

# Privacy Enhancing Technologies and the Privacy Arms Race: A Review

DAVID SIDI, The University of Arizona, USA

The literature on privacy-enhancing technologies (PETs) is vast, and evolving; narrower reviews of PETs from a variety of perspectives allow broader trends to be explored without being overwhelmed by an unworkable volume of articles. This article considers as an organizing principle the history of escalation in privacy attacks and defenses, the *privacy arms race*. From this perspective, several examples from the recent literature on anonymity, transparency, and usability are discussed.

Additional Key Words and Phrases: transparency, anonymity, usability

## ACM Reference Format:

David Sidi. 2019. Privacy Enhancing Technologies and the Privacy Arms Race: A Review. *ACM Comput. Surv.* 1, 1, Article 42 (December 2019), 39 pages. <https://doi.org/0000001.0000001>

## 1 INTRODUCTION

The advent of cheap personal computers in the 1980's and 1990's brought computing into private life for the first time for millions of Americans. Pioneers of computing heralded a revolution, the beginning of a new age: now the computer would now serve the individual, inverting its reputation as a tool for institutions seeking to reduce the individual to a "number."<sup>1</sup> Less than ten years later, similar themes would be revisited as the Internet began connecting networks of computers to one another, democratizing access to information resources that had previously been available only to powerful organizations.<sup>2</sup>

The individual user would become a prolific producer of personal information in this newly connected context, moving more and more of the activities of everyday life online. The personal information associated with activity online would in turn become available to a broad range of actors, including both private firms and public governments, and soon, information privacy came to occupy a central role on the Internet.

In practical terms, one result of individual users seeking to control their personal information online was an arms race between the development of privacy-enhancing technologies (PETs) and technologies for their circumvention.<sup>3</sup> This article considers the history of escalation in privacy attacks and defenses, the *privacy arms race*, as an organizing principle for the landscape of privacy technologies, selecting a few examples from the recent literature to highlight broader themes.

<sup>1</sup>See generally Nelson [60]. The idea of a person as a "number" was prominent in the Free Speech Movement at UC Berkeley, a west coast source of many early developments in personal computing. See Markoff [56].

<sup>2</sup>See Barlow [8].

<sup>3</sup>For an engaging popular treatment of the arms race in the case of ad-blockers on the web, see Cory Doctorow, 'Adblocking: How About Nah?' available at <https://www.eff.org/deeplinks/2019/07/adblocking-how-about-nah>.

Author's address: David Sidi, The University of Arizona, School of Information, P.O. Box 210076, Harvill Building, Tucson, AZ, 85721, USA, [dsidi@email.arizona.edu](mailto:dsidi@email.arizona.edu).

2019. 0360-0300/2019/12-ART42 \$15.00  
<https://doi.org/0000001.0000001>

Section 2 revisits previous surveys of privacy technology from the arms race perspective, revealing new relationships between and within them, and exposing several holes in coverage.<sup>4</sup> Sections 3, 4, and 5 focus on technologies of anonymity, transparency, and usability, respectively.<sup>5</sup> The entirety of the review is biased toward new work.<sup>6</sup>

## 2 RELATED WORK

This section reviews two previous surveys adopting distinct, but related, perspectives on the landscape of PETs. The first, Le Métayer [53], divides PETs into “hard” and “soft” to mark a difference in underlying trust assumptions; while the second, Danezis and Gürses [27] presents a categorization that distinguishes both hard and soft technologies from a third category, “privacy as practice.” The distinction is characterized here in terms of strategic and reactive technologies, which are, in short, those technologies that contribute to the privacy arms race, and those that undermine it.<sup>7</sup>

### 2.1 Hard and Soft Technologies: PETs organized by trust assumptions

Le Métayer [53] surveys *practical* privacy technologies, intentionally avoiding those that are still considered “research challenges.”<sup>8</sup> PETs are organized by “type of trust that they provide.” ‘Trust’ is not intended as a technical feature of a system, but as a feature of privacy technology in the context of its application:<sup>9</sup> in particular, the central concern is with which stakeholders are trusted in the deployment of the technology.

PETs are divided into two categories: “hard” technologies for minimization of disclosure, and “soft” technologies for enforcement of rights when disclosures occur. Hard technologies minimize trust as a technical design goal, and often make use of cryptography; while soft technologies have their effect on a trusted party who is assumed present—either building their privacy guarantees around the trusted party, or targeting the trusted party for accountability—and often make use of results from the social sciences. Accordingly, a useful way to think of these categories is as divided by the act of trusting: hard technologies provide *ex ante* protections against trusting, while soft provide *ex post* remedies once trust is assumed.<sup>10</sup>

The sequel in this section will primarily focus on the understanding the idea of minimization for hard technologies, and subject rights over personal information for soft technologies.

<sup>4</sup>As such, the treatment of these works does not follow the format of an analysis and evaluation of a article reporting a study. The goal in the section on related work is to build a recent picture of the PET landscape that includes technologies seen from the new perspective of the privacy arms race.

<sup>5</sup>Obfuscation appears partially as a subsection within the section on anonymity. This limitation of the broader area of obfuscation allows treatment of usability; however, future work could profitably expand obfuscation into a section of its own.

<sup>6</sup>In particular, the review covers the following ten articles in detail (with many others mentioned in support): Cooper [23], Danezis and Gürses [27], Diaz and Gürses [32], Dwork [38], Englehardt and Narayanan [42], Le Métayer [53], Luguri and Strahilevitz [54], Mathur et al. [57], Venkatadri et al. [77], Wagner and Eckhoff [78].

<sup>7</sup>The reasons for the choice of term “practice” are not fully clear, though one of the authors (Gürses) has elsewhere considered the integration of privacy into the actual development practices used in software engineering [51].

<sup>8</sup>Of course, today’s research challenge is (sometimes) tomorrow’s practical technology. For this reason, others have provided surveys with a similar bent towards practicality at regular intervals. For example, Ian Goldberg provided a survey of PETs every five years from 1997 to 2007 [45–47].

<sup>9</sup>A typical technical treatment of trust is in an information flow diagram with trust boundaries [70].

<sup>10</sup>As noted in the survey, this division is close to one drawn in the OECD privacy guidelines between the data minimization principle, and the rights of a subject to enforce their rights to personal data [53, note 8].

**2.1.1 Data Minimisation.** Four categories of functionality structure the discussion of minimization in Le Métayer [53]. Each is introduced with an argument that PETs must balance privacy against a distinctive countervailing value.<sup>11</sup>

*Communications services.* Communications services are described informally, as “the user of the system just [wanting] to communicate information to another user (or a group of users).”<sup>12</sup> Here the user wishes to minimize disclosure to either first- or third-party attackers while exchanging messages across a network.<sup>13</sup>

For first party attackers, the aim is to avoid a design in which the identity of a sender can be recovered from the information needed to route messages across a network.<sup>14</sup> Put more concretely, defense against first party attackers should permit the user both to receive a message anonymously, and to reply to the anonymous sender. For example, SecureDrop allows whistleblowers to submit documents anonymously to news organizations, and for reporters to follow-up with the source, who remains anonymous even to them, in order to ensure the story’s veracity [25].

For third party attackers the aim is to avoid disclosure of information to any party who is not part of a communication. Here the capabilities of the parties are left open, so attackers who control intermediate nodes to mount active attacks are countenanced alongside occasional curious observers. Again put more concretely, “avoiding disclosure of information” means it should be possible to communicate without allowing your messages to be linked to you or your communicant, or indeed even without allowing communication to be observed as occurring at all.<sup>15</sup> An example of a third-party is a provider of communication services: an Internet service provider, or an online social network, for example. PETs for limiting disclosure to third-parties without hindering communication include Hummingbird, which encrypts posts, interests, and hashtags to be visible only to a group the user specifies—which does not include Twitter itself—and Scramble!, which works similarly for a range of other social networks.<sup>16</sup>

<sup>11</sup>One section, on “database exploitation,” present in the survey is not covered, as its content overlaps with sections 3.2.2 and 3.2.3.

