An Improved Entropy Measure for Web Browser Fingerprinting Risk

Alex Berke
Google
USA
aberke@google.com

Enrico Bacis
Google
Switzerland
enricobacis@google.com

Umar Syed
Google
USA
usyed@google.com

Abstract

Browser fingerprinting is the practice of tracking users across the Web by collecting attributes from their devices and combining them to create unique identifiers. This practice poses major privacy risks to users, and more than a decade of research has quantified fingerprinting risks due to various attributes, leading browser developers to implement many privacy-enhancing changes.

Early work used Shannon entropy to quantify risks. However, Shannon entropy can grow with dataset size, limiting the ability to compare datasets and results. Researchers then introduced normalized entropy as a measure for comparing browser fingerprinting datasets of different sizes and numerous works followed using normalized entropy for this purpose.

We identify and address a resulting problem in the fingerprinting literature. We show normalized entropy is ill-suited to compare datasets of different sizes — it decreases as dataset size increases. We show this both analytically and empirically, leveraging a recently published dataset of browser attributes commonly used for fingerprinting.

Given the unmet need for a better fingerprinting risk measure, we define a minimal set of desired properties for such a measure: scale-invariance, monotonicity and estimability. We then propose to use Tsallis entropy as a more interpretable fingerprinting risk measure. We evaluate Shannon, normalized, and Tsallis entropy with respect to the properties, and prove that only Tsallis entropy satisfies all of them.

Keywords

Entropy, Privacy, Browser fingerprinting, Empirical study, Web tracking

1 Introduction

Browser fingerprinting is a method used to covertly track web users across websites, posing major privacy risks to users. Fingerprinting is achieved by collecting attributes from users' browsers which are related to their configuration, hardware, or installed software. Each attribute alone may be shared by many users, but when combined they can create a unique *fingerprint*, identifying a user's device [16]. While tracking cookies can be easily deleted or reset by users, fingerprints cannot, providing a more stable tracking identifier.

Similar strategies are used to uniquely identify and track mobile device users across apps, and devices more generally [26, 38]. Although this work is motivated by prior browser fingerprinting studies, our contributions extend to the study of device fingerprinting more broadly.

Browser fingerprinting mechanisms have been studied since 2009 [32], and continue to evolve [30]. Researchers have used entropy to quantify fingerprinting risks, and the marginal risks posed by various device attributes, and to make comparisons over time and populations [28]. In this work we describe problems with these previous measurements, and present an opportunity for improvement, by proposing Tsallis entropy as an example of a more interpretable and mathematically robust way to measure fingerprinting risk.

1.1 Background

The first large browser fingerprinting study was conducted by Eckersley in 2010 via the Panopticlick[1] website [17]. Browser attributes were collected from more than 470,000 informed users who visited the website. This study introduced *Shannon entropy* to measure fingerprinting risk, where higher entropy indicates higher risk, and published the entropy computed for various browser attributes.

The Panopticlick study had an important impact on user privacy. Major browser vendors addressed risks highlighted by the study by prioritizing attributes that had higher entropy with developments to reduce their entropy, and hence associated risk, by reducing the variation in values they could return [13, 14, 37, 42]. The Panopticlick study thus demonstrated the impact a metric can have on improving user privacy, underscoring the importance of establishing a robust metric to quantify fingerprinting risk.

In 2016, Laperdrix et al. published a study of browser attributes collected from 118,934 browsers via the AmIUnique[2] site [29]. Among their goals was to compare the AmIUnique dataset to the Panopticlick dataset. The different dataset sizes posed an obstacle for such a comparison because Shannon entropy is unbounded and can grow with dataset size (this is further explained in Section 4). To overcome this obstacle, Laperdrix et al. introduced a normalized version of Shannon entropy, termed *normalized entropy*, which is limited to the [0,1] range. Normalized entropy divides the Shannon entropy computed for each attribute, X, by the greatest possible Shannon entropy for X, which represents the worst case where all values of X are unique. After its definition, normalized entropy was adopted and used in numerous highly cited studies to quantify fingerprinting risk [1, 2, 9, 10, 21, 30, 51]. As stated by Laperdrix et al. in both [29] and [28], the goal of this normalization is to "compare two datasets despite their different sizes" as "the advantage of this

Proceedings on Privacy Enhancing Technologies 2026(3), 703–720

https://doi.org/10.56553/popets-2026-0102

[1] http://panopticlick.eff.org
[2] https://amiunique.org

measure is that it does not depend on the size of the anonymity set but on the distribution of probabilities". However, neither a theoretical nor empirical analysis of this property was provided.

1.2 Contributions

In this work we show how normalized entropy can yield misleading results, through both theoretical and empirical analysis. The ready adoption of normalized entropy, despite its limitations, shows that the research community needs a better way to measure fingerprinting risk, which we provide. Our contributions can be summarized as follows.

- We introduce *Tsallis entropy* as a more interpretable alternative to Shannon and normalized entropy for measuring fingerprinting risk.
- We show, via simulations and experiments on real data, that Shannon entropy and normalized entropy have high sensitivity to sample size, and also have large estimation errors, but Tsallis entropy does not.
- We define a set of formal properties that an entropy should satisfy — scale-invariance, monotonicity and estimability — and prove that, unlike Shannon and normalized entropy, Tsallis entropy satisfies all of them.

1.3 Outline

This paper is organized as follows. Section 2 provides a review of related work, highlighting how entropy has been used to study browser fingerprinting. Section 3 provides formal definitions of the entropy measures and related terms used in the paper. Section 4 then demonstrates problems with the common entropy measures used in prior work through illustrative examples. Section 5 addresses these problems by describing a desired set of properties for a fingerprinting risk measure, analytically evaluating each entropy measure against these properties, proving Tsallis entropy satisfies them all while the commonly used measures (Shannon and normalized entropy) do not. Section 6 follows with an empirical evaluation of the entropy measures, using a publicly available dataset of fingerprinting attributes. Finally, section 8 discusses the findings and their implications.

2 Related work

This section provides a brief review of how entropy has been used to study browser fingerprinting. The earliest study did not use entropy at all and instead quantified uniqueness. In 2009, Mayer described how Web 2.0 features could be used to uniquely identify users and track their web activities. He conducted a small study where he asked study participants to visit the website scoop.princeton.edu and collected data from n=1,328 site visitors over a two week period. He then reported on how 1278 (96.23%) could be uniquely identified via a limited set of browser attributes [32].

Entropy was later introduced in the 2010 Panopticlick study by Eckersley and the Electronic Frontier Foundation [17]. This study was much larger (n=470,161) and also quantified the percentage of browsers with unique fingerprints, finding a smaller result of 83.6% unique. Both Mayer and Eckersley described the difficulty of using their samples to reach conclusions about the global uniqueness of browsers. Eckersley then used Shannon entropy to quantify, in bits, the distribution of fingerprints in his sample overall, and computed entropy for each browser attribute in isolation, which he termed "mean surprisal".

Studies then followed using entropy to quantify the ability for specific attributes or fingerprinting strategies to identify browsers. For example, in 2012, Mowery and Shacham described how to fingerprint browsers via the HTML5 canvas element [36], and in 2014, Bojinov et al. described how to create reliable hardware fingerprints from sensors on a smartphone [5], where in each of these studies the authors used Shannon entropy to quantify the efficacy of their fingerprinting techniques. Entropy has since been incorporated into web specifications that highlight the need to minimize entropy to reduce browser fingerprinting risk [49].

"Normalized entropy" was introduced in a 2016 study by Laperdrix et al [29]. Similarly to Eckersley's Panopticlick study, they asked participants to visit their AmIUnique website in order to collect browser attributes and quantify associated fingerprinting risk. They collected n=118,934 fingerprints composed of 17 attributes, including attributes collected by Eckersley. Among the authors' goals was to compare the AmIUnique and Panopticlick datasets. Noting the datasets were different sizes, the authors then introduced their normalized variant of Shannon's entropy, which divides Shannon entropy by $\log(n)$ (the worst-case maximum entropy value that can only be reached when the values of all n users are unique.) The authors stated: *"The advantage of this measure is that it does not depend on the size of the anonymity set but on the distribution of probabilities. We are quantifying the quality of our dataset with respect to an attribute's uniqueness independently from the number of fingerprints in our database. This way, we can qualitatively compare the two datasets despite their different sizes."* The authors did not provide further analysis to validate these statements. They then used their normalized entropy to make comparisons between the datasets as well as between their dataset's mobile and desktop users.

Normalized entropy was then adopted in studies that followed. Researchers used normalized entropy to measure entropy for specific attributes, often making comparisons between their datasets and previously published normalized entropy statistics, such as from the Panopticlick and AmIUnique studies [9, 10, 21, 30].

Given that entropy has been established as a way to measure fingerprinting risk, other researchers have then used entropy to study browser fingerprinting beyond the direct comparisons of attributes. For example, Bacis et al. measured entropy across various websites by observing real user sessions and website API calls, and then made entropy comparisons across website categories to evaluate the relative fingerprinting risks [3]. Boussaha et al. used normalized entropy measurements to classify web fingerprinting activities to provide a component of their taint-tracking technique [6]. Other fingerprinting researchers have used entropy to simply provide descriptive statistics for the datasets used in their work [4, 41].

These prior works demonstrate researchers need a consistent metric that can quantify fingerprinting risk and make comparisons across datasets, browser attributes, and website surfaces more generally. Although Shannon and normalized entropy were introduced and readily adopted for such purposes, to our knowledge researchers have not provided analysis regarding how or why these entropy measures are suitable. Some recent work, such as by Rocher et al. [40], have studied alternatives to entropy, but did not provide

an analytical justification for their use. This work provides such analysis and identifies resulting problems that should be addressed to improve future fingerprinting research, and provides an entropy variant to address these problems.

3 Entropy definitions

Fingerprinting risk is the extent to which device-specific information poses a threat to the users' anonymity and allows for their re-identification as they browse the web. When studying fingerprinting risk, we typically focus on a particular *value* associated with each device. A value may be a single attribute (e.g. User agent or Screen resolution) or the concatenation of multiple attributes which together form a fingerprint. If the values are widely dispersed (*i.e.*, a typical user shares their value with only a few other users), then the fingerprinting risk is large. If the values are concentrated (*i.e.*, a typical user shares their value with many other users), then the fingerprinting risk is small. Since entropy quantifies the dispersal of values [43], entropy is typically used to quantify the level of identifying information in a fingerprint [28], hence serving as a proxy for fingerprinting risk.

Here we provide formal definitions for the terms we use throughout this work. A *sample* is a vector

$$X = (x_1, \ldots, x_n),$$

where each x_i is the *value* for user $i \in [n]$. For each possible value x let

$$n_x = |\{i : x_i = x\}|$$

be the number of users in sample X who have value x, such that the sample size is $n = \sum_x n_x$. Let the *domain size* of sample X be the number of distinct values in X. In other words, the domain size of X is

$$k = |\{x : \exists i \in [n] \text{ such that } x_i = x\}|.$$

The most widely used definition of entropy is Shannon entropy.

Definition 3.1. The *Shannon entropy* of sample X is

$$H(X) = -\sum_x \frac{n_x}{n} \log \frac{n_x}{n}.$$

The logarithm in the definition of Shannon entropy typically has base 2, where the units of Shannon entropy are called *bits*. Shannon entropy is always non-negative, and is maximized when each user has a unique value, *i.e.*, when $k = n$. In this case, $H(X) = \log n$. Laperdrix et al. [29] proposed a variant of Shannon entropy that is normalized by its maximum value, ensuring that the entropy is always contained in the interval $[0, 1]$.

Definition 3.2. The *normalized entropy* of sample X is

$$\bar{H}(X) = \frac{H(X)}{\log n}.$$

In this paper we also consider an entropy definition that was proposed by Tsallis [46] in the context of statistical physics. It has also been applied in ecology, political science and economics [23, 27, 44], but to the best of knowledge has not been used in the study of web fingerprinting.

