



International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

The Data Exhaust: What Everyday Digital Activities Reveal About Us

Assistant Professor Smita Mangesh Junnarkar¹, MV Tamil Eelan²

Department of Data Science, SIES COLLEGE OF ARTS, SCIENCE AND COMMERCE (EMPOWERED AUTONOMOUS), Sion (W), Mumbai.

Smita Mangesh Junnarkar, smitaj@sies.edu.in

Department of Data Science, SIES COLLEGE OF ARTS, SCIENCE AND COMMERCE (EMPOWERED AUTONOMOUS), Sion (W), Mumbai. MV

Tamil Eelan, mvtamileelan1819@gmail.com

ABSTRACT—

Every search, tap, payment, and journey made with a digital device leaves behind a trace that the person never set out to create. This by-product, commonly called data exhaust, is collected to keep services running, yet it can later be reused to describe who a person is, what they believe, where they go, and what they are likely to do next. This paper examines what everyday digital activities reveal. It follows a trace from generation through logging, linkage, modelling, and use, and it classifies eight families of activity: web browsing and search, location and movement, social media, shopping and payments, streaming, smart homes and wearables, smartphone apps and sensors, and device and interaction behaviour. For each family the paper explains what is recorded, what can be inferred, and why the trace is revealing. An assessment framework is then proposed, with an activity inventory, an information-based measure of identifiability, and metrics for inference, linkage, and retention, followed by safeguards at the personal, organisational, technical, and legal levels. The work is analytical and reports no new empirical results. Its purpose is to help students, researchers, and developers judge digital exposure and design services that respect the people who use them.

Keywords—*Data exhaust, digital footprint, behavioural data, privacy, re-identification, inference, metadata, data protection, surveillance*

1. INTRODUCTION

Most people think of data as something they create on purpose: a photograph they upload, a form they complete, a message they type. Yet the larger share of what digital services hold about a person was never offered in that sense. It is a by-product of ordinary use. A phone notes which cell towers it connected to, a browser logs which pages loaded and for how long, a payment app stores the time and place of a purchase, and a streaming service remembers where a viewer paused. The phrase data exhaust has been used in discussions of big data for this residue, by analogy with the fumes a vehicle emits while it travels towards some other goal [1].

Exhaust has turned out to be valuable because behaviour is harder to disguise than self-description. People may answer a survey the way they wish to be seen, but their clicks, routes, and purchases record what they actually did. Companies learned that logs collected to run a service can be reused to predict what a person will do next, and the resulting trade in predictions has been described as a new economic logic built on behavioural data [2].

The difficulty is that any single trace looks harmless. A timestamp, an address on a network, or a list of installed applications does not feel like a confession. Research over the past two decades has nevertheless shown that traces of this kind, once combined, can reveal personality, health, beliefs, routines, and even identity [3], [4], [5]. The person who agreed to share a trace rarely foresees the conclusions that will later be drawn from it.

This paper approaches the problem through three questions. First, through which everyday activities is exhaust produced, and what does each leave behind? Second, what can be inferred from each kind of trace, and how do traces from different activities strengthen one another when they are put together? Third, which safeguards at the personal, organisational, technical, and legal levels reduce the exposure? The answers are developed through a description of the data pipeline, a classification of eight activity families, an assessment framework, and a set of countermeasures.

The paper is organised as follows. Section 2 summarises earlier studies, and Section 3 explains how a trace travels from a device to a decision. Section 4 classifies the everyday activities that produce exhaust. Section 5 presents the assessment framework, Section 6 explains how the evidence should be interpreted, and Section 7 lists safeguards. Section 8 discusses limitations and ethical issues, Section 9 suggests future directions, and Section 10 concludes.

2. BACKGROUND AND RELATED STUDIES

2.1 Origins of the Idea

Mayer-Schönberger and Cukier described datafication as the conversion of more and more aspects of life into records that can be tabulated, and they observed that data gathered for one purpose are frequently reused for another [1]. Zuboff pushed the argument further. In her account of surveillance capitalism, behavioural traces that exceed what a service needs in order to work, which she calls behavioural surplus, are collected and turned into predictions offered to third parties [2]. Whatever one makes of the wider thesis, the observation that the by-product has become the product is a useful

starting point for a technical discussion.

