IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026

DOI: 10.17148/IJARCCE.2026.1510150

Measuring Web Tracking and Browser Fingerprinting Across the Popularity Spectrum: A Stratified Crawl of 20,911 Websites from the Tranco Top Million

Krushna Sanjay Mohod[1], Atharva Sunil Raut[1], Atharv Rajesh Kakade[1], Shravani Sachin Yawale[1], S. R. Gupta[2]

Student, Department of Artificial Intelligence and Data Science, Prof. Ram Meghe Institute of Technology & Research, Badnera-Amravati, India[1]

Head, Department of Artificial Intelligence and Data Science, Prof. Ram Meghe Institute of Technology & Research, Badnera-Amravati, India[2]

Abstract: Third-party tracking and browser fingerprinting are routinely measured on the most popular websites, yet most of the web lives in the long tail of the popularity distribution. We present a stratified measurement of 20,911 websites drawn from the Tranco top-one-million list in three rank bands (1–10k, 10k–100k and 100k–1M). Each landing page was loaded in an instrumented Chromium browser that routes every connection through a destination-validating safety proxy, records all requests and cookies, matches them against the DuckDuckGo Tracker Data Set, and counts calls to canvas, WebGL and audio fingerprinting APIs. 13,045 pages loaded successfully. Re-weighted to the whole list, 84.1% of websites contact at least one known tracker and 27.0% exhibit at least one fingerprinting signal. Tracker presence is flat across popularity, but invasive practices are not: session replay (17.6% vs 9.6%), fingerprinting (33.3% vs 26.8%) and third-party cookies (55.0% vs 44.2%) are significantly more common at the head of the list than in the tail. We also find that 3,660 sites (17.5%) answered the automated browser with an error or bot-challenge page, and that these challenge pages are themselves heavy finger printers. The resulting feature dataset (107 attributes per site) is used to train the Ovael malicious-tracker detection system.

Keywords: Web tracking, browser fingerprinting, web measurement, stratified sampling, Tranco, privacy, bot detection, dataset.

I. INTRODUCTION

Almost every page load on the modern web triggers requests to companies other than the site the user chose to visit. These third parties set cookies, read device characteristics and, in some cases, record every mouse movement and keystroke on the page. Large measurement studies have documented this ecosystem [1], [7], [8], and a parallel line of work has shown that a browser can be re-identified without any cookie at all by fingerprinting its canvas rendering, graphics hardware and audio stack [2]–[5].

Most measurement studies, however, either crawl only the top few thousand sites or crawl a full top-million list without reporting how behaviour changes with popularity. A user's browsing is dominated by a handful of popular sites, but most *distinct* sites a user encounters, and most sites that a protection tool must classify, sit far down the long tail. Whether the head of the list is representative of the tail is therefore a practical question for anyone building tracker-detection tools. This paper makes four contributions:

- **A stratified, re-weightable sample.** We draw 20,911 sites from three popularity strata of the Tranco list [6] in a shuffled order, so that any prefix of the crawl remains a stratified random sample and band-level results can be re-weighted into unbiased estimates for the whole list.
- **A safety-first instrumented crawler.** All browser traffic passes through a proxy that resolves every hostname itself and refuses non-public destinations, eliminating DNS-rebinding and internal-network exposure during large crawls.
- **Popularity-resolved prevalence estimates** for 13 tracking and fingerprinting behaviours, with significance tests between the head and the tail of the list.

 This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026

DOI: 10.17148/IJARCCE.2026.1510150

- **Evidence that bot-protection pages fingerprint the visitor,** together with an account of how such pages bias crawl-based measurements.

The dataset is the training and evaluation corpus for Ovael, our AI-based malicious web tracker detection system, which combines a Random Forest classifier [14] with an Isolation Forest novelty detector [15] and SHAP explanations [16].

II. RELATED WORK

Mayer and Mitchell [7] surveyed third-party tracking technology and policy. Lerner et al. [8] reconstructed two decades of tracking from archived pages and showed its steady growth. The OpenWPM study of Englehardt and Narayanan [1] crawled the top one million sites and measured stateful tracking as well as canvas, WebRTC, font and AudioContext fingerprinting; it reported that fingerprinting scripts were far more common on popular sites. Acar et al. [2] documented persistent tracking mechanisms such as canvas fingerprinting and evercookies, building on Eckersley's demonstration that browser configurations are highly identifying [3]. Laperdrix et al. [4] survey the fingerprinting literature, and Iqbal et al. [5] train a machine-learning classifier over JavaScript execution traces to detect fingerprinting scripts.