Definition 3.3. The *Tsallis entropy* of sample X is

$$T(X) = 1 - \sum_x \left(\frac{n_x}{n}\right)^2.$$

Like Shannon entropy and normalized entropy, Tsallis entropy is maximized when each user has a unique value. In this case, $T(X) = 1 - \frac{1}{n}$. Shannon entropy, normalized entropy and Tsallis entropy are all minimized when all users have an identical value. In this case each entropy is 0. Therefore, like normalized entropy, Tsallis entropy is always contained in the interval $[0, 1]$.

We note there is a more general way to define Tsallis entropy, where the exponent is a free parameter q. Our definition of Tsallis entropy, with $q = 2$, is motivated by the resulting interpretation: The Tsallis entropy of a sample X is the probability that two users chosen independently and uniformly at random from X have different values. See Section 5.3 for more details.

3.1 Estimating entropy

We can model the entire user population as a vector $P = (x'_1, \ldots, x'_N)$, and the sample $X = (x_1, \ldots, x_n)$ as being formed by drawing n elements from P uniformly at random without replacement. Our goal is to use X to estimate $\phi(P)$, where ϕ is one of the entropy measures defined above. The most common and straightforward approach is to use the *plug-in estimator*, which takes $\phi(X)$ to be the estimate of $\phi(P)$. We use the plug-in estimator in all analyses in the paper, except in Section 6.3. There we also present results using advanced Shannon entropy estimation procedures.

4 Examples illustrating problems with common fingerprinting measures

Previous browser fingerprinting studies have relied on Shannon entropy and normalized entropy to measure fingerprinting risk, computing these measures over their sample data. Ideally these measures would be independent of sample size. Here we provide illustrative examples of problems with these commonly used measures.

First, we use simulated data to more clearly illustrate these problems (4.1). We then use data from prior studies to provide an empirical example (4.2). To address these problems, in Section 5 we articulate what is desired from a fingerprinting risk measure and demonstrate how Tsallis entropy provides an improvement.

4.1 Simulated examples

For the following examples we simulate data. For many of these examples we use the power law distribution, $f(x) = x^{-1}$, to represent the long-tail distributions commonly found in the empirical data (Section 6) and other natural phenomena [15, 31]. Section 6 complements these simulated examples with empirical data, again illustrating the problems described here. Note that in the following examples we use the *plug-in estimator* for Shannon entropy, which is used again in Section 6.2, while Section 6.3 evaluates advanced Shannon entropy measures.

Example 1: Shannon entropy increases with sample size. A problem with using Shannon entropy for browser fingerprinting studies is that as the sample size of X grows, the entropy, $H(X)$, often grows. This can happen because fingerprinting studies use samples that are relatively small compared to the full population of browsers. These small samples may not include all the values in the domain of X, which would be observed in the full population.

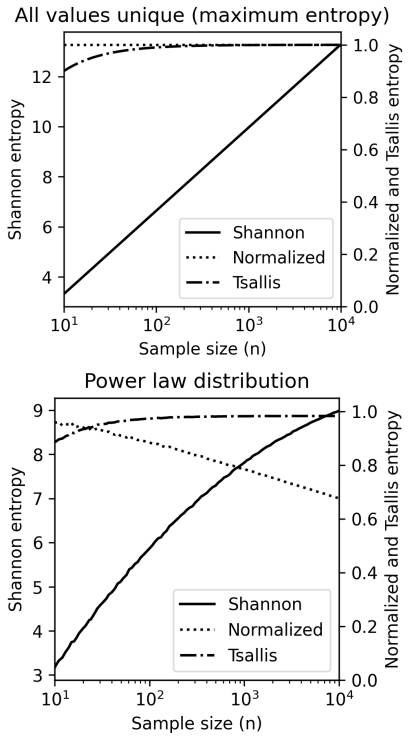

Figure 1: Shannon entropy increases with n (example 1). Top shows the extreme case where all values are unique. Bottom shows the case where values are distributed by the power law.

Thus as larger samples are collected, more new values of X are seen, resulting in a larger domain and larger entropy.

Figure 1 illustrates this problem. The top plot shows the extreme case, where all values of X are unique, exhibiting the highest possible entropy: For a sample of size n, $\frac{n_x}{n} = \frac{1}{n}$ for all x in X, and therefore $H(X) = -\sum_x \frac{1}{n} \log \frac{1}{n} = -n \left[\frac{1}{n} \log \frac{1}{n}\right] = \log n$. Clearly in this case, as n grows, the observed domain grows, and Shannon entropy grows. However, normalized entropy is designed to normalize for this worst case, since $\tilde{H}(X) = \frac{H(X)}{\log n} = 1$. Here, normalized entropy intuitively shows how entropy is at the maximum value, and this is consistent as n increases.

The bottom plot of Figure 1 shows a more common case, where values are distributed by the power law. We use X with domain size 10,000 and draw samples of increasing size, n, ranging from n=10 to 10,000, and for each n we plot the sample entropy averaged over 100 trials. In this case, Shannon entropy again increases as n grows. However, normalized entropy is not consistent either.

Example 2: Normalized entropy decreases with sample size. Despite the fact that normalized entropy was conceived with the goal to compare samples of different sizes, normalized entropy decreases with sample size because the denominator is $\log n$.

Consider the case where the domain of X can be observed from a small sample. Since Shannon entropy, $H(X)$, is computed

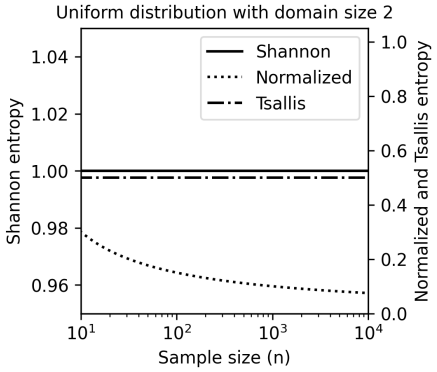

Figure 2: Normalized entropy decreases with n (example 2).

over the distribution of values, $H(X)$ is then constant as the sample size, n, grows. Yet normalized entropy, $\frac{H(X)}{\log n}$, decreases as n grows, since the numerator remains constant while the denominator grows. We illustrate the simplest case of the uniform distribution in Figure 2: X has domain size 2, such that for any x in X, $\frac{n_x}{n} = \frac{1}{2}$, regardless of sample size. For Shannon entropy, $H(X) = -\sum_x \frac{1}{2} \log \frac{1}{2} = -\log \frac{1}{2} = 1$. Yet for normalized entropy, $\tilde{H}(X) = \frac{H(X)}{\log n} = \frac{1}{\log n}$.

Example 3: Error grows with domain size. A problem for any fingerprinting measure is that sample estimates are subject to estimation error. This particularly affects long-tail distributions where the domain grows with sample size.

For this example, we assume a population of size $N = 1,000,000$ distributed by the power law. We compute the population entropy, H, and draw random subsamples of increasing size, n, up to N. For each n, we estimate the entropy, $\hat{H}_n$, compute relative error as $\frac{|H - \hat{H}_n|}{H}$, and take the average over 100 trials. We repeat this procedure for domain sizes 100, 1000, 10,000. Results are shown in Figure 3 which shows how for each population, relative error declines as the sample size, n, nears the full population. However, for any given n, the error is larger for Shannon and normalized entropy when values are distributed over larger domains. Yet this is not a problem for Tsallis entropy. Section 5 further describes this difference: estimation error for Shannon and normalized entropy is dependent on domain size, but this is not the case for Tsallis entropy. Section 6 serves to validate this claim through empirical analysis.

In contrast to Shannon entropy and normalized entropy, Tsallis entropy demonstrates robustness to changes in domain and sample size in all of these examples.

4.2 Empirical example

Here we use entropy data published in prior works to provide an example of how using Shannon and normalized entropy in fingerprinting research can lead to confusing findings, since these measures change with sample size. We consider a case study of the User agent attribute. The User agent is sent in HTTP request headers to inform a server about the type of browser requesting content, which helps the server deliver the most suitable web page

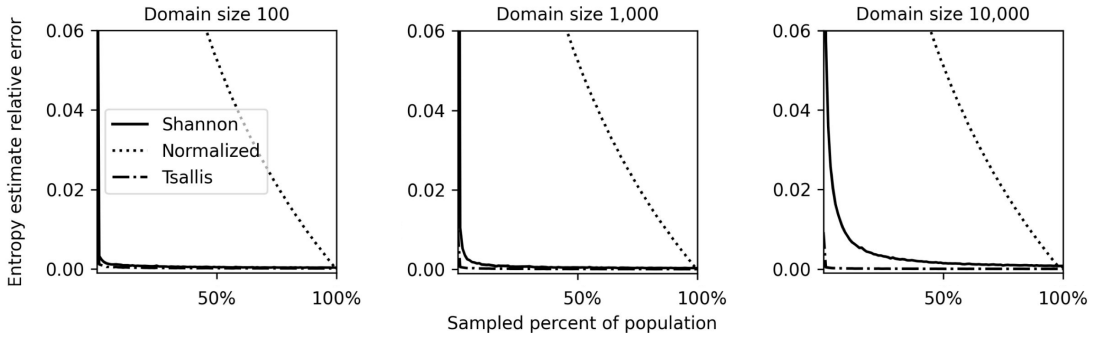

Figure 3: Relative error grows with domain size (example 3).

to the browser [18, 33]. Early fingerprinting research identified the User agent string as one of the highest entropy values contributing to fingerprinting risk [17, 21, 24, 29]. Major browsers have since made changes to reduce User agent entropy, such as by providing less detailed information in the User agent string (this is an example of putting into practice the *monotonicity* property, defined in Section 5.1). This effort is often referred to as "User-agent reduction" [7, 13, 34, 50]. We thus use User agent in our case study because we expect User agent entropy to decrease over time due to these efforts by browser developers. We also use User agent because it is the most consistently published browser attribute in prior work, contributing the data necessary for this case study.

4.2.1 Case study data. To compile the data used in this case study, we reviewed fingerprinting papers in web privacy and security related conferences and journals, where the authors published entropy metrics for browser attributes they studied. After determining User agent as the most consistently published attribute with entropy, we restricted the set of papers to those that included User agent entropy. To improve data reliability, we further excluded papers where the authors were not involved in the data collection and borrowed data from other sources. The resulting User agent entropy data is compiled in Table 1 with the year of publication and sample size. These include data from Eckersley (2010) [17], Laperdrix et al. (2016) [29], Cao et al. (2017) [9], Gómez-Boix et al. (2018) [21], Chalise et al. (2022) [10], Lin et al. (2023) [30], and Berke et al. (2025) [4]. Some of these papers only published either Shannon or normalized entropy, from which we then calculated normalized or Shannon entropy by dividing or multiplying by $\log(n)$, respectively.

4.2.2 Case study result. Figure 4 shows the Shannon and normalized entropy reported by each paper, plotted by publication year. As we would expect due to User-agent reduction efforts, Shannon entropy consistently decreases over time, demonstrating privacy enhancing improvements due to fingerprinting research and changes by browser developers. However, this downward trend in Shannon entropy values is noisy. This can be due to many factors, including differences in the dataset sample sizes. As described in Section 4.1, Shannon entropy often grows with sample size due to its dependence on domain size. This problem could explain the slight increase between the 2023 (n=1,848) and 2025 (n=8400) data. Hence, this data could be less noisy and easier to evaluate if Shannon entropy were easier to estimate and less dependent on sample size.

Table 1: User agent entropy published in prior work, ordered by publication date.

Publication	Year	Sample size (n)	User agent entropy	
			Shannon	Normalized
[17]	2010	470,161	10.0	0.531
[29]	2016	118,934	9.779	0.570
[9]	2017	3,615	7.234	0.612
[21]	2018	2,067,942	7.150	0.341
[10]	2022	2,093	6.466	0.586
[30]	2023	1,848	4.449	0.41
[4]	2025	8,400	4.613	0.354

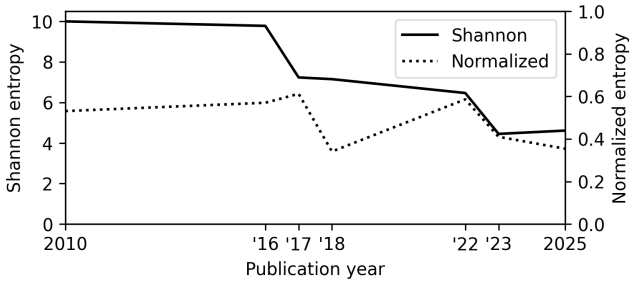

Figure 4: Shannon and normalized entropy reported in prior papers, by publication date.