<sup>12</sup>A slightly more formal account of the communications setting for exchange of messages via a computer network appears in the review of terminology in Anonymity, in section 3.1.1.

<sup>13</sup>A first-party attackers is someone with whom you intend to communicate, while a third-party attacker is not an intended communicant. Accordingly, the distinction between first- and third-parties inherits from the communicants any unclarity about the parties with whom they intend to communicate. This point is overlooked by Le Métayer [53].

<sup>14</sup>This creates a particular problem for replies. In the history of mixnets, which are a communications service with strong anonymity guarantees, a major development was the development of anonymous replies. See the history of single-use reply blocks, culminating in the Sphinx packet format Danezis and Goldberg [26].

<sup>15</sup>“The main objectives of the user in terms of privacy are to ensure that:

- The recipient gets the expected information from the sender and nothing more. For example, she does not get the identity of the sender if the latter does not wish to disclose it [...] (anonymity of the sender ). [...]
- Third parties (e.g., parties controlling the intermediate nodes in the communication chain or spies listening to the communication channel, etc.) cannot learn anything from this communication, which means that third parties should not be able to observe the content of the information sent or disclosed to the recipient ( confidentiality ) but also that they should not be able to guess that two messages were sent (or received) by the same user ( unlinkability ) or even that a message has been sent (or received) by a given user (unobservability).” (399)

<sup>16</sup>See De Cristofaro et al. [30], and <https://cosic.esat.kuleuven.be/scramble/about.html>.

Categorizing *anonymity*, *unlinkability*, and *unobservability* based on whether the attacker is first- or third-party, as in Le Métayer [53], is invidious.<sup>17</sup> Failures of anonymity may be the result of a third party as much as a first party attack; similarly, linkage attacks can be launched by first parties as much as third parties. It would be preferable to consider minimization in communications straightforwardly in terms of protecting *contents* and *metadata* in communications, with allowance made for the broadness inherent in those terms.

*Authentication and authorization services.* The next context for data minimization is authentication and authorization. Two aspects of authentication mechanisms may be observed. First, in the offline world, authentication technologies often reveal more than is strictly necessary for their purpose: a photo identification card reveals a central piece of biometric information, a face, along with the attributes that it is used to check, such as age when entering a bar.

*Binding* the attributes listed on an identification card to a person using face biometrics has several important drawbacks, including irrevocability, which is noted by Le Métayer [53], and observability, which is not. Together these drawbacks represent a bigger problem than either does separately, since the measurement of a biometric is used as an authenticator, and this measurement can be reproduced from a publicly observable original. Once compromised, a biometric authenticator cannot be “revoked” and replaced by a new one.<sup>18</sup>

To isolate the trust required for authentication, a centralized approach is to rely on “identity providers:” trusted parties who certify that attributes hold or do not hold for a given identity. For example, a service running on a server might certify that it has followed a procedure to verify that the public key belongs to an individual with another identifier, such as the aforementioned government-issued identification card.

Two important omissions of specific examples by Le Métayer [53] are, first, for the centralized model, the Certificate Authorities that provide TLS certificates [76]. TLS is the most widely used encryption scheme today, implemented on every major web browser. Second, Le Métayer [53] missed a straightforward distributed version of centralized identity providers, the “web of trust,” which has been in use since the mid-1990s, and has spawned new work in the area.<sup>19</sup>

More generally, the discussion fails to identify the general “key binding problem” of ensuring that a cryptographic key used to ensure confidentiality is controlled by exactly the person who is intended to be included in the communication. In practice, many “man in the middle” (MiTM) attacks rely on failures of key binding, substituting keys under the attackers control for both the sender’s and receiver’s keys when the communications is being established in order to view their communications.<sup>20</sup>

*Computation.* Minimization in the context of remote computation means that the user can run computations on a remote service without the results or the input being made

<sup>17</sup>These terms are now standard terms in the anonymity literature. They are discussed in section 3.

<sup>18</sup>It was precisely this problem led Ratha et al. to add a layer of cryptography to produce “cancellable biometrics” Ratha et al. [65].

<sup>19</sup>See the conference “Rebooting the Web of Trust,” at <https://www.weboftrust.info/>. On WoT generally, see section 4, “How it works,” of the Gnu Privacy handbook, available at <http://www.pa.msu.edu/reference/pgpdoc1.html#section-4>.

<sup>20</sup>Perhaps the simplest example of this form of attack occurs for the simple Diffie-Hellman key exchange protocol Diffie and Hellman [33]. An active attacker Mallory intercepts the secret value sent by the sender Alice, replacing it with Mallory’s own value, which is sent on to the recipient Bob. The same substitution is then performed in the other direction. Alice and Bob both receive encrypted communications, but the keys in use are bound to Mallory, not their intended communicants.

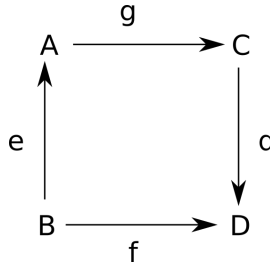


Fig. 1. A simple commutative diagram can be used to illustrate homomorphic encryption. A function  $f$  cannot be computed locally on  $B$ . Instead, values from  $B$  are encrypted with  $e$ , and a function  $g$  is computed on the encrypted output. The output of  $g$  is then decrypted by  $d$ , so that  $f = d \circ g \circ e$ .

available to the service provider. One approach to this question is homomorphic encryption, which allows computations to be performed directly on encrypted data in such a way that the results can be subsequently decrypted to yield the same result as if a desired operation had been performed on *unencrypted* data. Homomorphic encryption allows, for example, sensitive data to be shared with a powerful computing cluster that is made available as a service, without the need to trust all parties with access to the memory of any of the cluster nodes. The idea is given a short description in Le Métayer [53], but a diagram is a helpful addition (see Figure 1).<sup>21</sup>

### 2.1.2 Subject Rights over Personal Information.

*Decision support.* In the context of consent decisions, transparency PETs provide meaningful notice about the handling of personal information.<sup>22</sup> A simple kind of transparency is provided by notice via privacy policies and terms of service; however, this mechanism is burdensome for the user, who must read many such notices, and who must interpret them correctly. Such interpretation is increasingly difficult, as “a wealth of personal information is disclosed piece by piece to a variety of stakeholders, sometimes directly, sometimes indirectly,” with many stakeholders unknown to the subject. Moreover, in addition to collection and sharing, there is the further question of the uses to which data will be put. As the authors note, “[a]ssessing the consequences is particularly challenging with respect to profiling, which can have long-term effects on different aspects of the life of the subject (refusal of a credit, exclusion from a job selection procedure, refusal of a visa, etc.).”

From the perspective of an “arms race,” an adversarial aspect to transparency should be highlighted. Enforcement for deceptive practice turns on the representations made in privacy policies and terms of service; the construction of these documents by lawyers to avoid enforcement action means that many “notices” are dense legal documents that commit to very little. Attempts to lift the burden of privacy policy interpretation from the user with a machine-readable policy format was met with very low adoption [24], supporting the interpretation that service providers do not wish to be transparent about their practices.

<sup>21</sup>Another important technique used in PETs for minimization in remote computation is zero-knowledge proof (ZKP). A simplified, concrete demonstration is provided in the appendix as an aid to understanding.

<sup>22</sup>In the General Data Protection Regulation (GDPR), consent is defined as “any freely given, specific, informed and unambiguous indication of his or her wishes by which the data subject, either by a statement or by a clear affirmative action, signifies agreement to personal data relating to them being processed.” See GDPR Article 7.

