Triangulated Rank-ordering of Web domains

Jeffery Kline*, Avram Aelony†, Brian Carpenter†,Paul Barford‡
*American Family Insurance
Madison, WI USA
Email: jklin1@amfam.com
†Hitwise
Santa Monica, CA USA
Email:aaelony@catasys.com, briancarpenter@live.com
‡University of Wisconsin – Madison
Department of Computer Sciences
Madison, Wisconsin USA
Email: pb@cs.wisc.edu

Abstract—**The relative popularity of web sites, as expressed in published rankings, is of fundamental value in many contexts including search, advertising and research. In this paper, we consider the surprisingly challenging problem of *generating consistent and reliable web site rankings based on unique visitors per day*. We illustrate the challenge this represents using data from three large and independently-sourced Internet user panels. We begin by showing that generating a website ranking based simply on the observed unique daily visitors produces highly inconsistent rankings–even among the most popular sites. To mitigate the problems of bias and measurement error, we introduce a general methodology that identifies "canonical panelists": an abstract class of user that exhibits consistent behavior across panels. Our definition is based on the epistemological technique of triangulation, which refers to observing the same object from multiple perspectives at the same moment in time. We show that panelists in the canonical class exhibit desirable characteristics including improved persistence. Most significantly, we show that defining a domain's rank as a function of the aggregate behavior of canonical panelists improves overall alignment of rankings across all three of our panels.**

I. Introduction

Online advertising generates more than $100 billion in revenues annually. Revenue for individual web sites is a function of visits by users (resulting in *ad impressions*) and fees paid for ad placements. The former depends on issues such as general reputation and how sites are listed in search results [4]. The latter depends on a complex set of issues related to ad serving infrastructure such as Google's Ad Exchange. One way to boost both user visits and ad placement fees is by being highly placed in *web rankings*.

The objective of a web ranking is to arrange web sites in an ordered list based on visitation metrics such as *unique visitors* or *page views* over some period of time. Such rankings are typically produced by third parties such as Alexa [2] or Comscore [5], and the rankings are inferred from measurements of *user panels i.e.,* a group of people who are compensated to allow their web browsing behavior to be tracked.[1] Standard methods for panel data collection include JavaScript tags, packet tracing, client-side browser toolbars and data from installed software such as VPN or custom software. Panels are common, and they have been an industry norm within the traditional and digital media realms for decades. It is hard to underestimate the importance of metrics used to inform web rankings: they are cited in SEC filings of online companies and they are closely followed by investment firms and industry analysts (*e.g.,* [13]). Audience growth, an audience's composition and its current size are fundamental to the market's assessment of media and advertising companies valuation and future prospects. And yet the basic task of rank-ordering web domains remains an open problem.

To the uninitiated, domain ranking might seem as if it could be reduced to one or more elementary arithmetic operations that get applied to web server logs, for example. However, due to measurement bias and the fact that no standard calibration mechanisms that quantify measurement accuracy exist, *more data is not necessarily better*. Additionally, experience shows the counter-intuitive fact that simple aggregation of disparate data sources often does not improve stability of fundamental metrics. As evidenced by the many commercial ranking products on the market, each supported by a team of developers, data scientists and statisticians, *a meaningful domain ranking scheme requires sophisticated measurement infrastructure, an effective deployment method, an analysis framework that is aware of the roles that specific domains and subdomains play within organic web browsing behavior, and every component must adapt as technologies and markets evolve.* In short, generating a daily rank-ordering of domains that is robust to bias, corrupted data, human and measurement error, works at web scale, and does this in an efficient fashion is a significant challenge.

There are a myriad of issues that must be addressed in order to create useful rankings from user panels. Companies that maintain panels and produce rankings are aware of these issues and go to great lengths to assure that their projections are reasonable. However, a simple examination of publicly visible rankings shows that while they may be internally consistent, they are often not well aligned [14]. To illustrate concretely,

[1] While specifics vary, companies are very clear in their terms of service about the nature of tracking and users can opt out at any time.

wikipedia.org was ranked 5th, 17th and 150th according to Alexa, Comscore, and Quantcast, respectively in May, 2019. Cursory inspection of these lists finds that similar examples abound. We argue that the most significant challenge in producing consistent web site rankings from user panels is the inherent uncertainty in the data itself. Observing traffic from a uniform random sample of internet users is not technically feasible, and measuring and removing bias from panel data is a major challenge. Indeed, classical and sophisticated debiasing methods, such as Multilevel Regression and Poststratification (MRP) [7], are difficult to apply due to differences in panel distribution/composition and how each panel captures and defines essential metrics including User ID, unique visitor and page view.

