Yale Journal of Law & Technology
Volume 28, Symposium Issue

Anonymity, Consent, And Other Noble Lies: An Empirical Study of The Data Economy

Joel Reardon,[*] Serge Egelman,[**] Kenneth A. Bamberger,[***] Laurel E. McGrane[****]

While legal scholars have cited decades of computer science research that demonstrates why anonymity is hard (and that datasets should not be labelled as "anonymous" cavalierly), industry and legal practitioners have not heeded those warnings: many organizations trafficking in consumer data continue to assert to customers, courts, and regulators, that their data is anonymous or "deidentified." We acquired datasets from multiple data brokers to demonstrate empirically why this is false. Using publicly available email addresses found in data breaches posted on the Internet, we trivially reidentified 88% of the hashed email addresses that we obtained; using modern password-cracking techniques, we were able to reidentify 97% of the 6 million email addresses that we collected. Reidentifying hashed email addresses need not rely on illicit data or specialized hardware: by constructing rainbow tables with synthetic data representative of typical email addresses, we reidentified most of the hashed email addresses. In all cases, the hashed email addresses were linked to other device-based identifiers (e.g.,

---

[*] Associate Professor of Computer Science, University of Calgary.
[**] Director, Usable Security & Privacy Group at the International Computer Science Institute (ICSI); Research Scientist, Department of Electrical Engineering and Computer Sciences (EECS), University of California, Berkeley.
[***] The Rosalinde and Arthur Gilbert Professor of Law, UC Berkeley; Co-Faculty Director, Berkeley Center for Law and Technology.
[****] UC Berkeley, School of Law, JD '26.
Acknowledgements: This work was supported by the U.S. National Science Foundation under grant CCF-2217771 and the KACST-UCB Center of Excellence for Secure Computing. The authors would like to thank Trinity Chung, Joshua Cordeiro-Zebkowitz, Sam Croley (Chick3nman), Simon Giang (blazer), Maximilian Golla (m33x), Martin Guillen, Troy Hunt, Tin Le, Paul D. Ouderkirk (pdo), Angelina Rochon, Alex Sanchez, Michael Sprecher (hops), and Royce Williams (TychoTithonus).

mobile device advertising IDs, IPs, etc.), demonstrating why device-based identifiers have long been considered personally identifiable information.

Relatedly, organizations trafficking in this data make another assertion, that this data was collected from consumers with their consent. To evaluate this claim, we performed a survey (n=369), in which we emailed a subset of the reidentified individuals in our datasets to recruit them to participate. This survey asked participants about their recollections of having provided consent (99% had no recollection) and their feelings about the sale of their information (94% were opposed, while 77% said they planned to submit deletion requests). Overall, our study shows that hashed email addresses and device identifiers do not come close to meeting commonly understood definitions of "anonymous" or "deidentified" data, and that any notion of "consent" must also involve a similarly tortured definition. We argue that this industry and its defenders are not simply misinformed or indifferent to the veracity of their statements, but that this is an example of Plato's "noble lie": their entire social order relies on these demonstrably untrue statements being believed by courts, regulators, policymakers, and the public.

Article Contents

Introduction

> "It is the business of the rulers of the city, if it is anybody's, to tell lies, deceiving both its enemies and its own citizens for the benefit of the city; and no one else must touch this privilege."
>
> – Plato, *The Republic*

Data is considered "anonymous" when "[d]irect and indirect identifiers have been removed or technically manipulated to prevent reidentification."[1] This is in contrast to "deidentified" data, where "[d]irect and known indirect identifiers have been removed."[2] The distinction is that in the case of the former, reidentification is impossible, whereas in the latter, reidentification is *reasonably unlikely*. In computer science, it is well known that correctly anonymizing data can be a very difficult problem.[3] Almost 15 years ago, Paul Ohm warned that many claims of anonymity in industry and law are likely misstated: "[industry relies on claims of anonymity] to justify sharing data indiscriminately and storing data perpetually, all while promising their users (and the world) that they are protecting privacy. Advances in reidentification expose these promises as too often illusory."[4]

Yet, nearly 15 years later, these claims are pervasive within the data economy. Website operators and software developers post privacy policies informing users that only "anonymous" or "deidentified" data collected from their products/services will be shared with third parties. This data then makes its way to data brokers, who resell that data, while describing it as being "anonymous" or "deidentified" and having been collected with data subjects' consent (due to the mere presence of a privacy policy).

---

[1] PETER P. SWIRE & DEBRAE KENNEDY-MAYO, U.S. PRIVATE SECTOR PRIVACY: LAW AND PRACTICE FOR INFORMATION PRIVACY PROFESSIONALS 10 (3d ed. 2020).
[2] *Id.*
[3] *See, e.g.*, Arvind Narayanan & Vitaly Shmatikov, *Robust De-Anonymization of Large Sparse Datasets*, *in* PROCEEDINGS OF THE IEEE SYMPOSIUM ON SECURITY AND PRIVACY 111, 111–25 (2008).
[4] Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701, 1704 (2010).

Data brokers sell that data to businesses that use it to augment their own marketing and customer lists, so that those businesses can learn how specific individuals use specific websites and apps and the locations they frequent in the physical world.

Online identifiers—seemingly random strings of text that identify an individual Internet user—are the fuel that drives this economy. In web browsers, cookies and "fingerprints" are used to track consumers over time and across websites.[5] In mobile apps, both major platforms created "mobile advertising identifiers" (MAIDs) to be used specifically for this purpose. Yet, whole marketplaces exist in which data brokers sell lists of these identifiers tied to those individuals' app usage, web browsing history, physical locations, and other identifiers (that can then be used to query other data brokers who may have collected other data from the same data subjects). Many of the companies trafficking in this data make two specific claims: (i) that the data is anonymous (or deidentified) and that (ii) it is collected and resold with the consent of the data subjects. These same claims are being made in court filings by defense counsel, when online services are sued for violating their users' privacy by releasing this information to third parties.

We performed an empirical analysis of data brokers' anonymity and consent claims by examining data being sold by multiple data brokers. We approached multiple firms that were claiming to sell device identifiers paired with cryptographically hashed email addresses. All claimed that this data was gathered with consent. They also claimed that, despite the FTC's public warnings over a decade ago to the contrary,[6] cryptographically hashing email addresses somehow rendered them—and the data paired with them—anonymous. We received over six million data points from four different data brokers. To evaluate the purveyors' claims of anonymity, our first

---

[5] Gunes Acar et al., *The Web Never Forgets: Persistent Tracking Mechanisms in the Wild*, *in* PROCS. ACM SIGSAC CONF. ON COMPUT. & COMMC'NS SEC. (CCS '14) 674, 674–89 (2014).

[6] Ed Felten, *Does Hashing Make Data "Anonymous"?*, FED. TRADE COMM'N (April 22, 2012), https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2012/04/does-hashing-make-data-anonymous [https://perma.cc/9HNR-UHCQ].

experiment involved re-identifying the data subjects using publicly available data. We performed this experiment in three parts. First, we amassed a corpus of Internet users' email addresses by mining data breach forums so that we could perform a "dictionary attack": we built a corpus of approximately 1.7 billion email addresses found on the Internet (obtained from a single website). We hashed each address and then compared it to the hashes being sold by the data brokers. Second, we applied modern password-cracking techniques previously documented in the literature, [7] since cracking passwords similarly relies on reversing hashes through repeated guessing. Finally, to show that this can be done without any breach data or specialized hardware, we built rainbow tables, another classical attack on hash functions, containing synthetic email addresses for the most popular email providers.

Overall, we were able to reverse 88% of the hashed email addresses from the data brokers using data found online from data breaches and more than 50% without using data breach data (i.e., using only the rainbow tables). When we applied password-cracking techniques, we were able to reveal 97% of the more than six million hashed email addresses.

Using the unhashed email addresses, we performed a second experiment to evaluate data brokers' claims of consent: we emailed a random sample of the email addresses, inviting individuals to participate in a survey. We informed participants (n=369) that we contacted them because we received their email address from a data broker who was selling their information. We then asked them about their familiarity with the data broker in question, their recollections of having granted consent for the sale of their data, their opinions of this, and whether they would like to opt out of future sales (we provided relevant opt out links to survey respondents).

Our results not only shed light on this industry and the veracity of common anti-privacy arguments being made in

---

[7] Alexandra Nisenoff et al., *A Two-Decade Retrospective Analysis of a University's Vulnerability to Attacks Exploiting Reused Passwords*, *in* PROCEEDINGS OF THE 32D USENIX SECURITY SYMPOSIUM (USENIX SECURITY '23) 5127, 5127–44 (2023).

legal filings, but also have implications for regulations that aim to grant consumers privacy rights. For example, we empirically show why hashing an identifier does not impart anonymity upon it. We also show that mobile advertising identifiers are clearly being associated with personally identifiable information in ways that appear to systemically defy platform policies. [8] We also show that those trafficking in this information are operating with a tortured definition of "consent" that appears at odds with consumer understanding (and, frankly, lexicology). Worse, given that many respondents were unaware of these data transfers, the subsequent sales, or the companies selling their data, it is unreasonable to expect them to be able to exercise their rights under various privacy laws. For example, California residents can demand that regulated data brokers delete their data or discontinue sales of it,[9] but that requires that consumers be able to identify data brokers and know that sales are even taking place.

In the remainder of this paper, we provide background on notions of identifiability, linkability, and the difficulty of anonymization. We describe how dominant platforms and other technology and content firms continue to claim to courts that hashing and device-based identifiers render data anonymous for purposes of governing law. As a primer to understanding dictionary attacks, rainbow tables, and cryptographic hash functions, we also explain password storage and management.

Next, we introduce our first experiment, in which we obtained data from multiple data brokers and measured the proportion of hashed email addresses we could reidentify using data found publicly on the Internet, including applying modern dictionary attacks used to crack passwords.

We then introduce Rainbow Warrior, a tool that we created to build rainbow tables for reidentifying hashed email

---

[8] Both major mobile platforms prohibit the sharing of mobile advertising identifiers alongside other personally identifiable data without explicit user consent. *See, e.g.*, Google Play Policy Ctr., *Ads*, GOOGLE, https://support.google.com/googleplay/android-developer/answer/9857753 [https://perma.cc/C9S8-R2M8]; *App Store: User Privacy and Data Use*, APPLE INC., https://developer.apple.com/app-store/user-privacy-and-data-use/ [https://perma.cc/G45S-6WCM].

[9] CAL. CIV. CODE §§ 1798.105, 1798.120.

addresses. Using our tool, we further demonstrate why hashed email addresses and the device identifiers associated with them are readily identifiable.

Finally, we present the results of our survey and conclude with discussion of the implications for our findings.

I. Background

In this section, we provide necessary background information about concepts like "anonymity" and "deidentification," as well as well-known pitfalls. Despite clear technical definitions and best practices, we demonstrate that industry and its defenders regularly misuse these terms in legal filings and in disclosures to consumers.

A. Notions of Identifiability

Recall that data is "deidentified" when reidentification is reasonably unlikely (in contrast to "anonymous" data, which is theoretically impossible to reidentify).[10] Of course, what is considered "reasonably unlikely" may be subject to interpretation. NIST published a report about deidentification of personal information.[11] It defines *deidentification* as referring to a collection of approaches, algorithms, and tools that organizations can use to remove personal information from data that they collect, use, archive, and share with other organizations. The report points out inconsistencies among the uses of terms such as "anonymized," "pseudonymized," and "deidentified" data—even in official standards.

A key issue in defining these terms is when "deidentified data" can later become "reidentified." The report from NIST uses the term deidentified exclusively with the caveat that sometimes deidentified data can be reidentified, and sometimes it cannot. However, we disagree with this definition. Once previously deidentified data is *known* to be identifiable, it should no longer be considered deidentified; data that is

---

[10] SWIRE & KENNEDY-MAYO, *supra* note 1, at 10.

[11] NAT'L INST. OF STANDARDS & TECH., U.S. DEP'T OF COM., NIST REP. NO. 8053, DE-IDENTIFICATION OF PERSONAL INFORMATION 2–3 (2015), https://csrc.nist.gov/pubs/ir/8053/final [https://perma.cc/KXK2-DB3W] [hereinafter NIST REPORT].

simultaneously both identifiable and deidentified is an oxymoron.

Our view on deidentification is shared by various privacy laws. Beyond industry guidance, definitions of "anonymous" and "deidentified" data appear in privacy statutes and regulations. Under the California Consumer Privacy Act (CCPA), data is only considered deidentified if it "cannot reasonably be used to infer information about, or otherwise be linked to, a particular consumer." [12] Under the Health Information Portability and Accountability Act (HIPAA), data can only be considered deidentified when "there is no reasonable basis to believe that the information can be used to identify an individual."[13] Because of this, HHS requires that "device identifiers and serial numbers," "account numbers," and "any other unique identifying number, characteristic, or code" be removed from datasets for those datasets to be considered deidentified.[14]

Identifiers. The NIST report differentiates between two types of identifiers:[15] *direct* identifiers and *quasi*-identifiers. Direct identifiers are data values that directly identify a single individual. This includes names, Social Security Numbers, and email addresses. Quasi-identifiers, also called *indirect* identifiers, are values that, individually, do not directly identify an individual, but together serve as a unique "fingerprint" and can be linked with other datasets in which the same quasi-identifiers appear alongside direct identifiers, to reidentify an individual. Sweeney's seminal reidentification attack on medical records, for example, used the combination of birthday, ZIP code, and gender to link medical records to direct identifiers in another dataset.[16] She estimated that 87% of the U.S. population could be uniquely identified using only these indirect identifiers.

Pseudonymization. The consistent replacement of an identifier with an alternative value is called pseudonymization. Provided that the output value cannot be readily associated to

---

[12] CAL. CIV. CODE § 1798.140(m).
[13] 45 C.F.R. § 164.514(a).
[14] 45 C.F.R. § 164.514(b)(2)(i).
[15] NIST REPORT, *supra* note 11, at 19–22.
[16] Latanya Sweeney, *Simple Demographics Often Identify People Uniquely 2* (Carnegie Mellon Univ., Data Privacy Working Paper No. 3, 2000).

the original input, the resulting identifier *may* be deidentified. Pseudonymization is detrimental to privacy in two key ways: (i) it allows linkability among data *within* a dataset that pertains to the same individual, i.e., linking all of a particular person's "rows" in the dataset, and, (ii) it allows linkability *across* different datasets, even those managed by different entities, when they use the same pseudonymization method.

Linkability within a Dataset. Linkability among data entries within a dataset means that, while the identity of the data subject is not revealed, all their entries in the dataset are known to be about the same individual. This is also called *linkably anonymous*. Attempts at deidentifying data that nonetheless keep it linkable are often the crux of privacy disasters. Among notable ones, AOL released a dataset of its users and their searches, but with usernames consistently replaced by an arbitrary number.[17] Many users' identities were nonetheless revealed due to vanity searches or by using a collection of different searches, which individually were not identifying, but taken together identified the individual. Each such search thus reduced the entropy, a measure of uncertainty,[18] about whom the record pertained.

Linkability across Datasets. Linkability across multiple datasets can occur when there is a field common to those datasets (in database terms, this is known as a "join key"). This is trivial with direct identifiers but can still happen with pseudonymous values in two important ways. First, the same deterministic pseudonymization method can be performed on the identifying fields of multiple datasets resulting in consistent pseudonyms. Second, some arbitrary identifier can be ubiquitously used to index records during data collection by disparate data collectors. In particular, the hashed email address implements the former, and the MAID implements the latter. Linkability allows disparate data collectors to combine and share their datasets among themselves to generate more complete portraits of their data subjects.