For example, section 4.1.1 below discusses Facebook’s collection of phone numbers under the pretext of increasing account security via two-factor authentication, only to use the numbers for advertisement targeting. In that case, the notice on the web form for submitting a phone number makes only general, anodyne statements about the consequences of inputting a phone number, none of which would lead a consumer to understand their use for advertising purposes (if indeed the consumer read the notice at all, a significant cost in time while already performing an ancillary setup task not directly related to the use of the service).

One response to this failure of trusted parties to provide adequate transparency—and even to act as an adversary by misleading users, or gaming a notice system—is to seek to increase transparency without the product or service provider’s coöperation. Two such strategic approaches to transparency are discussed in section 4.

*Consent.* Mechanisms for consent allow the user’s decisions regarding control of personal information to be efficacious: the mechanism should allow user decisions to be represented precisely and specifically by the data controller. The description of tools for consent in Le Métayer [53] is overbroad. The description includes domain-specific languages for privacy policy creation, which is a tool in support of transparency better placed with “Decision support,” but also includes “specific tools or services allowing users to manage their privacy protections” which includes ad blocker extensions, and dedicated browser bundles such as the Tor Browser (or its Android equivalent at the time, Orweb).

Technologies adopting “a different strategy” are also included, and from the perspective of an arms race, “strategic technologies” are of particular interest (see 2.2). The only tool included in this category by Le Métayer [53] is TrackMeNot, and much of the discussion focuses on criticism of the work. TrackMeNot submits automated queries to search engines to thwart profiling based on what users reveal by their true search history. The aim of the project is to actively waste resources expended by a search engine operator to profile its users. Minimizing the profiling information that is given to a search engine is a poor approach, since accuracy in queries is directly related to the quality of the search results.<sup>23</sup> Approaches like TrackMeNot instead attempt to flood the service with falsified profiling information.

Despite an overinclusive discussion, some aspects of consent are omitted. One such is the role of user interfaces in influencing a consent decision. This issue is taken up in greater detail in discussion of “dark patterns” in section 5.

*Enforcement.* Transparency and consent mechanisms must be enforced. As in earlier sections, there is an approach to enforcement that seeks to minimize trust, which is accomplished without the service’s coöperation,<sup>24</sup> and an approach that seeks the best solution for working coöperatively with a service. For the latter, there is a requirement for privacy infrastructure—for example, a means to associate data to a privacy policy, however it is transformed and moved by a data controller.

Le Métayer [53] focuses on DRM technologies, which wrap data in an executable to enforce policies on data (for example, that the movie file can only be played for one day while it is “rented”). The problems with a DRM model for enforcement of privacy center on the fact that providing data to a controller is a form of trust, given the way data is handled by computers—rather than erase the data/process distinction in general in an uncontrolled

<sup>23</sup>Davidowitz [29] makes use of this fact in his research, highlighting the way in which search data on sensitive topics does not agree with data gathered in conventional research settings, such as surveys.

<sup>24</sup>Le Métayer [53] characterizes this as “local” enforcement, but the technologies of homomorphic encryption already introduced shows that untrusting remote enforcement is equally possible.

computing environment, a “soft” alternative to DRM seeks to align privacy interests with the interests of the controller. For example, a privacy infrastructure for a smart building might not only notify users of visibility to a camera, but help to limit incidental collection of sensitive information, which could lead to costly disclosures Das et al. [28].<sup>25</sup>

*Accountability.* Accountability ensures the performance of a system for acquiring consent, and for enforcing its terms. Le Métayer [53] describes the work of Colin Bennett on three types of accountability, namely, *accountability of policy*, *accountability of procedures*, and *accountability of practice* [9]. Gray [50] similarly describes a distinction between *technical* audits and *attestation* audits, with the latter category encompassing accountability of both policy and procedures. Accountability via attestation audits relies on the assertions of management, missing completely any testing of how a system actually functions, by contrast to how it is intended to function. A concrete example cited by Gray [50] is access controls. “One component of access control security is a strong password policy. An assessment would check to see if the organization has a strong password policy while a security audit would actually attempt to set up access with a weak password to see if the control actually has been implemented and works as defined in the policy.”<sup>26</sup> An attestation audit would ensure that the written policy was appropriate, whereas a technical audit would test whether the policy was actually followed.

Le Métayer [53] articulates the parts of a technical audit, which include a “logging” component recording the history of interaction with data, and automated approaches to using logs to conduct the audit itself. However, “logging” is the wrong designation for the maintenance of historical interaction with data: logging is not specific to data, but can encompass all kinds of events on a computer system. A better term is *data provenance*. The point is not only semantic: tools for handling logs are designed to handle a broader range of events, whereas tools for data provenance are designed more narrowly. For example, data provenance tools often include special features for ensuring that the records kept do not disclose sensitive data.

The attestation/technical audit distinction maps well onto the trust models described in 2.1, with attestation demanding trust in the assertions of management, while technical audits adversarially test the basis for trust. A further elaboration of the relations between these models—in particular, what sort of development might be described by which technical audit gives way to attestation, is an area for future work that has received no attention that we are aware of, but that is highlighted by the arms race paradigm.

**2.1.3 Limitations of the categories of “hard” and “soft”.** Le Métayer [53] initially provides an equitable account of the differing trust assumptions underlying “hard” and “soft” PETs, yet elsewhere reveals a strong stance on their unequal importance, stating that the “main benefit” of the use of PETs is to reduce the perimeter of trust, an *ex ante*, “hard” role:

The ideal situation for data subjects should be to have sufficient guarantees about the design of the technologies so that it is not necessary for them to trust any third party. (396)

<sup>25</sup>This perspective was adopted for a presentation by the author on such a system, which is currently under development. See <https://www.lightbluetouchpaper.org/2018/05/24/security-and-human-behavior-2018/#respond>. (The author is first author on this work, despite the presentation by Anderson).

<sup>26</sup>See note 19.

This view is related to another idea that is endorsed by Le Métayer [53], that most “privacy-enhancing technologies” could actually be called “data protection and management technologies.” From what section 2.2 calls a “reactive perspective,” encompassing “privacy as confidentiality” and “privacy as control” in Danezis and Gürses [27], this makes sense: trust is understood as an assumption about some part of a system that, if violated, violates the system’s privacy or security properties.<sup>27</sup> Accordingly, hard privacy technologies are seen in the best light as data protection and management technologies. Soft privacy technologies, by contrast—as indicated in the choice of nomenclature—appear weaker in this context. From the perspective of the privacy arms race, however, PETs that seek to strategically influence incentives are most often “soft,” failing to protect an individual subject’s personal data in order to create a cumulative effect that is costly to a privacy attacker.

## 2.2 Strategic and reactive technologies: PETs organized by their relation to new privacy attacks

Danezis and Gürses [27] and Diaz and Gürses [32] present a division of PETs into three classifications: privacy as “confidentiality,” privacy as “control,” and privacy as “practice.” The first two align with the “hard” and “soft” privacy technology distinction discussed in section 2.1, while the third contrasts with it, emphasizing transparency and the development of community-specific values around privacy.

Privacy as “confidentiality” to some extent overlaps with “hard” privacy technologies, taking information disclosure as the main paradigm for privacy violation, and setting PETs against centralization of trust: “placing [...] high levels of trust in organizations should be avoided whenever possible, as they leave individuals vulnerable to incompetent or malicious organizations” (2).<sup>28</sup>

Privacy as “control” similarly aligns with “soft” privacy technologies, taking information misuse—repurposing and sharing of information—as paradigms for privacy violation. From the perspective of privacy as “control,” PETs work with centralized parties to set policies for control of a subject’s information, to detect violations of those policies, and to facilitate enforcement.

The category of privacy as “practice” is introduced in with the both privacy as “confidentiality” and privacy “control.” The contrast can be seen as part of a broader one between strategic and reactive technologies. Reactive technologies are engineered to respond defensively to a privacy attack—such as tracking customers using mobile device requests for WiFi access points, or using an email address provided during registration to link activities across websites.