A novel aspect of our approach is that it is founded upon the assumption that a ground-truth rank-ordering of Web domains simply does not exist. An important corollary to this is that we do not expect the union of disparate data sources to yield a more accurate rank ordering than a single data source. This is because "accurate rank ordering" presupposes the existence of a ground-truth ordering. Instead, our aim is *alignment* of rank-orders across multiple data sources. This approach has several benefits. First, *alignment* is easy to quantify using standard statistical tools. Second, relying on multiple data sources provides the opportunity to be robust to skew, measurement artifacts and other errors that are certain to exist in all data sources, but not in the same way.

Indeed, observations of the same object from multiple perspectives at the same moment in time, *i.e., triangulation*, is a powerful epistemological technique. In this paper we consider the problem of generating consistent, reliable web site rankings by using the observations from multiple web user panels. Our objectives are to *(i)* develop an empirical understanding of the underlying issues that lead to misalignment and *(ii)* develop a principled method that will objectively improve alignment of rankings generated by multiple and diverse panel sources.

Our work is based on data provided by three large commercial internet user panels. The three populations of panelists are essentially disjoint, and the data are measured from different perspectives. While privacy concerns and terms-of-use-agreements prohibit us from releasing data, our methods are designed to be data agnostic, and our results may be tested and reproduced with data having similar form, and which can be acquired from an academic research institution or from one of many commercial data vendors. We consider data provided by each panel during the full months of February and March 2019.

We begin by analyzing rankings based on the total number of unique visitors in the untreated data sources. Similar to prior studies [11], [14], we find that rankings are not well aligned. Specifically, we find misalignment between rankings from all three panels in terms of ordering and inclusions/omissions even among the top 10 ranked sites. Beyond the top 10, we consider the top 100 and 1000 sites and see significant differences in rankings from all three panels.

Next, we propose a first principles approach for web site ranking based on the concept of a *canonical panelist*. We define a canonical panelist as a panelist whose behavior is consistent across panels and, when the population of canonical panelist is considered in aggregate, it will result in consistent site rankings. ***Our approach does not require that an individual user be tracked across panels.*** Rather, we develop a straightforward, panel-agnostic method that identifies a class of panelist that is based on a very simple set of rules related to panelist browsing activity. We show that this parsimonious approach improves alignment of site rankings across data sources for the top 10, 100 and 1000 domains. Finally, we contrast the rankings that result from our panels with those of Tranco [14], a recently-announced ranking methodology that combines rankings from diverse sources with the goal of mitigating the effects of adversarial manipulation of long-tail site rankings to facilitate reproducible academic research. Unsurprisingly, we find differences between rankings, highlighting the difficulty of identifying a single definitive notion of domain rank.

In summary, our paper makes the following contributions: we present an empirical evaluation of web site rankings using data from three large Internet panels, we develop a framework for defining *canonical panelists* across panels and we show improved web site ranking consistency. While our approach leads to improvements in ranking consistency, our results highlight the challenges and opportunities for further improvements in web site ranking alignment, web measurement and extracting useful information from diverse and independent panel data sources.

II. Data

We use three sources of web panel data that contain a portion of the HTTP requests issued from desktop (or laptop) devices. The dates spanned by these data sources cover 8 full weeks starting February 1, 2019. While each data source has international reach, we restrict our view, using IP geolocation, to US traffic only. All three of these data sets were acquired from third-party commercial web panel data vendors, which we believe to be representative of this space. For convenience, we label the data sets A, B and C. Figure 1 contains high-level descriptive statistics about each data source. Corrupted records and records identified as being generated by fraudulent means are removed before our analysis. Sensitive and private information within these data sources was handled in accordance with the most stringent of: established best practice, internal policy, terms-of-use agreement, or legal requirement.

As the data we use for this research represents a comprehensive view of an individual's online behavior, the ethical considerations of such data collection methods are of paramount importance. We acquired, handled and processed all data used in this work in a manner that is consistent with legal opinions and negotiated contractual agreements that were developed by an internal legal team, as well as the legal teams of our third party data sources. In order to maintain viable operational businesses, all entities involved in this domain must adhere to the policies of online application stores, OEM manufacturers, and expected public norms, or they risk being suspended and losing their user-base. Additionally, as we are processing third-party data, we must take care not to expose identities or behaviors that can identify either individuals or our data partners. Finally, the new regulatory environment (*e.g.*, GDPR [6] and CCPA [3]) establishes broad and strict penalties for violations of user privacy.

Data sets A, B and C all measure web traffic by executing a dedicated process on the client-side machine and by occasionally reporting telemetry to a remote service. However, each data source reports from a unique perspective, and we now describe the significant ways that A, B and C differ.