---

[17] Michael Barbaro & Tom Zeller Jr., *A Face Is Exposed for AOL Searcher No. 4417749,* N.Y. TIMES (Aug. 9, 2006) https://www.nytimes.com/2006/08/09/technology/09aol.html [https://perma.cc/MRW9-UPU8].

[18] Claude E. Shannon, *A Mathematical Theory of Communication*, 27 BELL SYST. TECH. J. 379, 379–423 (1948).

Hashed Email Address. A hashed email address is the output of a hash function on an email address. Hash functions are functions that produce a characteristic *random-looking* value for any arbitrary input and cannot, in general, be efficiently "run backwards" to provide the input given an output. Crucially, hash functions are deterministic: the same input will always produce the same output. They are also publicly available, so everyone can determine the hash of a particular email address. Two entities that pseudonymized their datasets by hashing the email address could later link their datasets by using the "anonymous identifier"[19] as the join key to know that their records pertain to the same individual.

Risks with Hashing. The fact that hash functions cannot be run backwards, also known as the "one-way" property, is leveraged by data collectors to dismiss concerns that a pseudonym can be reidentified. Beyond the risk that data collectors can link across datasets, there is another compelling reason not to use hash functions to pseudonymize personal information. We explore this at length in this article but exemplify it here with the New York City taxicab dataset.[20] Trip data for New York taxi rides were made public, noting pickup and drop-off location and time alongside how much the riders tipped. Taxicab medallion numbers—conspicuously visible on top of the vehicle as well as on its license plate—were included "pseudonymized" by computing a hash of its value. Journalists matched images of celebrities entering a taxi with its medallion number visible and computed the deterministic hash to find their "pseudonym" alongside trip details in the published dataset.

Furthermore, one can compute the hash values of all 19,000 potentially-valid medallion numbers—not just the ones observed in the wild—and store the results in a small lookup table to allow immediate "reversing" of the hashed value.

---

[19] This term is in quotes because it should be obvious at this point that such a dataset cannot reasonably be considered "anonymous" (much less "deidentified").

[20] Marie Douriez et al., *Anonymizing NYC Taxi Data: Does It Matter?*, *in* PROCEEDINGS OF THE 2016 IEEE INTERNATIONAL CONFERENCE ON DATA SCIENCE AND ADVANCED ANALYTICS 140, 140–48 (2016).

While hash functions cannot be "run backwards," that is not the only means to determine the input for a given hashed value. It is for this reason that in 2012 the FTC warned that hashing "often fails to provide effective anonymity."[21]

Arbitrary Identifiers. These attacks on privacy would not have been possible were the pseudonymous function *truly* arbitrary, for example were each medallion number consistently replaced with a random value and the mapping of medallion number to random value were deleted.

There is a simple test to assess whether an identifier is arbitrary: when an identifier is being generated for something it will identify—e.g., a pseudonym for identifying information—*all possible* valid values for the identifier must be equiprobable. Hash functions fail this test, because their output is dictated by the input, thus only one value is eligible. This test also excludes sequential identifiers, which would allow identifiers' values to reveal temporal information, like when they were created and which of the two is older. In fact, up until recently, Social Security Numbers were issued sequentially; when researchers demonstrated that this allows them to be inferred, the Social Security Administration changed its issuance policy to randomize them.[22]

Mobile Advertising Identifiers. Both Android and iOS mobile platforms offer mobile advertising identifiers (MAIDs): Android calls theirs the "Android Advertising ID" (AAID) and iOS calls theirs the "ID for Advertisers" (IDFA). Both are arbitrary identifiers that are simply large randomly generated numbers. They can be *reset* through a mobile device's system

---

[21] Felten, *supra* note 6. After a draft of this article was presented at the 2024 Privacy Law Scholars' Conference, the FTC felt the need to restate this guidance: Office of Technology Staff. "No, Hashing Still Doesn't Make Your Data Anonymous." Office of Tech. Staff, *No, Hashing Still Doesn't Make Your Data Anonymous*, FED. TRADE COMM'N (July 24, 2024), https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2024/07/no-hashing-still-doesnt-make-your-data-anonymous [https://perma.cc/M3QR-U22S].

[22] Alessandro Acquisti & Ralph Gross, *Predicting Social Security Numbers from Public Data*, 106 PROC. NAT'L ACAD. SCI. 10975, 10975–80 (2009); *Social Security Number Randomization*, U.S. SOC. SEC. ADMIN. (2011), https://www.ssa.gov/employer/randomization.html [https://perma.cc/PGD4-4W3E].

privacy settings. This is an operation that deletes the old value and generates a new value.[23]

The resettability of MAIDs and their arbitrary nature are key components of claims made about their privacy properties. Yet data brokers are selling databases with billions of rows of people's hashed email addresses (HEMs) paired with their MAIDs. This means that users with multiple devices tied to the same email account can be linked by the unchanged HEM, as well as link multiple values of MAIDs that result from *resetting it* on the same device. This gives sobering evidence that the purported privacy properties of MAIDs are falling woefully short in practice. While MAIDs are arbitrary and non-identifying in and of themselves, they get linked with users' personal information. Ads and analytics code from third parties running in mobile apps often include the user's MAID along with personal information, such as a user's geolocation. MAIDs are also unique per device, allowing them to act as a primary key across different datasets. Data collectors who obtain different pieces of information—perhaps innocuous on its own—can combine these datasets trivially after collection, effecting a privacy harm.

Failure of Anonymization. In *Broken Promises of Privacy*, Ohm observes that a substantial body of academic work, which he calls collectively the "easy reidentification result," proves that the assumption that data can be anonymized and thereby avoid harms to subjects is deeply flawed.[24] Ohm writes that anonymization has become privacy theatre and that "bad anonymization" is redundant. Ohm also encourages the word "scrubbing" in lieu of anonymization, because scrubbing implies *effort* that can vary in intensity and because scrubbing does not mislead one into the belief that a particular result is successfully achieved. Given that some privacy laws provide safe harbor exemptions for data that is "anonymized," the easy reidentification result makes such provisions appear unreasonable if "anonymized" data does not actually provide privacy. These failures are due to the ubiquity of quasi-identifiers and both the amount and ease of access to auxiliary

[23] Though in practice, consumers rarely do this.
[24] Ohm, *supra* note 4, at 1706–07.

public data available on the Internet linked to personal information.

Ohm's article surveys a variety of reidentification examples to buttress his argument. Each involved publicly released "anonymized" datasets and some creativity on the part of researchers using other sources of data. One such case is the Netflix prize: Netflix used to publish customer viewing preferences to crowdsource research and development into better algorithms by offering a prize—provided that the winning recipient give the source code and a license to Netflix. In 2008, Narayanan and Shmatikov published their results identifying users from the dataset.[25] They leveraged the Internet movie database (IMDb) to find users' movie reviews—suggesting that they had viewed them—and matched such users based on these movies to corresponding ratings in the "anonymized" Netflix data; rough timestamps of reviews and viewing behavior aided this effort. They found that it only took a small number of such matches to become unique in the dataset, meaning that many attributes, when linkable, can serve as a quasi-identifier.

Uniqueness from a surprisingly small sample manifests in many contexts. De Montjoye et al. looked at fifteen months of human mobility data for one and a half million individuals and found that for 95% of them, it was sufficient to have four random location observations at the granularity of the nearby cell-phone tower to be unique in the dataset.[26] Golle and Partridge looked at the uniqueness of work–home location pairs using U.S. Census data[27] and observed that revealing where one lives and works at the granularity of a census tract is unique for 5% of the population and that for half the population there are at most twenty other people with the same pair. Narayanan and Shmatikov found that patterns of "followers" in Twitter could be mapped to similar patterns of "contacts" in Flickr allowing them to recognize the same user

---

[25] *See* Narayanan & Shmatikov, *supra* note 3, at 111–25.

[26] Yves-Alexandre de Montjoye et al., *Unique in the Crowd: The Privacy Bounds of Human Mobility*, 3 SCI REP. 1376, 1 (2013).

[27] Philippe Golle & Kurt Partridge, *On the Anonymity of Home/Work Location Pairs*, *in* PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON PERVASIVE COMPUTING, 2009 at 395.

across different services.[28] Given these results it is not reasonable to consider, for example, one's linkable location history to be non-identifying.

B. Arguments About Anonymity Made Before Courts

Contravening this reality, a range of dominant platforms and other technology and content firms have claimed to courts that these technologies do, in fact, render data anonymous for purposes of governing law. Specifically, they have presented three arguments.

First, they have claimed that hashing anonymizes data and prevents re-identification of the hashed information—although Meta, one of the firms making this argument, has elsewhere also (correctly) argued the opposite: that hashing does *not* anonymize data (*see infra* Section I.B.2). Second, they have contended that device-based identifiers are anonymous data. Third, they have argued that re-identification is hypothetical, and therefore that data will remain functionally anonymous.

1. Arguments that hashing anonymizes data and companies cannot re-identify the hashed information.

Google, Meta, Amazon, and Disney have all argued that the use of hashing anonymizes their data.

In *Doe I et al v. Google LLC*,[29] Google users filed a class action complaint in the federal district court for the Northern District of California, alleging that Google was unlawfully tracking, collecting, and monetizing Americans' private health information.[30] The plaintiffs also argued that Google uses this information for their own advertising purposes.[31] The class of plaintiffs claimed relief under the Electronic Communications Privacy Act, violation of California's Invasion of Privacy Act, California Constitutional Invasion of Privacy, intrusion upon seclusion, violation of California's Comprehensive Computer

---

[28] Arvind Narayanan & Vitaly Shmatikov, *De-anonymizing Social Networks*, *in* PROCEEDINGS OF THE IEEE SYMPOSIUM ON SECURITY AND PRIVACY 173, 173–87 (2009).

[29] *See* Defendant Google LLC's Opposition To Plaintiffs' Motion For Preliminary Injunction; Motion To Dismiss ("Google Defs' Opp. Motion") at 8, Doe I et al v. Google LLC, No. 3:23-cv-02431-VC, (N.D. Cal. May 17, 2023).

[30] *Id.* at 3.

[31] *Id.* at 3.

Data Access and Fraud Act, violation of California's Unfair Competition Law, trespass to chattels, statutory larceny, breach of express and implied contract, and good faith and fair dealing.[32] Google disagreed, and argued that the "Google Advertising ID . . . is a pseudonymous, device-specific string of characters that is neither tied to a user's identity nor used to personally identify a user."[33] Its brief contended in several places that hashing anonymizes its data, arguing, for instance that "Google ensures their data still cannot be 'joined' or combined with any pseudonymous data," and that it "does so by removing overlapping identifiers, creating slight errors or 'fuzziness' in the data, and encrypting the logs using separate, short-term decryption keys."[34] Google further claimed that it does not use health information for advertising, but that the company's "developers may analyze patterns in user engagement using identifiers" which are "not personally identifiable."[35] (Yet Google's website helpfully explains to its advertising customers that they can use MAIDs to "target your ads to people based on the device they're using.")[36]

Ultimately, the district court granted Google's motion to dismiss. The court ruled that "Google has admonished health care providers not to use the source code in a way that causes users' personal health information to be transmitted to Google, . . . the allegations in the complaint are too vague to support an inference that the providers have . . . caused Google to receive the plaintiffs' personal health information . . . (and) the complaint does not adequately allege that Google intends to receive this information, or that Google intends to feed the information into its own advertising machinery.[37]"

---

[32] *Id.*
[33] *Id.* at 7.
[34] *Id*. at 8.
[35] *Id.* at 6.
[36] *About Device Targeting*, Google Ads Help, https://support.google.com/google-ads/answer/1722028?hl=en [https://perma.cc/9T6L-XP26].
[37] Order Granting Motion to Dismiss at 1–2, Doe I v. Google LLC, No. 3:23-cv-02431 (N.D. Cal. May 17, 2023).

Meta has also argued that their use of hashing anonymizes user data.[38] In *In re Meta Pixel Healthcare Litigation*, a class action complaint alleges that healthcare providers sent "sensitive information about [the plaintiffs] to Meta through a common Internet tool" in violation of both HIPAA and Meta's Terms of Service.[39] The plaintiffs, patients of medical providers using Meta's Pixel,[40] sought relief under theories of breach of contract, violation of good faith and fair dealing, invasion of privacy, violation of the Electronic Communications Privacy Act, the California Invasion of Privacy Act, negligent misrepresentation, and unfair competition.[41]

Meta responded to plaintiff's allegations, by arguing that it only sends "de-identified information back to the web developer, which [uses] the data to improve its online services."[42] Further, Meta's privacy policy says that it "will hash Contact Information that you send to [it] via a Facebook JavaScript pixel for matching purposes prior to transmission," indicating that Meta uses hashing to obfuscate personal information that it receives. [43] (Once the hashed contact information is received by Meta's servers, Meta attempts to match it to the hashed contact information it stores alongside its users' Facebook profiles, thereby personally identifying them.[44]) This case is ongoing with filings as of Nov. 11th 2024. In September, the court denied Meta's motion to dismiss on the plaintiffs' Electronic Communications Privacy Act

---

[38] *See* Defendant Meta Platforms, Inc.'s Opposition to Plaintiffs' Motion for Preliminary Injunction at Exhibit D, *In re* Meta Pixel Healthcare Litig., No. 3:22-cv-03580 (N.D. Cal. June 17, 2022).
[39] *Id.* at 8.
[40] *Id.* at 9.
[41] First Amended Class Action Complaint, *In re* Meta Pixel Healthcare Litig., No. 3:22-cv-03580, at 1 (N.D. Cal. July 15, 2022).
[42] *Id.* at 1.
[43] Defendant Meta Platforms, Inc.'s Opposition To Plaintiffs' Motion For Preliminary Injunction at Exhibit D, *In re* Meta Pixel Healthcare Litigation, No. 3:22-cv-03580, (N.D. Cal. June 17, 2022).
[44] *Get Started*. Meta (last visited April 7, 2026), https://developers.facebook.com/docs/meta-pixel/get-started/ [https://perma.cc/ZB3Z-GTDU].

(ECPA)[45] and California Invasion of Privacy Act (CIPA)[46] claims.[47]

In *Hryniewicki v. Amazon Web Services*, a class of AWS users alleged that AWS violated Illinois' Biometric Information Privacy Act (BIPA) by storing "data and information that is generated as a result of the capture, collection, and processing of biometric identifiers" without complying with BIPA's notice and consent requirements.[48] The plaintiff class also alleged that AWS violated BIPA by failing to develop and make publicly available a biometric data retention and destruction policy or providing notice concerning the purpose for which her biometrics were being stored.[49] Amazon disagreed and, in an interview with the court, discussed how AWS uses hashing to protect biometric information. A hash function is described as "gibberish, [and] encrypted."[50] Ultimately this case was dismissed by the plaintiff class after their motion to remand the claim was denied.[51] AWS alleged in its final court filing that the plaintiffs failed to allege violation of BIPA.[52]

---

[45] 18 U.S.C. § 2510 *et seq.*

[46] CAL. PENAL CODE § 630 *et seq.*

[47] Order on Motion to Dismiss at 26, *In re* Meta Pixel Healthcare Litig., No. 3:22-cv-03580 (N.D. Cal. June 17, 2022). https://www.bloomberglaw.com/product/blaw/document/X2EI39AC9EH97IR6P6ST0ID6ABP [https://perma.cc/WV7G-3A8A].

[48] Notice of Removal at 1, Hryniewicki v. Amazon Web Services, Inc., No. 1:19-cv-07569 (N.D. Ill. Nov. 15, 2019). https://www.bloomberglaw.com/product/blaw/document/X3SV99M6E6S818O2E6VMSPQSUOJ [https://perma.cc/7K6S-9WYB].