PETs for achieving privacy as “confidentiality” and privacy as “control” are reactive, though they define the nature of privacy attack differently. Strategic PETs contrast with reactive technologies in seeking to proactively establish conditions favorable to privacy. The design of strategic PETs need not appeal to a particular sort of privacy attack; instead, the aim may

<sup>27</sup>See Anderson [5].

<sup>28</sup>‘Confidentiality’ is not defined in this work. A useful definition is given in the Guidebook for the Institutional Review Board (IRB): “Confidentiality pertains to the treatment of information that an individual has disclosed in a relationship of trust and with the expectation that it will not be divulged to others in ways that are inconsistent with the understanding of the original disclosure without permission” [61]. The Federal Committee on Statistical Methodology adds a point of contrast in tension with the usage under discussion: “Confidentiality differs from privacy because it applies to business as well as individuals. Privacy is an individual right whereas confidentiality often applies to data on organizations and firms” Confidentiality and Data Access Committee [20].



be to align the incentives of a service provider with consumer privacy interests by making privacy practices transparent, advantaging pro-privacy competitors.<sup>29</sup> Transparency is the focus of privacy as “practice:”

The technologies in the two previous paradigms [privacy as “control” and privacy as “confidentiality”] have a strong security focus. Their goals are to either allow individual users to prevent information disclosure, or organizations to enhance the security of the personal data that they hold and prevent its abuse for illegitimate purposes. Privacy is however not just an individual matter – it has an important social dimension, as users often make privacy decisions not in isolation but based on how their communities make privacy decisions ... . In this paradigm, the focus is not only on concealing and controlling information but also on improving transparency and enabling identity construction.

Danezis and Gürses [27] emphasizes the centrality of cryptography to reactive privacy technology, beginning with a short history of the technical field of privacy technology from this perspective. This emphasis contributes to an oversight, as censorship circumvention technologies do not fit neatly into technologies of “confidentiality” or “control.” For example, PETs for publishing information in a way that resists removal by compulsion have a long history Anderson [4], Clarke et al. [16], Dingledine et al. [34], Wilcox-O’Hearn and Warner [81], but combine aspects of “hard” and “soft” PETs by avoiding centralized trust while seeking to prevent misuse—in particular deletion—of information that has been published.

Danezis and Gürses [27] is explicit about its orientation toward the broader context in which PETs are embedded:

[...] our aim is to explore the extent to which influences other than the actual end-user requirements were responsible for shaping [PETs] as they are today, as well as the resulting privacy properties. These include the established methodologies of cryptographers and computer security researchers, the attempt to integrate them within government mandated data protection frameworks, or the will to use them to make technologies that are intrinsically intrusive, from both privacy and surveillance perspectives, acceptable to the public. (2)

The argument that PETs must balance privacy against a countervailing value presents a negative face for PET research, but there are examples of privacy technology that attempt to align privacy interests with other interests, including transparency technology to redress information asymmetries in markets, and censorship circumvention services providing increased robustness and availability.

### 3 ANONYMITY

Section 3.1 revisits the standard terminology for anonymity in Pfitzmann and Kohntopp [63], but relates them to one another in a new way. The section ends by highlighting the importance of pseudonymity to strategic privacy technologies in the context of a privacy arms race.

Next, section 3.2 selects privacy metrics relevant to anonymity from Wagner and Eckhoff [78], again seeking to relate the terms to one another wherever possible. The task of justifying imposition of costs is essential to strategic PETs. Anonymity metrics provide a useful way

<sup>29</sup>Strategic PETs may be developed in response to new privacy attacks, but in that case they are not defensive, but offensive. For example, a strategic PET may impose new costs to redress the competitive advantage of engaging in bad privacy practices surreptitiously. Offensive privacy technologies are not considered in these reviews, and the broader category of strategic privacy technologies is nowhere articulated in full generality.

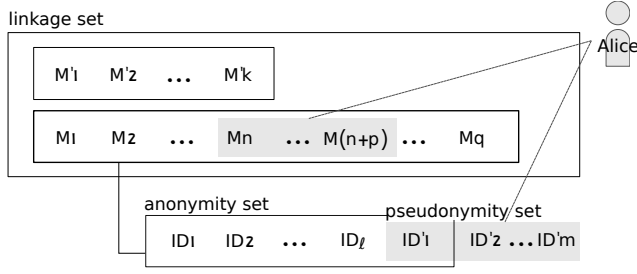


Fig. 2. Relations between *linkage*, *anonymity*, and *pseudonymity* sets. An attacker's linkage set contains the sets of messages that are possibly from the same sender (or to the same receiver). The messages  $M_i$  are a mixture of Alice's true messages  $M_n, \dots, M_{(n+p)}$ , indicated with a shaded box, and several others. If some of the  $M_i$  are "dummys" belonging to no one, then unobservability is achieved by the system against the attacker. The attacker's anonymity set for an identifier contains the possible sender identities for a bundle of messages (see 3.1 for a definition of 'bundle'). The attacker's anonymity set for Alice includes one of her true identities,  $ID'_1$ , as well as several more. Finally, Alice's *pseudonym* set contains her identities, indicated with a shaded box.

to compare the practices of product or service providers, and quantify harms in the course of specifying a response.

### 3.1 Terminology

Surveys of anonymity that discuss terminology sometimes mention a "consensus" in the anonymity literature on the usage of the terms *anonymity*, *unlinkability*, *unobservability*, and *pseudonymity*, based on the definitions provided in Pfizmann and Kohntopp [63]. These definitions will fix what it intended by subsequent uses of the terms, serving as a foundation for the remaining discussion.

The discussion will restructure the definitions to highlight the relations they bear to one another. Unlinkability is discussed first; as the most central, it is subsequently used to define 'anonymity,' 'unobservability,' and 'pseudonymity.'

**3.1.1 Setting.** Following Pfizmann and Kohntopp [63], communications over a network is taken to be the paradigmatic setting for anonymity terminology: there are *senders* who send *messages* to *receivers* over the network. More general settings are countenanced as well, by considering *runs* of a *system* involving *items of interest* (IOI) Pfizmann and Kohntopp [63, p. 1]. The remainder of this section places more emphasis on the broader context where it is helpful to highlight relations between terms, but otherwise treats communications settings as paradigmatic.

**3.1.2 Unlinkability.** 'Unlinkability' is in several ways the most central of the terms discussed: 'anonymity' can be defined in terms of 'unlinkability,' 'pseudonymity' comprises degrees of anonymity, and 'unobservability' can be defined as a kind of anonymity for a message.<sup>30</sup>

Unlinkability is in the first instance a property of sets of messages: let a *message bundle* be messages from the same sender or receiver.<sup>31</sup> Messages  $M$  are *unlinkable* if they are not part of any message bundle that can be identified by an attacker. What an attacker

<sup>30</sup>A generalization of the definition of unlinkability making its connection to unobservability would be worthwhile, but is not pursued here.

<sup>31</sup>This coinage is introduced here for the first time. Message bundles with a single message are trivial, in the present setting, so they may be excluded at a slight cost to the naturalness of the definition.

is capable of, and what is to count as “identified” is left open by the definition; however, ‘identification’ can be made precise with a variety of measures to be described in section 3.2, as can ‘attacker’ via adversary models.<sup>32</sup>

*Anonymity* is a variety of unlinkability.<sup>33</sup> A user with identifier *Alice* is anonymous to the extent that ‘Alice’ is unlinkable to her status as a sender/receiver by an attacker. That is, if the IOI are varied from messages to send/receiver status, and these IOI are unlinkable to a subject identifier (here, ‘Alice’) [63, p. 3 - 4], then anonymity is a kind of unlinkability.<sup>34</sup> Alice’s *anonymity set* comprises the set of senders/receivers who are possible senders, for a given attacker.