Methodological work has shown that crawler configuration strongly affects what is observed. Ahmad et al. [9] found that the choice of crawling tool changes measured third-party prevalence, and Jueckstock et al. [10] showed that headless and automated browsers see a different web from real users. Le Pochat et al. [6] introduced the Tranco list to make top-site rankings reproducible and resistant to manipulation. Our work follows this tradition but contributes a popularity-stratified design with explicit sampling weights, and examines the bot-challenge pages that automated crawls increasingly encounter.

III. METHODOLOGY

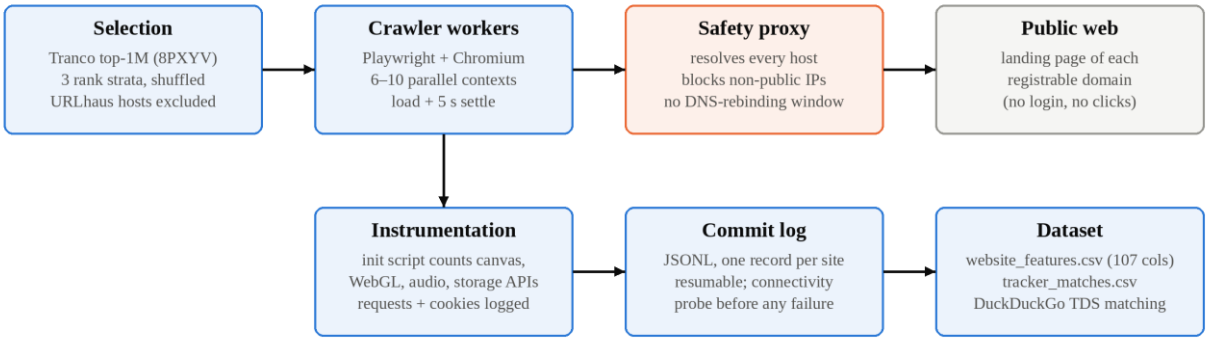


Fig. 1 Measurement pipeline

A. Population and Stratified Sampling

The population is the Tranco list 8PXYV generated on 10 October 2026, which combines the CrUX, Farsight, Majestic, Cloudflare Radar and Umbrella rankings over the preceding 30 days [6]. We verified the local copy against the list published on tranco-list.eu. Rows that were not parseable as domain names or were bare IP addresses (11) and domains listed in the URLhaus malware host list [12] (5) were excluded before sampling.

The remaining domains were divided into three rank bands. The top 10,000 were taken as a census, and 45,000 domains were sampled uniformly at random from each of the two lower bands with a fixed seed (2026). The combined selection was then shuffled. Because the crawl visits sites in this shuffled order, the set of sites attempted when the crawl is stopped is itself a stratified random sample. We stopped after 20,911 sites; Table I gives the resulting design.

TABLE I SAMPLING DESIGN (TRANCO 8PXYV, ELIGIBLE DOMAINS)

Rank band	Eligible	Attempted	Loaded	Weight w_b
1–10000	10,000	2,136	1,216	4.68
10001–100000	89,998	9,498	5,727	9.47
100001–1000000	899,986	9,277	6,102	97.01
Total	999,984	20,911	13,045	–

 This work is licensed under a Creative Commons Attribution 4.0 International License

DOI: 10.17148/IJARCCE.2026.1510150

B. Crawler and Measurement Protocol

The crawler is written in Python on top of Playwright [13] and drives headless Chromium with a 1366×768 viewport and the en-US locale. Each site is visited in a fresh browser context at the URL https://<domain>/. The crawler waits for the load event (up to 15 s, with a 30 s navigation timeout and a 150 s overall cap per site), then lets the page settle for 5 s so that deferred scripts can run. Network errors are retried up to twice with exponential back-off, and a server's Retry-After header is honoured up to 30 s. The crawl ran with ten parallel workers, reduced to six later in the run to limit memory use. No links are clicked, no forms are filled and no consent banners are dismissed, so all observations reflect what a first-time visitor's browser does before any interaction.