[49] *Id.*

[50] Defendant Amazon Web Services, Inc.'s Memorandum Of Law In Support Of Its Rule 12(B)(2) And 12(B)(6) Motion To Dismiss at Exhibit C, Hryniewicki v. Amazon Web Services, Inc., (No. 1:19-cv-07569) (N.D. Ill. Nov. 15, 2019). https://www.bloomberglaw.com/product/blaw/document/X2JTN3BEOGN9HGB5FT473OQSDQR [https://perma.cc/8EGV-AAYD].

[51] Plaintiff's Notice of Voluntary Dismissal of Action Pursuant to Fed. R. Civ. P. 41(a)(1)(a)(i), Hryniewicki v. Amazon Web Services, Inc., No. 1:19-cv-07569 (N.D. Ill. Nov. 15, 2019). https://www.bloomberglaw.com/product/blaw/document/X4ANR88VIEN8L4BVPTVRRUK29NK [https://perma.cc/Z4Y3-9MBD].

[52] Defendant Amazon Web Services, Inc.'s Amended Response To Plaintiff's Motion To Remand, Hryniewicki v. Amazon Web Services, Inc., No. 1:19-cv-07569 (N.D. Ill. Nov. 15, 2019).

Finally, in *Robinson v. Disney Online,* Disney argued that hashed information cannot be used to re-identify individuals because the third party receiving the data in the case, Adobe, was not capable of such reidentification.[53] (Adobe, meanwhile, publicly advertises its ability to link device identifiers with email addresses and "anonymous" IDs for the purpose of "identity resolution"—a term of art in the advertising industry that means reidentifying people. [54]) That case involved allegations by a class of plaintiffs who stream Disney programming using Roku devices that "each time they use the Disney Channel to watch Disney videos or television shows, Disney discloses their personally identifiable information—including a record of every video clip viewed by the user (collectively, 'PII')—to unrelated third parties."[55] This pattern of behavior, plaintiffs contended, violated the Video Privacy Protection Act (VPPA),[56] which generally prohibits certain content providers from knowingly disclosing, to a third-party, "personally identifiable information concerning any consumer."[57]

Disney argued to the court that "plaintiff's conclusory allegation that Adobe is 'capable' of using hashed Roku device serial numbers from the [Disney Interactive (DI)] Channel to personally identify DI Channel users should be disregarded because . . . there is no basis for Plaintiff's allegation that Adobe is 'capable' of personally identifying DI Channel users by 'linking' hashed Roku device serial numbers from the DI Channel to actual Roku device serial numbers and associated

---

https://www.bloomberglaw.com/product/blaw/document/X3PUARATA4U8IAOVNML2ITU0732 [https://perma.cc/FDB5-8SRG].

[53] Memorandum Of Law In Support Of Defendant's Motion To Dismiss First Amended Complaint at 15, Robinson v. Disney Online d/b/a Disney Interactive, No. 1:14-cv-04146 (S.D.N.Y. June 09, 2014). https://www.bloomberglaw.com/product/blaw/document/X3237DPIMTV84KP4ARQCKUTBUGA [https://perma.cc/5GHE-ML9V].

[54] *Identity and Identity Graphs Overview*, ADOBE EXPERIENCE PLATFORM TUTORIALS, ADOBE EXPERIENCE LEAGUE (last updated Feb. 24, 2025), https://experienceleague.adobe.com/en/docs/platform-learn/tutorials/identities/understanding-identity-and-identity-graphs [https://perma.cc/A99M-PB7B].

[55] Robinson v. Disney Online, 152 F. Supp. 3d 176, 177 (S.D.N.Y. 2015).

[56] *Id.* at 15.

[57] 18 U.S.C. § 2710(b).

personal information in Adobe's databases."[58] Simultaneously, however, Adobe maintains a marketing website that advertises its capabilities as follows:

> Devices don't buy products. People do. Each day, you receive countless visits from customers seeking a relationship with your brand. With Adobe Experience Platform Identity Service, you can get to know the people behind the devices. Start recognizing familiar customers across unfamiliar devices — so you can deliver personal experiences every time.[59]

Indeed, instead of furthering the farce argued before the court that Adobe is not "capable" of linking device identifiers, Adobe's marketing materials detail the information it uses to specifically do just that, including through the use of device identifiers (Figure 1). (We draw specific attention to the fact that they claim to use "anonymous IDs" in their effort to personally identify people.)

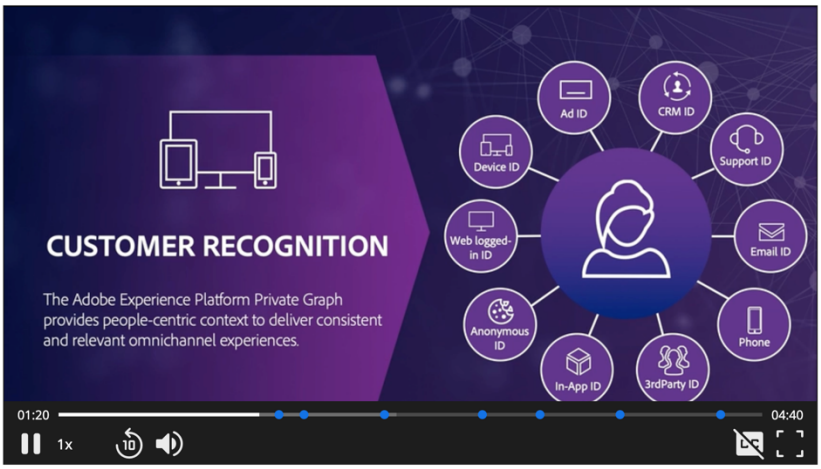

Figure 1: Adobe's marketing materials that explain how device identifiers (as well as "anonymous" identifiers—an obvious misnomer) are used to personally identify individuals. Source:

---

[58] *Supra* note 54.

[59] *Harmonized Data Access*, ADOBE FOR BUSINESS, https://business.adobe.com/products/experience-platform/identity-service.html [https://perma.cc/6H5J-TTRS] (last visited May 23, 2026).

https://experienceleague.adobe.com/en/docs/platform-learn/tutorials/identities/understanding-identity-and-identity-graphs [https://perma.cc/7PBH-7LWR].

The court dismissed Robinson's case, saying that "Robinson's allegations, as measured against the definition of personally identifiable information adopted by the Court, fail to show that he is entitled to relief."[60] However, the court also noted that "in dismissing this action, the Court is sensitive to the policy implications posed by the increasing ubiquity of digital technologies, which, as Robinson ably alleges, have dramatically expanded the depth, range, and availability of detailed, highly personal consumer information."[61]

2. Meta's arguments that hashing does not anonymize data

While many companies speciously argue that hashing anonymizes data, Meta has (correctly) made the argument, in a case it brought against ad consulting firm BrandTotal, that the use of hashing does *not* anonymize data.[62] BrandTotal engaged in "scraping"[63] as a means to collect data gathered from Meta, which Meta alleged violated Facebook's and Instagram's Terms and Policies as well as the Computer Fraud and Abuse Act, and the California Comprehensive Computer Data Access and Fraud Act.[64]

Meta raised the argument BrandTotal had violated the law in part because the hash function used on personally identified information did not adequately anonymize that data.[65] Specifically, an expert witness for Meta testified:

---

[60] Opinion & Order at 1, Robinson v. Disney Online d/b/a Disney Interactive, No. 1:14-cv-04146 (S.D.N.Y. June 09, 2014).

[61] *Id.*

[62] Complaint at 1, Meta Platforms, Inc. v. BrandTotal Ltd., No. 3:20-cv-07182 (N.D. Cal. Oct. 14, 2020).

[63] *See* Daniel J. Solove & Woodrow Hartzog, *The Great Scrape: The Clash Between Scraping and Privacy*, 113 CAL. L. REV. (forthcoming 2025) (discussing the practice of scraping, and the tension with privacy).

[64] *Id.* Meta also alleged breach of contract and unjust enrichment.

[65] Meta's Opposition to Defendants' Motion to Exclude Certain Opinions of Dr. Prowse And Dr. Thaw at Exhibit 2, Meta Platforms, Inc. v. BrandTotal Ltd., No. 3:20-cv-07182 (N.D. Cal. Oct. 14, 2020).

> I understood from our investigation that BrandTotal's claim that it anonymized user data was misleading because it used a hash function on user IDs that did not fully or securely anonymize the information collected. Even with a properly hashed user ID, for example, it is possible to correlate information from different entries bearing the same hashed user ID to discover the identity of the user.[66]

Meta also claimed that "BrandTotal's Technology has at least the effect—if not also the design—of creating a corpus of data describing individuals' preferences, attributes, or activities (a 'User Data Corpus'). BrandTotal's use of insecure hashing algorithms to 'anonymize' user IDs, . . . creates the risk that the User Data Corpus created by Brand Total could be re-associated with identifiable individual users." [67] Meta was ultimately granted an injunction and the counter claim against it was dismissed.[68] BrandTotal is now permanently enjoined from scraping, using, or selling any of Meta's data.[69]

3. Arguments that device-based identifiers anonymize data

In *Robinson v. Disney Online,* discussed above, Disney further argued that their disclosure of device-based identifiers to third parties satisfied requirements of anonymity. In particular, a brief in support of Disney argues that "Disney's alleged disclosure of Mr. Robinson's Roku device serial number along with his video viewing history did not constitute [PII, because] a hashed Roku device serial number is not PII under the VPPA because it does not without more, itself link an actual person to actual video materials." [70] The brief continued to argue that "an anonymous device identifier does

---

[66] *Id.* at Exhibit 2.
[67] *Id.*
[68] Stipulation and Order Regarding Permanent Injunction and Dismissal, Meta Platforms, Inc. v. BrandTotal Ltd., No. 3:20-cv-07182 (N.D. Cal. Oct. 14, 2020).
[69] *Id*.
[70] *Supra* note 54.

not, without more, itself identify a person, and therefore does not constitute PII under the VPPA."[71]

Accordingly, the court ultimately held that "PII is solely limited to information which, in and of itself, identifies a person. Because the anonymized Disclosures here do not themselves identify a specific person . . . they are not prohibited."[72] (We note that under this definition of PII, phone numbers, postal addresses, fine-grained location data, and even Social Security Numbers would not be considered PII.)

In *Hoge v. Southern Theaters LLC*, a class of online visitors sued Southern Theatres, alleging that "through its websites, Southern Theaters is sharing its customers' private video viewing information without obtaining the legally required consent."[73] The plaintiff class claimed that Southern Theatres "violated the VPPA through the Meta Pixel, a piece of computer code that Meta Platforms, Inc. ("Meta") provides website owners to collect information about user activity on their sites."[74] While visiting Southern Theatres' website, they claimed, the Meta Pixel transmitted the web addresses of pages visitors had made to Facebook, and sought relief under the VPPA.[75]

In response, Southern Theatres argued that "because the ordinary recipient would not readily understand the Facebook ID in the c_user[76] cookie as identifying a specific person, the information Plaintiff alleges Southern Theatres disclosed to Facebook does not constitute PII."[77] Southern Theatres claims that "to the ordinary recipient, an IP address or a digital code in a cookie file would likely be of little help in trying to identify

---

[71] *Id.* at 11.
[72] Robinson v. Disney Online, 152 F.Supp.3d 176, 184 (S.D.N.Y. 2015).
[73] Original Class Action Complaint at 1, Hoge v. Vss-Southern Theatres LLC., No. 1:23-cv-00346 (M.D.N.C. April 26, 2023).
[74] *Id.*
[75] *Id.* at 1.
[76] The "c_user" cookie stores each Facebook user's numeric Facebook profile identifier. To identify a user's Facebook profile, anyone (an "ordinary person") can just visit https://www.facebook.com/<c_user value> [https://perma.cc/QUB2-KNKE], which is likely to yield the individual's name and profile picture (which are publicly available at this URL by default, if they have a Facebook account).
[77] *Id.* at 12.

an actual person."[78] Southern Theaters then notes that both the "Third and Ninth Circuits have found 'static digital identifiers' like 'IP addresses, browser fingerprints, unique device ID numbers, and cookie identifiers' do not constitute PII."[79] This September, the defendant's motion to dismiss was granted when the district court agreed that the plaintiff failed to state a claim under the VPPA, as "Congress did not include movie theaters within the purview of the VPPA."[80]

In *Eichenberger v. ESPN*, the U.S. Court of Appeals for the Ninth Circuit held that the device-based identifiers used for reidentification purposes are not PII unless they readily permit an "ordinary person" to identify a specific individual.[81] *Eichenberger* again involved a class action brought by Roku device users—this time those using the using Roku to stream ESPN.[82] As the court described, a plaintiff in this class would have "downloaded the WatchESPN Channel on his Roku device and used it to watch sports-related news and events. He did not consent to Defendant's sharing his information with a third party. "But every time Plaintiff watched a video, Defendant knowingly disclosed to a third party, Adobe Analytics: (1) Plaintiff's Roku device serial number and (2) the identity of the video that he watched."[83] The plaintiffs argued that even though the serial number was not personally identifiable, it could be reidentified to him.[84] The court considered reidentification with a device based identifier and held that "personally identifiable information" means only that information that would "readily permit an ordinary person to identify a specific individual's video-watching behavior."[85]

The court acknowledged that "today's technology may allow Adobe to identify an individual from the large pool by

---

[78] *In re* Nickelodeon Consumer Privacy Litigation, 827 F.3d 262, 283 (3d Cir. 2016).
[79] Defendant's Brief In Support of Its Motion to Dismiss the Complaint Under Rule 12(B)(6) at 13, Hoge v. Vss-Southern Theaters LLC., No. 1:23-cv-00346 (M.D.N.C. April 26, 2023).
[80] Order at 3, Hoge v. Vss-Southern Theaters LLC., No. 1:23-cv-00346 (M.D.N.C. April 26, 2023).
[81] Eichenberger v. ESPN, Inc., 876 F.3d 979, 985 (9th Cir. 2017).
[82] *Id.* at 981.
[83] *Id.*
[84] *Id.*
[85] *Id.* at 985.

using other information" but ultimately decided this was not something the ordinary person could do.[86] The court thus dismissed the case "because the information described in Plaintiff's complaint does not constitute "personally identifiable information" under the VPPA."[87]

The "ordinary person" standard for PII is widely criticized because it treats identifiability as if it were obvious and human-readable—like a name or a Social Security number—rather than something that emerges from data processing. By asking whether a hypothetical layperson could identify someone from the disclosed information alone, courts applying this test ignore how identification actually works in modern data ecosystems, where companies routinely combine device identifiers, browsing histories, and other metadata to pinpoint individuals. This creates a gap between legal doctrine and technical reality: data that is trivially linkable in practice may be deemed "non-identifying" in law. As a result, the standard can under-protect privacy, incentivize superficial "de-identification" practices (like replacing names with persistent IDs), and allow extensive tracking to fall outside statutory protections simply because the identifying step is performed by machines rather than an "ordinary" human observer.

U.S. courts are divided on this issue. The Second and Third Circuits align with the Ninth Circuit's "ordinary person" approach, holding that information qualifies as PII only if it would readily identify a specific individual to a typical person without additional tools or data.[88] In contrast, the First Circuit applies a broader, more functional test, asking whether the information is "reasonably linkable" to an individual—especially in the hands of the recipient.[89] Under this alternative

---

[86] *Id.*

[87] *Id.* at 986.