On the other hand, normalized entropy, which was designed for the purpose of comparing datasets, shows no trend at all. This is likely due to how normalized entropy is also dependent on sample size and yields large estimation errors (Figure 3).

In contrast, Tsallis entropy can be estimated with error independent of sample size, which may contribute to better comparisons of future datasets and clearer evaluations of trends. This is demonstrated in Sections 5 and 6.

5 Analytical evaluation

In this section we propose a set of mathematical properties that an entropy measure ought to satisfy. We select the properties based

on the principle that an entropy measure should be useful for comparing samples, ignoring differences between the samples that are irrelevant to fingerprinting risk, while being sensitive to differences that may indicate increased fingerprinting risk.

We do not claim that our proposed properties completely characterize a good fingerprinting risk measure. Rather, we argue that each property is basic enough that any reasonable fingerprinting risk measure should satisfy it. We then show the entropies that are most commonly used to measure fingerprinting risk, Shannon entropy and normalized entropy, do not meet even this low bar, while Tsallis entropy does.

Beyond satisfying these formal mathematical properties, we argue that an entropy measure should also be interpretable, providing a meaningful relationship to anonymity, such that its numerical value offers an easy-to-understand quantification of the fingerprinting risk of the users in the sample. We provide mathematical analysis supporting the claim that Tsallis entropy is more interpretable than both Shannon and normalized entropy.

5.1 Mathematical properties

In the definitions of each of the properties below, we use ϕ to denote an entropy measure.

Our first property states that the entropy measure should be invariant to sample size changes. A sample X' is a *scaling* of sample X if there exists a $c > 0$ such that $c \cdot n_x = n'_x$ for all values x, where n_x and n'_x are the number of users in samples X and X' (respectively) who have value x. In other words, a scaling changes the size of a sample without changing its distribution of values. If two samples are identical up to scaling, then they contain no indication that the underlying populations from which they are drawn differ in any way, so their entropy values should be the same.

Property #1: Scale-invariance. If X' is a scaling of X then $\phi(X) = \phi(X')$.

Our next property states that the entropy measure should increase after changes to the sample that clearly increases fingerprinting risk. A *refinement* of a sample X is a sample X' where any pair of users with the same value in X' have the same value in X, but the converse is not true for at least one pair of users. This can occur under a few scenarios. For example, when an attribute is updated to yield new values, or when a fingerprint is constructed by combining attributes together and a new attribute is added to the fingerprint. In either case, some users who previously were indistinguishable may now have different fingerprints.

More formally, a sample $X' = (x'_1, \ldots, x'_n)$ is a refinement of sample $X = (x_1, \ldots, x_n)$ if $x'_i = x'_j$ implies $x_i = x_j$ for all $i, j \in [n]$ and there exists $i, j \in [n]$ such that $x'_i \neq x'_j$ and $x_i = x_j$.

Property #2: Monotonicity. If X' is a refinement of X then $\phi(X) < \phi(X')$.

Our final property is motivated by practical considerations. We want to accurately estimate the entropy of a population from a relatively small random sample of its users. More precisely, we want an algorithm A that inputs a sample X, drawn from a population P, and that outputs an estimate of $\phi(P)$. The expected error of the estimate should decrease as the size of X increases.

Property #3: Estimability. There is an algorithm A and a decreasing function $f : \mathbb{N} \to \mathbb{R}$ with $\lim_{n\to\infty} f(n) = 0$ such that for any population P we have $\mathrm{E}[|A(X) - \phi(P)|] \leq f(n)$, where $X = (x_1, \ldots, x_n)$ is a sample drawn uniformly at random without replacement from P, and $\mathrm{E}[\cdot]$ denotes expected value.

When the estimability property holds, if a data analyst wants to estimate $\phi(P)$ to within expected error ε, it suffices for her to collect a sample of size n that satisfies $f(n) \leq \varepsilon$. She does not need to know anything about P.

5.2 Analysis of properties

We now analyze which of the above properties are satisfied by the entropy measures defined in Section 3. Our first theorem states that Shannon entropy and Tsallis entropy are scale-invariant, but normalized entropy is not. Essentially, because normalized entropy is Shannon entropy divided by $\log n$, where n is the size of the sample, it cannot always be insensitive to sample size.

Theorem 5.1. *Shannon entropy and Tsallis entropy satisfy scale-invariance. Normalized entropy does not satisfy scale-invariance.*

Proof. From Definitions 3.1 and 3.3, we see that Shannon entropy and Tsallis entropy depend only on $p_x = \frac{n_x}{n}$ for all values x, *i.e.*, the distribution of values in sample X. The p_x's are clearly scale-invariant, and thus so are Shannon entropy and Tsallis entropy.

Now to analyze normalized entropy, $\bar{H}(X)$, let X and X' be samples with sizes n and m respectively, where $n < m$, and assume that X' is a scaling of X. We have

$$\bar{H}(X) = \frac{H(X)}{\log n} = \frac{H(X')}{\log n} > \frac{H(X')}{\log m} = \bar{H}(X')$$

where the second equality follows from the scale-invariance of Shannon entropy. Thus normalized entropy does not satisfy scale-invariance. □

Our next theorem states that all of the entropy measures satisfy monotonicity.

Theorem 5.2. *Shannon entropy, normalized entropy and Tsallis entropy all satisfy monotonicity.*

Proof. Let $X = (x_1, \ldots, x_n)$ and $X' = (x'_1, \ldots, x'_n)$ be samples such that X' is a refinement of X. For any value x let

$$V_x = \bigcup_{i:\, x_i = x} \{x'_i\}$$

be the set of values belonging to the users in X' who have value x in X. Observe that the V_x's are a partition of the values in X', *i.e.*, every value in X' is contained in exactly one set V_x. This holds because, by the definition of refinement, if two users have the same value in X', then they have the same value in X. Also observe that $n_x = \sum_{y \in V_x} n'_y$ for each value x, where $n'_y = |\{i : x'_i = y\}|$ is the number of users in sample X' who have value y.

Consider a strictly subadditive function $f : [0, 1] \to \mathbb{R}$, *i.e.*, we have $f(a + b) \leq f(a) + f(b)$ for all $a, b \in [0, 1]$ and $f(a + b) < f(a) + f(b)$ for all $a, b \in (0, 1]$. Let $p_x = \frac{n_x}{n}$ and $p'_x = \frac{n'_x}{n}$ for each

$x \in X$. For any value x, by subadditivity we have

$$f(p_x) \le \sum_{y \in V_x} f(p'_y),$$

and further if $|V_x| > 1$ then by strict subadditivity we have

$$f(p_x) < \sum_{y \in V_x} f(p'_y).$$

By the definition of refinement, there must exist at least one value x such that $|V_x| > 1$. Fix any $c \in \mathbb{R}$ and define $g(X) = c + \sum_{x \in X} f(p_x)$. We have

$$\begin{aligned} g(X) &= c + \sum_x f(p_x) \\ &< c + \sum_x \sum_{y \in V_x} f(p'_y) \\ &= c + \sum_x f(p'_x) \\ &= g(X') \end{aligned}$$

where the inequality uses the (strict) subadditivity of f, and the second equality uses the fact that the V_x's are a partition of the values in X'.

Now observe that if $c = 0$ and $f(z) = -z \log z$ then $g(X) = H(X)$. Also, if $c = 1$ and $f(z) = -z^2$ then $g(X) = T(X)$. In either case, we have that f is strictly concave and $f(0) = 0$, which sufficies to show that f is strictly subadditive. Therefore Shannon entropy and Tsallis entropy satisfy monotonicity. Finally

$$\bar{H}(X) = \frac{H(X)}{\log n} < \frac{H(X')}{\log n} = \bar{H}(X')$$

where the inequality follows from the monotonicity of Shannon entropy. Thus normalized entropy satisfies monotonicity. □

Our final theorem states that Tsallis entropy is estimable, but Shannon and normalized entropy are not. Estimability requires that the estimation error is upper bounded by a function $f(n)$, where n is the sample size. For Shannon entropy, we show that estimation error is lower bounded by a function $f(n, k)$, where k is the domain size of the underlying population P. The existence of a lower bound that depends on k rules out the possibility of an upper bound that is independent of k. If k is unknown (and it often is, since we seldom observe the entire population), we cannot know how many samples we need to collect to achieve a desired error. This is the case for normalized entropy as well, since it is defined by Shannon entropy. Thus Shannon and normalized entropy, unlike Tsallis entropy, are impractical to estimate.

Theorem 5.3. *Tsallis entropy satisfies estimability. Shannon entropy and normalized entropy do not satisfy estimability.*

Proof. First we prove that Tsallis entropy is estimable. Let $P = (y_1, \ldots, y_N)$ be a population of N users, where y_i is the value for user i in the population. Let $N_x = |\{i : y_i = x\}|$ be the number of users in population P with value x. Clearly $N = \sum_x N_x$. Let $X = (x_1, \ldots, x_n)$ be a sample drawn uniformly at random without replacement from P. Assume for simplicity that n is even. For each $i \in [n/2]$ let $Z_i = \mathbf{1}\{x_{2i} = x_{2i-1}\}$ be a Boolean variable indicating whether the ith consecutive pair of values in X are equal. We have

$$\begin{aligned} \mathrm{E}[Z_i] &= \Pr[x_{2i} = x_{2i-1}] \\ &= \sum_x \Pr[x_{2i-1} = x] \Pr[x_{2i} = x \mid x_{2i-1} = x] \\ &= \sum_x \frac{N_x}{N} \frac{N_x - 1}{N - 1} \\ &= \frac{N}{N-1} \sum_x \left(\frac{N_x}{N}\right)^2 - \sum_x \frac{N_x}{N(N-1)} \\ &= \frac{N}{N-1} \sum_x \left(\frac{N_x}{N}\right)^2 - \frac{1}{N-1} \\ &= 1 - \frac{N}{N-1} T(P) \end{aligned}$$

where the third equality uses Lemma B.1 in Appendix B. Consider an estimation algorithm A that, given sample X as input, outputs the estimate $\hat{T} = \frac{N-1}{N}\left(1 - \frac{2}{n}\sum_{i \in [n/2]} Z_i\right)$. Clearly

$$\mathrm{E}[\hat{T}] = \frac{N-1}{N}\left(1 - \frac{2}{n}\sum_{i \in [n/2]} \mathrm{E}[Z_i]\right) = T(P)$$

and $|\hat{T} - \mathrm{E}[\hat{T}]| \le 1$. Therefore for any $\varepsilon > 0$

$$\begin{aligned} \mathrm{E}[|A(X) - T(P)|] &= \mathrm{E}[|\hat{T} - \mathrm{E}[\hat{T}]|] \\ &\le |\hat{T} - \mathrm{E}[\hat{T}]| \cdot \Pr[|\hat{T} - \mathrm{E}[\hat{T}]| \ge \varepsilon] \\ &\quad + \varepsilon \cdot \Pr[|\hat{T} - \mathrm{E}[\hat{T}]| \le \varepsilon] \\ &\le \Pr[|\hat{T} - \mathrm{E}[\hat{T}]| \ge \varepsilon] + \varepsilon \\ &\le 2\exp(-n\varepsilon^2/4) + \varepsilon \end{aligned}$$

where the last inequality follows from Lemma B.2 in Appendix B (with $t = \frac{N}{N-1}\frac{n}{2}\varepsilon$). Plugging $\varepsilon = \sqrt{\frac{2\log n}{n}}$ into the above inequality we have

$$\mathrm{E}[|A(X) - T(P)|] \le \frac{2 + \sqrt{2\log n}}{\sqrt{n}}.$$

Observe that the right-hand side is decreasing in n and has a limit of 0 as $n \to \infty$, which shows that Tsallis entropy satisfies estimability.