An initialisation script is injected into every frame before any page script runs. It wraps the relevant Web APIs and counts calls to them: canvas read-back (HTMLCanvasElement.toDataURL and toBlob, CanvasRenderingContext2D.getImageData), WebGL parameter queries for UNMASKED_VENDOR_WEBGL and UNMASKED_RENDERER_WEBGL, construction and rendering of OfflineAudioContext objects, localStorage and sessionStorage reads and writes, and document.cookie access. The crawler also records every network request (URL host, resource type, status) and every cookie (domain, lifetime, Secure, HttpOnly and SameSite attributes), together with the redirect chain and the iframes present in the final document.

C. Safety and Ethics

Crawling a million-site list means visiting domains whose operators are unknown. Every browser context therefore sends all of its traffic (navigations, redirect hops, sub-resources, iframes and WebSockets) through its own instance of a destination-validating forward proxy. Chromium does not resolve proxied hostnames itself, so the proxy is the only component that performs DNS resolution. It rejects a connection if any resolved address is loopback, private (RFC 1918), link-local, carrier-grade NAT, multicast or reserved, and it then connects to the exact address it validated, so there is no window for DNS rebinding. QUIC is disabled and WebRTC is restricted to proxied UDP so that no traffic can bypass the proxy. During the crawl, 58 sites were blocked because they resolved to non-public addresses.

Each site receives a single landing-page visit, with a one-second delay between sites per worker, which is a load comparable to one human visitor. No credentials are used and no personal data is collected: the dataset stores hostnames, counts and cookie metadata, request paths are truncated to 200 characters, and cookie values are never stored, in keeping with the data-minimisation principle of India's Digital Personal Data Protection Act [20]. Before any failure is recorded, the crawler probes public DNS resolvers. If local connectivity has been lost, the site is left pending instead of being recorded as unreachable, so outages at the measurement vantage point do not contaminate the failure statistics.

D. Tracker Attribution and Features

Request hosts are matched against the DuckDuckGo Tracker Data Set [11] (version of 21 September 2026; 999 tracker domains owned by 760 entities). Each match records the owning entity and its categories (for example Advertising, Analytics, Session Replay). A request is counted as third-party only when its registrable domain belongs to a different entity from the visited site, so first-party analytics hosted by the same company are not counted as cross-site tracking. Each website is summarised by 107 attributes covering navigation, request mix, tracker categories, cookies, storage, fingerprinting counters, redirects and iframes. A page is said to exhibit a *fingerprinting signal* if it performs a canvas read-back, queries the unmasked WebGL renderer or vendor, or constructs an OfflineAudioContext. An iframe is *hidden* if it is not rendered, is at most one pixel wide or high, or lies entirely off-screen.

E. Estimation

Let N_b be the number of eligible domains in band b and n_b the number attempted. Every site in band b receives the weight $w_b = N_b / n_b$ (Table I). For a binary property y, the population estimate over successfully loaded sites is the ratio estimator $\Sigma\, w_i\, y_i / \Sigma\, w_i$ [21]. Band-level proportions are reported with 95% Wilson score intervals [19], and the head band (1–10k) is compared with the tail band (100k–1M) using Pearson's χ^2 test with one degree of freedom.

IV. DATASET OVERVIEW

A. Crawl Outcomes

Table II summarises the outcome of the 20,911 visits. 13,053 sites (62.4%) returned a successful page, 3,660 (17.5%) returned an HTTP error status, 4,140 (19.8%) could not be reached, and 58 were blocked by the safety proxy. Most unreachable domains no longer resolve: DNS failures (1,703) and DNS timeouts (899) together account for 62.9% of failures, followed by connection timeouts (857). Popular domains fail more often than tail domains. This reflects the fact that many top-ranked entries are infrastructure hostnames (CDNs, API and telemetry endpoints) that are popular in DNS-based rankings but serve no web page.