[88] *See* In re Nickelodeon, 827 F.3d 262, 284 (3rd Cir. 2016) ("[O]ur review of the legislative history convinces us that Congress's purpose in passing the Video Privacy Protection Act was quite narrow: to prevent disclosures of information that would, with little or no extra effort, permit an ordinary recipient to identify a particular person's video-watching habits."); Solomon v. Flipps Media, Inc., 136 F.4th 41, 51 (2d Cir. 2025) (2d Cir. 2025) ("[W]e adopt the Third and Ninth Circuits' ordinary person standard.").

[89] *See* Yershov v. Gannett Satellite Info. Network, Inc., 820 F.3d 482, 486 (1st Cir. 2016) (defining PII in this contaxt as "information reasonably and foreseeably likely to reveal which ... videos [a person] has obtained.").

approach, data like unique device identifiers or viewing histories may count as PII if they can be combined with other datasets to identify someone, even if they appear anonymous in isolation. This circuit split reflects a deeper disagreement about whether privacy law should focus on how information appears on its face or on what can realistically be done with it in context.

Similarly in *Ellis v. Cartoon Network, Inc.*, the U.S. Court of Appeals for the Eleventh Circuit also held that device-based viewing records are not PII, and therefore that their disclosure did not violate the VPPA.[90] This case involved a class action lawsuit against Cartoon Network by viewers who used the Network's free CN app on Android smartphones to watch video clips. [91] Without consent, Cartoon Network kept full records of the videos watched and shared those records with a third-party data analytics company called Bango.[92] Cartoon Network provided Bango both with users' Android IDs, and with their video viewing records. [93] Nonetheless, the Court rejected Ellis' claim for relief under the VPPA, in part, on the grounds that "Mr. Ellis' Android ID and video viewing records were not 'personally identifiable information' . . . because they did not, 'in [their] own right, without more, link an actual person to actual video materials.'"[94] The court reasoned that someone would have "to take additional steps to match the Android ID to Mr. Ellis" and "[a]lthough the district court acknowledged that an Android ID is 'unique to each user and device,' it was not akin to a name."[95]

Because the plaintiff in this case did not actually re-identify the information, the court determined this was evidence that the information was not PII.[96] This leaves open the door for re-identification by sophisticated actors.[97] The court also ruled that the plaintiff was not a "subscriber" because he simply

---

[90] Ellis v. The Cartoon Network, Inc., 803 F.3d 1251, 1254 (11th Cir. 2015).
[91] *Id.*
[92] *Id.*
[93] *Id.*
[94] *Id.* at 1254–55
[95] *Id.* at 1255.
[96] *Id.*
[97] *See id.*

downloaded a free app to watch free content.[98] Because the information was not PII and Mr. Ellis was not a "subscriber under the VPPA," the case was dismissed.[99]

Finally, in *Frasco et al. v. Flo Health et al.,*[100] defendant Flo Health, the developer of a period and pregnancy tracking app, was accused of sharing users' personally-identifiable usage information with third-party advertisers. At trial, Flo argued that when they transmitted users' device identifiers—specifically, mobile advertising identifiers (MAIDs)—alongside users' health information to various advertising companies, this data was somehow de-identified:[101] "It was sent not with a name or contact information, but with that device ID [which is] not tied to a profile."[102] "They didn't send any personally-identifiable information…they used a device identifier to shield a user from that information…[they] uniquely identify a device, but that's all they uniquely identify."[103]

Of course, Meta, one of the recipients of this data, helpfully explains to their advertising customers that they "can create ads *targeting people* by customer lists, and one way to make such a list is *by using mobile advertiser IDs* [emphasis added]."[104] Indeed, Meta's documentation specifically states that "there are several technologies [Meta's customers] can use to identify a person," including the "Apple Advertising Identifier (IDFA)" and "Android Advertising ID."[105]

---

[98] *Id.* at 1258.
[99] *Id.*
[100] Frasco et al. v. Flo Health, Inc. et al., No. 3:21-cv-00757-JD, 2022 WL 21794391 (N.D. Cal. June 6, 2022).
[101] *See* Flo Health, Inc.'s Motion to Dismiss at 15, Frasco et al. v. Flo Health, Inc. et al., No. 3:21-cv-00757-JD (N.D. Cal. Jan. 29, 2021).
[102] Trial Transcript Vol. 1 at 133:5–7, July 21, 2025. Frasco et al. v. Flo Health Inc. et al., No. 3:21-cv-00757-JD (N.D. Cal.).
[103] *Id.* at 664:16–665:2.
[104] *Targeting by Mobile Advertiser IDs*, META (last visited Mar. 13, 2025), https://developers.facebook.com/docs/app-ads/targeting/mobile-advertiser-ids/ [https://web.archive.org/web/20250405061323/https://developers.facebook.com/docs/app-ads/targeting/mobile-advertiser-ids/].
[105] *Id.*

4. Arguments that theoretical possibility of re-identification should not prevent treating data as anonymous

Finally, the issue of how to treat the possibility of re-identification of "anonymized" data, discussed above in *Ellis* and *Eichenberger*, has been echoed in a number of other contexts in which parties, including Google and NBC, have argued to courts that re-identification should be treated as merely hypothetical, and therefore de-identified data should continue to be treated as anonymous.[106] For example, in *C.A.F. v. Viacom, Inc.*, a class action of Viacom users raised a claim alleging that Viacom's practice of sharing information about users with Google violated the VPPA.[107] In a brief opposing *certiorari* from the Third Circuit's dismissal of these statutory claims, Google contends that "petitioners' allegations thus cannot support liability under VPPA because [t]he allegation that Google will assemble otherwise anonymous pieces of data to unmask the identity of individual children is, at least with respect to the kind of identifiers at issue here, simply too hypothetical."[108] *Certiorari* was denied.[109] (Meanwhile, Google continues to advertise its ability to use its Analytics product to track individual users across the web for the purpose of targeting individuals with advertisements.[110])

State v. Johnson & Johnson, where Washington state sued Johnson & Johnson over consumer-protection violations related to their role in the opioid crisis, raised the issue differently.[111] In discovery, Washington produced 11 years of data from a database of all Medicaid claims, and Johnson & Johnson "moved the court to compel the State to supplement the Medicaid claims data with the month and day of the

---

[106] *See* Brief in Opposition at 10, C.A.F. v. Viacom, Inc., sub nom. *In re* Nickelodeon Consumer Privacy Litig., 827 F.3d 262 (3d Cir. 2016), *cert. denied*, 580 U.S. 1048 (2017).
[107] *Id.* at 5.
[108] *Id*. at 10.
[109] C. A. F. v. Viacom Inc., 580 U.S. 1048 (2017).
[110] *Enabling Remarketing with Google Analytics Data*, GOOGLE (last visited Oct. 31, 2025), https://support.google.com/analytics/answer/9313634?hl=en [https://perma.cc/VH6Q-FEKQ].
[111] State v. Johnson & Johnson, 536 P.3d 204 (Wash. Ct. App. 2023).

services and prescriptions."[112] Washington objected on the grounds that the information could then be re-identified by Johnson & Johnson, which would violate HIPAA.[113] Johnson & Johnson disagreed and submitted an expert witness declaration that said there was virtually no risk of reidentification as reidentification is merely hypothetical.[114] The expert testified that "[the re-identification] risk . . . is de minimis, if indeed any nonzero risk exists at all. . . ."[115]

The state brought its own expert witness who empirically proved reidentification possible: "through progressive experiments Dr. Sweeney was able to demonstrate how 191 hospice patients in the Medicaid dataset uniquely matched 191 named records."[116] A state appeals court agreed with the State,[117] concluding that "releasing full dates created a risk of reidentifying Medicaid patients that was not small enough to be acceptable under HIPAA."[118]

Finally, in *Afriyie et al v. NBC Universal Media*, NBC and Adobe similarly argued that the hypothetical possibility of re-identification of PII by a third party should not result in a determination that data was no longer anonymous[119] (notwithstanding the fact that the hypothetical possibility of re-identification definitionally means that data *cannot* be deemed "anonymous"). This case, like others, involved allegations that a defendant (here NBC) transmitted data, including video viewing history and personally identifiable information, to third parties such as Adobe.[120] In a motion to dismiss, defendants point to the speculative nature of reidentification, arguing that "plaintiffs don't allege whether Adobe can link

---

[112] *Id.* at 206–07.
[113] *Id.* at 207.
[114] *Id.* at 208.
[115] *Id.*
[116] *Id.* at 210.
[117] *Id*. at 206.
[118] *Id.*
[119] *See*, *e.g.*, Part I.
[120] Class Action Complaint at 3, Afriyie v. NBC Universal Media, LLC., No. 1:23-cv-09433 (S.D.N.Y. Oct. 26, 2023).

Video IDs to specific videos—only that 'someone' can do it in some undescribed way."[121] The case is ongoing.[122]

C. Claims About Anonymity and Deidentification Made to the Public

Erroneous claims of data anonymity or deidentification are not just limited to legal filings. We identified several online services that transmitted hashes of various types of personal information, while making claims about that data being "anonymized" or "deidentified." We offer several in this section as examples, in no particular order.

Merriam Webster. We tested the behavior of the Merriam-Webster webpage by providing an email address to enroll in the "Word of the Day" newsletter in 2025.[123] We observed that the hash of the entered email address was promptly shared with multiple third parties. While assessing whether these behaviors were disclosed in its privacy policy, under the heading "advertising and interest-based, behavioral advertising," we found this telling passage:

> We also may rely on third-party service providers to do things like: take user-level information, anonymize it through hashing (a hashed email address might look something like this: e820bb4aba5ad74c5a6ff1aca16641f6) and match it against other anonymized, people-based identifiers – doing this helps us serve you personalized, relevant advertisements.[124]

Merriam Webster is saying the quiet part out loud: it admits to "anonymizing" email addresses by hashing them and then using those hashes as master keys to link them to other

---

[121] Defendants' Motion to Dismiss Class Action Complaint at 2, Afriyie v. NBC Universal Media, LLC., No. 1:23-cv-09433, (S.D.N.Y. Oct. 26, 2023).
[122] Response re: 47 Notice (Other) of Supplemental Authority at 1, Afriyie v. NBC Universal Media, LLC., No. 1:23-cv-09433, (S.D.N.Y. Oct. 29, 2024).
[123] *See, e.g.*, Merriam Webster, https://www.merriam-webster.com/ [https://perma.cc/VXL9-HZXT].
[124] *Privacy Policy*, MERRIAM WEBSTER, https://www.merriam-webster.com/i/privacy-policy, 2024 [https://perma.cc/LF2H-B46T].

"people-based identifiers" for targeted advertising. Of course, definitionally, these hashes cannot be deemed deidentified, much less anonymous: not only is it not "reasonably unlikely" that they will be used to reidentify individuals, but they are in fact being collected and shared for that explicit purpose! Thus, this data cannot be called anonymous either (i.e., not only is it possible that individuals will be reidentified, it is the *raison d'être* for this data collection).

This Merriam Webster example was found and used before we created Rainbow Warrior or even embarked on this study. In fact, it was somewhat of a motivation: the claim of anonymity is being made by an entity presumably in a reasonable position to understand accurate definitions of words. While a dictionary attack against Merriam Webster's claims was considered, we opted instead to reverse the above hash (quoted as an example in their privacy policy) by guessing. While this approach may be impractical at scale, we reversed it on our very first guess: example@gmail.com.

WebMD. We visited the website for WebMD,[125] a health information provider and found juxtaposed statements that it both cared about our privacy and that it shared our unique identifiers to process personal information with 57 third parties. Among those third parties that receive personal information from WebMD is TikTok Insights. On the WebMD homepage, users are given the option to enroll in an email newsletter providing "[d]octor-approved health and wellness information."[126] If this is done, TikTok becomes aware of the email address that was entered: TikTok "automatically identifies form fields on pages where the Pixel is installed, and hashes and collects the customer information entered on those pages."[127] This is done alongside the website that is visited, i.e., the URL, which can include things like user search queries or the medical information being sought.

ByteDance's website explains that WebMD data is collected to reidentify website visitors (by comparing the email address with email addresses associated with TikTok profiles):

---

[125] *See* WEBMD, https://www.webmd.com/ [https://perma.cc/3ZSG-PL5D].
[126] *Id.*
[127] *About Advanced Matching for Web,* TIKTOK (last visited April 7, 2026) https://ads.tiktok.com/help/article/advanced-matching-web [https://perma.cc/94E7-4BKE].

"automatically find customer information and match it with people on TikTok" and "TikTok will use hashed information to link event information to people on TikTok."[128] Of course, WebMD can simply not automatically disclose its users' medical curiosities to TikTok by not including TikTok's tracking tools on its webpages.

TikTok is not the only party that learns the WebMD browsing and search history tied to a particular email address. The hostname bh.contextweb.com is sent both the hash of the user's email and every webpage they visit on WebMD, including the free-form search feature. Simply visiting contextweb.com, as well as investigating Internet site ownership information[129] reveals that this data collection is done by PulsePoint. PulsePoint asserts that by "leveraging proprietary datasets and methodology, PulsePoint targets healthcare professionals and consumer populations with an unprecedented level of accuracy."[130]

Taco Bell. A recent visit to the Canadian version of Taco Bell's website allowed us to observe that users can sign in with their email addresses. Users who do so unknowingly send the SHA256 hash of their email address to www.facebook.com through Meta's Pixel.[131] Versions of the website that we visited on prior dates allowed visitors to sign up with phone numbers, which were also hashed and transmitted to Meta via the Pixel.

Unlike emails, phone numbers are limited worldwide to 15 digits,[132] making constructing comprehensive hash-reversing

---

[128] *FAQs for Advanced Matching for Web*, TIKTOK (last visited Dec. 4, 2025), https://ads.tiktok.com/help/article/faqs-for-advanced-matching-for-web [https://perma.cc/WC3F-BWE8].

[129] *See* WHOXY, https://www.whoxy.com/company/10693758 [https://perma.cc/EUJ6-8S6Y].

[130] *PulsePoint Achieves Exponential Growth Propelled by Pipeline of Innovation and 100% Customer Satisfaction,* PULSEPOINT (Sept. 6, 2023), https://www.pulsepoint.com/blog/pulsepoint-achieves-exponential-growth-propelled-by-pipeline-of-innovation-and-100-customer-satisfaction [https://perma.cc/AQZ3-BCEX].

[131] *See* TACO BELL (last visited April 7, 2026), https://www.tacobell.ca/ [https://perma.cc/U39Q-NN2S].

[132] INT'L TELECOMM. UNION, *The International Public Telecommunication Numbering Plan*, ITU-T Recommendation E.164, at 5 (Nov. 2010), https://www.itu.int/rec/T-REC-E.164-201011-I/en [https://perma.cc/2RE2-PM37].

dictionaries trivial. For example, most North American phone numbers begin with a one and are followed by ten digits, meaning that there are only 10 billion possible values.[133] It is easy to list the entire domain of phone numbers. We did this for SHA256 and the entire lookup dictionary for North America is 420 GiB: ten billion entries of length 42 where 10 bytes are for the number and 32 bytes are for the hash.

Any claim that the hash of a small number cannot be reversed through a brute-force attack is an outright falsehood. (Though again, this should not be news, as the FTC publicly explained this over a decade ago.)[134] It is difficult to reconcile that a well resourced technology company would incorrectly use basic cryptographic primitives, particularly giving the clear guidance from the FTC, unless providing secrecy for the phone numbers and email addresses was not the motive behind hashing them.