**3.1.3 Unobservability.** Unobservability for a message *m* defined relative to an *unobservability set* of messages that cannot be distinguished from *m* by an attacker. Thus, in particular, noise messages are indistinguishable from real messages: the size of the anonymity set is then controllable by the system, and is not bounded by the number of subjects. Pfizmann and Kohntopp [63] note that this property is a point of contact between steganographic and anonymity systems.

**3.1.4 Pseudonymity.** *Pseudonyms* are strings that are unique, and suitable to authenticate Alice and her IOIs (for example, her messages). Pseudonyms admit of degrees of linkage to their holders: public pseudonyms are those for which the linkage is publicly known, as for a published PGP public key, or telephone number. Initially unlinkable pseudonyms are not linkable by anyone to their holders, excepting possibly the holders themselves, when they are created. A subject holding multiple pseudonyms is *pseudonymous*. Just as unobservability concerns dummy messages, pseudonymity concerns dummy identifiers.

Pseudonymity is of special interest to strategic PETs, as they allow attribution without identification. A distributed tool imposing a cost on a service can avoid being coöpted by sybils—fake users multiplying individual users’ effect—by implementing careful use of pseudonymity.<sup>35</sup>

---

<sup>32</sup>On adversary models, see the discussion of personas in Shostack [70]. On identification, a simple approach is to appeal to a *linkage set*, which includes all possible message bundles for a subject: ‘identification’ of a message bundle then consists in the bundle being the only member of a linkage set for a given attacker. Despite the naturalness of this interpretation given similar appeal to sets of possibilities in ‘anonymity sets’ and ‘unobservability sets’ by Pfizmann and Kohntopp [63], the term ‘linkage sets’ does not appear in that work. More importantly, in contrast to the flexibility described here for the definition of ‘unlinkability,’ Pfizmann and Kohntopp [63] specifies a particular interpretation in terms of reduction in uncertainty:

With respect to the system of which we want to describe anonymity, unobservability, or pseudonymity properties, unlinkability of two or more items means that within this system, these items are no more and no less related than they are related concerning the a-priori knowledge. This means that the probability of those items being related stays the same before (a-priori knowledge) and after the run within the system (a-posteriori knowledge of the attacker).  
(2)

A simple interpretation of an adversary might appeal to a model representing the number of nodes whose messages the attacker can observe.

<sup>33</sup>The history of anonymity research precedes its formalization, beginning with Chaum [13], which introduces the notion of a *mixnet*.

<sup>34</sup>The survey of Diaz et al. [31] overlooks this connection between unlinkability and anonymity.

<sup>35</sup>This general point is not intended to minimize the difficulty of implementing a distributed tool resistant to sybil attacks; rather, the point is that pseudonymity can be a valuable aspect of such a tool’s design.

### 3.2 Anonymity measures

Wagner and Eckhoff [78] provides a recent review of measures of privacy, aiming for comprehensiveness. This section reviews the literature on privacy measures from the narrower perspective of anonymity, dividing the landscape, as Wagner and Eckhoff [78] does, into approaches based on uncertainty, information gain/loss, data-similarity, indistinguishability, and accuracy, and attempting to relate terms to one another wherever possible.

**3.2.1 Entropy-based metrics.** For this section, fix a discrete random variable  $X$  with domain  $A_X = \{a_1, a_2, \dots, a_n\}$  and distribution given by  $P(X = a_i) = p_i$ , for  $i = 1, 2, \dots, t$ .

*Shannon Entropy.* One understanding of the information content of an outcome is given in the context of Information Theory, as

$$h(x) := \frac{1}{\log(P(x))}$$

where  $P(x)$  is the probability of the outcome  $x$  [69].<sup>36</sup> The *Shannon entropy* of the random variable  $X$  with values from the set  $A_X$  is the expected value of  $h(x)$  :

$$H_\emptyset(X) := \sum_{x \in X} P(x)h(x) = \mathbb{E}_P[h(x)]$$

The definition of Shannon entropy satisfies several intuitively desirable properties of information content.  $H$  is

- nonnegative,
- maximized for uniform distributions,
- additive.

Nonnegativity is desirable because observing an outcome should never reduce the amount of prior information available. The nonnegativity of the logarithm in  $h$  ensures nonnegativity for it and  $H$ . Uniformly distributed variables are the least certain;<sup>37</sup> therefore, to learn the outcome of a uniformly distributed variable is to learn as much as possible. That is,

$$\operatorname{argmax}(H(X)) \sim \mathcal{U}(0, 1).$$

The max-entropy is equal to  $\log |A_X|$ , which depends only on the number of people in the anonymity set [69]. This value has been taken as an anonymity metric itself, but fails to capture the information present in the distribution over the anonymity set. A more subtle use of the max-entropy is to normalize the Shannon entropy,

$$\tilde{H}(X) = \frac{H(X)}{\log |A_X|}.$$

Whereas the max-entropy quantifies uncertainty over a set *without* the distribution over  $X$ , the *min-entropy* quantifies the information required for the easiest task to achieve given the distribution over  $X$ : identifying the most probable element. The definition of min-entropy is simply

$$H_\infty(X) = \log(\max_{x \in X} P(X))$$

<sup>36</sup>This value is variously termed the *information content* of  $x$  [55], or the *information surprisal* or *self-entropy* [14].

<sup>37</sup>This is proved using Jensen's inequality.

The max entropy is helpfully seen as the information needed to identify an element of a set without its true distribution, while the min entropy is the information needed to identify the most probable element of a set when the distribution is known.<sup>38</sup>

*Rényi Entropy: a generalization of Shannon Entropy.* Rényi entropy is defined as

$$H_{\alpha}(X) = \frac{1}{1-\alpha} \log \sum p(x)^{\alpha}$$

the max- and min-entropy are then bounds of the domain of  $\alpha$ : with  $\alpha \rightarrow 0$ ,  $H_{\alpha}(X)$  converges to the max-entropy, and as  $\alpha \rightarrow 1$ ,  $H_{\alpha}(X)$  converges to the Shannon entropy. Finally, as  $\alpha \rightarrow \infty$ ,  $H_{\alpha}(X)$  converges to the min-entropy [17].

Entropy measures “uncertainty” in a distribution most generally. In the context of anonymity, the relevant uncertainty ordinarily concerns the identity of a sender or receiver, or a relationship between them. For example, the Panopticlick project is a transparency PET that describes the bits of Shannon entropy associated with various measured properties of a user’s browser.<sup>39</sup>

**3.2.2 Data Similarity.** Measures of data similarity quantify the risk of disclosure for databases containing personal information records, or *microdata databases*. Microdata typically combine high-dimensionality with sparsity—that is, more of the data subject’s attributes are kept to support exploratory analysis, and the attributes are broadly distributed.<sup>40</sup>

For concreteness, we illustrate the database setting with Figure 3, not discussed in Wagner and Eckhoff [78]. The data in question is *de-identified*, meaning the names, social security numbers, and other identifying fields have been removed; the attack aims to *re-identify* the data, finding a unique identifier for the records (which are often sensitive).<sup>41</sup>

*k-anonymity.* *k-anonymity* measures re-identification risk with the *anonymity set size*, or *bin-size* parameter  $k$ , which guarantees that an attacker<sup>42</sup> cannot reduce the size of the set of possible subjects identified with a particular row to less than  $k$  [72].

To achieve *k-anonymization* a technique must include all fields that can be used in a linkage attack. The practical difficulties of meeting the *k-anonymity* standard is illustrated by several well-known reidentification attacks. Sweeney [72] provided evidence that 87% of the American population could be uniquely identified by only three common demographic attributes: five digit zip code, full date of birth (day, month, and year), and gender; while 53% is uniquely identifiable with the zip code replaced by the city, town, or municipality in which the person resides; and 18% is uniquely identifiable with the zip code replaced by the county.<sup>43</sup> Later attempts to reproduce these percentages found a more modest 61% identifiable, rather than 87% in the same dataset, with 63% identifiable in a version of data from a later year [49]. Narayanan and Shmatikov [59] showed that a dataset released by Netflix as part of a competition could be re-identified for 99% of participants with eight

<sup>38</sup>In Shannon [69], the set was an alphabet. In the context of anonymity, it is a set of messages, or a set of identifiers. See section 3.1.