 This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026

DOI: 10.17148/IJARCCE.2026.1510150

TABLE II CRAWL OUTCOMES BY RANK BAND (% OF ATTEMPTED SITES)

Outcome	1–10k	10k–100k	100k–1M	Sites
Success	56.9	60.4	65.8	13,053
HTTP error	20.4	18.9	15.4	3,660
Unreachable	22.2	20.5	18.6	4,140
Blocked (unsafe IP)	0.5	0.3	0.2	58

B. Volume

Analyses in Section V use the 13,045 successfully loaded pages whose instrumentation completed. Together these pages issued 1,494,209 requests, of which 464,954 (31.1%) went to a known tracker domain, and set 168,977 cookies, of which 86,142 (51.0%) were third-party. 817 distinct tracker domains belonging to 642 entities were observed. Table III shows that pages become lighter down the popularity list: the median top-10k page makes 93 requests and contacts 8 third-party domains, against 71 and 6 in the tail.

TABLE III MEDIAN PER-PAGE VALUES AMONG SUCCESSFULLY LOADED PAGES

Quantity	1–10k	10k–100k	100k–1M
Requests	93	81	71
Third-party requests	30	25	19
Third-party domains	8	7	6
Tracker domains	5	4	4
Tracker requests	13	11	9
Cookies	8	6	4
Third-party cookies	1	1	0

V. RESULTS

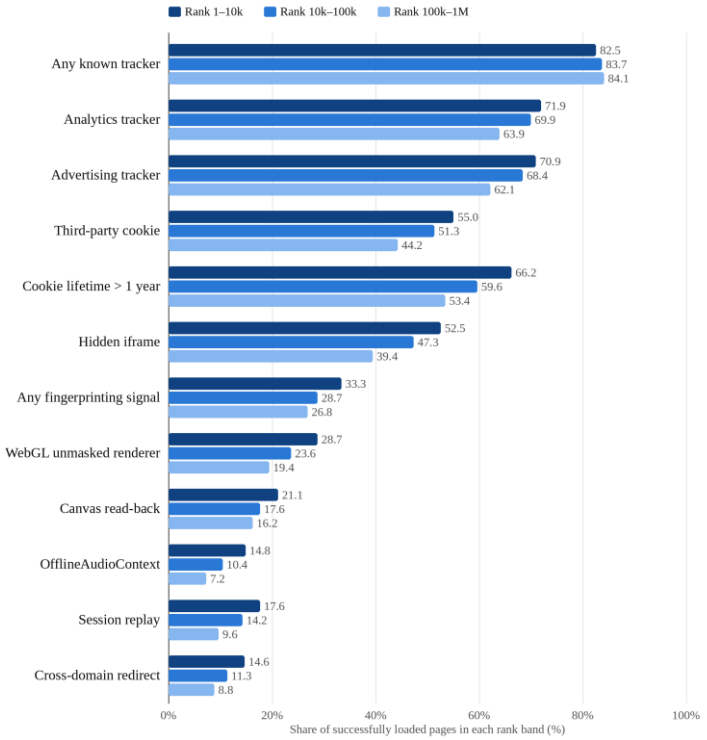


Fig. 2 Prevalence of tracking and fingerprinting behaviours by Tranco rank band (successfully loaded pages)

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026

DOI: 10.17148/IJARCCE.2026.1510150

A. Tracking Is Near-Universal at Every Popularity Level

An estimated 84.1% of successfully loaded sites in the Tranco top million contact at least one known tracker domain. This share hardly changes with popularity (82.5%, 83.7% and 84.1% in the three bands; head vs tail $\chi^2 = 1.9$, $p = 0.17$). Being a small website does not, by itself, mean that visitors are free of third-party tracking.

B. Invasive Techniques Concentrate at the Head

The *kind* of tracking, by contrast, depends strongly on popularity (Fig. 2, Table IV). Advertising trackers appear on 70.9% of top-10k pages but on 62.1% of tail pages ($\chi^2 = 33.6$, $p < 0.001$), and third-party cookies on 55.0% against 44.2% ($\chi^2 = 47.3$, $p < 0.001$). Session-replay services, which record the user's interaction with the page and can capture what the user types [17], are present on 17.6% of top-10k sites, almost twice the 9.6% seen in the tail ($\chi^2 = 66.4$, $p < 0.001$).

Fingerprinting follows the same gradient. At least one fingerprinting signal occurs on 33.3% (95% CI 30.7–36%) of top-10k pages, 28.7% of pages ranked 10k–100k and 26.8% (25.7–27.9%) of tail pages ($\chi^2 = 21.2$, $p < 0.001$). The steepest decline is for OfflineAudioContext, which falls from 14.8% to 7.2%. Of the 3,683 pages with any signal, 1,110 used all three techniques together, which is characteristic of commercial fingerprinting libraries rather than incidental use of the canvas or WebGL APIs. Re-weighted to the whole list, an estimated 27.0% of loaded sites fingerprint their visitors on the first page view.