2.2 Inferring Traits from Digital Records

Kosinski and colleagues analysed the Facebook Likes of more than 58,000 volunteers and showed that simple statistical models could predict attributes the users had never disclosed, among them sexual orientation, ethnic origin, and religious and political views, with accuracy that was high for several of them [3]. Youyou et al. later found that a modest number of Likes was enough for a model to judge personality more accurately than acquaintances such as work colleagues [6]. Onnela and Rauch proposed treating phone sensors and usage logs as a measuring instrument for behaviour and health, a programme they called digital phenotyping [7]. Together these studies show that exhaust is informative, and the property that benefits research and care is the same one that exposes the person.

2.3 Re-identification and the Limits of Anonymisation

The usual reply to such concerns is to remove names and share the remainder. Sweeney's early work on demographic fields showed that this offers little protection, because the combination of postcode, gender, and birth date singled out most residents in the data she examined [8]. Narayanan and Shmatikov demonstrated that records in a large, sparse collection of film ratings could be matched to named people using only a few outside facts [9]. For movement, de Montjoye and colleagues found that four approximate points in space and time were enough to single out most individuals among 1.5 million phone users [4], and a similar result followed for credit-card transactions [10]. Rocher et al. later estimated that a model using fifteen ordinary demographic attributes could correctly re-identify almost everyone in the United States within a sample of that kind [11]. Ohm drew the legal lesson that the confidence placed in anonymisation had been misplaced [12].

2.4 Privacy Theory, Fairness, and Regulation

Several frameworks help organise these concerns. Solove's taxonomy separates harms that arise from collection, processing, dissemination, and intrusion [13]. Nissenbaum's idea of contextual integrity holds that information flows carry norms tied to the setting in which they began, so a harm occurs when data move into a setting that breaks those norms, even if the data were never secret [14]. Acquisti and co-authors explained why individuals struggle to manage this alone: preferences are uncertain, depend on context, and are easily swayed by the way choices are presented [15]. Barocas and Selbst showed that models trained on behavioural data can produce discriminatory outcomes without any intention to discriminate [16]. On the regulatory side, the European General Data Protection Regulation defines personal data broadly enough to cover online identifiers [17], and India's Digital Personal Data Protection Act, 2023 places duties on those who decide why and how data are processed [18]. For this reason the present paper treats legal compliance as a floor and not as a complete answer.

3. HOW A TRACE TRAVELS FROM DEVICE TO DECISION

To see where exposure arises, it helps to follow one trace through a typical service. There are five steps: generation, logging, linkage, modelling, and use. The person influences only the first, and then only partly. The remaining four take place on infrastructure that the person cannot see.

3.1 Generation

Any interaction with a device produces events: a touch on the screen, a network request, a sensor reading, a transaction. Most events are created automatically and carry a timestamp, a device or account identifier, and often a coarse location. The user's intention concerns the task, such as finding a route or reading the news, and not the by-product that the task leaves behind.

3.2 Logging

Servers and applications write events into logs for debugging, billing, security, and analytics. Third-party components embedded in an app or a web page may send the same events to additional parties, so a single tap can be recorded by several organisations [19]. Retention periods are set by the operator, and logs kept for troubleshooting often outlive the purpose for which they were created.

3.3 Linkage

Linkage joins records that belong to the same person across services and over time, using an account number, an advertising identifier, an email address, or a device fingerprint. Eckersley showed that the combination of browser settings, fonts, and plug-ins can be distinctive enough to recognise a browser without any cookie [20]. Linkage is the step that turns many scattered facts into a single profile, and it is where the cost of apparently harmless data changes.

3.4 Modelling

Features are built from the linked record: counts, durations, frequencies, sequences, and relations between people. Classifiers or recommenders are then trained to predict a target such as a click, a purchase, a loan default, or a personal attribute. Because the target need not be something the person ever stated, the model can produce knowledge about them that existed in no single record.

3.5 Use and Feedback

The predictions steer what the person sees next, whether advertisements, prices, rankings, or offers, and the person's response becomes new exhaust. This loop means that a system does not merely observe behaviour but also shapes it. An experiment on a large social network, in which the emotional tone of users' feeds was adjusted, showed that such shaping can measurably change what people go on to post [21].