A large fraction of the records in A and C reflect HTTP requests that a person browsing the web is likely to see: the URLs in these two sources are likely to appear directly in a web browser's location bar. On the other hand, the URLs contained in B reflect this portion of HTTP requests well as the HTTP requests that occur out of direct sight of the user, such as requests to content delivery networks, advertisement requests, heartbeats, and so on.

Every record is guaranteed to include the timestamp of the HTTP request, a persistent user identifier (UID) and the domain portion of the URL appearing in the HTTP request. However, the UID can have several meanings, and persistence of the UID over time is highly desirable but not guaranteed. For example, if the data collection scheme is through an application that is installed on the client machine, the UID may be assigned at install time and is typically highly persistent. On the other hand, if the data collection method is through web browser extensions, the UID may be equivalent to a browser cookie, and therefore subject to cross-domain restrictions and other standard security policies, and it may not survive intact across multiple browsing sessions. Careful treatment of such technical concerns is absolutely critical to well-formed notions of browsing session, monthly active user, and other foundational Internet audience metrics.

Within the web site audience reporting ecosystem, many methods of ranking domains are used. Two of the simplest methods count the number of unique visitors to a domain and/or count the total volume of HTTP requests to a domain. In this paper, our focus is domain rank according to the number of distinct persistent unique identifiers (UID). Examples of UIDs include browser cookies and UIDs that are assigned at the time of software installation. An idealized assumption is that a one-to-one mapping exists between a UID reported in web traffic and an individual desktop machine. In practice, and for a multitude of reasons, this mapping is often many-to-1, and the manner in which the mapping diverges from a bijection is often not uniform across domains. A view that addresses persistence of UIDs within data sets A, B and C is contained in Figure 1. User persistence in A and B is high while C contains a large volume of users who are seen on just one day over the course of a month. We refer to such users as *ephemeral*.

Our ansatz is that data sources A, B and C each represent noisy and biased samples of observations about events that actually occurred in the real world. But we do not have enough information about the confounding issues that cloud our view of actual events. Developing a clear understanding of what really happened based on observations reported in A, B and C is challenging, as illustrated in Figure 2. In this figure, we observe the daily counts of distinct UIDs visiting www.mozilla.org and the domain of a top-tier publisher that has both large audience and national reach. It is clear that traffic to Mozilla's web site was highly correlated within the populations of A and C but not in B. On the other hand, UID volume to the publisher is more closely aligned in A and B than between A and C. The publisher has been anonymized, and the y-axis labels eliminated to reduce the information exposed about panel composition, size and details about publisher audiences that may have some impact on current business activities or contractual agreements with data partners, and naming the publisher is inessential to the present work. Neither The Mozilla Foundation nor the Wikimedia Foundation are commercial publishers, which is why some of their details are preserved.

We end this section with general comments about our processing infrastructure. Processing of our panel data was straightforward: each of the data sources was first reshaped into a normalized format and then placed at a location where it became accessible for subsequent analysis. The total storage requirements were under 1 TB. Most processing occurred on a Hive cluster that consists of several hundred nodes and a commercial cloud-hosted data warehouse.

III. Methods and Results

Our goal is to order domains according to the mean daily number of unique UIDs observed, and to do this in a way that is independent of the data source. It is far from clear that any reasonable treatment of A, B and C could align domain ranks. However, as demonstrated by Figure 2, multiple views of the same object at the same moment in time, *i.e., triangulation*, can provide substantial evidence in support of ground truth in internet measurement. Based on experience, drastic changes in the traffic of a single data source are, in fact, fairly common. Principled methodologies that produce web site rankings need to be robust against these common and unpredictable step changes. The common occurrence of unexpected–and often unexplained–large scale step-changes impels the community to take a closer look at the foundations of web site rankings, and to take steps to mitigate their effect.

A visualization of the challenge that this represents is provided in Figure 3 (right). This figure compares ranking between data sources B and C. Each dot in the figure represents a domain that appeared in both data sources. The (x, y)-location of each dot corresponds to the domain having rank x in B and y in C. Domains that have the same rank within both data sources live along the main diagonal. It is clear from this figure that large disagreements exists between the rankings within B and C, even for highly-ranked domains.

While many metrics are used to rank domains, we focus on the count of distinct UIDs, in part for its simplicity. Two

	Distinct UIDs		Total Records	
	daily median	total	daily median	total
A	1138k	2634k	215m	5777m
B	349k	929k	107m	2629m
C	73k	597k	6m	162m

	Num. distinct UIDs		
	exactly 1 day	2 – 28 days	ratio
A	488k	2147k	0.23
B	239k	690k	0.35
C	395k	202k	1.95

Fig. 1. *Left.* Descriptive metrics of data sources A, B and C for all 28 days of February 2019. *Right.* Data sources A and B have fairly persistent UID populations while C has a significant ephemeral UID population. A related analysis appears in Figure 9.