TABLE IV PREVALENCE AMONG SUCCESSFULLY LOADED PAGES (%) AND RE-WEIGHTED TOP-1M ESTIMATE

Behaviour	**1–10k**	**10k–100k**	**100k–1M**	**Top-1M est.**
Any known tracker	82.5	83.7	84.1	84.1
Advertising tracker	70.9	68.4	62.1	62.7
Analytics tracker	71.9	69.9	63.9	64.5
Session replay	17.6	14.2	9.6	10.0
Third-party cookie	55.0	51.3	44.2	44.9
SameSite=None cookie	66.4	60.6	51.3	52.2
Cookie lifetime > 1 year	66.2	59.6	53.4	54.0
Canvas read-back	21.1	17.6	16.2	16.4
WebGL unmasked renderer	28.7	23.6	19.4	19.8
OfflineAudioContext	14.8	10.4	7.2	7.5
Any fingerprinting signal	33.3	28.7	26.8	27.0
Hidden iframe	52.5	47.3	39.4	40.2
Cross-domain redirect	14.6	11.3	8.8	9.1

 This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026

DOI: 10.17148/IJARCCE.2026.1510150

C. Who Tracks

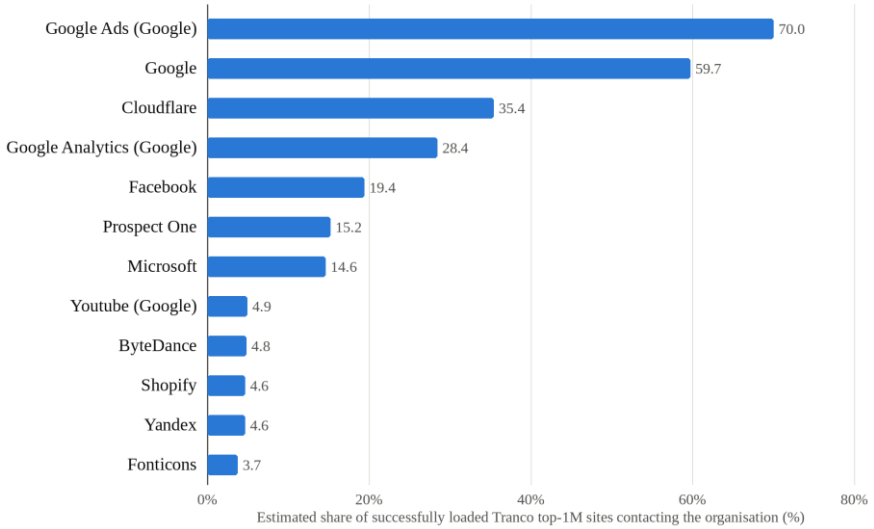


Fig. 3 Organisations contacted as third parties, as an estimated share of successfully loaded top-1M sites

Fig. 3 ranks the organisations that sites contact as third parties. Google entities dominate: the Tracker Data Set's "Google Ads" entity, which also covers googletagmanager.com and fonts.googleapis.com, appears on 70.0% of sites, and other Google domains on 59.7%. Cloudflare (35.4%), Facebook (19.4%) and Microsoft (14.6%) follow. The presence of an organisation reflects any request to its domains, including content delivery and fonts, and is not by itself evidence of tracking. Two contrasts are notable. Microsoft (25.8% vs 13.9%) and The Trade Desk (9.5% vs 2.9%) are concentrated at the head, consistent with enterprise advertising stacks, whereas Shopify (0.2% vs 4.9%) is almost entirely a tail phenomenon, reflecting small online shops built on hosted platforms.

D. Bot-Challenge Pages Fingerprint the Visitor

3,660 sites returned an HTTP error status instead of content, most often 403 Forbidden (2,856). The page titles show that the majority are bot-protection interstitials: "Just a moment..." (1,170) and "Attention Required! | Cloudflare" (548) are Cloudflare challenge pages [18], and "Access Denied" and "403 Forbidden" pages come from other CDN and web-application-firewall providers. These pages are not empty: 29.9% of error pages exhibit a fingerprinting signal, *more* than ordinary pages, and 1,156 of the 2,856 403 pages ran all three fingerprinting probes, 1,136 of them while loading Cloudflare scripts. Bot detection and tracking therefore use the same browser primitives, and a detector that flags fingerprinting behaviour alone will flag security vendors as well as trackers.