4. EVERYDAY ACTIVITIES AND WHAT THEY REVEAL

The activities below are sorted into eight families. For each one the discussion covers what is recorded, what can be inferred, and what makes the trace revealing. The reasoning rests on published studies and on how data pipelines are commonly built. It is meant to guide assessment, not to claim that every service draws every inference listed.

4.1 Web Browsing and Search

Search queries state intentions in plain words, including worries a person would not voice aloud. Browsing history points to interests, employer, health concerns, and financial stress, and the time of day at which pages are opened suggests sleep and work patterns. Third-party trackers extend the record beyond the sites a person knowingly visits: a measurement of one million websites found tracking to be widespread, including techniques that survive the deletion of cookies [19]. Even without an account, the configuration of a browser can serve as an identifier [20].

4.2 Location and Movement

A phone that is carried all day produces a nearly continuous record of where its owner sleeps, works, worships, studies, and seeks treatment. Human movement is regular enough to be highly predictable. One analysis of mobile-phone records estimated that, in principle, as much as 93 per cent of a person's whereabouts could be anticipated [22], and earlier work found that people keep returning to a small set of familiar places [23]. The same regularity makes location traces hard to anonymise, as the four-point result above indicates [4]. Two facts, such as a place where the phone rests at night and a place where it spends working hours, are often enough to identify its owner.

4.3 Social Media and Messaging

Likes, follows, reactions, and the timing of posts expose both the person and their contacts. Models built on such records predict sensitive attributes and personality [3], [6]. Even when the content of messages is encrypted, the metadata, meaning who contacted whom, when, how often, and for how long, remain visible to the carrier. A study of telephone records collected from volunteers found that call and text metadata alone could point to sensitive matters such as medical conditions [5].

4.4 Shopping and Payments

A receipt records what was bought, where, when, and for how much. Purchases hint at health, household composition, income, and religious observance. Card data are sparse yet distinctive: just four purchases, with approximate time and place, identified most people in a set of 1.1 million cardholders [10]. Loyalty schemes and digital wallets tie these records to a stable identity and so make linkage easy.

4.5 Streaming and Media Consumption

Viewing and listening histories reflect mood, language, politics, and stage of life. The work on film ratings shows that a list of preferences behaves like a fingerprint, because few people share the same combination of lesser-known titles [9]. Pausing, skipping, and rewatching add fine-grained signals of engagement that the viewer seldom regards as data.

4.6 Smart Homes and Wearables

Connected speakers, thermostats, cameras, plugs, and fitness bands record when people are at home, awake, active, or unwell. Apthorpe et al. showed that even when the traffic of smart-home devices is encrypted, its rate and timing can reveal when particular devices are in use, which discloses daily patterns to anyone able to observe the network connection [24]. Wearables generate heart-rate, sleep, and activity series that researchers use for health monitoring [7], so their exhaust is at once valuable for care and sensitive.

4.7 Smartphone Apps and Sensors

Lists of installed apps, screen-on times, battery state, accelerometer readings, and Wi-Fi scans are rarely seen by the user. An app list can hint at religion, health conditions, or finances, and motion sensors can indicate activity and the mode of transport. Because permissions are requested one at a time and described in terms of function, people do not easily see what the combination allows, and the choices they make are shaped by context and presentation [15].

4.8 Device and Interaction Behaviour

The way a person types, swipes, or moves a pointer is a behavioural signature. Keystroke timing was proposed as a means of authentication as early as the 1990s [25], which means that the same signal can in principle recognise a person across accounts. Network addresses, device models, language, time zone, and screen size are logged by default and, taken together, help to re-link sessions [20]. This family is the most invisible, because no content is involved and the person may not know that any measurement takes place.

5. FRAMEWORK FOR ASSESSING INFERENCE RISK

A fair comparison across activities needs a common yardstick. The framework below is independent of any one dataset or service, so it can be applied to a single app, to a whole organisation, or to a personal audit.

5.1 Scope and Participants

The first step is to define the unit of analysis, which may be a person, a household, or a device. The activities that unit performs in a typical week are listed together with the services involved and the parties that receive data, whether directly or through embedded components. For any empirical use, a sample of at least ten volunteers with different routines, devices, and occupations is advisable, and informed consent and a data-handling plan should be in place before anything is collected (see Section 8).