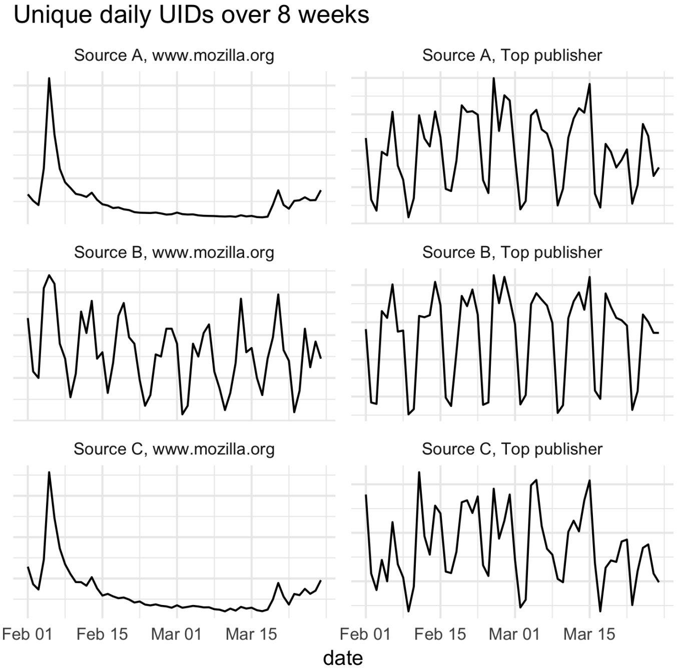

Fig. 2. Requests to www.mozilla.org in data sources A and C are highly correlated while daily visitors to a top-tier publisher are better aligned within A and B.

other widely-used methods of ranking domains are 1) the number of user sessions and 2) the number of monthly active users [13], [8]. Both of these pre-suppose the existence of a persistent unique identifier and apply additional heuristic rules (and possibly other data sources) to estimate or infer the final number. While each of these other metrics finds widespread use, each also introduces complexity and brings its own set of challenges.

It is common, when reporting on domain traffic to aggregate subdomain traffic to a parent domain. We also make no effort at domain normalization. Aggregation of this sort is important both from a technical standpoint as well for more strategic use cases. However, the problem of identifying the proper method of aggregation depends strongly on what the use-case is. The standard technical approach to aggregations of subdomains (*e.g.*, aggregations using the public suffix list [15]) is often in conflict with a view that better informs the more strategic use case. The main consequence of not aggregating is that a handful of highly fragmented groups of domains (*e.g.*, Yahoo!) may be accurately reported though the fragmentation causes the parent brand to assume a rank lower than would be found by aggregating domains using the public suffix [15] list. To make this more concrete, Figure 4 displays the number of February 2019 distinct UIDs to the Yahoo! family of domains. The domain www.yahoo.com is the dominant domain visited in each data source, but several other subdomains also exhibit large volume and also rank highly.

Measurement artifacts present themselves in other ways. This is highlighted in Figure 7. In this figure, domains are rank-ordered according to raw measurements. The perspective of measurement for each data source skews the type of domain present in the top few domains.

We now introduce the notion of a *canonical user*. The canonical user is a UID that is highly traceable within a panel (*i.e.,* behaviors of these UIDs can be traced across domains and over time better than most UIDs) and is said to represent "typical" online activity. This is a deliberately general statement, and our implementation described below is intentionally kept simple. Given the complex and dynamic nature of Internet data, we argue that simplicity is an essential

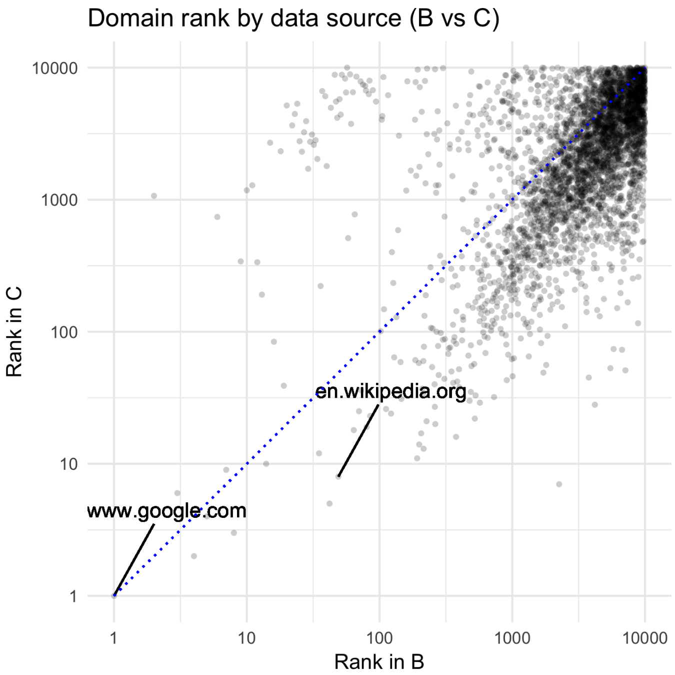

Fig. 3. Rank misalignment is shown for 10,000 domains within B and C. Note the log-log scale.