Gravatar. Gravatar, short for "globally recognizable avatar,"[135] is a service that converts email hashes into personal information curated at the user's discretion. Gravatar users create public profiles, which can include their full names, links to their social media profiles, email addresses and phone numbers, and a profile picture. The idea is that participating social media sites can hash a user's email to retrieve a relevant profile picture to use on their website (without that user having to register at each and every site). A number of popular websites do this, which is how we first discovered them: we were looking for hashes of our own email addresses while we were browsing the web to measure the scale of this problem in practice.

One of Gravatar's commonly asked questions is: "Why do I have a Gravatar account if I don't remember creating one?" Its answer: any user with a WordPress account—from a sister company—automatically gets a Gravatar account.[136] If you

---

[133] *About The North American Numbering Plan Administrator (NANPA),* NANPA (last visited DATE HERE), https://www.nanpa.com/about [https://perma.cc/ZT6N-YK75].

[134] *See supra* note 6.

[135] *See* GRAVATAR, https://gravatar.com/ [https://perma.cc/9DDC-YD5N].

[136] *Data Privacy FAQs*, GRAVATAR (last visited DATE HERE), https://web.archive.org/web/20250128101414/https://support.gravatar.com/privacy-and-security/data-privacy/.

have ever had a thumbnail icon of yourself effortlessly show up after registering on a website, you likely have a Gravatar account, too.

Users of this service may be unaware that the identifier Gravatar happens to use is the same that is trafficked by data brokers, meaning they are unwittingly helping reidentify themselves in entirely unrelated datasets. Astonishingly, Gravatar's own privacy policy states that they may "share a hashed version of your email address to facilitate customized ad campaigns on other platforms" under the heading of "deidentified information," all the while supporting a public service that maps email address hashes to photographs and other personal information, in effect, easily and readily reidentifying the user based on a hash of their email address.[137]

D. Password Storage

Cryptographic hashing is a reasonable security mechanism when implemented correctly. To better understand how hashing is supposed to be used, we now explain best practices for secure password storage. To provide background, we explain how hashing is used for password authentication. Passwords are ubiquitous for authenticating into computers.[138] This is achieved by having the authentication server—the computer that is being logged into—store information about all its valid users and their passwords. This information is stored in a password file and is consulted during authentication to assess the validity of a user's password.[139]

A primary concern about such systems—indeed one that has existed for nearly half a century[140]—is the secrecy of the password file. A typical threat is an adversarial *attacker* who compromises a system and is able to access the password file

---

[137] *Privacy Policy*, AUTOMATTIC, https://automattic.com/privacy/ [https://perma.cc/2NVP-HUQL].
[138] *See* Joseph Bonneau et al., *The Quest to Replace Passwords: A Framework for Comparative Evaluation of Web Authentication Schemes*, *in* PROCEEDINGS OF THE 2012 IEEE SYMPOSIUM ON SECURITY AND PRIVACY 553, 553–67 [https://perma.cc/2TFN-HDXC].
[139] *See generally* Robert Morris & Ken Thompson, *Password Security: A Case History*, 22 COMMUN. ACM 594, 594–597 (1979) [https://perma.cc/CSS5-TU2Q].
[140] *Id*.

on that system. Ideally, they should not learn usernames and passwords as a result of this compromise. Thus, storing the password file in plaintext is not suitable. But encrypting the file is also not useful: due to the continual need to authenticate users, either the encryption key or the decrypted password file must be available in memory, making it available to the attacker who can read the encrypted file.

Consequently, security best practices are that passwords are not only *hashed* but *salted* and further hashed with a *time-intensive* hash function before being stored in a password file.[141] Passwords should have sufficient complexity to not be readily guessed by an attacker.[142] Additionally, the ability to attempt authentication is *rate limited*, meaning that an attacker who has not compromised the password file can only make a few guesses before the account is frozen for some time.[143] Best practices also state that when there is evidence that the password file has been *compromised*, the passwords for all affected users should be *changed*.[144] This is because once the hash of a password is known, it is only a matter of time before the password is known. Password storage practices serve to give a longer buffer so the user has time to change their password after a breach. It is assumed a leaked hash will eventually be reversed; best practices only buy time.

Hash Functions. A hash function is an algorithm that deterministically transforms an arbitrarily long binary string to a fixed-length concise string, often called a digest or hash. A cryptographic hash function is a hash function that has a cryptographically large range, such as 256 bits. Hash functions have a few properties, but relevant for our discussion is "pre-image resistance," also known as the "one-way" property, and which informally means that given the output of the hash

---

[141] *See* Paul A. Grassi et al., NAT'L. INST. of STANDARDS & TECH., U.S. DEP'T OF COMMERCE, NIST SPECIAL PUBLICATION 800-63B, NIST DIGITAL IDENTITY GUIDELINES: AUTHENTICATION AND LIFECYCLE MANAGEMENT 16 (2017), https://nvlpubs.nist.gov/nistpubs/specialpublications/nist.sp.800-63b.pdf [https://perma.cc/6QRK-RS4J].
[142] *Id.* at 13.
[143] *Id.* at 25.
[144] *Id.* at 14.

function on some input, one cannot run it backwards to get the corresponding input.

Brute Force. Hash functions are public and can be easily run forwards. An obvious approach to finding hash function pre-images is to enumerate the domain of possibilities and hash each one until one is found that matches the desired value. A brute-force attack on known hash values is always effective when the domain of possible passwords is not *cryptographically large.*[145] For example, if one knows that a hash is the result of the outcome of a coin flip, one could hash "heads" to see whether the output matches and conclude whether it is heads or tails. Incidentally, Facebook claims that by hashing the values "m" or "f" to represent users' genders, this somehow preserves privacy.[146] More generally, the most time-intensive part of mounting a pre-image attack on a hash value is the time to enumerate and hash the entire domain of possible values. This strategy works even when there are many possible values.[147]

A key factor in the success of brute-force attacks is the rate at which an adversary can check possible pre-images—the faster the adversary can hash, the sooner they arrive at a pre-image. Password storers can slow the attacker down, for example, by hashing the input multiple times (e.g., thousands or millions). This makes mounting the brute-force attack more

---

[145] When we write "cryptographically many" or "cryptographically large," we precisely mean that the total number is so large that a brute-force attack on the entire domain is computationally infeasible.

[146] *See Advanced Matching*, META (last visited DATE HERE), https://developers.facebook.com/docs/meta-pixel/advanced/advanced-matching/ [https://perma.cc/U6HF-639T]. Instead of preserving privacy, hashing simply encodes male as 62c...c5a and female as 252...111, and is functionally equivalent to any other publicly-known encoding, indeed, the use of "m" and "f" in the first place is already such an encoding.

[147] *See* Patrick Gage Kelley et al., *Guess Again (and Again and Again): Measuring Password Strength by Simulating Password-Cracking Algorithms*, *in* 2012 IEEE SYMPOSIUM ON SECURITY AND PRIVACY 523, 523–37 [https://perma.cc/5JVQ-87MK]; Matt Weir et al., *Testing Metrics for Password Creation Policies by Attacking Large Sets of Revealed Passwords*, *in* ACM CONFERENCE ON COMPUTER AND COMMUNICATIONS SECURITY, 162, 162–175 (2010) [https://perma.cc/ESW6-7ZZ2].

time consuming, while keeping the user-perceived delay in logging in with a password imperceptible.[148]

Computational Asymmetries. Attackers can react to slower hash functions by increasing their computational power through specialized commercial graphics cards, widely used in AI training, that can run the hash function many times with different guesses simultaneously. Such computational improvements only bear benefit to the attacker, because the system trying to check whether a single user's password is correct gets no benefit from being able to compute multiple hashes at once.

The defense against special-purpose hardware is special-purpose hash functions designed for hashing passwords for password files. These make it hard to develop cheap single-purpose hardware to compute, reducing extreme differentials between the hash rates of the attacker and the defender. The winner of the password hashing competition in 2015 was Argon2 and its use remains the best practice for password hashing (as of 2025).[149] The need for time-intensive hashing algorithms extends to any circumstances in which one is hashing something from a small domain, such as emails, with the goal of delaying the inevitable success of a brute-force search.

Dictionary Attacks. With appropriate hash functions, the attacker cannot leverage specialized hardware and so breaking a single password can take a long time. Yet the attacker can "remember" the mapping from hash function input to output while working. Thus, while they still expend the effort required to compute an Argon2 hash, they can store the result to avoid expending the effort again in the future. This can be useful for

---

[148] OWASP Cheat Sheet Series, *Password Storage Cheat* Sheet, https://cheatsheetseries.owasp.org/cheatsheets/Password_Storage_Cheat_Sheet.html [https://perma.cc/N6GK-TGKJ].

[149] *See id.*; Alex Biryukov, Daniel Dinu & Dmitry Khovratovich, *Argon2: New Generation of Memory-Hard Functions for Password Hashing and Other Applications*, IEEE EUROPEAN SYMPOSIUM ON SECURITY AND PRIVACY 292, 292–302 (2016).

an attacker who is continually trying to break passwords, or to offer it as a service.[150]

Hash Chains. One problem with dictionary attacks is the storage requirements. A billion entries require gigabytes; a trillion, terabytes. Building dictionaries for the most plausible entries may have reasonable storage costs, but building complete dictionaries for large finite domains or extensive dictionaries for substantive coverage of vast domains may require infeasible storage.

In 1980, Hellman introduced the idea of a time–memory tradeoff for performing dictionary attacks, [151] effectively "compressing" the storage requirements for the dictionary at the cost of increasing the lookup time. Oechslin later proposed "rainbow tables" as a mechanism to mitigate practical inefficiencies with Hellman's design, and rainbow tables remain the best tool for this purpose.[152] In short, they create a "random walk" starting at a password, going to its hash, using the hash to deterministically select a new password, and so on. Only the start and end of the walk are stored, but all the passwords and their hashes visited on the walk are "stored" because one can always reconstruct the walk by going forwards and rewinding to the start once the end is reached.

Salt. The defense against rainbow tables and dictionary attacks in general is to salt the password before hashing it. A salt is simply a random value from a cryptographically large space that is concatenated to a password. Salts are not meant to be secret and are typically stored alongside the hash value in the password file, where each user has their own unique salt.

Salting passwords has a number of benefits. First, two people with the same password will not have the same salt, so the resulting hashes will differ. This means that an attacker who compromises a large number of passwords cannot look for sets

---

[150] *See, e.g.*, HASHES.COM, https://www.hashes.com [https://perma.cc/3YKF-F46E] (last visited Dec. 2, 2025) (an example of a website that breaks hashes).

[151] Martin Hellman, *A Cryptanalytic Time-Memory Trade-Off*, IEEE TRANSACTIONS ON INFORMATION THEORY, 26(4): 401, 401–406 (1980) [https://perma.cc/8WEH-3S8S].

[152] Philippe Oechslin, *Making A Faster Cryptanalytic Time-Memory Trade-Off*, ANN. INT'L CRYPTOLOGY CONF. 617, 617–630 (2003).

of users that seem to share a password under the assumption that they independently chose the same common password that will be easier to guess. This also means that two different entities computing a salted hash will arrive at a different value, unlike how a hashed email serves as a join key. Second, it means that dictionaries or rainbow tables built for unsalted passwords cannot be used to expedite brute forcing of the password. Given a particular salt, however, it is still possible to construct a dictionary or rainbow tables to reverse passwords that use that salt.

Salting precludes the utility in creating rainbow tables. Provided every user gets a random, cryptographically large salt, a rainbow table for a particular salt will only ever be useful to find a pre-image for that single value, and it is infeasible to create rainbow tables for every salt due to the cryptographically large space of salts. Since building a rainbow table is simply a brute-force attack with the results stored to expedite the attack in the future, it makes little sense to store the results once a pre-image is obtained, because the salt should never be reused. Thus, salting hashes forces the attacker to expend the same effort as a brute-force attack to find a pre-image every time a pre-image is sought, giving time to the administrator to observe the breach of hashed values and forcing the users to adopt new passwords.

Summary. According to best practices, passwords should be hashed before they are stored in order to prevent someone who obtains the password file from determining the original password. [153] Password storage best practices state that passwords must be hashed using hash functions that are designed to be slow to compute.[154] This is because the space of plausible passwords is relatively small, so brute-force attacks are feasible in practice. Slow hash functions make attacks more time consuming for the attacker and preclude specialized hardware that can be leveraged to greatly increase the attacker's advantage.

Password storage best practices also require salting passwords. This is because salting passwords prevents dictionary and rainbow table attacks where the cost of

---

[153] *See* Morris & Thompson, *supra* note 139.
[154] *See* Grassi, *supra* note 141.

computing the slow hash function is only borne once and can later be used to perform arbitrarily many brute-force attacks. Note that the notion of salting passwords appears in the original description of a password-based authentication system—nearly fifty years ago[155]—because it was obvious that passwords themselves would not be sufficiently random to thwart brute-force attacks.

Given these best practices, it would be unreasonable to store unsalted passwords hashed with unsuitable hash functions, while holding out that your hashed passwords are not reasonably likely to be reversed (i.e., the definition of deidentified data). Yet, the hashed email addresses freely shared and trafficked by data brokers are exactly that: unsalted and hashed with hash functions unsuitable for use with passwords. The mere act of *sharing* password files is a violation of best practices: password files, containing only (salted) hashes, are never intentionally shared or published online except in the context of cyber security training, e.g., hacking exercises meant to demonstrate how feasible brute-force attacks are to perform in practice.

Moreover, email addresses are often *intentionally* chosen by users to be *memorable* and *identifying*. Emails tend to match usernames or are permutations of letters and names. Unlike passwords, users selecting email addresses are not advised to make them "unguessable" or forced to include numbers and special characters in a bid to make them harder to brute force from their hashes. These familiar "password requirements" *only exist* to defeat a brute-force attack *after the compromise of a password file*: without compromise, the attacker can only make guesses at the rate that the service will allow them, and it is best practice to rate limit online guesses after some number of failed attempts or to require a second factor. This is all to say that while hashing is instrumental in keeping users' passwords secure, just because data is hashed does not magically imbue it with security and/or privacy properties. These factors conspire to render the entire endeavor of hashing an email address purposeless from a privacy and security standpoint.

There is no basis to expect hash functions to magically achieve the one-way property for email addresses while doing

---

[155] *See* Morris & Thompson, *supra* note 139.

none of the best practices required to achieve *the very same* goal for passwords. The exact same requirements that apply to password storage to prevent easily reversing them apply to hashed emails and they apply for the exact same reasons. This observation is so otherwise obvious that no one has bothered publishing it in the computer science literature—aside from marginal remarks about how it goes without saying.[156] The extravagances of surveillance capitalism regarding the use of hashed emails and other identifiers, as well as the depth that it has pervaded privacy discussions, has compelled us to comprehensively refute it.

II. Evaluating Privacy Claims

Despite decades of research and warnings from experts,[157] entire marketplaces exist for trafficking in consumers' sensitive data, all personally identifiable by way of hashed email addresses. As an example, Figure 2 depicts a screenshot of a product listing that was posted by a data broker on Datarade,[158] a marketplace for data brokers.[159] This particular listing is for a database of 650 million email addresses that are paired with consumers' first and last names, physical addresses, IP addresses, phone numbers, and birthdates. Each of these records also includes the hashes of the email address in three different formats: MD5, SHA1, and SHA256.[160] Thus, an ordinary person who has obtained only a hashed email address may compare it against this dataset (or others similarly being sold by numerous other data brokers). If the hashes match—

---

[156] *See, e.g.*, Steven Englehardt, Jeffrey Han & Arvind Narayanan, *I Never Signed Up for This! Privacy Implications of Email Tracking*, 2018 PROC. PRIVACY ENHANCING TECHS. 109, 109–26, https://doi.org/10.1515/popets-2018-0006 [https://perma.cc/9CGG-GXCE].