Finally, we prove that Shannon entropy and normalized entropy do not satisfy estimability. Valiant and Valiant [47] showed that there exist positive constants c and k_0 such that for any $k \ge k_0$ and any estimation algorithm A it holds that

$$\mathrm{E}[|A(X) - H(P)|] \ge c \cdot \frac{k}{n \log k}$$

for some population P with domain size k, where X is a random sample of size n drawn from P. This lower bound on estimation error, which depends on both n and k, refutes the existence of an upper bound that depends only on n. Thus Shannon entropy is not estimable. By the same argument we have

$$\mathrm{E}[A(X) - \bar{H}(P)] \ge c \cdot \frac{k}{n \log n \log k},$$

since Shannon entropy and normalized entropy differ only by a factor $\log n$. Thus normalized entropy is not estimable. □

5.3 Interpretability

A major drawback of common entropy measures is that they lack a clear relationship to the anonymity of the users in a sample. Other researchers have helped address this by proposing alternative

models that link anonymity to entropy [40]. Here we focus on improving interpretability when using entropy, which is a problem for common entropy measures except in special cases. For example, if the values in sample X are uniformly distributed, and the Shannon entropy of X is b, then each user in X shares her value with $\frac{n}{2^b}$ users in X, where n is the sample size. However, this relationship breaks down when the values in X are not uniformly distributed.

An inconvenience of Shannon entropy is that its maximum value is $\log n$, so it is difficult to know whether the Shannon entropy of a sample is small or large without comparing it to the sample size. Normalized entropy was introduced in an attempt to remedy this problem. Since it is always contained in the range $[0, 1]$, it appears easier to determine how close the entropy of a sample is to its minimum and maximum values. However, since normalized entropy is not scale-invariant (Theorem 5.1), this value is impacted by sample size, again making it difficult to know if the value should be considered small or large. Furthermore, like Shannon entropy, normalized entropy has no intuitive interpretation when values are not uniformly distributed.

We argue that Tsallis entropy is more interpretable than Shannon entropy and normalized entropy. Like normalized entropy, Tsallis entropy is always contained in the interval $[0, 1]$. Tsallis entropy also has a much closer connection to fingerprinting risk. The next theorem states that the Tsallis entropy of X is the probability that two users chosen independently and uniformly at random from X have different values, which means that higher entropy indicates a higher likelihood that their values are identifying.

Theorem 5.4. *Let $X = (x_1, \ldots, x_n)$ be a sample. Let users $I, J \in [n]$ be chosen independently and uniformly at random. $T(X) = \Pr[x_I \neq x_J]$.*

Proof. Recall that $n_x = |\{i : x_i = x\}|$ is the number of users with value x. We have

$$\begin{aligned}\Pr[x_I = x_J] &= \sum_x \Pr[x_I = x \mid x_J = x] \Pr[x_J = x] \\ &= \sum_x \Pr[x_I = x] \Pr[x_J = x] \\ &= \sum_x \frac{n_x}{n} \frac{n_x}{n} \\ &= \sum_x \left(\frac{n_x}{n}\right)^2\end{aligned}$$

Note that $\Pr[x_I \neq x_J] = 1 - \Pr[x_I = x_J]$ and $T(X) = 1 - \sum_x \left(\frac{n_x}{n}\right)^2$, which completes the proof. □

An even more direct connection between Tsallis entropy and anonymity is established by the following theorem, which states that the Tsallis entropy of X is inversely related to the average size of the anonymity groups of the users in X. Anonymity group sizes have been used in addition to entropy to quantify fingerprinting risk [17, 28]. Thus researchers interested in the fingerprinting risk faced by the average user will be interested in calculating Tsallis entropy.

Theorem 5.5. *Let $X = (x_1, \ldots, x_n)$ be a sample. Let a_i be the fraction of users in X that share their value with user i. $T(X) = 1 - \frac{1}{n}\sum_i a_i$.*

Proof. Let $m_i = |\{j : x_i = x_j\}|$ be the number of users who share their value with user i. We have

$$\begin{aligned}\frac{1}{n}\sum_i a_i &= \frac{1}{n}\sum_i \frac{m_i}{n} \\ &= \frac{1}{n}\sum_x \sum_{i:\, x_i = x} \frac{m_i}{n} \\ &= \frac{1}{n}\sum_x \sum_{i:\, x_i = x} \frac{n_x}{n} \\ &= \frac{1}{n}\sum_x \frac{n_x^2}{n} \\ &= \sum_x \left(\frac{n_x}{n}\right)^2\end{aligned}$$

Noting that $T(X) = 1 - \sum_x \left(\frac{n_x}{n}\right)^2$ completes the proof. □

The proofs of Theorems 5.4 and 5.5 depend on the fact that an exponent of 2 is used in the formula for Tsallis entropy (see Definition 3.3), which was our motivation for making that choice in the definition.

6 Empirical evaluation

6.1 Dataset

The empirical analysis uses an open dataset of web browser attributes that are commonly used for fingerprinting [4]. The data was collected in December 2023 from N=8400 US browser users with their informed consent. See Appendix Tables 3-4 and [4] for details on the devices, users, and data collection, respectively. These browser attributes were previously collected and analyzed in prior fingerprinting studies [2, 17, 21, 28, 29] which used Shannon entropy or normalized entropy to quantify the associated fingerprinting risks. They are also used in practice by the prevailing fingerprinting library, FingerprintJS [19]. Table 2 shows the browser attributes with the number of distinct values (domain size), percent of users with a unique value, and the Shannon, normalized, and Tsallis entropy, computed from the dataset. See Appendix Figure 11 for a visual comparison between the attributes' uniqueness and entropy measures. The overall "Fingerprint" for each user is the hashed concatenation of the other attribute values, computed similarly to FingerprintJS. Note that the ordering of attributes according to their entropy is largely invariant to the choice of entropy definition.

6.2 Empirical results

We empirically evaluate how easily the entropy measures can be estimated from a small sample, and how the estimates vary with sample size. We consider the full dataset ($N = 8400$) as the population, compute entropy for the population (shown in Table 2), and estimate entropy for smaller subsamples of size n ranging up to N. For each n in {200, 400, 600, ...8400}, we use repeated random sampling (without replacement) to compute entropy for the subsample, averaged over 100 trials. We compute the average relative error for each n, as the difference between the population entropy, $H(X_N)$, and the subsample entropy estimate: $\frac{|H(X_N) - H(\tilde{X}_n)|}{H(X_N)}$.

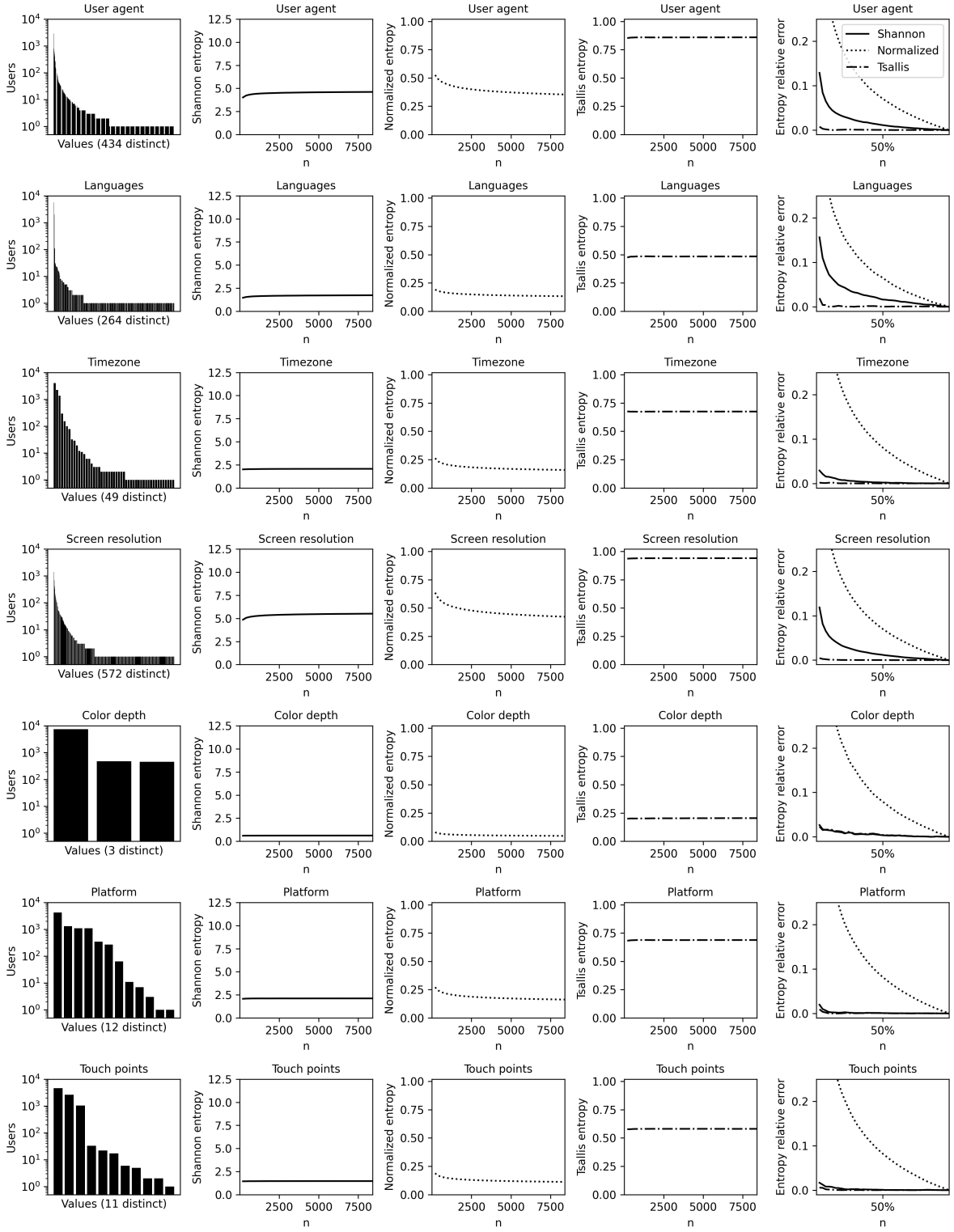

Figure 5: Attribute value distributions and entropy estimates for all attributes shown in Table 1. Left: Histograms show the (logscale) number of users for the given attribute values. Middle: Plots show the mean entropy values estimated over 100 trials for subsamples ranging from n=200 to 8400. Right: The entropy estimate sample error computed as the difference between the mean estimate for the subsample versus population (N=8400).