	A		B		C	
rank	domain	UIDs	domain	UIDs	domain	UIDs
1	www.yahoo.com	355k	www.yahoo.com	33k	www.yahoo.com	23k
2	mail.yahoo.com	335k	mail.yahoo.com	31k	search.yahoo.com	19k
3	search.yahoo.com	305k	search.yahoo.com	22k	mail.yahoo.com	14k
4	finance.yahoo.com	73k	guce.yahoo.com	19k	images.search.yahoo.com	4k
5	answers.yahoo.com	71k	answers.yahoo.com	18k	finance.yahoo.com	4k
-	yahoo.com	5k	yahoo.com	43(*sic*)	yahoo.com	0

Fig. 4. A common pattern of subdomain usage is that a flagship domain is the dominant domain related to the "classical" use case while subdomains are used for some other purpose such as applications (*e.g.* email) or to host supplementary content (*e.g.*, images).

feature of any framework that provides both robustness and reproducibility.

As our starting point, we identify N domains with the property that each domain has, within each data source, broad reach, *i.e.,* a nontrivial fraction of UIDs have been observed visiting said domain. The list of domains we selected appears in Figure 6. The length of this list aims to balance being small enough to facilitate manual review (*e.g.,* maintenance, sanity checks and tractability) while being large enough to provide diversity. Two other well-established rules of thumb suggest a rough lower-bound of 25 is sufficient to achieve a certain kind of diversity, namely "the rule of 30" of classical statistics, and the lower-bound of 20 for a well-diversified financial asset mix [10], [9].

An excess of precision in both the size and composition of this list is unwarranted here: extending the list from 25 to 50, modifying the thresholds that define what "nontrivial fraction" means, and so on, do not fundamentally change the composition of the canonical user population. Note too that our aim is organic web browsing behavior. Since *.google.com and facebook.com (distinct from www.facebook.com) serve multiple unusual high-volume use-cases (social media, online application hosting, CDN infrastructure and search), these domains have been manually excluded.

Post-hoc analysis of this list reveals that each of the selected domains has high business reputation[2], and we find that each domain is reported in the top 100 of at least two third-party publicly available online media ranking lists. Based on available evidence, we argue that, with overwhelming likelihood, each member of the US online population interacts with some positive number of these websites at least once over the course of 4 weeks. Our central thesis is that this population of users, in the aggregate, behaves similarly across data sets, and that rankings derived from this population will be better

[2] The *business reputation* of a site is high if brands have faith that their paid-for placements will be delivered and contract terms honored.

List of Canonical domains	
www.amazon.com	www.paypal.com
www.cnn.com	www.pinterest.com
www.ebay.com	www.quora.com
www.etsy.com	www.reddit.com
www.facebook.com	www.twitter.com
www.foxnews.com	www.walmart.com
www.imdb.com	www.washingtonpost.com
www.instagram.com	weather.com
www.linkedin.com	www.wsj.com
www.msn.com	www.yahoo.com
www.netflix.com	www.yelp.com
www.nytimes.com	www.youtube.com
	www.zillow.com

Fig. 5. The 25 domains that are used to identify canonical users.

	A		B		C	
k	UIDs	%	UIDs	%	UIDs	%
0	677k	25.7	334k	36.0	309k	51.7
1	490k	18.6	184k	19.8	190k	31.8
2	313k	11.9	102k	11.0	46k	7.7
3	232k	8.8	66k	7.0	20k	3.4
4	186k	7.0	49k	5.3	11k	1.8
5	155k	5.9	39k	4.2	7k	1.1

Fig. 6. The number of users who visited exactly k of the selected domains. A significant fraction of UIDs in every data source have visited at least 1 of the domains.

aligned than the unrestricted populations. Additionally, and critically, this set of users comprises a relatively large subset of the entire corpus of data. Thus we argue it can be used to report on traffic for domains that are not within the selected set of prominent high-reputation web sites mentioned above.

To make the process of assembling a list of canonical domains more formal, we present the following. A domain may be admissible if:

1) (reach) a nontrivial fraction of all UIDs visit the domain in each data source,
2) (stability) the daily volume of visits to the domain is consistent over multiple days in each data source,
3) (heterogeneity) the domain is one of at least 25 domains in each data source that exhibits both stability and reach,
4) (viability) the domain should represent a viable commercial or not-for-profit entity that has, as part of its mission, a large online presence.