[157] *See, e.g.*, *supra* notes 6 & 4.

[158] *See* DATARADE, *Stirista*, https://datarade.ai/data-products/email-address-data-email-database-us-consumers-564-milli-stirista (last visited April 6, 2026) [https://perma.cc/3CKU-BT5B].

[159] *See* DATARADE, https://www.datarade.ai/ [https://perma.cc/6SQV-26RC].

[160] This suggests that the data broker has the actual email address to compute these values, as there were no gaps or missing entries in the dataset would one expect were the HEMs being matched opportunistically.

because the underlying email address is one of the 650 million contained in this dataset—the underlying email address is trivially revealed (along with the owner's real name, date of birth, phone number, physical address, and other personal information).

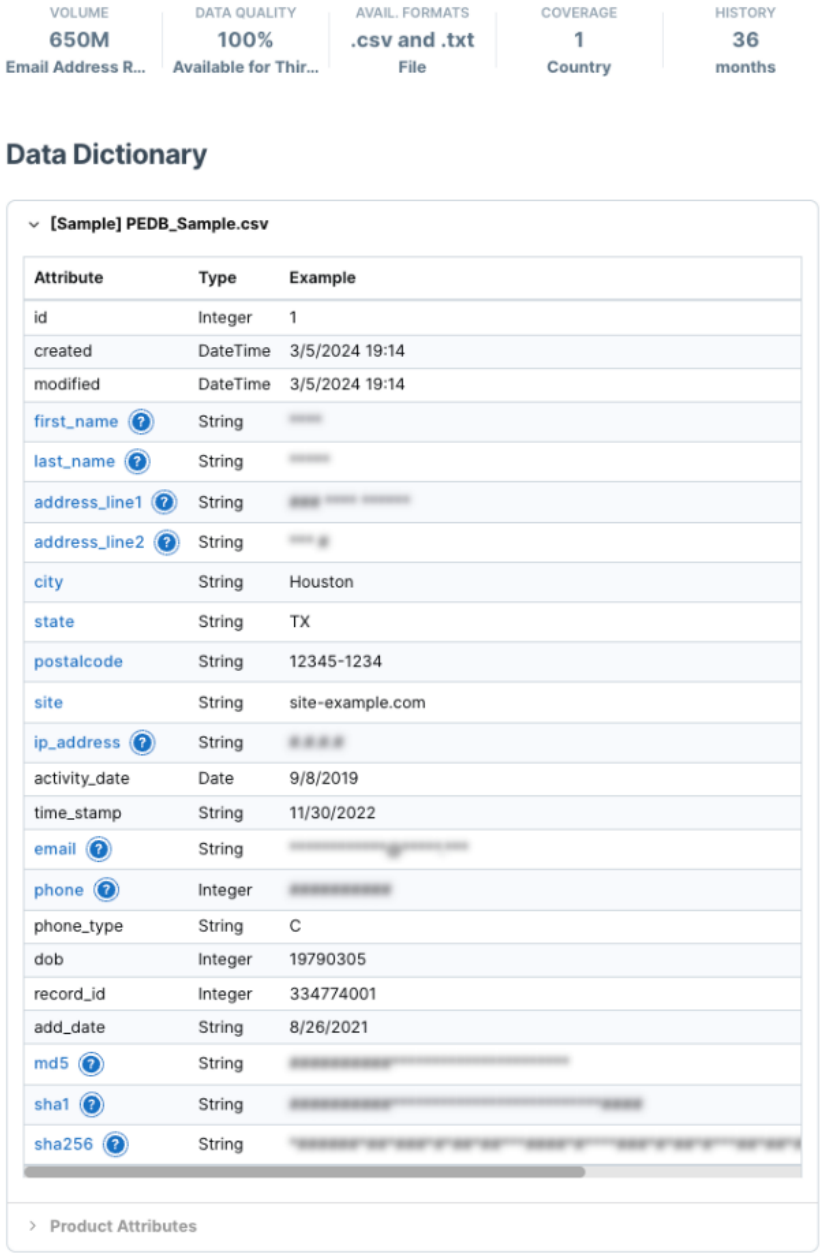

Figure 2: Screenshot of a dataset being sold on the data broker marketplace, Datarade, in which email address hashes are

paired with other personal information and identifiers (names, email addresses, physical addresses, birthdates, IP addresses, etc.). Source: https://datarade.ai/data-products/email-address-data-email-database-us-consumers-564-milli-stirista (last visited April 6, 2026) [https://perma.cc/3CKU-BT5B].

We performed an initial experiment to quantify the proportion of hashed email addresses that can be reidentified using only information found publicly on the Internet. To that end, we conducted a dictionary attack: we built a corpus of 1.7 billion email addresses found publicly on the Internet and then hashed each using multiple algorithms (i.e., SHA1, SHA256, and MD5), in both uppercase and lowercase. We then examined the hashed email addresses separately received from data brokers to see if any matched the hashes of email addresses in our dictionary. Overall, we observed that 42% of the hashed email addresses received from data brokers could be found in our dictionary, thereby revealing the plaintext email addresses. Spending more time and effort to build a larger dictionary would no doubt yield a greater reidentification rate.

A. Methodology

We identified data brokers selling datasets that included hashed email addresses and mobile device identifiers by perusing Datarade,[161] AWS Marketplace,[162] and Snowflake Marketplace.[163] Many of the data brokers trafficking in this type of data offer free samples, which we discovered varies from hundreds to millions of rows. Separately, we downloaded publicly available copies of data breach datasets found on the Internet to compile a corpus of email addresses (i.e., the "dictionary"). In this section, we describe how we used this corpus of email addresses to reverse the hashes of the email addresses found in the data brokers' datasets.

---

[161] *See supra* note 159.
[162] *Amazon Web Services*, AMAZON, https://aws.amazon.com/marketplace [https://perma.cc/BPA9-U7DG] (last visited Dec. 4, 2025).
[163] *See* SNOWFLAKE, https://www.snowflake.com/en/data-cloud/marketplace/ [https://perma.cc/65YX-AHAB].

1. The Hashed Email Dictionary

The website, "have i been pwned?"[164] tracks recent data breaches and allows website visitors to search and see whether their email addresses were included in any. Using this list of recent breaches, we searched the Internet for the underlying data. In an hour or two of searching, we found data revealed by the following data breaches:

- Adobe:[165] 152,472,418 email addresses.
- Collection #1:[166] 160,279,158 email addresses.
- People Data Labs:[167] 180,538,242 email addresses.
- RaidForums:[168] 1,570,223,764 email addresses.
- Sony:[169] 801,685 email addresses.
- Twitter:[170] 16,059,992 email addresses.

In total, this accounted for 1,654,410,807 unique email addresses. We estimate that this is roughly 30–40% of all email

---

[164] *See* HAVE I BEEN PWNED?, https://haveibeenpwned.com/ [https://perma.cc/7AEK-FL5G].

[165] *See* Troy Hunt, *Adobe Credentials and the Serious Insecurity of Password Hints*, TROY HUNT (Nov. 12, 2013), https://www.troyhunt.com/adobe-credentials-and-serious/ [https://perma.cc/6MSS-XB4L].

[166] *See* Troy Hunt, *The 773 Million Record "Collection #1" Data Breach*, TROY HUNT (Jan. 17, 2019), https://www.troyhunt.com/the-773-million-record-collection-1-data-reach/ [https://perma.cc/4REC-XLJS].

[167] *See* Troy Hunt, *Data Enrichment, People Data Labs and Another 622M Email Addresses*, TROY HUNT (Nov. 23, 2019), https://www.troyhunt.com/data-enrichment-people-data-labs-and-another-622m-email-addresses/ [https://perma.cc/JA8L-62ME].

[168] *See* Lawrence Abrams, *New Hacking Forum Leaks Data of 478,000 RaidForums Members*, BLEEPINGCOMPUTER (May 29, 2023, 9:55 PM), https://www.bleepingcomputer.com/news/security/new-hacking-forum-leaks-data-of-478-000-raidforums-members/ [https://perma.cc/EG2C-N3RS].

[169] *See* Troy Hunt, *A Brief Sony Password Analysis*, TROY HUNT (June 6, 2011), https://www.troyhunt.com/brief-sony-password-analysis/ [https://perma.cc/Q3L3-VE6H].

[170] *See* Lawrence Abrams, *200 Million Twitter Users' Email Addresses Allegedly Leaked Online*, BLEEPINGCOMPUTER (Jan. 4, 2023), https://www.bleepingcomputer.com/news/security/200-million-twitter-users-email-addresses-allegedly-leaked-online/ [https://perma.cc/EM8J-6HD5].

addresses in existence.[171] Using this list of email addresses, we computed six different cryptographic hashes for each: every email address was hashed using the SHA1, SHA256, and MD5 algorithms, both as lowercase and uppercase characters. Thus, we built a lookup table allowing us to map these hashes back to plaintext email addresses.

2. Data Sets Containing Hashed Emails

Shortly after we began this project, the FTC filed its initial complaint against Kochava,[172] a data broker accused of unfairly collecting and trafficking in consumers' mobile location data. While not directly related to our study, many data brokers suddenly required stringent contractual terms (including confidentiality and indemnification) prior to sharing sample data. Thus, we limited our study to four data brokers who provided us with a cumulative 6,040,775 rows of free sample data—consisting of mobile advertising identifiers (MAIDs) paired with hashed email addresses—without us having to sign any agreements. While these are clearly not a random sample of all data brokers (i.e., they are ones willing to provide free samples without us having to sign legal agreements), the types of data that they offer are clearly representative of the types of data that other data brokers are selling (i.e., nearly all of the data brokers we encountered post schemas describing the data on offer). We describe the collected datasets as follows.[173]

Dataset 1. This dataset consisted of 100,000 rows and included hashed email addresses alongside the following mobile device information (suggesting that it was collected

---

[171] *See* The Radicati Group, *Number of E-mail Users Worldwide from 2018 to 2028*, STATISTA https://www.statista.com/statistics/255080/number-of-e-mail-users-worldwide/ [https://perma.cc/KT97-GPB2].

[172] *See* Press Release, FED. TRADE COMM'N, *Agency Alleges that Kochava's Geolocation Data from Hundreds of Millions of Mobile Devices Can Be Used to Identify People and Trace Their Movements*, (Aug. 29, 2022) https://www.ftc.gov/news-events/news/press-releases/2022/08/ftc-sues-kochava-selling-data-tracks-people-reproductive-health-clinics-places-worship-other [https://perma.cc/L5SL-ZGEF].

[173] To preserve their privacy, we have hashed all data broker names.

either by a mobile app or code from third parties embedded within a mobile app):[174]

- Timestamp device was first seen
- Timestamp data point was collected
- Mobile advertising ID
- Device type (Android vs. iOS)

Dataset 2. This dataset consisted of 10,692 rows and included hashed email addresses alongside the following mobile device information (again suggesting it was collected from a mobile app):[175]

- Timestamp data point was collected
- Mobile advertising ID
- Device type (Android vs. iOS)
- Device make/model
- GPS coordinates
- IP address

Dataset 3. This dataset consisted of 4,930,083 rows of hashed email addresses paired with mobile advertising identifiers (MAIDs), suggesting that the data was collected from mobile devices.[176]

Dataset 4. This dataset consisted of 1,000,000 rows and included hashed email addresses alongside the following mobile device information (suggesting it was collected from a mobile app):[177]

- Timestamp data point was collected
- Mobile advertising ID
- Device type (Android vs. iOS)

---

[174] The dataset was provided by a company uniquely identified as "0d35b601b619836e857975f713adbb997c38188a."
[175] The dataset was provided by a company uniquely identified as "1c5d990a196d52668e67edd3db1cede9c700b584."
[176] The dataset was provided by a company uniquely identified as "31a597dc19c6f84372c914d10ad217d9f302ce30."
[177] The dataset was provided by a company uniquely identified as "6f5267d993df44e3d4a77a12ca69d609e5b28313."

- GPS coordinates
- IP address

B. Results

Dataset	Total Emails	Matched (%)
1	100,000	53,630 (53.6%)
2	10,692	5,402 (50.5%)
3	4,930,083	2,011,151 (40.8%)
4	1,000,000	473,946 (47.4%)
Total:	6,040,775	2,544,129 (42.1%)

Table 1: For each of the four datasets, the number of hashed email addresses contained in that dataset and the proportion that were reidentified using our dictionary attack.

Overall, across all four datasets (

Table *1*), we were able to reidentify 42% of the roughly six million hashed email addresses that we received, with minimal time invested. This number obviously scales with the size of the dictionary and is consistent with our initial estimate that our dictionary includes roughly 30–40% of all email addresses.

Spending more than a few hours, one could presumably amass a much larger dictionary and therefore a larger proportion of hashed email addresses could be reidentified. The data above suggests that a corpus of 5 billion unique email addresses is likely sufficient to reidentify the vast majority of the data we received from data brokers.[178]

Despite the fact that all four datasets came from different data brokers, they all had fairly comparable reidentification rates. While there is not 100% overlap (as would be the case if they were all reselling the same dataset), reidentification rates ranged from 40–54%. We would expect the reidentification rate to grow linearly with the size of the dictionary (e.g.,

[178] E-mail from Troy Hunt, *Creator of have i been pwned?* (June 29, 2024) (on file with authors). Hunt confirmed that "have i been pwned?" stores 5.7 *billion* unique email addresses found across all reported data breaches. As described in the next paragraph, because a dictionary of that size contained 88% of the hashes we provided, this suggests that a corpus of 6.5 billion email addresses is likely sufficient to reidentify nearly every email address.

spending additional effort to double the size of the dictionary, one would expect to see reidentification rates double).

After performing this initial analysis, we contacted Troy Hunt, the proprietor of the "have i been pwned?" website,[179] and asked whether he would be willing to search his entire dataset for our hashes. While he graciously agreed, he was only able to use Datasets 1, 2, and 4, as Dataset 3 used deprecated MD5 hashes ("have i been pwned?" only stores SHA-1 hashes). Nonetheless, of the more than 1.1 million hashed email addresses that we provided, he reported a match rate of 88%. Given that our initial dictionary attack showed comparable hit rates across all four datasets, we have no reason to doubt that the match rate for Dataset 3 would be comparable to the others.

To recap, because of the likelihood of an individual's email address being involved in a data breach, we have shown that most hashed email addresses can be trivially reidentified simply by downloading these publicly available datasets. While there are obviously ethical issues with the use of this type of data,[180] this data is nonetheless readily available for download by members of the public (i.e., an "ordinary person"). Nonetheless, in the next section, we demonstrate that hashed email addresses can be trivially reidentified without relying on illicit (albeit publicly available) data.

C. Applying Modern Cracking Techniques

The previous section is not meant to be a comprehensive stress test of the crackability of email hashes: indeed, our point is in demonstrating that even a cursory examination using public resources yields a substantial proportion of the underlying email addresses. A more robust systematic approach is likely to yield many more—conclusively demonstrating the ridiculousness of anonymity claims.

---

[179] *Who, What, and Why*, HAVE I BEEN PWNED?, (last visited Oct. 27, 2025), https://haveibeenpwned.com/About [https://perma.cc/PS6F-AFLG].

[180] *See* Serge Egelman, et al., *It's Not Stealing If You Need It: A Panel on the Ethics of Performing Research Using Public Data of Illicit Origin*, *in* FINANCIAL CRYPTOGRAPHY AND DATA SECURITY: FC 2012 WORKSHOPS, USEC AND WECSR 2012, KRALENDIJK, BONAIRE, MARCH 2, 2012, REVISED SELECTED PAPERS 124, 124–125, (Jim Blythe, et al., 2012).