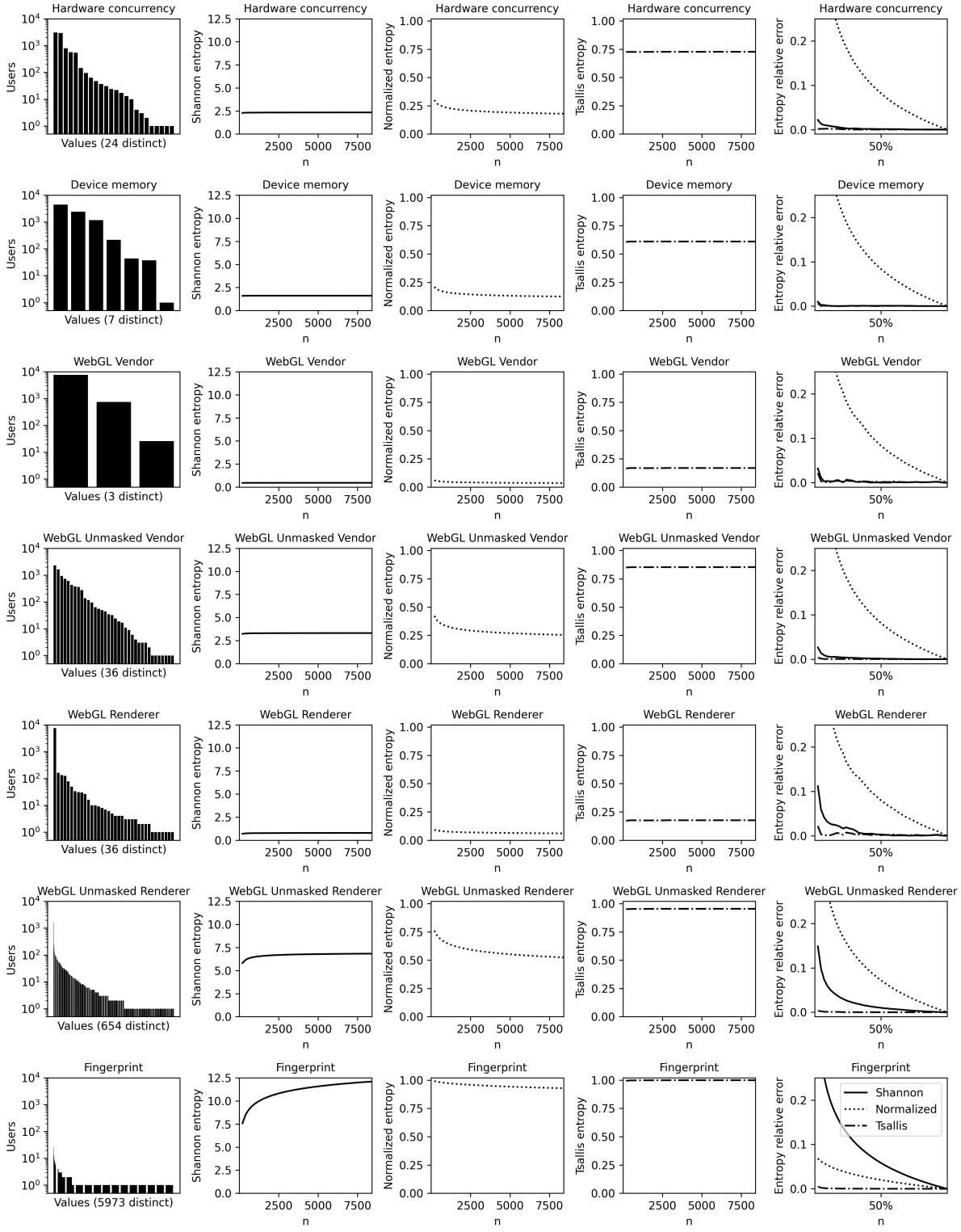

Figure 6: Figure 5 continued. Shannon entropy increases, and normalized entropy decreases, with sample size, especially for attributes with more distinct values, while Tsallis entropy is constant. Estimation error is also lowest for Tsallis entropy.

Attribute	Distinct values	% Unique	Shannon entropy	Norm. entropy	Tsallis entropy	Example value (most frequent)
User agent	434	2.8	4.613	0.354	0.859	Mozilla/5.0 (Windows NT 10.0; Win64; x64) Apple...
Languages	264	2.4	1.730	0.133	0.484	en-US,en
Timezone	49	0.2	2.064	0.158	0.674	AmericaNew_York
Screen resolution	572	4.5	5.510	0.423	0.940	[1920,1080]
Color depth	3	0.0	0.616	0.047	0.204	24
Platform	12	0.0	2.114	0.162	0.688	Win32
Touch points	11	0.0	1.463	0.112	0.581	0
Hardware concurrency	24	0.1	2.340	0.180	0.727	4.0
Device memory	7	0.0	1.611	0.124	0.610	8.0
WebGL Vendor	3	0.0	0.465	0.036	0.169	WebKit
WebGL Unmasked Vendor	36	0.1	3.313	0.254	0.853	Google Inc. (Intel)
WebGL Renderer	36	0.1	0.782	0.060	0.176	WebKit WebGL
WebGL Unmasked Renderer	654	3.2	6.833	0.524	0.954	Apple GPU
Fingerprint	5973	60.2	12.101	0.928	0.999	3226705957305930625

Table 2: Browser attributes and fingerprinting measures (N=8400).

Results are shown in Figure 5. (Note these results use the plug-in estimator for Shannon entropy, while Section 6.3 evaluates advanced entropy measures). To make the results more intuitive, we also illustrate the distribution of values for each attribute, shown as a histogram with a logscale y-axis (number of users). These results are consistent with the examples and analysis in Sections 4 and 5. For attributes with small domains (e.g. Platform and Hardware concurrency), Shannon entropy is relatively consistent as n increases. Yet for attributes with long-tail distributions, as the sample size, and hence domain size, increases Shannon entropy increases. (Section 4 Example 1). For normalized entropy, we can see the problem described in Example 2: estimated entropy decreases as sample size increases, and this is the case for each of the attributes. In contrast, the Tsallis entropy estimates are more consistent regardless of sample size. The relative error plots in Figure 5 illustrate why these consistencies are important: larger differences in estimates results in larger error. These plots report on the *estimability* property from Section 5. Normalized entropy has the highest error for all attributes, except for the combined fingerprint, where more than 60% of the values in the sample are unique (Table 2), and where Shannon entropy estimates have higher error. This aligns with how normalized entropy is designed to handle the case of maximum entropy, i.e. unique values. For all of the attributes, Tsallis entropy maintains the lowest relative error, empirically demonstrating its superior estimability which was analytically described in Section 5.

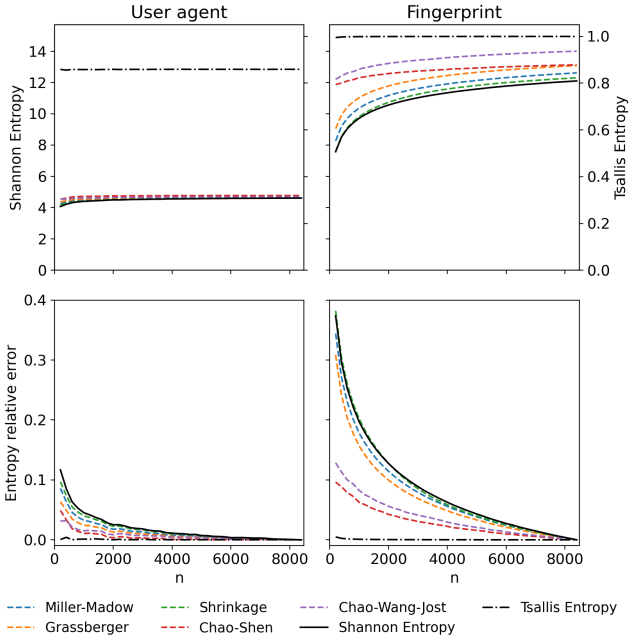

Figure 7: Analysis from Figure 5 repeated using advanced Shannon entropy estimators, with comparisons to Tsallis entropy. Values are computed over increasing sample sizes, averaged over 100 trials. Left: User agent. Right: Overall "Fingerprint". Top: Variation in mean entropy values. Bottom: Variation in relative error, computed as the difference between the mean estimate for the subsample versus population (N=8400).

6.3 Comparing Tsallis to advanced Shannon entropy measures

Prior research has addressed some of the limitations of the plug-in estimator for Shannon entropy. While every estimator of Shannon entropy is biased [39], the Miller-Madow [35] and Grassberger [20] estimators have lower bias than the plug-in estimator, at the expense of higher variance. Alternatively, the variance of the plug-in estimator can be reduced via regularization, for example by employing the James-Stein technique and shrinking the plug-in estimate towards the uniform distribution [22]. Other methods use Good-Turing frequency estimation to account for missing values in the sample, such as the Chao-Shen [11] and Chao-Wang-Jost [12] estimators. An important question is whether one of these advanced estimators for Shannon entropy can better serve as a fingerprinting risk measure than the plug-in estimator. To answer this question, we repeat the analysis from Section 6.2 using these advanced estimators with the

User agent and overall "Fingerprint". (We compute these entropy estimators by leveraging the infomeasure open source library [8].) Namely, we compute the entropy estimates for increasingly large sample sizes, averaged over 100 repeated random sampling trials, and then plot the resulting estimates and average relative error. This is shown in Figure 7, with comparisons to the plug-in estimators for Shannon entropy and Tsallis entropy.

These results show the limitations of Shannon entropy persist. Tsallis entropy still maintains more consistent values and smaller relative error compared to corrected estimators, across increasing sample sizes. Our analytical analysis (Section 5) identifies the reason: Shannon entropy (and its estimators) is dependent on domain size, which can grow with sample size, while Tsallis entropy is not.

7 Limitations of Tsallis Entropy

A fingerprinting risk measure is unlikely to accurately reflect the risk faced by every user in a population, as each user will have a different fingerprint value and thus experience a different level of anonymity. Tsallis entropy is equivalent to the average anonymity across all users (Theorem 5.5). But one might instead prefer to focus on users with the least possible anonymity. A user is said to be *unique* if they share their fingerprint value with no one else in the population.

Uniqueness is sensitive to small changes in the distribution of fingerprint values in a population, as a user can go from unique to non-unique by changing a single value. Therefore we might expect fingerprinting risk measures that can detect small changes in a distribution to be better correlated with uniqueness. From this perspective, Tsallis entropy is worse than Shannon entropy, a limitation that can be established both formally and empirically.

Let the *total variation distance* between samples X and X' be defined

$$d_{\mathrm{TV}}(X, X') = \frac{1}{2} \sum_{x} \left| \frac{n_x}{n} - \frac{n'_x}{n'} \right|.$$

In other words, total variation distance is the sum of the differences between the frequencies of each value in the two samples (divided by 2 to compensate for double counting). The next theorem shows that a small change to the distribution of values in a sample can only lead to a small change to its Tsallis entropy, but may produce an arbitrarily large change to its Shannon entropy. The proof is in Appendix C.

Theorem 7.1. *Let $\varepsilon > 0$. Let X and X' be samples satisfying $d_{TV}(X, X') \leq \varepsilon$. Then $|T(X) - T(X')| \leq 6\varepsilon$, but for any $\delta > 0$ it is possible that $|H(X) - H(X')| \geq \delta$.*

Empirically, we also find that both Shannon entropy and normalized entropy are better correlated with uniqueness than Tsallis entropy by using data from Table 2 (see Appendix F)

Previous work by Rocher et al. [40] developed a sophisticated Bayesian method for estimating the percent of a population with a unique value. In Appendix F we show that their method is better at estimating uniqueness than either Tsallis or Shannon entropy. Ultimately, the choice of estimator will depend on which aspect of fingerprinting risk a researcher is most interested in estimating: how much attributes contribute to the average user's fingerprinting risk, or the fraction of users facing the worst-case risk. For example, focusing exclusively on uniqueness may underestimate the risks faced by users in small anonymity groups of size larger than one (see Appendix Figure 13).

8 Discussion

This work identifies and addresses an overlooked problem in the browser fingerprinting literature, namely the persistent use of normalized entropy to evaluate and compare browser fingerprinting data. Normalized entropy was designed to compare samples of differing sizes, yet this work shows how it is actually dependent on sample size and ill-suited for this task. We show this both analytically (Section 5) and empirically (Section 6). We also provide an empirical case study in Section 4.2 (Figure 4) showing why this is a problem – normalized entropy can lead to confusing findings when evaluating privacy-enhancing browser changes over time. In particular, in Section 4.2 we evaluate entropy measurements published for the User agent attribute in prior works over time, where we expect entropy to decrease due to "User-agent reduction" efforts [7, 13, 34, 50] by browser developers (corresponding to the *monotonicity* property defined in Section 5.1). We show how normalized entropy measurements do not reflect these changes due to how the prior measurements used samples of varying sizes, and normalized entropy does not satisfy the *scale-invariance* property.

The fact that researchers persistently use normalized entropy to describe their data and compare samples, despite its limitations, indicates the research community needs a better measure to quantify fingerprinting risk and make comparisons across samples.

8.1 Tsallis entropy can improve fingerprinting risk measurement

To address the research community's needs, we propose the use of Tsallis entropy. Tsallis entropy has long been used in other scientific fields.In ecology, it is a metric for species diversity [44]. In political science, it quantifies the effective number of parties contesting an election [27]. In economics, it is a measure of market competition among firms [23]. To justify the use of Tsallis entropy for measuring fingerprinting risk, we described desired properties for a fingerprinting risk measure, and showed how Tsallis entropy satisfies these properties, while the commonly used measures of Shannon entropy and normalized entropy do not. Beyond our mathematical analyses, we argue that a good fingerprinting risk measure should be interpretable. The way it is defined should communicate real fingerprinting risk for users, and Section 5 describes how Tsallis entropy provides this benefit over the other measures.