The terms "nontrivial fraction" and "consistent" could be stated more precisely, as a threshold, or some other type of rule. But doing so would merely cloak arbitrary decisions with an artiface of precision. The guiding principle is that whatever thresholds and domains are selected, the output of the ranking process should not be overly sensitive to them.

Membership to the canonical population of UIDs depends both on the set of canonical sites selected and the number of sites visited. We hold fixed the selected sites, but several of our analyses vary by the number of sites visited, k. Figure 6 shows the population sizes for a range of k. In this figure, the population shown describes the number of UIDs who visited exactly k of the canonical domains while the definition of the canonical population requires that the UID visit *at least* k domains. Interestingly, and consistent with our hypothesis, in all three data sources a very large fraction of observed UIDs have visited at least 1 of the canonical domains.

To quantify performance of rank alignment, we use Spearman's rank correlation coefficient, also known as Spearman's ρ. Given n pairs of observations, (u_i, v_i), assign ranks to u_i and v_i. The Spearman rank correlation coefficient between u and v is the correlation between the ranks. That is, the Spearman correlation coefficient is the standard correlation coefficient except one uses the ranks in place of the actual observations [18]. Figure 8 displays Spearman's ρ across data sources as a function of the number of domains visited, k. It is clear that as k increases, the correlation across data sources also increases. The results of this figure show alignment for the top 10, 100 and 1000 domains.

Churn of the canonical user population is described in Figure 9. The analysis here is identical to the analysis shown in Figure 1, except that the population considered is the canonical UIDs. Based on a comparison of these two figures, it is clear that the canonical population is far more likely to have multi-day visits than the general population. Next, we explore in detail the rank by data source over time for a single domain.

Figure 10 displays the rank of en.wikipedia.org for the first 7 days of February for both raw and canonical populations. One interesting feature is that even though en.wikipedia.org is not on the selected canonical list (see Figure 6), it still has a very high rank. Compared to the raw view, rankings of this domain by the canonical population are more stable and align better with general expectations based on publicly available reports on that domain's typical daily volume [17]. As discussed above, data source B contains a large fraction of traffic that is associated with follow-on requests, such as advertising and CDN traffic. As a result, it is expected that domains of this type of content will rank highly, and several of them outrank en.wikipedia.org.

Finally, we compare our raw and canonical ranks of data source C to the most recent results from Tranco [14], a recently-announced ranking method that combines rankings from four sources: browser toolbar data, DNS lookups, crawl data, and ISPs. Although Tranco is informed by multiple data sources just as our method is, the two methods differ substantially. As one illustration of this, we note that our method relies on persistent identifiers, while Tranco, which includes crawl data, does not. To make our un-aggregated domain rankings compatible with Tranco's, subdomains are aggregated to match Tranco's nonstandard notion of "pay domains." A comparison of the top domains is shown in Figure 10. We speculate that the Tranco methodology gives more weight to the raw count of HTTP requests than our methodology, which is focused on unique users.

	A		B		C	
r	Raw	Canonical	Raw	Canonical	Raw	Canonical
1	www.google.com	www.google.com	www.google.com	www.google.com	www.google.com	www.youtube.com
2	www.youtube.com	www.youtube.com	www.youtube.com	www.youtube.com	www.youtube.com	www.google.com
3	mail.google.com	www.amazon.com	docs.google.com	www.facebook.com	mail.google.com	www.amazon.com
4	www.facebook.com	www.facebook.com	www.facebook.com	www.amazon.com	www.amazon.com	en.wikipedia.org
5	www.amazon.com	mail.google.com	mail.google.com	*(blank)*	www.facebook.com	mail.google.com
6	en.wikipedia.org	en.wikipedia.org	drive.google.com	contacts.google.com	en.wikipedia.org	www.ebay.com
7	docs.google.com	twitter.com	contacts.google.com	mail.google.com	www.pornhub.com	www.facebook.com
8	drive.google.com	www.ebay.com	*(blank)*	coin.amazonpay.com	drive.google.com	www.reddit.com
9	twitter.com	docs.google.com	eus.rubiconproject.com	hangouts.google.com	www.ebay.com	www.imdb.com
10	www.ebay.com	www.reddit.com	classroom.google.com	docs.google.com	docs.google.com	www.walmart.com

Fig. 7. Domains rank-ordered by the number of total UIDs observed during the month of February, 2019 for both canonical and raw populations.