These pages also bias crawl-based measurement. The share of sites served an error or challenge page is highest at the head of the list (20.4% of top-10k sites against 15.4% in the tail), and challenge pages carry almost no advertising (1.2%) or session replay (0.2%). Including them would lower the measured prevalence of invasive tracking, most of all for popular sites. We therefore report them separately and exclude them from Section V-A to V-C.

E. Weak Labels for the Ovael Detector

For training Ovael, each successfully loaded page receives a documented *weak label*. A page is labelled **invasive tracker** if it shows any fingerprinting signal, contacts a session-replay service or contacts a domain tagged as malware. It is labelled **tracker** if it contacts advertising, analytics or ad-motivated tracking domains, or receives third-party cookies from a tracker domain. Otherwise it is labelled **benign**. Table V gives the resulting distribution. Invasive pages fall from 42.5% at the head to 32.1% in the tail, while benign pages rise from 21.4% to 29.3%. Because these labels are derived from the same features the classifier sees, they serve as a bootstrap: in Ovael, reviewer-confirmed labels replace them over time, and retrained models are promoted only if their macro F1 does not decrease and their false-alarm rate rises by no more than two percentage points.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026

DOI: 10.17148/IJARCCE.2026.1510150

TABLE V WEAK-LABEL DISTRIBUTION AMONG SUCCESSFULLY LOADED PAGES (%)

Label	1–10k	10k–100k	100k–1M	Top-1M est.
Invasive tracker	42.5	37.1	32.1	32.6
Tracker	36.1	39.4	38.6	38.6
Benign	21.4	23.5	29.3	28.8

VI. DISCUSSION AND LIMITATIONS

The central finding is that popularity predicts *how* a site tracks rather than *whether* it does. Nearly every site, popular or obscure, contacts a known tracker, typically for analytics or tag management. The heavier techniques, namely fingerprinting, session replay, long-lived third-party cookies and real-time-bidding advertising stacks, are concentrated on popular sites with the commercial incentive and engineering capacity to deploy them. Measurements restricted to the top few thousand sites therefore overstate the prevalence of invasive tracking on the web as a whole, while understating how many distinct sites leak visits to third parties at all.

Several limitations apply. **Vantage point:** all visits originate from a single residential connection in India with an en-US locale, and sites may serve different content, consent banners and advertisers to other regions. **Automation:** a headless browser is detectable [10], and about one site in six served a challenge or error page instead of content. **Interaction:** we observe only the first landing page before any consent decision or interaction, so behaviour behind cookie banners or on inner pages is not captured. **Attribution:** tracker matching depends on the coverage of the DuckDuckGo Tracker Data Set, and fingerprinting signals are counted per page, not attributed to individual scripts, so legitimate uses of canvas or WebGL (for example charts or games) can produce false positives. **Sample size:** the crawl covers 2.1% of the list; band-level sample sizes of 1,216 to 6,102 loaded pages give 95% intervals of roughly ±1–3 percentage points, adequate for the comparisons made but not for rare entities.

VII. CONCLUSION AND FUTURE WORK

We measured tracking and fingerprinting on 20,911 websites sampled from three popularity strata of the Tranco top million, using a crawler that routes all traffic through a destination-validating safety proxy. An estimated 84.1% of loaded sites contact a known tracker and 27.0% fingerprint the visitor on first view. Invasive practices decline significantly from the head of the list to its tail, while basic tracker presence does not. Bot-protection pages are a large and fingerprinting-heavy part of what an automated crawler sees, and they must be handled explicitly in both measurement and detection. Future work will extend the crawl toward the full 100,000-site selection, add script-level attribution of fingerprinting calls, crawl from additional vantage points, and use the dataset to train and evaluate the Ovael Random Forest and Isolation Forest models on the long tail of the web.
The crawler, analysis scripts and feature dataset are available from the authors on request.