Although we are proposing a shift in how researchers measure fingerprinting risk, a shift towards Tsallis entropy need not be disruptive. Table 2 shows how the relative ordering of the various entropy measures, computed over the empirical data, are largely consistent. Attributes with higher Tsallis entropy tend to have higher Shannon entropy and normalized entropy. The web browser development community commonly refers to efforts to reduce "entropy", to improve user privacy, without specifying the type of entropy [48]. Tsallis entropy can also be referred to as "entropy" and ongoing efforts to reduce entropy will reduce Tsallis entropy as well.

8.2 Establishing a metric

Given the prevalence and problems with Shannon entropy and normalized entropy in the fingerprinting literature, it is worth asking: How did we get here? The answer may be that the authors who introduced these previous measures had varying goals, and the authors who later adopted these measures simply needed a measure to evaluate their work, without the time to evaluate the measure. For example, the early fingerprinting research aimed to demonstrate a problem with Web 2.0 features, namely that the features that made the web a richer experience could also be abused to identify and track users. To this end, both Mayer and Eckersley quantified how many browsers in their datasets could be uniquely identified, and then described the difficulty of scaling their methods to measure the global uniqueness of browsers [17, 32]. In Eckersley's canonical Panopticlick study, an additional goal was to compare how various attributes contributed to uniqueness. For this purpose, Eckersley introduced Shannon entropy and compared attributes within a single dataset; the fact that Shannon entropy can grow with dataset size was not an issue. Normalized entropy was introduced to contend with this issue because the authors of the AmIUnique study wanted to compare their dataset to the Panopticlick dataset [29]. The AmIUnique study authors may not have had access to the raw Panopticlick data, which would have made such a comparison possible. What they did have was the published Shannon entropy measures and the total dataset size. Given that normalized entropy is defined by only Shannon entropy and total dataset size, it is possible normalized entropy was then defined out of convenience. While the authors provided their motivation and goals for normalized entropy, they did not provide an evaluation of whether normalized entropy satisfies these goals.

Our work begins to rectify this issue. Similar to prior work in the Panopticlick and AmIUnique studies, we provided motivation and goals for a fingerprinting measure, and introduced a new measure (Tsallis entropy). Unlike prior work, the measure we provided is interpretable with respect to unique identification risk and we formally evaluated whether the measure satisfies the goals. The goals were described and evaluated as mathematical properties, which we consider a minimal set that future work can build upon.

8.3 Generalizability

This work focuses on measuring browser fingerprinting risk, yet our contributions can extend to the study of fingerprinting more broadly. Device fingerprinting strategies collect attributes from users' devices in order to identify and track device users across both websites and web browsers [9] and the mobile applications they use [26, 38]. Although our empirical analysis leverages browser fingerprinting data, our analytical analyses and discussion of interpretability apply to measuring fingerprinting risks for device attributes more generally.

8.4 Limitations and future work

While our work shows how Tsallis entropy can improve fingerprinting risk measurements, we do not claim that it is the optimal measure, and we hope future work will develop even better measures. For example, a limitation of our analytical analysis (Section 5) is that the set of desired properties is small. This set of properties could be expanded to identify the best measure, in much the same way that Shannon entropy has been axiomatically characterized as the best measure of information [25]. A specific limitation of Tsallis entropy is that it is not well-suited to estimate uniqueness, as discussed in Section 7.

An additional limitation of this work is that our empirical analysis uses a relatively small dataset. Future work would benefit from more openly published data.

Overall, this work aims to encourage the research community to align on better metrics, by first articulating goals for a metric and then demonstrating how a metric satisfies the goals. This work presents a step in that direction, with a higher level goal of improving fingerprinting risk measurement, to help improve user privacy.

9 Conclusion

In this work we identify an oversight and resulting problem in the fingerprinting risk measurement literature, and propose a solution. We show that although Shannon entropy and normalized entropy have been adopted to measure fingerprinting risk in numerous works, they are ill-suited for this purpose due to their dependence on sample size and domain size. In particular, normalized entropy was designed with the explicit purpose to make comparisons across datasets, and we demonstrate how it does not serve this purpose through both analytical and empirical analyses. In order to address the need for a fingerprinting measure that can make better comparisons across datasets, we first define a minimal set of desired properties for such a measure. We then show that better measures are available, such as Tsallis entropy, by demonstrating how Tsallis entropy satisfies the desired properties while Shannon entropy and normalized entropy do not. Our demonstrations include mathematical proofs, empirical analyses with real data, and a discussion of interpretability. We intend for our contributions to improve future fingerprinting risk research, and especially to spur the development of better measures of fingerprinting risk, in order to help facilitate future privacy enhancing developments and assess their impact.

Code availability

The data and analysis code for this work are available in an open repository: https://github.com/google/fingerprinting-risk-measures

Ethics

This work leverages an openly published dataset of fingerprinting signals collected from US web users' browsers. The data was published for anonymous users yet could potentially be used to reidentify their browsers. However, the data was collected with users' informed consent, as described in [4], this work complies with the dataset terms of use, and this work does not raise any new ethical issues for the users.

Acknowledgments

All authors conducted this research as employees at Google. We thank Robin Lassonde and Hamzeh Zawawy for their support.

References

[1] Nampoina Andriamilanto, Tristan Allard, and Gaëtan Le Guelvouit. 2021. "Guess Who?" Large-Scale Data-Centric Study of the Adequacy of Browser Fingerprints for Web Authentication. In *Innovative Mobile and Internet Services in Ubiquitous Computing* (Cham), Leonard Barolli, Aneta Poniszewska-Maranda, and Hyunhee Park (Eds.). Springer International Publishing, 161–172. https://doi.org/10.1007/978-3-030-50399-4_16

[2] Nampoina Andriamilanto, Tristan Allard, Gaëtan Le Guelvouit, and Alexandre Garel. 2021. A Large-scale Empirical Analysis of Browser Fingerprints Properties for Web Authentication. *ACM Transactions on the Web* 16, 1 (2021), 4:1–4:62. https://doi.org/10.1145/3478026

[3] Enrico Bacis, Igor Bilogrevic, Robert Busa-Fekete, Asanka Herath, Antonio Sartori, and Umar Syed. 2024. Assessing Web Fingerprinting Risk. In *Companion Proceedings of the ACM on Web Conference 2024*. 245–254.

[4] Alex Berke, Badih Ghazi, Enrico Bacis, Pritish Kamath, Ravi Kumar, Robin Lassonde, Pasin Manurangsi, and Umar Syed. 2025. How Unique is Whose Web Browser? The role of demographics in browser fingerprinting among US users. *Proceedings on Privacy Enhancing Technologies* (2025). https://doi.org/10.56553/popets-2025-0038

[5] Hristo Bojinov, Yan Michalevsky, Gabi Nakibly, and Dan Boneh. 2014. Mobile device identification via sensor fingerprinting. *arXiv preprint arXiv:1408.1416* (2014).

[6] Soumaya Boussaha, Lukas Hock, Miguel Bermejo, Ruben Cuevas Rumin, Angel Cuevas Rumin, David Klein, Martin Johns, Luca Compagna, Daniele Antonioli, and Thomas Barber. 2024. FP-tracer: Fine-grained Browser Fingerprinting Detection via Taint-tracking and Entropy-based Thresholds. *Proceedings on Privacy Enhancing Technologies* (2024).

[7] Brave Browser. 2022. Enable reduce-user-agent-minor-version by default (Issue #23938). GitHub Issue. https://github.com/brave/brave-browser/issues/23938 Accessed: 2025-02-11.

[8] Carlson Moses Büth, Kishor Acharya, and Massimiliano Zanin. 2025. infomeasure: a comprehensive Python package for information theory measures and estimators. *Scientific Reports* 15, 1 (2025), 29323. https://doi.org/10.1038/s41598-025-14053-5

[9] Yinzhi Cao, Song Li, and Erik Wijmans. 2017. (Cross-)Browser Fingerprinting via OS and Hardware Level Features. In *Proceedings 2017 Network and Distributed System Security Symposium* (San Diego, CA). Internet Society. https://doi.org/10.14722/ndss.2017.23152

[10] Shekhar Chalise, Hoang Dai Nguyen, and Phani Vadrevu. 2022. Your Speaker or My Snooper?: Measuring the Effectiveness of Web Audio Browser Fingerprints. In *Proceedings of the 22nd ACM Internet Measurement Conference* (Nice France). ACM, 349–357. https://doi.org/10.1145/3517745.3561435

[11] Anne Chao and Tsung-Jen Shen. 2003. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environmental and ecological statistics* 10, 4 (2003), 429–443.

[12] Anne Chao, Yun Tao Wang, and Lou Jost. 2013. Entropy and the species accumulation curve: a novel entropy estimator via discovery rates of new species. *Methods in Ecology and Evolution* 4, 11 (2013), 1091–1100.

[13] Chromium. 2021. *User-Agent Reduction*. The Chromium Projects. https://www.chromium.org/updates/ua-reduction/#reduced-navigatorplatform-values-for-all-versions

[14] Chromium. 2024. *Dom_plugin_array.Cc - Chromium Code Search*. https://source.chromium.org/chromium/chromium/src/+/main:third_party/blink/renderer/modules/plugins/dom_plugin_array.cc;l=187

[15] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. 2009. Power-law distributions in empirical data. *SIAM review* 51, 4 (2009), 661–703.

[16] Alissa Cooper, Hannes Tschofenig, Bernard D. Aboba, Jon Peterson, John Morris, Marit Hansen, and Rhys Smith. 2013. Privacy Considerations for Internet Protocols. https://doi.org/10.17487/RFC6973

[17] Peter Eckersley. 2010. How Unique Is Your Web Browser?. In *Privacy Enhancing Technologies* (Berlin, Heidelberg) *(Lecture Notes in Computer Science)*. Springer, 1–18. https://doi.org/10.1007/978-3-642-14527-8_1

[18] Roy Fielding and Julian Reschke. 2014. *Hypertext Transfer Protocol (HTTP/1.1): Semantics and Content*. Technical Report 7231. Internet Engineering Task Force (IETF). Section 5.5.3 cited. Available at: https://datatracker.ietf.org/doc/html/rfc7231#section-5.5.3.

[19] FingerprintJS Inc. 2024. *Fingerprintjs/Fingerprintjs*. Fingerprint ®. https://github.com/fingerprintjs/fingerprintjs

[20] Peter Grassberger. 2003. Entropy estimates from insufficient samplings. *arXiv preprint physics/0307138* (2003).

[21] Alejandro Gómez-Boix, Pierre Laperdrix, and Benoit Baudry. 2018. Hiding in the Crowd: An Analysis of the Effectiveness of Browser Fingerprinting at Large Scale. In *Proceedings of the 2018 World Wide Web Conference* (Republic and Canton of Geneva, CHE) *(WWW '18)*. International World Wide Web Conferences Steering Committee, 309–318. https://doi.org/10.1145/3178876.3186097

[22] Jean Hausser and Korbinian Strimmer. 2009. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research* 10, 7 (2009).

[23] Orris C Herfindahl. 1997. *Concentration in the steel industry*. Columbia university.

[24] Jean Luc Intumwayase, Imane Fouad, Pierre Laperdrix, and Romain Rouvoy. 2023. UA-Radar: Exploring the Impact of User Agents on the Web. In *Proceedings of the 22nd Workshop on Privacy in the Electronic Society*. 31–43.

[25] A Ya Khintchine. 1957. *Mathematical foundations of information theory*. Dover.

[26] Andreas Kurtz, Hugo Gascon, Tobias Becker, Konrad Rieck, and Felix Freiling. 2016. Fingerprinting mobile devices using personalized configurations. *Proceedings on Privacy Enhancing Technologies* (2016).

[27] Markku Laakso and Rein Taagepera. 1979. "Effective" number of parties: a measure with application to West Europe. *Comparative political studies* 12, 1 (1979), 3–27.