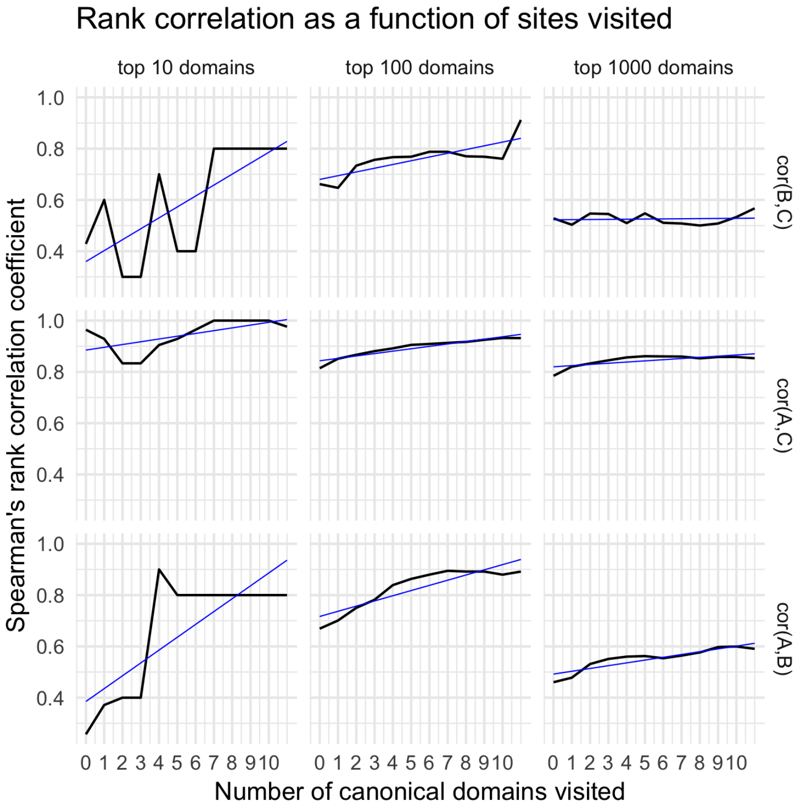

Fig. 8. Spearman's ρ between pairs of data source as a function of k, for the top 10, 100, and 1000 domains. Generally, ρ increases with k.

	Num. distinct UIDs (canonical)		
	exactly 1 day	2 – 28 days	ratio
A	147,083	1,809,974	0.08
B	60,212	534,738	0.11
C	140,737	147,585	0.95

Fig. 9. Persistence of the canonical population by data source. The canonical UIDs are significantly more likely to be observed on more than 1 day. The analysis here is similar to the analysis of the general UID population presented in Figure 1.

IV. Related Work

Methods for web site rankings have been described in several prior studies. Lo and Sedhain analyze the similarities and differences between the top 100 ranked websites from six publicly available lists toward the goal of assessing their reliability [11]. They use four different metrics to show that while membership in the top 10 is relatively consistent, beyond that there can be large divergences. They conclude by suggesting that a more reliable list might be generated by combining several different rankings. More recently, Le Pochat *et al.* consider the issue of website rankings from an adversarial perspective, recognizing that "traffic-based rankings" such as page view or unique visitors can be subject to manipulation [14]. They show the lack of correspondence between four web ranking lists, their vulnerability to manipulation and propose a list called *Tranco* that combines four commercial rankings, and is meant to improve agreement on domain popularity and stability over

	Rank of en.wikipedia.org	
Date	Raw pop. A–B–C	Canonical pop. A–B–C
2019-02-01	11–49–8	6–15–6
2019-02-02 (Sat)	11–48–8	6–13–6
2019-02-03 (Sun)	10–45–8	6–12–6
2019-02-04	11–47–8	6–13–6
2019-02-05	12–47–8	6–15–6
2019-02-06	12–49–8	6–15–6
2019-02-07	12–50–8	6–15–6

Domain	Tranco	C	Canonical C	Alexa
google.com	1	1	1	1
youtube.com	2	2	2	2
netflix.com	3	25	23	21
facebook.com	4	6	5	6
microsoft.com	5	16	30	32
wikipedia.org	7	5	4	9
yahoo.com	12	3	7	11
amazon.com	15	4	3	14

Fig. 10. *Left.* Rank of en.wikipedia.org using the population of UIDs that visited more than 9 canonical sites. *Right.* Ranks of selected domains (aggregated) for Tranco [14], data source C the ranking derived from the canonical users of C and Alexa.

time. Our work is complementary to these studies in that we consider the problem of generating consistent rankings from panel-based measurements.

Commercial entities that publish website rankings typically publish only a very high-level description of the methodology that they use. For example, Comscore uses a method called Unified Digital Measurement (UDM), which combines panel and site-based traffic measurements [5]. Amazon's Alexa, which is frequently cited in research studies, generates rankings based on panel data using a proprietary measure of unique visitors and page views [2]. Without providing details, Alexa states that it accounts for bias in their panel through "data normalization" [1]. In general, the notions about what rankings represent are surprisingly vague. Companies such as Comscore aggregate traffic to "web properties" using proprietary and highly complex rules based on decades of experience and complicated by business logic. On the other hand, Alexa aims to heap the observations from large volumes of results into a single, very-high level statistic. However, rich web pages, background requests that commonly occur in web traffic, and that are irrelevant to measuring an atomic page view event on a publisher's web page, cannot be filtered out.