ACKNOWLEDGMENT

The authors thank the Department of Artificial Intelligence and Data Science, Prof. Ram Meghe Institute of Technology & Research, Badnera-Amravati, for providing the facilities used in this work, and the maintainers of the Tranco list, the DuckDuckGo Tracker Radar and URLhaus for making their data openly available.

REFERENCES

[1]. S. Englehardt and A. Narayanan, "Online tracking: A 1-million-site measurement and analysis," in *Proc. ACM SIGSAC Conf. Computer and Communications Security (CCS)*, pp. 1388–1401, 2016.
[2]. G. Acar, C. Eubank, S. Englehardt, M. Juarez, A. Narayanan and C. Diaz, "The web never forgets: Persistent tracking mechanisms in the wild," in *Proc. ACM SIGSAC Conf. Computer and Communications Security (CCS)*, pp. 674–689, 2014.
[3]. P. Eckersley, "How unique is your web browser?," in *Proc. Privacy Enhancing Technologies Symposium (PETS)*, LNCS vol. 6205, pp. 1–18, 2010.
[4]. P. Laperdrix, N. Bielova, B. Baudry and G. Avoine, "Browser fingerprinting: A survey," *ACM Transactions on the Web*, vol. 14, no. 2, pp. 1–33, 2020.
[5]. U. Iqbal, S. Englehardt and Z. Shafiq, "Fingerprinting the fingerprinters: Learning to detect browser fingerprinting behaviors," in *Proc. IEEE Symposium on Security and Privacy (SP)*, pp. 1143–1161, 2021.

 This work is licensed under a Creative Commons Attribution 4.0 International License



International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026
DOI: 10.17148/IJARCCE.2026.1510150

[6]. V. Le Pochat, T. Van Goethem, S. Tajalizadehkhoob, M. Korczyński and W. Joosen, "Tranco: A research-oriented top sites ranking hardened against manipulation," in *Proc. Network and Distributed System Security Symposium (NDSS)*, 2019.
[7]. J. R. Mayer and J. C. Mitchell, "Third-party web tracking: Policy and technology," in *Proc. IEEE Symposium on Security and Privacy (SP)*, pp. 413–427, 2012.
[8]. A. Lerner, A. K. Simpson, T. Kohno and F. Roesner, "Internet Jones and the raiders of the lost trackers: An archaeological study of web tracking from 1996 to 2016," in *Proc. USENIX Security Symposium*, 2016.
[9]. S. S. Ahmad, M. D. Dar, M. F. Zaffar, N. Vallina-Rodriguez and R. Nithyanand, "Apophanies or epiphanies? How crawlers impact our understanding of the web," in *Proc. The Web Conference (WWW)*, pp. 271–280, 2020.
[10]. J. Jueckstock, S. Sarker, P. Snyder, A. Beggs, P. Papadopoulos, M. Varvello, B. Livshits and A. Kapravelos, "Towards realistic and reproducible web crawl measurements," in *Proc. The Web Conference (WWW)*, pp. 80–91, 2021.
[11]. DuckDuckGo, "Tracker Radar," https://github.com/duckduckgo/tracker-radar, accessed Oct. 2026.
[12]. abuse.ch, "URLhaus: Malware URL exchange," https://urlhaus.abuse.ch/, accessed Oct. 2026.
[13]. Microsoft, "Playwright: Fast and reliable end-to-end testing for modern web apps," https://playwright.dev/, accessed Oct. 2026.
[14]. L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[15]. F. T. Liu, K. M. Ting and Z.-H. Zhou, "Isolation forest," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, pp. 413–422, 2008.
[16]. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
[17]. S. Englehardt, "No boundaries: Exfiltration of personal data by session-replay scripts," *Freedom to Tinker*, Center for Information Technology Policy, Princeton University, Nov. 2017.
[18]. Cloudflare, "Cloudflare challenges," https://developers.cloudflare.com/cloudflare-challenges/, accessed Oct. 2026.
[19]. E. B. Wilson, "Probable inference, the law of succession, and statistical inference," *Journal of the American Statistical Association*, vol. 22, no. 158, pp. 209–212, 1927.
[20]. Ministry of Law and Justice, Government of India, "The Digital Personal Data Protection Act, 2023," Act no. 22 of 2023.
[21]. W. G. Cochran, *Sampling Techniques*, 3rd ed. New York: John Wiley & Sons, 1977.

 This work is licensed under a Creative Commons Attribution 4.0 International License