[28] Pierre Laperdrix, Nataliia Bielova, Benoit Baudry, and Gildas Avoine. 2020. Browser Fingerprinting: A Survey. *ACM Transactions on the Web* 14, 2 (2020), 8:1–8:33. https://doi.org/10.1145/3386040

[29] Pierre Laperdrix, Walter Rudametkin, and Benoit Baudry. 2016. Beauty and the Beast: Diverting Modern Web Browsers to Build Unique Browser Fingerprints. In *2016 IEEE Symposium on Security and Privacy (SP)* (San Jose, CA). IEEE, 878–894. https://doi.org/10.1109/SP.2016.57

[30] Xu Lin, Frederico Araujo, Teryl Taylor, Jiyong Jang, and Jason Polakis. 2023. Fashion faux pas: Implicit stylistic fingerprints for bypassing browsers' anti-fingerprinting defenses. In *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 987–1004.

[31] Aniket Mahanti, Niklas Carlsson, Anirban Mahanti, Martin Arlitt, and Carey Williamson. 2013. A tale of the tails: Power-laws in internet measurements. *IEEE Network* 27, 1 (2013), 59–64. https://doi.org/10.1109/MNET.2013.6423193

[32] Jonathan R Mayer. 2009. "Any Person... a Pamphleteer:" Internet Anonymity in the Age of Web 2.0. https://jonathanmayer.org/publications/thesis09.pdf

[33] MDN Contributors. 2025. *User-Agent: HTTP header*. Mozilla Developer Network (MDN). Accessed: October, 2025.

[34] MDN Web Docs. 2026. User-agent reduction. https://developer.mozilla.org/en-US/docs/Web/HTTP/Guides/User-agent_reduction. Accessed: 2026-02-09.

[35] George Miller. 1955. Note on the bias of information estimates. *Information theory in psychology: Problems and methods* (1955).

[36] Keaton Mowery and Hovav Shacham. 2012. Pixel perfect: Fingerprinting canvas in HTML5. *Proceedings of W2SP* 2012 (2012).

[37] Mozilla Support. 2025. *Firefox's protection against fingerprinting*. Mozilla. https://support.mozilla.org/en-US/kb/firefox-protection-against-fingerprinting

[38] Gerald Palfinger and Bernd Prünster. 2020. Androprint: analysing the fingerprintability of the android api. In *Proceedings of the 15th International Conference on Availability, Reliability and Security*. 1–10.

[39] Liam Paninski. 2003. Estimation of entropy and mutual information. *Neural computation* 15, 6 (2003), 1191–1253.

[40] Luc Rocher, Julien M Hendrickx, and Yves-Alexandre de Montjoye. 2025. A scaling law to model the effectiveness of identification techniques. *Nature Communications* 16, 1 (2025), 347.

[41] Elif Ecem Şamlioğlu, Sinan Emre Taşçi, Muhammed Fatih Gülşen, and Albert Levi. 2025. ThresholdFP: enhanced durability in browser fingerprinting. *IEEE Access* 13 (2025), 141509–141524. https://doi.org/10.1109/ACCESS.2025.3597821

[42] Asuman Senol and Gunes Acar. 2023. Unveiling the Impact of User-Agent Reduction and Client Hints: A Measurement Study. In *Proceedings of the 22nd Workshop on Privacy in the Electronic Society* (Copenhagen Denmark). ACM, 91–106. https://doi.org/10.1145/3603216.3624965

[43] Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal* 27, 3 (1948), 379–423.

[44] Edward H Simpson. 1949. Measurement of diversity. *nature* 163, 4148 (1949), 688–688.

[45] StatCounter. 2023. Operating System Market Share United States Of America. https://gs.statcounter.com/os-market-share/all/united-states-of-america/.

[46] Constantino Tsallis. 1988. Possible generalization of Boltzmann-Gibbs statistics. *Journal of statistical physics* 52 (1988), 479–487.

[47] Gregory Valiant and Paul Valiant. 2017. Estimating the unseen: improved estimators for entropy and other properties. *Journal of the ACM (JACM)* 64, 6 (2017), 1–41.

[48] W3C. 2025. *Mitigating Browser Fingerprinting in Web Specifications*. https://w3c.github.io/fingerprinting-guidance/

[49] W3C Privacy Working Group. 2025. *Mitigating Browser Fingerprinting in Web Specifications*. Technical Report. World Wide Web Consortium (W3C). Section "Identifying Fingerprinting Vectors" cited. Available at: https://www.w3.org/TR/fingerprinting-guidance/#identifying.

[50] WebKit Team. 2026. Tracking Prevention. https://webkit.org/tracking-prevention/. Accessed: 2026-02-09.

[51] Shujiang Wu, Pengfei Sun, Yao Zhao, and Yinzhi Cao. 2023. Him of Many Faces: Characterizing Billion-scale Adversarial and Benign Browser Fingerprints on Commercial Websites.. In *NDSS*. https://dx.doi.org/10.14722/ndss.2023.24394

A Extended examples

A.1 Extended Example 1

Section 4 Example 1 shows how Shannon entropy can increase with sample size due to observing larger domains. Figure 8 replicates Figure 1 (bottom plot) using the same data, but with the observed domain size on the x-axis instead of sample size.

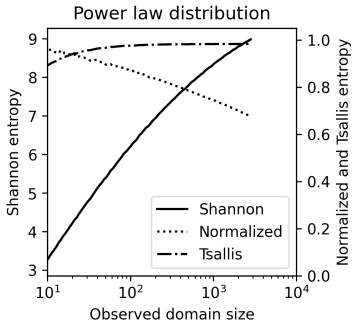

Figure 8: Figure 1 (bottom plot) replicated with observed domain size on the x-axis.

A.2 Extended Example 3

Section 4 Example 3 shows how the error for Shannon and normalized entropy grows with the domain of the value distribution. In Figure 3, values are distributed by the power law, $f(x) = x^{-1}$. Figure 9 replicates Figure 3 using the same approach but with different value distributions. Values are drawn from Zipf distributions, $f(x) = x^{-a}$, with varying a.

B Lemmas for proof of Theorem 5.3

Lemma B.1. *Let sample $X = (x_1, \ldots, x_n)$ be drawn uniformly at random without replacement from population $P = (y_1, \ldots, y_N)$. For every possible value x and all $i \in \{2, \ldots, n\}$*

$$\Pr[x_i = x] = \frac{N_x}{N}$$
$$\Pr[x_i = x \mid x_{i-1} = x] = \frac{N_x - 1}{N - 1}$$

Proof. Note that $x_1, \ldots, x_n$ are exchangeable random variables, so the marginal distribution over any $x_{i_1}, \ldots, x_{i_k}$ is independent of how we choose indices $i_1, \ldots, i_k \in [n]$. Therefore $\Pr[x_i = x] = \Pr[x_1 = x] = \frac{N_x}{N}$. Also

$$\begin{aligned}\Pr[x_i = x \mid x_{i-1} = x] &= \frac{\Pr[x_{i-1} = x \wedge x_i = x]}{\Pr[x_{i-1} = x]} \\ &= \frac{\Pr[x_1 = x \wedge x_2 = x]}{\Pr[x_1 = x]} \\ &= \frac{\frac{N_x}{N}\frac{N_x - 1}{N-1}}{\frac{N_x}{N}} \\ &= \frac{N_x - 1}{N - 1}\end{aligned}$$

□

Lemma B.2. *Let sample $X = (x_1, \ldots, x_n)$ be drawn uniformly at random without replacement from population P. Assume n is even. Let $C = \sum_{i=1}^{n/2} \mathbf{1}\{x_{2i} = x_{2i-1}\}$ be the number of collisions between consecutive values in X. For all $t > 0$ we have*

$$\Pr[|\operatorname{E}[C] - C| \geq t] \leq 2\exp(-t^2/n)$$

Proof. For each $i \in \{0, \ldots, n/2\}$ let

$$Y_i = \operatorname{E}[C \mid x_1, \ldots, x_{2i}]$$

be the expected number of collisions after observing the first i pairs of values. Thus $Y_0 = \operatorname{E}[C]$ and $Y_{n/2} = C$. Note that Y_i is a martingale (known as a Doob martingale), since

$$\begin{aligned}\operatorname{E}[Y_i \mid x_1, \ldots x_{2(i-1)}] &= \operatorname{E}[\operatorname{E}[C \mid x_1, \ldots, x_{2i}] \mid x_1, \ldots x_{2(i-1)}] \\ &= \operatorname{E}[C \mid x_1, \ldots, x_{2(i-1)}] \\ &= Y_{i-1}\end{aligned}$$

where the second equality is the law of iterated expectation. We also have $|Y_i - Y_{i-1}| \leq 1$, since the expected number of collisions in the sequence cannot change by more than 1 after drawing one additional pair of values. Therefore

$$\Pr[|\operatorname{E}[C] - C| \geq t] = \Pr[|Y_0 - Y_{n/2}| \geq t] \leq 2\exp(-t^2/n)$$

where the inequality is Azuma's inequality. □

C Proof of Theorem 7.1

We first show that for all $\varepsilon > 0$ and samples X, X' if $d_{\mathrm{TV}}(X, X') \leq \varepsilon$ then $|T(X) - T(X')| \leq 6\varepsilon$. Assume without loss of generality that $T(X) \leq T(X')$. It suffices to prove $T(X) \geq T(X') - 6d_{\mathrm{TV}}(X, Y)$. For brevity let $p_x = \frac{n_x}{n}$ and $q_x = \frac{n'_x}{n'}$. Define $\delta_x = p_x - q_x$. We have

$$\begin{aligned}T(X) &= 1 - \sum_x p_x^2 \\ &= 1 - \sum_x (q_x + \delta_x)^2 \\ &= 1 - \sum_x q_x^2 - 2\sum_x q_x\delta_x - \sum_x \delta_x^2 \\ &= T(X') - 2\sum_x q_x\delta_x - \sum_x \delta_x^2 \\ &\geq T(X') - 2\sum_x q_x|\delta_x| - \sum_x |\delta_x| \\ &\geq T(X') - 2\sum_x |\delta_x| - \sum_x |\delta_x| \\ &= T(X') - 6d_{\mathrm{TV}}(X, X').\end{aligned}$$

We now show that for all $\varepsilon, \delta > 0$ there exist samples X, X' such that $d_{\mathrm{TV}}(X, X') \leq \varepsilon$ and $|H(X) - H(X')| \geq \delta$. We construct samples X and X' as follows. The users in the two samples will each have one of $k + \ell$ possible values. Assume for convenience that the possible values are ordered. Let the users in sample X be uniformly distributed on the first k values. Let $(1 - \varepsilon)$ fraction of the users in sample X' be uniformly distributed on the first k values, and the remaining users be uniformly distributed on the last ℓ values. Thus if p and q are vectors representing the distributions of values in X and X', respectively, we have by this construction that

$$p = \left(\frac{1}{k}, \ldots, \frac{1}{k}, 0, \ldots, 0\right)$$

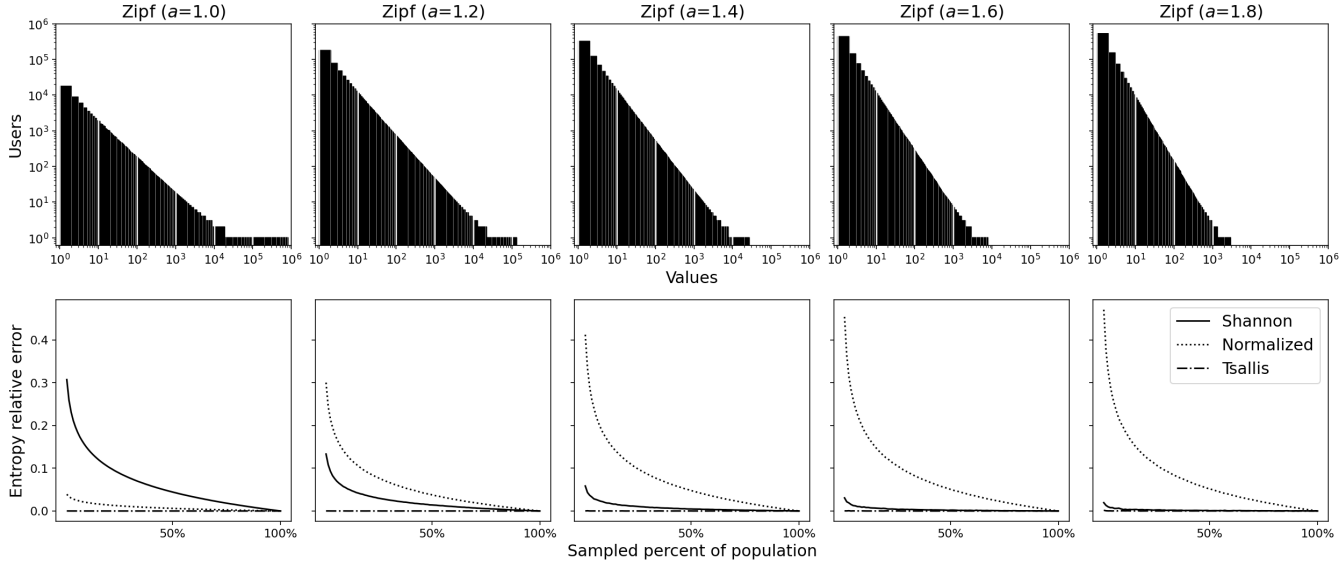

Figure 9: Figure 3 replicated using a Zipf distribution with different a coefficients.