5.2 Activity Inventory

Table 1 lists the typical traces for each family. For every trace the assessor should note how long it is retained, which identifiers are attached to it, and whether it leaves the device.

Table 1. Typical traces recorded for each activity family

Activity family	Typical traces recorded
Web and search	Queries, pages visited, time spent, referring sites, cookies, and tracker requests
Location	Coordinates, cell and Wi-Fi identifiers, visits to places, travel routes and times
Social media	Likes, follows, posts, reactions, contact lists, message metadata
Shopping and payments	Items, amounts, merchants, places, times, loyalty or wallet identifiers
Streaming	Titles, play and pause events, skips, repeat views, search within the service
Smart home and wearables	Device on and off events, voice requests, heart rate, sleep, step counts
Apps and sensors	Installed apps, screen-on periods, battery state, motion and Wi-Fi scan data
Device behaviour	Typing rhythm, touch patterns, network address, device model, language, time zone

5.3 Quantifying Identifying Information

Identifiability can be expressed in bits. If an attribute value a occurs with probability $P(a)$ in a population, its surprisal is defined as follows.

$$I(a) = -\log_2 P(a)$$

Rare values carry more bits. A person's gender gives about one bit, whereas the day of the year on which they were born gives at most about $\log_2 365 \approx 8.5$ bits. To single out one person among N individuals, about $\log_2 N$ bits are needed, which is roughly 33 bits for a population of eight billion. If attributes were independent, their bits would add up, and a profile holding K attribute values a_1 to a_K would identify a person once the total B satisfied the following.

$$B = \sum_k I(a_k) \geq \log_2 N$$

Real attributes are correlated, so the sum overstates the information, and the expression serves as a rough guide and not as an exact count. A direct empirical alternative is unicity, the share of people in a dataset who are the only ones matching p randomly chosen points of their own trace [4]. With $M_p(u)$ denoting the set of users whose traces contain the p points drawn from user u , and D the set of users, unicity is

$$U_p = |\{u \in D : M_p(u) = \{u\}\}| / |D|$$

5.4 Inference Models

To measure what can be inferred, standard classifiers are trained on the traces to predict a sensitive attribute, using only the users who disclosed that attribute for evaluation. Logistic regression, random forests, and gradient-boosted trees are suitable baselines because they are quick and easy to inspect. Evaluation should use users who were not seen in training, otherwise the memorising of individuals would be mistaken for genuine inference. Comparing

several models helps separate the exposure created by the data from the skill of one particular algorithm.

5.5 Metrics

Table 2 defines the measures. Information bits and unicity describe identifiability, the area under the ROC curve describes inference, and linkage rate, retention time, and egress share describe how far and for how long a trace travels. Using all of them gives a fuller picture than any single figure.

Table 2. Assessment metrics

Metric	Definition	Purpose
Information bits	Sum of the surprisal values of the attributes found in a trace	Rough size of identifying content
Unicity	Share of users uniquely matched by a few sampled points of their own trace	Risk of re-identification
Inference AUC	Area under the ROC curve when predicting a sensitive attribute from traces	Strength of inference
Linkage rate	Share of records in one source that can be matched to a record in another	Ease of building a profile
Retention time	Period for which a trace is stored	Duration of exposure
Egress share	Share of events that leave the device	Reach of the trace

Every reported value should carry a confidence interval, for example from bootstrap resampling over participants. Because the same people are measured under several settings, such as raw, coarsened, and noised data, differences between settings should be checked with a paired test such as the Wilcoxon signed-rank test, which does not assume normally distributed scores.

6. INTERPRETING THE EVIDENCE

6.1 Finding Where Exposure Begins

When a profile turns out to reveal too much, the first question is where in the pipeline the trouble begins. Three kinds of exposure can be told apart. In collection exposure, more is gathered than the service needs. In linkage exposure, records that were harmless apart are joined into one profile. In inference exposure, conclusions are drawn that the person never stated. Location and browsing mostly create collection and linkage exposure, social and shopping traces mostly create inference exposure, and device behaviour mainly strengthens linkage.

The distinction matters because the remedies differ. Collection exposure calls for minimisation, linkage exposure for separation of identifiers and short retention, and inference exposure for limits on purpose, aggregation, and audits. Table 3 summarises the connections.