In doing such a systematic investigation, one could look to advances in password cracking research. Over the past fifty years, several papers have examined methods for more efficiently cracking hashed passwords.[181] In our case, email addresses can simply be thought of as simple passwords in that they are primarily alphanumeric, with limited symbols and generally written as username@hostname. To demonstrate this approach, the authors of a recent password cracking study offered to apply their tools towards reversing our collected email hashes.[182] They were able to reverse over 97% of the more than six million hashed email addresses that we provided them in a matter of days just by repeated guessing.

With nearly all of the email addresses revealed, we looked at some of the less prominent email providers found within our datasets. While Gmail was indeed the most prevalent, we found non-zero results for emails from the following, potentially very sensitive domain names, and whose email usernames appear to include their users' real names: @us.army.mil (457), @navy.mil (418), @nasa.gov (85), @usdoj.gov (51), @house.xx.gov and @senate.xx.gov (10), @ftc.gov (6), @nsa.gov (5), and @international.gc.ca (3). Additionally, there were 246 Proton Mail users (either @protonmail.com or @pm.me).[183] The most prevalent .edu address was @umich.edu (560), nearly double the second place of @osu.edu. It also seems that data is being collected from work devices used by healthcare workers: @kp.org (305), @providence.org (122), @sutterhealth.org (87), among others. Finally, a few transnational organizations appeared, including the @worldbank.org (57) and @redcross.org (51). Keep in mind that these numbers are only from mere millions of sample data from brokers claiming to possess billions of rows. Of course, access to cutting-edge research tools is not necessary to apply these techniques to

---

[181] *See, e.g.*, Morris & Thompson, *supra* note 139.
[182] *See* Alexandra Nisenoff, et al., *A Two-Decade Retrospective Analysis of a University's Vulnerability to Attacks Exploiting Reused Passwords, in* PROCEEDINGS OF THE 32ND USENIX SECURITY SYMPOSIUM 5127 (2023).
[183] Proton Mail (https://proton.me/mail) advertises itself as being more privacy protective than other email providers, and thus its users likely have heightened privacy concerns.

reidentify hashed email addresses: commercial services already exist, allowing an "ordinary person" to also do this.[184]

To reiterate: as of 2025, an ordinary person has access to tools and techniques to trivially reverse hashed email addresses, and therefore hashing email addresses does absolutely nothing to preserve data subjects' privacy.

III. Rainbow Warrior

We built an open-source tool called Rainbow Warrior to show how easy it is reidentify email hashes with no additional information (e.g., email addresses found in data breaches), in practice. It consists of a few thousand lines of C++ code that does the following: (i) defines and uses a syntax to describe the input domains, e.g., email addresses, by concatenating random selections from a set of possibilities, (ii) builds either a full dictionary or a rainbow table of arbitrary length to cover the input domain, (iii) provides a database format to store the data, and (iv) uses the database to reverse provided hashes if they are stored in the database. In essence, it automatically builds rainbow tables of strings that look like email addresses based on a provided format.

A. Reidentification Results

In this section, we discuss the rainbow tables that we built to match email addresses to hash values and their measured efficacy among the datasets we obtained from multiple data brokers. Our rainbow table designs were partially based on personal experience with email address formats, as well as observations we made about the formats of email addresses disclosed through data breaches.

All of our tables focus on the username component of an email address, that is, the component before the 'at' symbol. We constructed plausible email addresses for each by appending several popular domains: gmail.com, outlook.com, hotmail.com, yahoo.com, icloud.com, live.com, yopmail.com, and protonmail.com. These were nearly all selected based on popularity to increase the reidentification potential; the

[184] *See, e.g.*, HASHES.COM, https://hashes.com/en/decrypt/hash [https://perma.cc/A65K-VBFG] (last visited Dec. 2, 2025) (an example of a website that breaks hashes).

exception is protonmail.com, which was included out of curiosity as to whether people willing to pay for an email service that respects their privacy "consented" to having their personal information sold to data brokers. After reidentifying users from the data brokers' data using the breach data (§3), we added the following list of domains to our list: aol.com, comcast.net, msn.com, me.com, mac.com, sbcglobal.net, verizon.net, and att.net.[185]

Rainbow Table Name	**Description**	**Match Rate (MR)**	**Unique MR**	**Non-Breach Unique MR**
all	one from names, words, and number	23.5%	3.0%	6.1%
namelettername	<first name> <letter> <last name>	12.2%	1.1%	2.0%
allall	two from names, words, and number	12.1%	2.7%	5.5%
namename	<any name> <any name>	9.5%	0.2%	0.6%
commonfirstlastnumber	<firstname> <lastname> <0--2 numbers>	7.0%	1.2%	2.3%
letterletterlastnumbernumber	<0--2 letters> <lastname> <0--2 numbers>	6.8%	0.7%	1.4%
8letters	7 or 8 letters and optional number	5.1%	0.6%	1.2%
7something	5 letters numbers and punctuation	3.8%	0.5%	1.0%

[185] Note that not all our rainbow tables and dictionaries use this expanded set.

numbercommonfirstlast	<0--2 numbers> <firstname> <lastname>	3.4%	0.0%	0.1%
6something	5 letters numbers and punctuation	3.2%	0.2%	0.4%
allextradomains	one from names, words, and numbers for a wealth of domains	2.3%	0.2%	0.2%
wordnumber	<dictionary word> <number>	1.8%	0.1%	0.3%
6letters	1 to 6 letters	1.4%	0.0%	0.0%
alfnumname2alfnum	<alphanumeric> <name> <0--2 alphanumeric>	1.2%	0.1%	0.2%
prefixnamesuffix	<common prefix> <name> <common suffix>	1.0%	0.1%	0.1%
nameletterletter	<any name> <letter> <letter>	0.9%	0.0%	0.0%
lettername	<letter> <name>	0.9%	0.1%	0.1%
lettersnumbers	<1--3 letters> <0--4 numbers>	0.8%	0.1%	0.1%
5something	5 letters numbers and punctuation	0.7%	0.1%	0.1%
letterwordnumbers	<letter> <word>	0.6%	0.0%	0.0%
namenumbernumber	<any name> <number> <number>	0.6%	0.0%	0.0%

namenumber	<any name> <number>	0.2%	0.0%	0.0%
numberscharlname	<numbers> <letter> <last name>	0.2%	0.1%	0.1%
numbersfnamechar	<numbers> <first name> <letter>	0.2%	0.1%	0.1%
nameyear	<any name> <last 64 years>	0.2%	0.0%	0.0%
nametheword	<any name> "the" <any word>	0.1%	0.0%	0.1%

Table 2: Description of rainbow tables and dictionaries. The match rate (MR) is the percent of hashes that the tables are able to compute from a random subset of each dataset. The unique match rate is the percent of elements that were matched by this and no other rainbow table or dictionary. The non-breach unique match rate is the percent of elements that were only matched by this rainbow table and no other methods. In the descriptions, we use the following shorthand: <first name> is a random name from a large list of given names, <last name> is analogous for surnames, and <any name> is a random choice from both sets; <letter> is a letter from the English alphabet, <number> a single digit from 0 to 9, and, for example, <0—2 numbers> is either nothing, a single number, or two numbers; <alphanumeric> is either a letter or a number; <dictionary word> is a word from the English language. Finally, <common prefix> and <common suffix> are sets of characters often seen before a name, such as "ms" or "dr", or after, such as "jr" or "sr," respectively.

Table 2 gives the performance of the rainbow tables that we used. We took a random sample (n=1,000) of each of the data brokers' datasets (with the exception of one[186]) and computed three statistics. The first is the percentage of entries that were

[186] One dataset was received after we initially performed this analysis, and while we planned to redo the analysis to include this additional dataset, upon reversing 97% of the hashed email addresses using the methods described previously, we no longer felt the need.

matched by that rainbow table's synthetic data. The second is the percentage of data matched by that rainbow table that was not matched by any other dataset, including data breach data. Finally, we list the percent that was only matched by this rainbow table but not another rainbow table.

Note that some of the rainbow tables' sizes do not scale according to the size of the input space and the chain length. This is because a rainbow table can have any number of chains and the creation and use of these tables was done in an exploratory and evolving process, to assist us in trying to find useful email formats that are either not-yet reversed or as yet only reversed through data breach data. Our preference for the latter is useful to illustrate that completely synthetic data is sufficient proof that simply hashing emails is ineffective at deidentifying them: one need not be a victim of a data breach nor have access to such data. Instead, we used the data breach data to guide insights into suitable email formats; certainly others can derive insights into the structure and format of email addresses using other means.

IV. Data Subjects and "Consent"

As a third experiment in this study, we conducted an online survey. Our goal was to examine data brokers' claims that the data they are pedaling has been collected (and resold) with data subjects' consent. We sent recruitment emails to a random sample of the email addresses that we had reidentified, inviting recipients to participate in our survey.[187]

A. Methodology

Using the methods previously outlined, we were able to reidentify almost all of the hashed email addresses being sold by data brokers. From the millions of email addresses that this yielded, we randomly sampled from them to recruit for a survey. We recruited participants via an email message, informing them that we were contacting them because a data broker was selling their information and that we would like to ask them some questions about that. We sent this message in

[187] This survey was approved by the Committee for the Protection of Human Subjects (CPHS), the University of California, Berkeley's Institutional Review Board (IRB).

batches, eventually obtaining 369 completed survey responses. We told participants that we anticipated the survey to take approximately ten minutes to complete and that all completed surveys will be compensated with a $5 Amazon gift card. We also informed recipients that this would be the only email that we would send them (i.e., we did not repeatedly bug them to participate). For those interested in participating, we instructed them to click a link to open our survey on the Qualtrics platform.

After viewing the consent form and agreeing to participate, we again informed participants that we received their data from a data broker and our survey is about their awareness and opinions of this. The survey started by asking participants how surprised they were to learn that a data broker was selling their data, how concerned that made them, as well as their best guess as to how their information was obtained by the data broker. Next, we asked them whether they recalled granting consent for the collection and resale of their data, as well as whether they believed it was illegal. For those that stated that they did believe it was illegal, we asked them to name any specific laws that they believed were violated.

On another page of the survey, we asked participants to state whether they had previously heard of various companies that we listed. This list included the four data brokers from whom we obtained data, two larger data brokers (Experian and Acxiom, which we reasoned might have the best chance of being recognized, and can therefore serve as baselines), and two that we made up (Discount Data Warehouse and Eklesia). We asked participants whether they believed their email addresses could be found on the Internet, as well as several questions about their smartphone ownership and usage. As part of the latter set of questions, we asked participants whether they have ever interacted with their phone's privacy settings (e.g., to reset the advertising identifier or to opt out of tracking altogether), their usage of new app store privacy labels, their reading of privacy policies, and their beliefs about how app stores enforce privacy rules (both platform policies and relevant laws).

In the last part of the survey, we informed participants how they can send a deletion request to the specific data broker selling their data and included a link to the data broker's

privacy policy. We then asked participants if they planned to submit one, or if they plan to use any services that submit such requests on their behalf to multiple data brokers. Finally, we asked participants to complete the IUIPC[188] privacy attitudes scale, a question about their prior experiences being notified about data breaches, and then demographic information.

B. Results

Demographics. Overall, our survey respondents were fairly diverse. While every question was optional, we were nonetheless able to collect demographic information from around 85% of respondents. Upon removing data from one respondent who claimed to be under 18,[189] we observed the following age distribution:

- 18-24 years old (1%)
- 25-34 years old (11%)
- 35-44 years old (20%)
- 45-54 years old (21%)
- 55-64 years old (22%)
- 65 years and older (24%)

Regarding gender, 161 reported male (51%), 143 report female (45%), whereas 3 reported non-binary or other (1%). Education levels skewed higher than reported national statistics for bachelor's and graduate degree attainment:[190]

- High school or less (9%)
- Trade/technical/vocational training (9%)
- Bachelor's degree (40%)
- Master's degree (24%)
- Professional degree (5%)

---

[188] *See* Naresh K. Malhotra, Sung S. Kim, & James Agarwal. *Internet Users' Information Privacy Concerns (IUIPC): The Construct, the Scale, and a Causal Model*, 15 INFO. SYS. RES. 336, 351–53 (2004).

[189] If accurate, this also indicates that at least one data broker is trafficking in data collected from minors.

[190] *See* Melanie Hanson, *Education Attainment Statistics*, EDUCATION DATA INITIATIVE (last updated Jan. 14, 2025), https://educationdata.org/education-attainment-statistics [https://perma.cc/J7VM-THU6].

- Doctorate (11%)

Awareness, Appropriateness, and Consent. On the first page of the survey, we informed respondents again[191] that we were surveying them because we identified their email address in data being sold by a data broker. Thus, the first three questions asked respondents how surprised and concerned they felt about this, and whether they believed it was appropriate. Each of these questions was answered using a 5-point Likert scale (from "not at all <surprised/concerned>" to "very <surprised/concerned>" for the first two questions, and from "very inappropriate" to "very appropriate" for the latter). We observed that while respondents cynically were "not at all surprised" (this was the median response, selected by 59% of the 358 respondents answering this question; only 8% and 11% were "surprised" or "very surprised," respectively), they were nonetheless "very concerned" about this. That is, while the median concern level was "concerned," the mode was "very concerned" (27% and 31% of 356 were "concerned" or "very concerned," respectively). Only 6% of respondents indicated they were "not at all concerned."

In terms of appropriateness, the median response was that respondents felt these types of sales were "very inappropriate" (selected by 69% of respondents); only 3% of respondents indicated that they thought this was either "somewhat" or "very" appropriate.

With regard to consent, of the 353 respondents who answered this question, only 3 recalled having provided consent (<1%), whereas 92% stated they had no recollection of consenting (the rest were unsure). While 94% (331 of 352 respondents) stated that they would *prefer* that their email address not be sold by data brokers, a majority went so far as to say that they believed that such sales without their informed consent are illegal (54% of 353; 26% were unsure, and only 20% correctly recognized that, while objectionable, such sales are likely not illegal).

For the 190 respondents who indicated that they believed that such sales are illegal, we asked an open-ended follow-up question: "Can you name any specific laws that you think may

[191] We initially informed them of this in our recruitment email.

have been violated?" We received 167 responses that ran the gamut. While a plurality of respondents conceded that they had no idea, a substantial minority cited ambiguous "privacy laws" (e.g., responses included, "data privacy laws," "privacy law," and "FTC privacy law"). Others cited laws that are likely not applicable here: "California Consumer Privacy Act" (which does not require that data be sold with informed consent, only that regulated entities allow data subjects to opt out), the "CAN-SPAM Act," "HIPAA," "GLBA," and "GDPR." This corroborates prior research that observed that "most people either aren't sure how they are protected by current laws, or they [incorrectly] assume they are."[192]

Disclosures and Understanding. On a subsequent page of the survey, we asked respondents whether "it is likely that a privacy policy you never read disclosed that your email address would be sold." A majority of respondents agreed that this was likely the case (63% of 347 respondents), whereas 26% were unsure.

