us/articles/360001631538

[104] Similarweb. 2020. Our Data. *Our Data | Similarweb*. Retrieved October 4, 2020 from https://www.similarweb.com/corp/ourdata/

[105] SimilarWeb. 2020. What does connecting my Google Analytics account with SimilarWeb mean? – Knowledge Center - SimilarWeb. Retrieved October 5, 2020 from https://support.similarweb.com/hc/en-us/articles/208420125-What-does-connecting-my-Google-Analytics-account-with-SimilarWeb-mean-

[106] SimilarWeb. 2020. SimilarWeb Category. *Knowledge Center - SimilarWeb*. Retrieved October 8, 2020 from https://support.similarweb.com/hc/en-us/articles/360000810469

[107] SimilarWeb. 2020. SimilarWeb All Categories. *SimilarWeb.com*. Retrieved October 8, 2020 from https://www.similarweb.com/category/

[108] Himani Singal and Shruti Kohli. 2016. Trust necessitated through metrics: estimating the trustworthiness of websites. *Procedia Computer Science* 85, (2016), 133–140.

[109] Himani Singal and Shruti Kohli. 2016. Mitigating Information Trust: Taking the Edge off Health Websites. *International Journal of Technoethics (IJT)*. Retrieved October 4, 2020 from www.igi-global.com/article/mitigating-information-trust/144824

[110] Anna S. Smoliarova and Tamara M. Gromova. 2019. News Consumption Among Russian-Speaking Immigrants in Israel from 2006 to 2018. In *Digital Transformation and Global Society* (Communications in Computer and Information Science), Springer International Publishing, Cham, 554–564. DOI:https://doi.org/10.1007/978-3-030-37858-5_47

[111] Tawatchai Suksida and Lalita Santiworarak. 2017. A study of website content in webometrics ranking of world university by using similar web tool. In *2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP)*, 480–483. DOI:https://doi.org/10.1109/SIPROCESS.2017.8124588

[112] The SaaS Report. SimilarWeb | The Software Report. Retrieved October 2, 2020 from https://www.thesoftwarereport.com/top-companies/similarweb/

[113] Mike Thelwall. 2006. Interpreting social science link analysis research: A theoretical framework. *Journal of the American Society for Information Science and Technology* 57, 1 (2006), 60–68. DOI:https://doi.org/10.1002/asi.20253

[114] Mike Thelwall, Liwen Vaughan, and Lennart Björneborn. 2005. Webometrics. *Annual Review of Information Science and Technology* 39, 1 (2005), 81–135. DOI:https://doi.org/10.1002/aris.1440390110

[115] Times Internet. 2020. Insights - competition benchmarking tools in the internet industry | Times Internet. Retrieved October 2, 2020 from https://timesinternet.in/advertise/marketing/insights/competition-benchmarking-tools-in-the-internet-industry/

[116] Todd Spangler. 2019. U.S. Households Have Average of 11 Devices. 5G Will Push That Higher - Variety. Retrieved October 6, 2020 from https://variety.com/2019/digital/news/u-s-households-have-an-average-of-11-connected-devices-and-5g-should-push-that-even-higher-1203431225/

[117] Tyler Horvath. 2020. 8 Most Accurate Website Traffic Estimators. *Ninja Reports*. Retrieved October 2, 2020 from https://www.ninjareports.com/website-traffic-estimators/

[118] Edward Upton. 2018. 88% of Shopify stores have Google Analytics set up incorrectly. *Econsultancy*. Retrieved October 9, 2020 from https://econsultancy.com/shopify-stores-google-analytics-set-up-incorrectly/

[119] Liwen Vaughan and Rongbin Yang. 2013. Web traffic and organization performance measures: Relationships and data sources examined. *Journal of Informetrics* 7, 3 (July 2013), 699–711. DOI:https://doi.org/10.1016/j.joi.2013.04.005

[120] Amy Vecchione, Deana Brown, Elizabeth Allen, and Amanda Baschnagel. 2016. Tracking User Behavior with Google Analytics Events on an Academic Library Web Site. *Journal of Web Librarianship* 10, 3 (July 2016), 161–175. DOI:https://doi.org/10.1080/19322909.2016.1175330

[121] Suzanne Vranica. 2014. A “Crisis” in Online Ads: One-Third of Traffic Is Bogus. *Wall Street Journal*. Retrieved October 4, 2020 from https://online.wsj.com/article/SB10001424052702304026304579453253860786362.html

[122] Chaitanya Vyas. 2019. Evaluating state tourism websites using Search Engine Optimization tools. *Tourism Management* 73, (August 2019), 64–70. DOI:https://doi.org/10.1016/j.tourman.2019.01.019