Table 3. Activities, typical inferences, and main safeguards

Activity	Typical trace	Likely inference	Main safeguard
Web and search	Queries, pages, timing	Interests, health concerns, daily schedule	Tracker blocking, short log retention
Location	Place and time points	Home, workplace, routines	Coarse location, on-device processing
Social media	Likes, follows, posts	Personality, beliefs, social ties	Audience controls, purpose limits
Shopping	Items, place, amount	Health, income, household	Data minimisation, tokenised identifiers
Streaming	Titles, pauses, skips	Mood, language, politics	Separate profiles, history deletion
Smart home and wearables	Device events, vital signs	Presence, sleep, illness	Local processing, traffic padding
Apps and sensors	App lists, sensor readings	Habits, finances, mobility	Permission review, sensor limits
Device behaviour	Typing rhythm, address, settings	Re-identification across sessions	Fingerprint resistance, uniform settings

6.2 Likely Order of Importance

The true ranking depends on the person and the service, yet the mechanisms suggest a general pattern. Location and payment traces tend to rank high for identification because they are sparse but highly distinctive, so a handful of points already singles out a person. Search and social activity rank high for inference because their content is closest to intent and belief. Streaming and smart-home traces are less distinctive but accumulate quickly and expose

routine. Device behaviour is the easiest to underestimate, since it allows everything else to be linked.

6.3 Combined Effects

Activities seldom stand alone. A location trace places a person at a clinic, a payment trace shows a pharmacy purchase an hour later, and a search that evening mentions a medicine. Each fact is unremarkable, but together they suggest a condition. The effect resembles the way bits add up in Section 5.3: independent sources contribute new information, so the total grows faster than any one stream would suggest. This explains how a service that collects only apparently non-sensitive data can still end up knowing something sensitive, and it is the reason the framework evaluates combinations as well as single sources.

6.4 Consent and Awareness

Notice and consent is the usual control, but it works poorly for exhaust. Users cannot weigh a consequence they cannot foresee, and inferences are produced long after the moment of agreement. The findings on shifting, context-dependent preferences [15] imply that the control a person actually has depends less on the length of a privacy policy than on the defaults, on whether choices can be revisited, and on whether processing outside the stated purpose is prevented. In the language of contextual integrity [14], a flow becomes problematic when data supplied in order to obtain directions reappear where prices or job offers are decided.

6.5 Exposure over Time

Exposure also grows with time. If a service gains new identifying attributes at an average rate of r bits per day, and B_0 bits are already known from other sources, the number of days needed to reach the identification threshold is roughly

$$T \approx (\log_2 N - B_0) / r$$

The estimate is crude, because attributes overlap and the rate differs from person to person, but it makes one point clearly: retention decisions matter. Deleting old records can lower the total in the same way that new records raise it.

7. SAFEGUARDS

7.1 Individuals

Personal measures help but cannot carry the whole burden. Reviewing app permissions, switching off precise location when it is not needed, blocking third-party trackers, using a separate browser profile for sensitive searches, and deleting old histories reduce what is recorded and for how long. Choosing services with strong defaults matters more than staying vigilant on every occasion.

7.2 Organisations and Data Governance

The guiding rule is to collect only what the function needs, to state the purpose, and to delete when the purpose ends. These principles of data minimisation, purpose limitation, and storage limitation appear in the GDPR [17] and are echoed in India's Digital Personal Data Protection Act [18]. Practical steps include data inventories that show where each identifier flows, separation of identifier tables from behaviour tables, short default retention for logs, access controls with audit trails, and a review of third-party components before they are shipped.

7.3 Technical Methods

The idea of k -anonymity requires every released record to be indistinguishable from at least $k-1$ others on the identifying fields [26]. It is simple to explain, but high-dimensional behavioural traces rarely satisfy it, and the results discussed in Section 2.3 show that sparse data resist this approach. Differential privacy offers a stronger guarantee. A randomised mechanism M is ϵ -differentially private if, for any two datasets D and D' that differ in the records of one person and any set of outputs S ,

$$\Pr[M(D) \in S] \leq e^\epsilon \cdot \Pr[M(D') \in S]$$

so that the presence or absence of any individual changes the distribution of outputs only slightly [27], [28]. A smaller ϵ means stronger protection and noisier results. Federated learning keeps raw data on the device and shares only model updates [29], although the updates themselves can leak information unless other protections are added. Processing on the device, coarsening location to a district, and grouping timestamps into wide intervals all reduce the bits carried by a trace in the sense of Section 5.3.