<sup>39</sup>See Eckersley [41], and <https://panopticlick.eff.org/about#about>. On transparency PETs generally, see section 4.

<sup>40</sup>High dimensionality and sparsity in microdata has been identified elsewhere Narayanan and Shmatikov [59], Rocher et al. [68], but without the explanation provided here.

<sup>41</sup>This is the so-called record linkage model. A more general model is attribute linkage, which is ignored here. See Fung et al. [44].

<sup>42</sup>The attacker is generic; a model of the attacker is not part of *k-anonymity*.

<sup>43</sup>This study, along with the work on differential privacy, informed the design of standards for privacy of health information, and for the release of Census, research, and statistical information [36, 39, 49].


	$X_1 :=$ Date of birth	$X_2 :=$ Height	...	$X_d :=$ Income	$X_{d+1} :=$ Diagnosis
$\mathbf{x}^{(1)}$					
$\mathbf{x}^{(2)}$					
$\vdots$					
 $\mathbf{x}^{(i)}$	01/09/1965	63	...	1.5M	Positive
$\vdots$					
$\mathbf{x}^{(n_D)}$					
$\vdots$					
$\mathbf{x}^{(s)}$	01/09/1965	63	...	1.5M	Negative
$\vdots$					
$\mathbf{x}^{(n)}$					

Fig. 3. Example database of  $n$  basketball players, with rows outside the sample of  $D$  rows shown in gray. The variables  $X_1, X_2, \dots, X_d$  are demographic attributes considered non-sensitive in isolation (they might also be considered values observable without the subject's cooperation), while  $X_{d+1}$  is a test result for a medical condition, and is considered sensitive. Attributes that together identify a data subject are *quasi-identifiers* [73]. If data sources can be combined that include the sensitive attribute and combine to provide a set of quasi-identifiers, then the bearer of the sensitive attribute is disclosed (this is an example of a *linkage attack*). The status of quasi-identifiers depends in part on the individual: in the database of professional basketball players shown, any set of attributes that includes height will be a quasi-identifier for Muggsy Bogues, who had the statistically unlikely value of 5'3" in height in the NBA. Sampling approaches that rely on the probability of an unobserved row in the dataset matching on quasi-attribute values are undermined by a version of this problem that applies much more generally when combinations of attributes (rather than only height, as in Muggsy's case) are considered [68].

movie rankings, while 68% could be reidentified with only two rankings. These and similar difficulties have led to the expression “anonymized data isn't” [52].<sup>44</sup>

*Using  $k$ -anonymity to measure reidentification risk in released data.* Sampling as a technique for achieving  $k$ -anonymization is frequently composed with other methods to provide a layer of plausible deniability in the event of a successful attack. The reasoning is that even if an attack appears to succeed in reidentifying a dataset, the attacker can never be sure that they have re-identified the right person, since a row that is unique in the sampled dataset might not be unique in a dataset inclusive of the entire population [35, 66].

This reasoning is challenged in recent research Rocher et al. [68], which presents a measure of identification risk in a full population that can be estimated very precisely for sampled datasets with common demographic attributes.

The broader risk model is built atop a method for modeling the joint distribution over  $X_1, X_2, \dots, X_d$ . The model for the distribution requires only the marginal distributions  $\Psi$  of the  $X_i$ , and the dependency structure  $\Sigma$ . Straightforward maximum likelihood estimation

<sup>44</sup>‘Anonymization’ of a dataset is another term for deidentification.

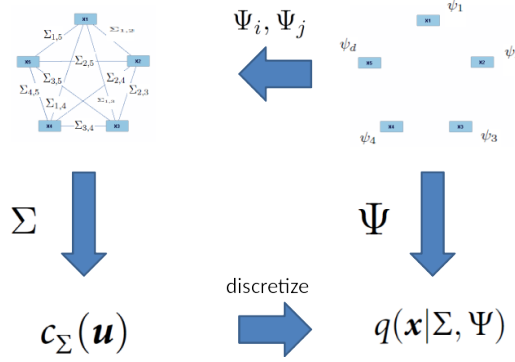


Fig. 4. High level structure of an approach to estimation of individual likelihood of identification Rocher et al. [68]. The individual marginal likelihoods  $\psi_i$  in the upper right of the figure can be estimated from limited sample data. The covariance matrix  $\Sigma$  can be computed from the pairwise matrices  $\Sigma_{i,j}$ , shown in the upper left of the figure, and represented in a separate copula distribution  $c_\Sigma(\mathbf{u}) = \frac{1}{\sqrt{\det \Sigma}} \exp(-\frac{1}{2} \Phi^{-1}(\mathbf{u})^\top \cdot (\Sigma^{-1} - I) \cdot \Phi^{-1}(\mathbf{u}))$ , as shown in the lower left. Finally, the model of the individual likelihood is given with  $\Sigma, \Psi$  as parameters:  $q(\mathbf{x}|\Sigma, \Psi) = \int_{F_1^{-1}(x_1-1|\Psi)}^{F_1^{-1}(x_1|\Psi)} \dots \int_{F_1^{-1}(x_d-1|\Psi)}^{F_1^{-1}(x_d|\Psi)} c_\Sigma(\mathbf{u})$ . This estimate is shown to perform well as a basis for estimating the individual likelihood of identification in a sampled dataset.

is used to estimate  $\Psi$  based on the available sample, while the estimation of  $\Sigma$  is computed by projecting the solution to the relaxed problem of estimating pairwise correlations. An important insight is that the separately-computed  $\Psi$  and  $\Sigma$  can be combined using a Gaussian Copula to accurately estimate the true joint distribution (see Figure 4).<sup>45</sup>

With a model of the joint distribution  $p(\mathbf{x}) = P(X = \mathbf{x})$ , the likelihood  $\xi_{x_i}$  that an individual  $\mathbf{x}^{(i)}$  is unique is simply the probability that  $\mathbf{x}^{(i)}$  is unique in all the rows, or  $(1 - p(\mathbf{x}^{(i)}))^{n-1}$ .  $\xi_{x_i}$  is in turn used to compute the probability that the number of elements in the anonymity set, or bin-size, for the individual is  $1, 2, \dots, n$ . Letting  $T$  be the binsize, we have a correctness score  $\kappa_x$  measuring the probability that a uniquely-matched row in a sampled database is identified.

$$\kappa \equiv \sum_{k=0}^{n-1} \frac{1}{k+1} P(T = k)$$

<sup>45</sup>Gaussian copulas were introduced by George Takeuchi [74].

since the sample records are assumed to be independently drawn,  $T$  follows the binomial distribution  $B(n-1, p(x))$ . Therefore the right hand side is equal to

$$\begin{aligned}
 & \sum_{k=0}^{n-1} \frac{1}{k+1} B(n-1, p(x)) \\
 &= \sum_{k=0}^{n-1} \frac{1}{k+1} \binom{n-1}{k} p(\mathbf{x})^k (1-p(\mathbf{x}))^{(n-1-k)} \\
 &= \frac{1}{np(\mathbf{x})} (1 - (1-p(\mathbf{x}))^n) \\
 &= \frac{1}{n} \left( \frac{1 - \xi_x^{n/(n-1)}}{1 - \xi_x^{1/(n-1)}} \right)
 \end{aligned}$$

The results of applying the model to the same source deidentified by early  $k$ -anonymity researchers were that “99.98% of Americans would be correctly re-identified in any dataset using 15 demographic attributes.” This result is a generalization of a more specific result, that the average likelihood of a user being unique in a particular census database with 15 demographic attributes<sup>46</sup> is 0.9998—generalization to “any dataset,” and any 15 attributes, follows from the application of the model to five other corpora, comprising “210 different datasets with differing levels of uniqueness and socio-demographic, survey, and health attributes that would be reasonable quasi-identifiers.”<sup>47</sup>

However, the evidence for this generalization is not entirely clear. Significant variation in estimated uniqueness  $\hat{\xi}_x$  was found when other attributes were added.<sup>48</sup> In addition, Rocher et al. [68] appeals to the low error of their estimates for *population* uniqueness  $\hat{\Xi}_x$ . In order to connect population uniqueness to individual uniqueness, the population uniqueness needs to be recoverable from individual uniqueness. But exactly this connection is the subject of a “calibration problem” reported in the supplementary material, which is left unexplained.