[123] W3Techs. 2020. Usage Statistics and Market Share of Google Analytics for Websites, October 2020. Retrieved October 5, 2020 from https://w3techs.com/technologies/details/ta-googleanalytics

[124] Peiling Wang, Michael W. Berry, and Yiheng Yang. 2003. Mining longitudinal web queries: trends and patterns. *J. Am. Soc. Inf. Sci. Technol.* 54, 8 (June 2003), 743–758. DOI:https://doi.org/10.1002/asi.10262

[125] Michael Weissbacher. 2016. These Browser Extensions Spy on 8 Million Users Extended. Retrieved October 4, 2020 from /paper/These-Browser-Extensions-Spy-on-8-Million-Users-Weissbacher/3fe57d1556158da7fe373fb577ac5cbbc3f1e84b

[126] Kirsty Williamson, Frada Burstein, and Sue McKemmish. 2002. Chapter 2 - The two major traditions of research. In *Research Methods for Students, Academics and Professionals (Second Edition)*, Kirsty Williamson, Amanda Bow, Frada Burstein, Peta Darke, Ross Harvey, Graeme Johanson, Sue McKemmish, Majola Oosthuizen, Solveiga Saule, Don Schauder, Graeme Shanks and Kerry Tanner (eds.). Chandos Publishing, 25–47. DOI:https://doi.org/10.1016/B978-1-876938-42-0.50009-5

[127] Craig E. Wills and Mihajlo Zeljkovic. 2011. A personalized approach to web privacy: awareness, attitudes and actions. *Information Management & Computer Security* 19, 1 (March 2011), 53–73. DOI:https://doi.org/10.1108/09685221111115863

[128] WRAL. 2009. A study of Internet users’ cookie and javascript settings. *smorgasbork*. Retrieved October 6, 2020 from http://www.smorgasbork.com/2009/04/29/a-study-of-internet-users-cookie-and-javascript-settings/

[129] Yassir Sahnoun. 2018. SimilarWeb Review: Know Your Audience, Win Your Market. *Monitor Backlinks Blog*. Retrieved October 2, 2020 from https://monitorbacklinks.com/blog/content-marketer/similarweb-review

[130] Yossi Wasserman. 2018. How SimilarWeb analyze hundreds of terabytes of data every month with Amazon Athena and Upsolver. *Amazon Web Services*. Retrieved October 2, 2020 from https://aws.amazon.com/blogs/big-data/how-similarweb-analyze-hundreds-of-terabytes-of-data-every-month-with-amazon-athena-and-upsolver/

[131] Zaiqing Nie, S. Kambhampati, and U. Nambiar. 2005. Effectively mining and using coverage and overlap statistics for data integration. *IEEE Transactions on Knowledge and Data Engineering* 17, 5 (May 2005), 638–651. DOI:https://doi.org/10.1109/TKDE.2005.76

[132] M. Zhang. 2015. Understanding the relationships between interest in online math games and academic performance. *Journal of Computer Assisted Learning* 31, 3 (2015), 254–267. DOI:https://doi.org/10.1111/jcal.12077

[133] Meilan Zhang. 2014. Who are interested in online science simulations? Tracking a trend of digital divide in Internet use. *Computers & Education* 76, (July 2014), 205–214. DOI:https://doi.org/10.1016/j.compedu.2014.04.001

[134] Zhiqiang (Eric) Zheng, Peter Fader, and Balaji Padmanabhan. 2011. From Business Intelligence to Competitive Intelligence: Inferring Competitive Measures Using Augmented Site-Centric Data. *Information Systems Research* 23, 3-part-1 (November 2011), 698–720. DOI:https://doi.org/10.1287/isre.1110.0385

[135] 2011. Social Network Sites and Its Popularity. *International Journal of Research and Reviews in Computer Science* 2, 2 (2011), 522–526.

[136] Building Brands Through Social Listening. Retrieved October 4, 2020 from https://web.b.ebscohost.com/abstract?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=1936203X&AN=134109920&h=M98IFv9KXJl5Uay3jL4WvedasHn808XKYjkjP6LkkGkiUJ1jnRUxkdNLuetcXgAbru%2bdod8zXP4olQveb%2bmgNw%3d%3d&crl=c&resultNs=AdminWebAuth&resultLocal=ErrCrlNotAuth&crlhashurl=login.aspx%3fdirect%3dtrue%26profile%3dehost%26scope%3dsite%26authtype%3dcrawler%26jrnl%3d1936203X%26AN%3d134109920