7.4 Platforms and Devices

Operating systems can limit passive collection through permission prompts that reflect actual use, approximate location options, randomised hardware addresses, visible indicators when sensors are active, and restrictions on identifiers shared between apps. Browsers can shrink the fingerprinting surface by reporting uniform values. These measures act on the generation and logging stages and therefore benefit everyone downstream, which makes them more effective than individual choices.

7.5 Interface and Policy

Interfaces should show people what has been collected, by whom, and what was concluded from it, in plain language. Dashboards that let people download, correct, and delete their data give practical meaning to rights that would otherwise remain on paper. Impact assessments before a new inference feature is launched, audits for discriminatory outcomes [16], and clear limits on sensitive categories complete the picture. Rules should cover inferred data as well as supplied data, because the conclusion is often more sensitive than any input.

8. LIMITATIONS AND ETHICAL ISSUES

The study has limits. It is analytical, so it contains no newly collected measurements, and its statements about the size of each effect rest on published studies whose populations, platforms, and years differ from those of any particular reader. Several of the cited results come from datasets of specific countries or services and may not transfer to every setting, including India. Practices change quickly: platforms alter their logging, operating systems add protections, and laws are amended. The eight families overlap in practice, and the bit-counting model assumes an independence that real attributes seldom satisfy. Finally, the paper cannot say what any particular company infers, since such models are generally not public.

Several ethical points also deserve attention. Research on inference can itself cause harm if it is turned into a tool for profiling, so evaluations should use consenting volunteers or properly protected data, and sensitive predictions should not be published at the level of individuals. Studies that re-identify people, even to demonstrate a risk, should pass ethics review. Errors in inference may fall unevenly across groups, a concern already raised for algorithmic decisions in general [16]. Children, patients, and people monitored by employers or authorities face higher stakes and may have less ability to refuse. Anyone collecting data should obtain informed consent, keep records only as long as needed, and store them securely. It would also be wrong to conclude from this analysis that privacy is a purely personal responsibility, since much of the exposure is decided by design choices that individuals never see.

9. FUTURE DIRECTIONS

Several directions follow from this analysis. The most immediate is to apply the framework to the data exports that many services let users download, working with consenting volunteers, and to publish the results with open protocols so that other groups can compare findings. A second is to build a personal exposure dashboard that estimates, from a user's own logs, the bits and inferences a profile may hold, and shows how these change when settings are altered. A third is to study how exposure accumulates across the services used by a single household. A fourth is to examine conditions specific to India, such as shared devices, multilingual usage, and the wide use of mobile payment systems, in the light of the Digital Personal Data Protection Act. Finally, evaluating differential privacy and federated learning on realistic behavioural logs would show how much protection survives at useful accuracy.

10. CONCLUSION

Everyday digital life produces a steady flow of data that nobody set out to create. This paper traced how eight families of activity, namely browsing, location, social media, shopping, streaming, smart devices, apps and sensors, and interaction behaviour, leave traces that are logged, linked, modelled, and reused. Location and payment records identify people through their distinctiveness, social and search activity reveal through their content, and device behaviour ties the rest together. Because traces combine, judging them one at a time understates what can be learnt.

The framework proposed here joins an activity inventory, an information-based measure of identifiability, standard inference models, and metrics that include linkage and retention, and so offers a structured way to judge exposure. The safeguards at the personal, organisational, technical, platform, and policy levels suggest that much of the risk can be reduced through modest changes in defaults and design. It is hoped that the paper encourages both users and developers to look at the by-product of digital services with the same care given to the product, so that convenience does not quietly become a record of a life.