Finally, the mean absolute error (MAE) is used repeatedly by Rocher et al. [68] as evidence of performance; however, as an evaluation metric, MAE does not balance false positives with false negatives. The  $F$ -score does a better job of this. Rocher et al. [68] reports  $F1$  scores only in the back of supplementary materials, and the results are mixed: when choosing an  $\xi$  cutoff to minimize false positives ( $\xi > 0.90$ ), the  $F$  score drops to 0.21 and 0.33. At the threshold giving the best  $F$  scores ( $\xi = 0.50$ ), the  $F$  score ranges from 0.78 to 0.86 across the corpora.<sup>49</sup>

Therefore, while the row-level estimation of reidentification risk presented is a valuable new tool, the contours of its generalization remain unknown, undermining the very general claims used to promote the research to some extent.

**3.2.3 Indistinguishability.** ‘Differential privacy’ was first defined in Evfimievski et al. [43], and generalized by Dwork [38],<sup>50</sup> and has spawned a subarea of privacy research. Differential

<sup>46</sup>The dataset was the 5% Public Use Microdata Sample (PUMS).

<sup>47</sup>See “Supplementary Information” for the article, available at [https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-019-10933-3/MediaObjects/41467\\_2019\\_10933\\_MOESM1\\_ESM.pdf](https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-019-10933-3/MediaObjects/41467_2019_10933_MOESM1_ESM.pdf).

<sup>48</sup>See Rocher et al. [68], Figure 3(c), which shows the differing effect on estimate uniqueness of adding an different attributes.

<sup>49</sup>Supplementary table 7.

<sup>50</sup>The provenance of the term often mistakenly begun with Dwork [36], despite Dwork et al. [39] correcting the record. Even a book-length treatment by Dwork herself does not mention the original definition Dwork et al. [40].



privacy can be intuitively thought of as measuring a worst-case distance between query response distributions for databases that differ only on an individual row. Example queries include histogram queries, which partition the database into some number of bins, and counting queries, which return the number of rows satisfying a predicate. Ideally, closeness in this setting, will mean that reversing the transformations to a database will require a large sequence of queries.

*Differential Privacy.* A database may be used in ways that harm the subject, for example, to establish that Alice is twice as likely to require expensive medical treatment later in life, making her a greater insurance risk. These harms are independent of Alice's presence in the database, however. The harms that flow directly from Alice's inclusion in a database are prevented by differential privacy.

The differential privacy guarantee is that the distribution over responses to a query is unchanged by whether Alice is present in the database, at least approximately, up to a public parameter  $\varepsilon$ . Thus, differential privacy provides a specific guarantee about inclusion of a subject's data in a database; namely, that inclusion will not cause harm to the data subject [38].

The setting for differential privacy includes a randomized function  $\mathcal{K}$  over queries to a database.<sup>51</sup> Let  $x, y$  be databases differing in exactly one row, and let  $\mathcal{S}$  be a set of rows in the range of the randomized algorithm  $\mathcal{K}$  (which may be thought of as an approximate answer to a query). The definition of  $\varepsilon$ -differential privacy is

$$P(\mathcal{K}(x) \in \mathcal{S}) \leq \exp(\varepsilon) P(\mathcal{K}(y) \in \mathcal{S})$$

or, to highlight the role of  $\varepsilon$  as a bound,

$$\log \frac{P(\mathcal{K}(x) \in \mathcal{S})}{P(\mathcal{K}(y) \in \mathcal{S})} \leq \varepsilon.$$

The Kullback-Leibler divergence (*KL divergence*, or *relative entropy*) between the neighboring databases  $x$  and  $y$  provides an expectation of the proportion on the left side;  $\varepsilon$ -differential privacy bounds not the average but the worst case.<sup>52</sup> Composition of two mechanisms  $\mathcal{K}_1$  and  $\mathcal{K}_2$  with respectives that are respectively  $\varepsilon_1$ - and  $\varepsilon_2$ -differentially private results in a mechanism  $\mathcal{K}_1 \circ \mathcal{K}_2$  that is  $(\varepsilon_1 + \varepsilon_2)$ -differentially private [37].

Generalizations of differential privacy followed the original definition. One such generalization adds a parameter  $\delta$  :

$$P(\mathcal{K}(x) \in \mathcal{S}) \leq \exp(\varepsilon) P(\mathcal{K}(y, \theta) \in \mathcal{S}) + \delta.$$

$\delta$  is chosen to be small, less than a polynomial in the size of the database. This relaxation ensures that with probability  $1 - \delta$ , the *privacy loss* or *leakage* associated with an observation  $\xi$  is bounded by  $\varepsilon$ .

To achieve differential privacy for a given database requires a transformation. To guide the choice of transformation, the sensitivity of a query is defined with the  $\ell_1$ -sensitivity,<sup>53</sup>

$$\max_{x, y} \|f(x) - f(y)\| \leq S(f)$$

<sup>51</sup>Dwork et al. [40] calls  $\mathcal{K}$  a randomized “mechanism” or “algorithm.” This leads to an unfortunate collision of terminology with Economics when discussing “mechanism design,” particularly since economic perspectives are sometimes explicitly invoked (see for example section 2.3.1, “What differential privacy promises: An Economic View”).

<sup>52</sup>This connection to KL divergence is not noted in [36, 38, 40].

<sup>53</sup> $\ell_2$  sensitivity is defined similarly, using the Euclidean distance.

[37, 39]. Noise can then be added, or other transformations applied, in a way that is scaled to the sensitivity of the function.

## 4 TRANSPARENCY

Transparency is central to undermining an escalating privacy arms race, since information about reduction in product or service quality via privacy attacks advantages competitors who are able to provide higher quality services, offsetting competitive gains incentivizing privacy attacks.

In this section three studies are discussed: one of a particular company, Facebook (FB), which is conducted to determine whether PII collected from contexts with a clear purpose were being used for ad targeting; another is a study of the user response to revelation of Google's new policy of sharing data between their more than 60 services (at the time); and a final study explores tracking on the web more generally with an automated transparency tool, which is used to investigate the prevalence of tracking across the web.

### 4.1 Reconstructing privacy practices by experimentation

**4.1.1 Facebook's repurposing of phone numbers collected for two-factor authentication.** The goal of Venkatadri et al. [77] is to infer whether personally identifiable information (PII) collected by Facebook (FB) in a variety of different contexts is repurposed for advertisement targeting. The study explores, in particular, which of the following seven sources of PII are used for targeting:

- (1) **PII provided to WhatsApp**
- (2) **PII uploaded by other advertisers to target customers via custom audiences**
- (3) **PII added directly to a user's profile**
- (4) **PII provided through FB Messenger**
- (5) PII shared with FB when sharing a phone's contacts
- (6) PII added to user accounts for two-factor authentication
- (7) PII added for login alerts

Bolded items above indicate that PII from the source was used for ad targeting. There are several notable results reported by this study. The central, high-level result is that techniques used to protect privacy, such as adding noise to statistical results, are used by FB to hide privacy attacks in the form of data repurposing. Data repurposing in this case involves collecting data from a source with a clearly defined purpose, such as account security, and using it for an unrelated purpose, such as ad targeting.