$$q = \left(\frac{1-\varepsilon}{k}, \ldots, \frac{1-\varepsilon}{k}, \frac{\varepsilon}{\ell}, \ldots, \frac{\varepsilon}{\ell}\right)$$

Clearly $d_{\mathrm{TV}}(X, X') = \varepsilon$ and $H(X) = \log k$. By a straightforward calculation we also have

$$H(X') = (1-\varepsilon)\log\frac{k}{1-\varepsilon} + \varepsilon \log\frac{\ell}{\varepsilon}.$$

Observe that we can make $H(X')$ arbitrarily large by increasing the value of ℓ (since $x \mapsto \log x$ is an increasing function), and doing so will not change $H(X)$. Therefore $|H(X) - H(X')|$ can be made larger than any value δ.

D Data representativeness

Tables 3 and 4 provide more information on the sample of users and devices from which the fingerprinting data used in Section 6 was collected. The data was collected from US users in December 2023. Table 3 compares the devices' operating system (OS) distribution to US market share estimates from StatCounter [45] for December 2023. Table 4 shows users' demographic attributes, with comparisons to US Census data. More information on how the data was collected and reported is available in [4].

Table 3: OS market share distribution.

	Sample		US [45]
OS	**n**	**%**	**%**
Windows	4214	50.17	35.29
iOS	1568	18.67	26.77
Android / Chrome OS / Linux	1526	18.17	22.38
OS X	1092	13	14.03
Other			1.53

Table 4: User sample demographic attributes.

	Sample (N=8400)		US Census
	N	%	%
Gender			
Female	3990	47.5	50.9
Male	4227	50.3	49.1
Other	183	2.2	
Age			
18 - 24 years	1302	15.5	11.9
25 - 34 years	2859	34	17.3
35 - 44 years	2024	24.1	16.8
45 - 54 years	1158	13.8	15.3
55 - 64 years	712	8.5	15.7
65 or older	345	4.1	22.9
Household income			
Less than $25,000	1097	13.1	15.5
$25,000 - $49,999	1842	21.9	17.9
$50,000 - $74,999	1692	20.1	18.7
$75,000 - $99,999	1267	15.1	12.1
$100,000 - $149,999	1411	16.8	15.1
$150,000 or more	951	11.3	20.7
Prefer not to say	140	1.7	
Hispanic origin			
Yes	923	11	19.1
Race			
White	5911	70.4	75.5
Black	938	11.2	13.6
Asian	842	10	6.3
American Indian	53	0.6	1.3
Native Pacific Islander	0	0	0.3
Other or mixed	656	7.8	3

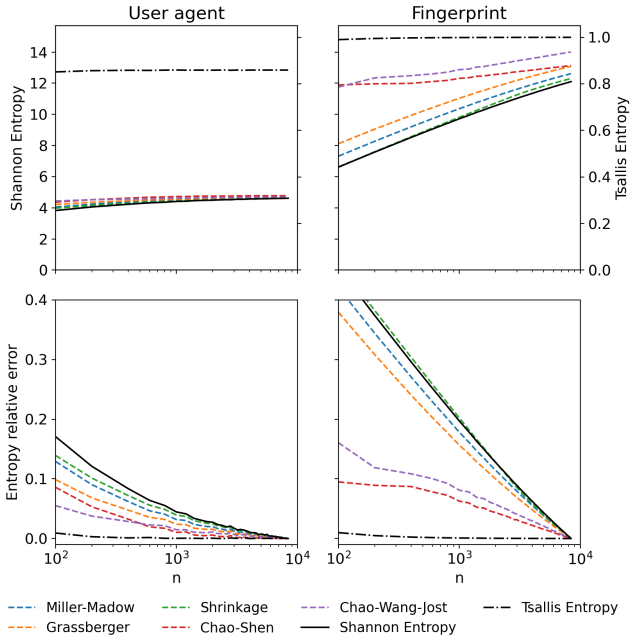

Figure 10: Figure 7 replicated with the x-axis on a log scale.

E Comparing Tsallis to advanced Shannon entropy measures with log scale

Figure 10 replicates Figure 7 with the x-axis on a log scale, for comparison to Figures 1 - 3.

F Entropy versus uniqueness

Table 5: Correlation coefficients comparing the entropy to % Unique values from Table 2. Note Shannon and normalized entropy are linearly correlated. All correlation coefficients are statistically significant at the p<0.05 level.

Entropy	Pearson	Spearman	Kendall
Shannon	0.808	0.756	0.604
Normalized	0.808	0.756	0.604
Tsallis	0.562	0.654	0.520

Figure 11 compares the % unique to the entropy metrics from Table 2 for each browser attribute. Table 5 shows the corresponding correlations computed between % unique and entropy. These show how Shannon and normalized entropy are more correlated with % unique versus Tsallis entropy. As discussed in Section 7, Tsallis entropy is better for estimating average anonymity set size, while Shannon entropy is better at estimating uniqueness, i.e. the fraction of users with an anonymity set size of 1. An even better method for estimating % unique for a population is demonstrated by the Pitman-Yor process (PYP) method of Rocher et al. [40]. Figure 12 applies their PYP method [3] to the User agent and Fingerprint attributes from Table 2, using the repeated random sampling method described in Section 6.2, showing how the PYP method can effectively estimate uniqueness.

Figure 13 compares entropy measures to the PYP method for estimating % unique, where metrics are computed over synthetic data representing the uniform distribution with varying domain sizes (and thus varying anonymity set sizes). It shows how when anonymity set sizes are any greater than 1, yet still small, such as size 2, entropy measures may better reflect risk. Yet if the goal is to measure the percent of a population that is unique, then the PYP method is superior.

[3] Code to use methods from [40]: https://github.com/synthetic-society/dataless

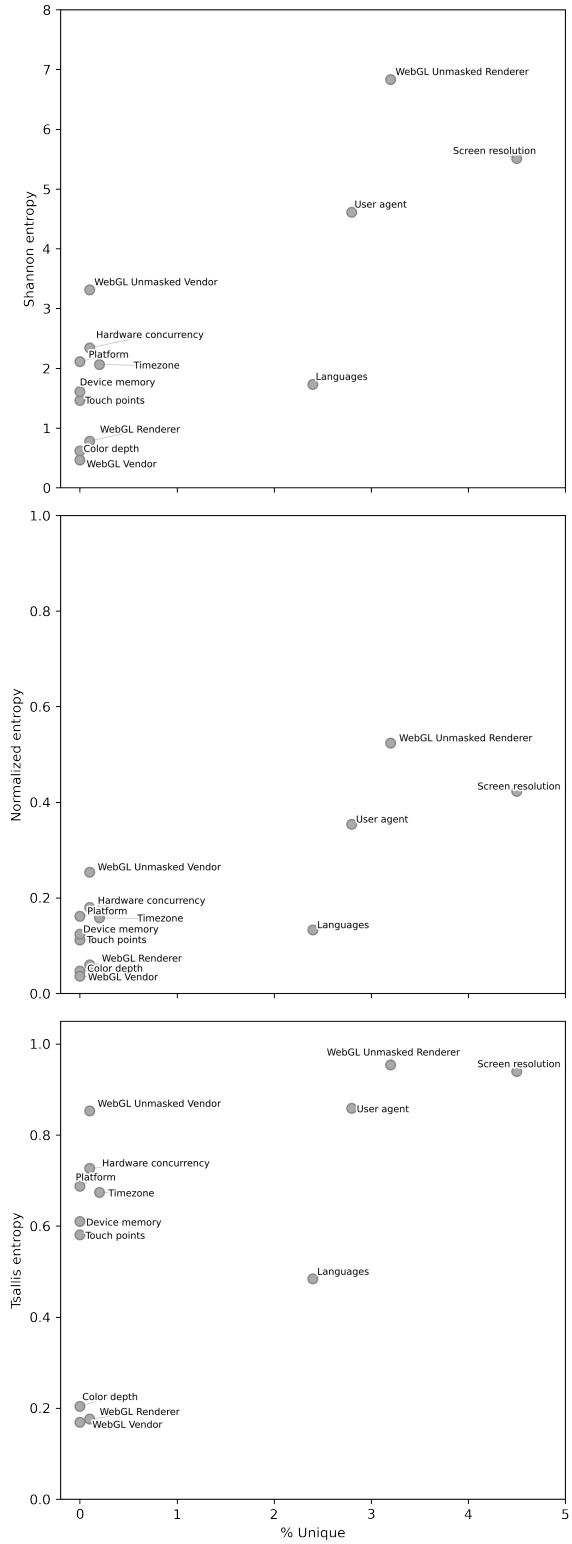

Figure 11: A comparison between % Unique and entropy, for each browser attribute in Table 2, for (top) Shannon entropy, (middle) normalized entropy, and (bottom) Tsallis entropy.

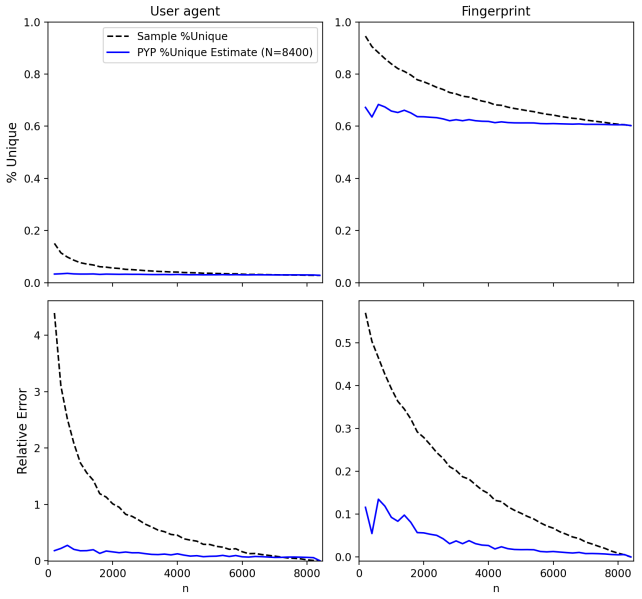

Figure 12: Results from applying the PYP method from [40] to estimate % unique for the (left) User agent and (right) Fingerprint values, using data from Table 2. Mean values were estimated over 100 trials for subsamples ranging from n=200 to 8400.

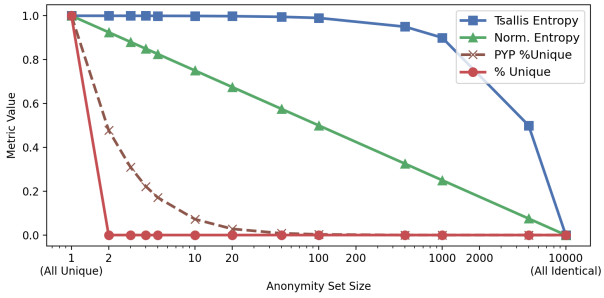

Figure 13: Entropy metrics compared to % unique and the PYP estimates for % unique, computed over synthetic data representing the uniform distribution with varying anonymity set sizes (domain sizes).