Scheitle, *et al.*, in [16] use top ranked lists as they are published, and the authors report on observed characteristics of said lists from multiple perspectives. The author's observations compare and contrast the composition of top n lists that are collected and processed by a wide variety of methods and collection techniques, and the authors state desirable goals of top lists for use in research: that they be generated with transparency, temporal stability and structural consistency.

Closely related to website ranking is information retrieval in web search, which became a major focus of research after the publication of the PageRank paper by Page *et al.* [12] and the rapid rise in popularity of the Google search engine. Cho and Roy highlight the symbiotic relationship between search results and the popularity of Web pages [4].

V. Conclusions

Web site rankings based on user panel data are fundamental to the large value the market has placed on online advertising, search results and also to academic research studies. In this paper we consider the problem of generating website rankings from multiple panels that are consistent in terms of sites listed and ordering. We present an empirical evaluation of web site rankings using data from three large Internet panels to illustrate how different panels produce different rankings. We develop a method for identifying *canonical users i.e.,* users that exhibit similar behavior on well known sites. We identify canonical users in each of our panel data sets and show that web site ranking based on these users improves consistency between rankings. While our results are encouraging, they point to further opportunities for improvement *e.g.,* through further refinements in canonical user identification, new ranking metrics and new methods for bias identification and removal. The importance that is placed on web site rankings by publishers, investors and advertisers, combined with significant misalignment amongst top-tier sites within prominent web-site ranking lists, shows that there is an urgent need to improve the state of this art.

References

[1] Amazon Alexa. How are Alexa's traffic rankings determined? https://support.alexa.com/hc/en-us/articles/200449744. Accessed: 2019-04-11.

[2] Amazon Alexa. The top 500 sites on the web. https://www.alexa.com/topsites. Accessed: 2019-04-11.

[3] CCPA. California Consumer Privacy Act. https://oag.ca.gov/privacy/ccpa. Accessed: 2019-05-08.

[4] Cho, J., and Roy, S. Impact of Search Engines on Page Popularity. In *Proceedings of The World Wide Web Conference* (May 2004).

[5] Comscore. Latest Rankings. https://www.comscore.com/Insights/Rankings. Accessed: 2019-04-11.

[6] EU-GDPR. The GDPR Information portal. https://eugdpr.org. Accessed: 2019-05-08.

[7] Gelman, A. Struggles with Survey Weighting and Regression Modeling. *Statistical Science 22*, 2 (2007).

[8] Google. How users are identified for user metrics. https://support.google.com/analytics/answer/2992042?hl=en. Accessed: 2019-05-07.

[9] Graham, B. *The Intelligent Investor.* Harper & Row, 1965.

[10] Hogg, R., and Tanis, E. *Probability and Statistical Inference.* Prentice Hall, 2006.

[11] Lo, B., and Sedhain, R. How Reliable Are Website Rankings? Implications for E-Business Advertising and Internet Search. *Issues in Information Systems 7*, 2 (2006).

[12] Page, L., Brin, S., Motwani, R., and Winograd, T. The pagerank citation ranking: Bringing order to the web. Tech. rep., Stanford InfoLab, 1999.

[13] Pinterest, I. Form S-1 REGISTRATION STATEMENT UNDER THE SECURITIES ACT OF 1933. https://www.sec.gov/Archives/edgar/data/1506293/000119312519083544/d674330ds1.htm. Accessed: 2019-05-07.

[14] Pochat, V. L., Goethem, T. V., Tajalizadehkhoob, S., Korczynski, M., and Joosen, W. Tranco, A Research-Oriented Top Sites Ranking Hardened Against Manipulation. In *Proceedings of Network and Distributed Systems Security Symposium* (February 2019).

[15] Public Suffix. Public Suffix. https://publicsuffix.org. Accessed: 2019-04-11.

[16] Scheitle, Q., Hohlfeld, O., Gamba, J., Jelten, J., Zimmermann, T., Strowes, S. D., and Vallina-Rodriguez, N. A long way to the top: significance, structure, and stability of internet top lists. In *Proceedings of the Internet Measurement Conference 2018* (2018), ACM, pp. 478–493.

[17] WikiMedia. Wikimedia Statistics. https://stats.wikimedia.org/v2/#/all-projects. Accessed: 2019-05-09.

[18] Zwillinger, D., and Kokoska, S. *CRC standard probability and statistics tables and formulae*. Crc Press, 1999.