In terms of *how* the email address might have been obtained by data brokers, responses were all over the place:

- 67% believed it may have been due to providing it to a retail store
- 59% believed it may have been due to them providing it to a mobile app
- 25% acknowledged that their email address is posted publicly online
- 19% suggested it may have been from providing it to a website as part of creating an account
- 8% acknowledged posting it on their own personal website, whereas another 8% acknowledged that it was posted on their employer's website

In this same section, we also asked about respondents' familiarity with each of the data brokers from whom we acquired their data. In addition to the four real names, we

---

[192] Nathan Malkin, et al., *"What Can't Data Be Used for?" Privacy Expectations about Smart TVs in the US*, *in* PROCEEDINGS OF THE 3RD EUROPEAN WORKSHOP ON USABLE SECURITY (EUROUSEC), LONDON, UK (2018).

included four additional choices, as noted earlier. We included Acxiom and Experian because we reasoned they may be the most commonly recognized names in this industry (and thus could provide a baseline); Discount Data Warehouse and Eklesia were fictitious names included as decoys. We observed that the decoys were "recognized" by 1.6% and 0.5% of respondents, respectively, which was not statistically different from the real data brokers: familiarity ranged from 0.8% to 3.2%. Experian was recognized by over 81%, whereas Acxiom was recognized by 9% of respondents.

Perceived Benefits. We asked respondents how they felt about advertisements that "use your information to target you personally" using a 5-point Likert scale (from "very opposed" to "very in favor"). We observed that 73% (of 340) were either "very" or "somewhat" opposed to this practice, and that only a single respondent was "very in favor" (8 were "somewhat in favor," which corresponded to 2%).

In terms of perceived benefits, very few participants felt that they received benefits from advertisements that target them personally: only 18% answered affirmatively, 63% answered negatively, and another 19% were unsure (of 339 responding).

Use of Privacy Controls, Reading of Privacy Policies. Because both Google's Android and Apple's iOS include privacy controls that theoretically allow users to reset or disable their device MAIDs, we included several questions to gauge respondents' use of these features. We included screenshots of the controls for either iOS or Android, depending on what type of mobile device participants reported owning (98% reported owning a smartphone; 59% claimed to use an iOS device, whereas 37% claimed to use an Android device). There were no statistically significant differences between the number of apps users of each operating system reported installing (a mean of 59 for iOS versus 50 for Android; $t=1.072$, $p<0.285$). Prior to showing screenshots of the privacy settings, iOS users were significantly more likely to claim that they have previously used their phone's privacy settings to opt out of tracking and/or targeted advertising (78% versus 62%; $X^2=9.550$, $p<0.008$).

After showing the screenshots and specifically asking whether respondents had used the interface to reset their

device's MAID, 11% of iOS users claimed to have done so, versus 13% of Android users, which was not a statistically significant difference. We also asked participants whether they have used these interfaces to "limit ad tracking" or "opted out of ads personalization." This was also not statistically significant: 50% of iOS users claimed to have done so versus 39% of Android users. Nearly 28% of iOS users and 43% of Android users claimed to have never seen these settings before, which was statistically significant (X^2=7.885, p<0.005). Of course, since prior research has demonstrated that MAIDs are frequently transmitted alongside other identifiers that cannot be reset, these controls are effectively useless.[193]

We asked whether respondents are likely to read the privacy policies of the apps that they install on their mobile devices (5-point Likert scale from "never" to "always"), as well as the "privacy labels" that both major app stores now require. Overall, we did observe that respondents were significantly more likely to report reading the new succinct privacy labels than traditional privacy policies (Wilcoxon Signed Rank Test; n=336, W=4,350, p<0.001), with a small-to-medium effect size[194] (r=0.221). We observed no differences between the two platforms: across all participants, the median responses were that respondents both "rarely" read privacy policies (40%; 28% said "never" and only 7% said they "often" or "always" read them) and "rarely" read the new privacy labels (38%; 23% said "never" and 16% said "often" or "always"). We similarly observed that respondents "rarely" read privacy policies prior to entering their email addresses into websites (37%; 27% said "never" and 7% again said "often" or "always").

We asked respondents, "knowing now that your personal information is being sold by data brokers, does that make you more or less likely to read privacy policies in the future?" The median response was that respondents would be "somewhat more likely to read privacy policies" (31% of 335); 27% said that they would be "much more likely," though 30% said that

---

[193] *See* Irwin Reyes, et al., *"Won't Somebody Think of the Children?" Examining COPPA Compliance at Scale*, in THE 18TH PRIVACY ENHANCING TECHNOLOGIES SYMPOSIUM (2018).

[194] *See* Mathias Jesussek, *Wilcoxon Signed-Rank Test*, NUMIQO, https://datatab.net/tutorial/wilcoxon-test [https://perma.cc/BHW7-LKRF].

it is "unlikely to make a difference." Responses to this question correlated with their level of surprise at learning that their personal information was being sold (Spearman's $\rho=0.340$, $p<0.001$), those expressing surprise indicated that they would be more likely to read privacy policies in the future.

Companies have made arguments in courts that device identifiers do not identify people, only devices. We took the opportunity to ask participants whether they share their mobile devices with anyone else: nearly 95% indicated that they had a mobile phone not shared with anyone else. At the same time, 90% (of 315) indicated that they do not share their email addresses with anyone else and only 16% lived alone (i.e., 84% shared a postal address with someone else). Thus, a mobile device is more likely to be uniquely identifying than either an email or physical address, data types that have been long understood to be personally identifying.[195]

Privacy Compliance and Policy Enforcement. Depending on whether respondents reported having an Android or iOS mobile device, we asked them to evaluate the following two statements as true or false:

- If an app is available for download from the <Apple App Store/Google Play Store> that means that <Apple/Google> has determined that the app complies with their app store policies.
- If an app is available for download from the <Apple App Store/Google Play Store> that means that <Apple/Google> has determined that the app complies with applicable privacy laws.

Because we observed no statistically significant differences between platforms, we analyzed all responses together. Overall, only 23% (of 319 respondents) correctly understood that neither platform exhaustively analyzes all of the apps that they distribute to assess compliance with all of their app store policies. Similarly, only 30% correctly understood that neither

[195] *See* 45 C.F.R. § 164.514 (2013) (the HIPAA Privacy Rule, which became effective in 2001, specifically identified telephone numbers, email addresses, and postal addresses as personally identifiable information).

platform ensures that the apps that they distribute comply with all applicable privacy laws.[196]

Exercising Data Rights and Data Breaches. We presented each participant with the name and URL for the specific data broker from whom we received their information, as well as a link to the company's privacy policy and instructions for sending a deletion request (all four data brokers claimed in their privacy policies to respond to deletion requests, regardless of the location of the data subject). We then asked respondents whether they planned to submit such a request and then why or why not. In total, 251 respondents (77% of 327) indicated that they would follow the instructions to submit a deletion request. Of the 76 who said they would not or were not sure, 60 left explanations. Two authors performed inductive coding by independently examining the responses, which resulted in 7 agreed-upon categories of responses:[197]

- "Unspecified": The response does not convey *why* the respondent did not want to submit a deletion request.
- "Helpless": The respondent indicated that they did not believe sending a deletion request would make any difference.
- "Unsure How": The respondent was unsure how they should submit a deletion request (despite the provided link to instructions on the data broker's website).
- "Effort": The respondent does not believe that the effort to submit a deletion request is worth the potential benefit.

---

[196] Prior research similarly found that many app developers incorrectly believe that the app stores evaluate apps for compliance with privacy laws prior to making them available to the public. *See* Noura Alomar & Serge Egelman, *Developers Say the Darnedest Things: Privacy Compliance Processes Followed by Developers of Child-Directed Apps*, 4 PROCEEDINGS ON PRIVACY ENHANCING TECHNOLOGIES (POPETS) (2022).

[197] The inter-rater agreement rate (IRR) was "almost perfect": Cohen's $\varkappa$=0.81. *See* J. Richard Landis & Gary G. Koch, *The Measurement of Observer Agreement for Categorical Data*, 33 BIOMETRICS 159, 159–74 (1977).

- "Benefit": The respondent does not want to submit a deletion request because they perceive a benefit in the data broker having it.
- "Indifferent": The respondent does not see a risk in their information being publicly disclosed.
- "Backfire": The respondent believed that submitting a deletion request may backfire (e.g., confirming their email address is real, revealing other personal information, etc.).

Fourteen responses were not responsive to the question ("unspecified"). Of the remaining 46, the most prevalent response—provided by 18 respondents (39% of 46)—fell under the "helpless" category: respondents did not believe it was worth the effort to submit a deletion request to any one data broker when there are countless other unknown data brokers selling the same information. Some examples of these responses included:

- "I don't think it would make a difference at this point."
- "I feel like asking one data broker to delete my data won't make a difference in the grand scheme of things, as the data will still be out there with others."
- "Isn't a guaranteed they will delete it."
- "Someone else will get it."
- "They are only one of a few dozen brokers, and I don't have time to track down all of them."

An additional 8 respondents indicated that submitting a deletion request would be too much of a hassle, three did not understand how they were supposed to send deletion requests, and another three thought that the request would "backfire" (i.e., the data used to submit the request would be used for additional objectionable purposes). Conversely, of the remaining 14 responses, 11 stated that they were indifferent to the consequences (e.g., "I don't really care that much"), whereas 3 believed they received benefits from data brokers' sales of their data (e.g., "my personal information may actually provide benefit for some organizations – especially those who

would like to ask me to volunteer for causes I care about")—less than 1% of our total sample.

Finally, we asked respondents about their awareness of prior data breach notifications to gauge the likelihood of their email addresses being freely available on the Internet. Over 91% of respondents indicated that they have previously been notified that their personal information was involved in one (15%) or more (76%) data breaches. Thus, the simplicity with which an ordinary person can reverse a hashed email address is further demonstrated by the fact that nearly everyone's email address has become publicly available as part of a data breach.

Conclusion

In sum, we acquired over six million hashed email addresses paired with mobile advertising identifiers (MAIDs) that were being sold by multiple data brokers. The fact that we were able to reidentify over 97% of the underlying email addresses demonstrably refutes any claims that hashing somehow anonymizes or deidentifies this data, or that MAIDs only identify devices and are not personally identifiable information. Similarly, only three participants (0.8% of 369) recalled consenting to the sale of their data and most wanted it deleted. This ought to rebut claims that this data is being collected and sold with data subjects' "consent."

Yet, claims of "consent" are pervasive in the data economy, as Frischmann and Vardi observe:

> The reasons why people remain uninformed, despite being given notice of and hyperlinked access to terms, are many, but in large part, it boils down to a simple fact: Any supposed decision is heavily influenced by design, namely, choice architecture designed to nudge people to click a button labeled agree/accept without deliberation, inquiry, or further consideration. Asymmetric friction-in-design encourages automatic clicking and discourages further inquiry, reading terms, and deliberation. It is frictionless in the direction of clicking-to-contract: just click the prominent button. Yet

> there is substantial friction stacked in the direction of further inquiry, reading terms, and deliberation. A person must find and click the hyperlink(s) to access proposed contractual terms. There may be more than one set of terms to consider, for example a Terms of Service and a Privacy or Data Use Policy. When those other pages load, the person must devote considerable time to reading and trying to understand the terms. There may be embedded hyperlinks, and a person might need to consult dictionaries, Wikipedia, or other external sources to make sense of the proposed terms. If a person has questions, often there is no one to ask. We could go on. The design practice of asymmetric friction stacking is powerful.[198]

Recognizing that so-called "consent" for data collection may not have been explicit nor informed, nascent privacy laws now recognize that consumers have a right to request that their data be deleted after the fact. For example, California consumers are allowed to submit deletion requests to entities regulated under CCPA; residents of other states have similar rights.[199]

Of course, exercising a deletion right requires knowing the identity of the entity to whom the request should be directed. Yet, we observed that nearly none of our survey respondents had ever heard of the data brokers selling their data (and therefore prior to our study did not possess the requisite knowledge to submit deletion requests).[200] Worse, even after

---

[198] Brett Frischmann & Moshe Y. Vardi, *Better Digital Contracts with Prosocial Friction-in-Design*, 65 JURIMETRICS J. 1, 25–26 (2025).

[199] *See* Müge Fazlioglu, *U.S. State Comprehensive Privacy Laws Report*, IAPP RESOURCE CTR. (2025), https://iapp.org/resources/article/us-state-privacy-laws-overview/ [https://perma.cc/2DX9-BG9A].

[200] Incidentally, during the IRB approval process, we were originally told that we should provide recommendations to participants for commercial services that send deletion requests to data brokers on their behalf (many such services now exist and most publish lists of the data brokers that they track). However, after we pointed out that none of the services that we could find include all of the data brokers in our study, our IRB agreed

we identified the names of the data brokers, many subjects indicated they were apprehensive about submitting deletion requests because they believed it would be futile or exacerbate the objectionable data collection.

One takeaway from this study is that while policymakers are paying more attention to consumers' privacy concerns by enshrining privacy rights into laws, they may not actually be serving consumers well. As it stands today, while consumers often do not consent to data collection, they are largely powerless to stop it.

During the course of this research, in internal discussions, we regularly used the phrase "bullshit anonymity claims" to describe the inappropriate labeling of hashed email addresses—and the device identifiers to which they are linked—as "deidentified" or "anonymized" data. However, we realize that this turn of phrase is inaccurate: we argue that industry's claims of "anonymity" and "consent" are not actually bullshit, as Frankfurt explains:

> [The bullshitter] does not reject the authority of the truth, as the liar does, and oppose himself to it. He pays no attention to it at all . . . . One who is concerned to report or to conceal the facts assumes that there are indeed facts that are in some way both determinate and knowable. His interest in telling the truth or in lying presupposes that there is a difference between getting things wrong and getting them right, and that it is at least occasionally possible to tell the difference.[201]

In this case, those making claims of anonymity and consent have a very strong vested interest in their claims being perceived as true: in many cases, their business could not legally operate or could only do so with lower profitability. In *The Republic*, Plato argued that it is the duty of the rulers to lie to the populace, when those rulers believe it is in society's best interest. This is known as the "noble lie" ("noble" as in nobility

---

with us that it would be unethical to recommend that participants sign up for services that we know will not actually solve the problem.

[201] Harry G. Frankfurt, *On Bullshit*, 6 RARITAN Q. REV., Fall 1986, at 81, 98–99.

and not in any "honorable" sense; to avoid this confusion, it is sometimes referred to as the "lordly lie"[202]). In his criticism of Plato, Popper explains that the noble lies that Plato exhorts the ruling class to tell the public are better thought of as "propaganda lies" ("nothing is more in keeping with Plato's totalitarian morality than his advocacy of propaganda lies").[203] Their goal is to assist in "controlling the behaviour of . . . the bulk of the ruled majority.[204]

It is through this lens that we must view industry's claims that hashed email addresses or device identifiers are somehow anonymous or deidentified or gathered with consent as examples of these propaganda lies. The type of data we examined in this study is sold specifically for the purpose of reidentifying individuals. At the same time, privacy laws in many jurisdictions now require either consent or anonymity to legally traffic it. Thus, profitability depends on courts, policymakers, and regulators believing a noble lie that is demonstrably and definitionally untrue.

Research going back nearly half a century on consumer privacy attitudes consistently shows that consumers are opposed to their personal information being shared for secondary purposes.[205] By telling the noble lie, companies are allowed to continue collecting and trafficking in consumer data in ways that the vast majority of consumers find objectionable.

---

[202] D. Dombrowski, *Plato's "Noble" Lie*, 18 HIST. POL. THOUGHT 565, 566 (1997).
[203] Karl Popper, THE OPEN SOCIETY AND ITS ENEMIES: THE SPELL OF PLATO 141 (4th ed. 1962).
[204] *Id.* at 139.
[205] *See* Ponnurangam Kumaraguru & Lorrie F. Cranor, *Privacy Indexes: A Survey of Westin's Studies*, INST. FOR SOFTWARE RSCH. INT'L (2005) (surveying studies of consumer privacy preferences going back to the 1970s).