REFERENCES

- [1] V. Mayer-Schönberger and K. Cukier, *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston, MA, USA: Houghton Mifflin Harcourt, 2013.
- [2] S. Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York, NY, USA: PublicAffairs, 2019.
- [3] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," *Proceedings of the National Academy of Sciences*, vol. 110, no. 15, pp. 5802–5805, 2013.
- [4] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Scientific Reports*, vol. 3, art. 1376, 2013.
- [5] J. Mayer, P. Mutchler, and J. C. Mitchell, "Evaluating the privacy properties of telephone metadata," *Proceedings of the National Academy of Sciences*, vol. 113, no. 20, pp. 5536–5541, 2016.
- [6] W. Youyou, M. Kosinski, and D. Stillwell, "Computer-based personality judgments are more accurate than those made by humans," *Proceedings of the National Academy of Sciences*, vol. 112, no. 4, pp. 1036–1040, 2015.
- [7] J.-P. Onnela and S. L. Rauch, "Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health," *Neuropsychopharmacology*, vol. 41, no. 7, pp. 1691–1696, 2016.
- [8] L. Sweeney, "Simple demographics often identify people uniquely," Carnegie Mellon University, Data Privacy Working Paper 3, 2000.
- [9] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proc. IEEE Symposium on Security and Privacy*, 2008, pp. 111–125.

- [10] Y.-A. de Montjoye, L. Radaelli, V. K. Singh, and A. Pentland, "Unique in the shopping mall: On the reidentifiability of credit card metadata," *Science*, vol. 347, no. 6221, pp. 536–539, 2015.
- [11] L. Rocher, J. M. Hendrickx, and Y.-A. de Montjoye, "Estimating the success of re-identifications in incomplete datasets using generative models," *Nature Communications*, vol. 10, art. 3069, 2019.
- [12] P. Ohm, "Broken promises of privacy: Responding to the surprising failure of anonymization," *UCLA Law Review*, vol. 57, no. 6, pp. 1701–1777, 2010.
- [13] D. J. Solove, "A taxonomy of privacy," *University of Pennsylvania Law Review*, vol. 154, no. 3, pp. 477–564, 2006.
- [14] H. Nissenbaum, "Privacy as contextual integrity," *Washington Law Review*, vol. 79, no. 1, pp. 119–157, 2004.
- [15] A. Acquisti, L. Brandimarte, and G. Loewenstein, "Privacy and human behavior in the age of information," *Science*, vol. 347, no. 6221, pp. 509–514, 2015.
- [16] S. Barocas and A. D. Selbst, "Big data's disparate impact," *California Law Review*, vol. 104, no. 3, pp. 671–732, 2016.
- [17] European Parliament and Council of the European Union, "Regulation (EU) 2016/679 (General Data Protection Regulation)," *Official Journal of the European Union*, L 119, pp. 1–88, 2016.
- [18] Government of India, "The Digital Personal Data Protection Act, 2023 (Act No. 22 of 2023)," 2023.
- [19] S. Englehardt and A. Narayanan, "Online tracking: A 1-million-site measurement and analysis," in *Proc. ACM Conf. Computer and Communications Security (CCS)*, 2016, pp. 1388–1401.
- [20] P. Eckersley, "How unique is your web browser?" in *Privacy Enhancing Technologies (PETS)*, *Lecture Notes in Computer Science*, vol. 6205, 2010, pp. 1–18.
- [21] A. D. I. Kramer, J. E. Guillory, and J. T. Hancock, "Experimental evidence of massive-scale emotional contagion through social networks," *Proceedings of the National Academy of Sciences*, vol. 111, no. 24, pp. 8788–8790, 2014.
- [22] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.
- [23] M. C. González, C. A. Hidalgo, and A.-L. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, pp. 779–782, 2008.
- [24] N. Apthorpe, D. Reisman, S. Sundaresan, A. Narayanan, and N. Feamster, "Spying on the smart home: Privacy attacks and defenses on encrypted IoT traffic," *arXiv preprint arXiv:1708.05044*, 2017.
- [25] F. Monroe and A. Rubin, "Authentication via keystroke dynamics," in *Proc. ACM Conf. Computer and Communications Security (CCS)*, 1997, pp. 48–56.
- [26] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557–570, 2002.
- [27] C. Dwork, "Differential privacy," in *Proc. International Colloquium on Automata, Languages and Programming (ICALP)*, 2006, pp. 1–12.
- [28] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.
- [29] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," *Proceedings of Machine Learning Research*, vol. 54, pp. 1273–1282, 2017.