Transformations performed on the data FB reports were modified to thwart methods described in previous academic research into reconstructing their practices. FB hid from users a reuse of their data not only by failing to provide adequate consent,<sup>54</sup> but by obstructing the progress of research that would increase their transparency.

**4.1.2 Methods.** A merchant with an interest in advertising to its customers on FB has the option of importing a list of PII imported from its own site. A *custom audience* comprises the FB users who match the PII provided by an advertiser to Facebook. Fifteen types of PII are allowed by FB, but the study only made use of *email address* and *phone number*. *Potential reach* is the number of active daily users in a custom audience, with several transformations performed to obscure the true number.

<sup>54</sup>Notice was either missing altogether or too general to be substantial, as will be discussed in the next section.

To generate ground truth data for testing, the authors of the study recruited 103 emails and phone numbers of known users of FB, and also produced “dummy” accounts with “impossible” PII that is known not to match any user: phone numbers with too many digits, and email addresses that use very large random strings.<sup>55</sup>

*High level.* To clarify the roles of the merchant, the user, and FB, the high-level description of the approach taken is, first, to find a test for the presence of PII on FB using potential reach. Once such a test is found, a before-after test can be conducted in which PII is first shown to be missing, then becomes available after a user under the researchers’ control adds it to FB using one of the sources listed in section 4.1.1 above. Furthermore, a result confirming that Alice’s PII has become available for ad targeting via a particular source is further verified by purchasing advertisements for the audience, and observing whether the ad is shown to Alice.

*More detailed description.* Previously, researchers had shown that FB was hiding the true value of potential reach by simple rounding. In addition to the existence of a greatest common denominator (*GCD*) above 1 for all potential reach values (what they call “granularity”), another consequence of rounding is monotonicity: when simple rounding is the only transformation, increasing the custom audience sizes should increase the reported potential audience.<sup>56</sup> Tests did not show monotonicity after the previous study was published, suggesting that noise is newly being added to (or subtracted from) the potential reach. Since a *GCD* of 10 was found for the potential reach estimates, the conclusion of this part of the reconstruction was that rounding was being combined with noise in the estimate.

*Noise .* Noise values can come from any distribution, and can be combined with the true potential reach by addition, multiplication, or any other function. The distribution would be recoverable if the researchers were able to receive new samples from the noise distribution by repeatedly querying for estimated potential reach while holding the custom audience fixed, so FB always reports the same (noisy) value for a given custom audience.

To circumvent this defense, the researchers first note that to repeatedly generate the same value from the noise distribution, a seed must be used. Since the value of the potential reach depends on the custom audience, the authors surmise that a hash of the custom audience is used to generate a seed.<sup>57</sup> Recovering the noise distribution is then a matter of forcing a new seed to be chosen for a custom audience while holding fixed the potential reach. The solution is to upload multiple custom audiences with a varying “dummy” record that is known not to affect the potential reach. This resets the seed for the noise, and produces a new estimate.

The authors use their ability to force new noise samples by uploading lists of 70 to 90 phone numbers, incrementing by a single phone number, and obtain 1,225 samples at each audience size. The samples are found to remain within a fixed range of thirty values or less, indicating that the noise component is bounded within a fixed range for all audience sizes.

<sup>55</sup>“We generate dummy phone numbers by adding a random sequence of 20 digits to the Italian country code (+39). Since Italian phone numbers do not exceed 12 digits, these dummy numbers cannot correspond to any Facebook user.” Similarly impossible email addresses are also generated, and their lack of match to any FB user validated by confirming the floor estimate of potential reach, 20, for 1000 dummy values. (231)

<sup>56</sup>The same will be true for any transformations that are, collectively, monotone increasing, when composed with rounding.

<sup>57</sup>The precise role of seeded random number generation in sampling from the noise distribution is not explained. In general, a plausible explanation is that a number selected uniformly at random from 0 to 1 is generated from the seed, and used to sample from the inverse-CDF of the noise distribution. In the particular case of FB, the noise distribution itself was the uniform distribution from 0 to 20, so the task was simplified.

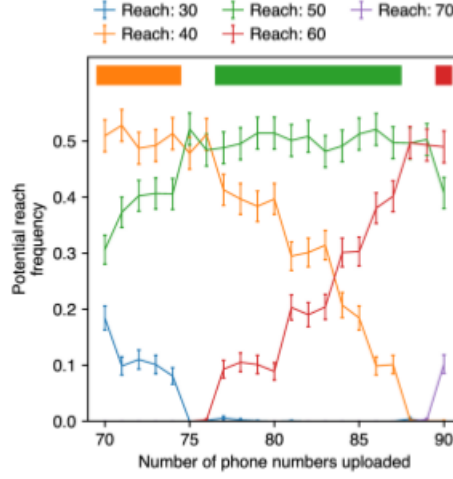


Fig. 5. The median of the potential reach reported by FB occurs with a constant frequency of 0.5 across lists of phone numbers of increasing size, indicating uniform noise is added. *Number of phone numbers uploaded*, as represented in the chart, includes only known FB users. The activity status was not tracked for the users associated with the phone numbers, so the true “potential reach” value cannot be directly compared with the estimated potential reach reported by FB, which is used to generate the *Potential reach frequency* shown in the chart. [77, 234]

Moreover, since the values are always multiples of 10, the estimates reported are rounded. For example, a list of 70 numbers might have estimated reach values from  $\{30, 40, 50\}$ .

The evidence so far does not determine the order in which rounding and noising occur. To determine which order is correct, the distribution of observed values versus the size of the custom audience is considered (see Figure 5). Shifts in the frequency of a given estimated potential reach value (say, 40) do not happen only when the custom audience reaches a multiple of 10, as would be expected if noising occurs after rounding; thus the noising must occur before rounding by 10.<sup>58</sup>

In addition, the frequency of a given estimated reach value (say, 40) is seen to change in uniform steps of about 0.1, consistent with additive noise—for multiplicative noise, the step size would increase with an increase in the custom audience.

A subtler observation from 5 is that, whereas the median estimated potential reach has a frequency of 0.5, the other two values observed for a given audience size change as the true potential reach grows. Since the expected frequency of occurrences of the mean potential reach is  $\frac{10}{m}$ , this shows that the distribution of noise is uniformly distributed between 0 and 20.

This result is inconsistent with the observed step size for the smallest observed value as the custom audience grows: since one fewer value can be mapped to the smallest potential reach as the custom audience size increases, the frequency of the smallest potential reach

<sup>58</sup>In addition, clearly if noise is added after rounding, then it must be sampled from a distribution over multiples of ten to be consistent with the values observed.

should decrease by 0.05, not 0.1 as observed.<sup>59</sup> To accomodate this observation, an additional rounding step must be included before noising, which rounds the true potential reach by 2.

Ultimately, the result of this stage of the reconstruction is a composition of transformations that FB performs on the precise potential reach value: rounding by 2, adding uniform noise from integers in  $[1, 20]$ , and rounding the result by 10.

*Going on offense in transparency research.* A database reconstruction attack can be used to identify subjects in a database, harming privacy. Conversely, attempts to prevent reconstruction by performing transformations such as adding noise or rounding can be used to protect against such harms. In the present context, however, the roles of attack and defense are reversed: database reconstruction is used to serve a privacy goal, ensuring that users of FB are informed of the repurposing of data collected from a source with a clear purpose, such as security, in order to serve a different purpose, targeting advertisements. Similarly, transformations are used to hide these practices.<sup>60</sup>

Going on offense in privacy research in this way contributes to conditions favorable to privacy, as consumer privacy preferences are better served by a market with greater transparency regarding the privacy practices of services like FB.

## 4.2 Measuring privacy practices on the web with instrumented browsing

Englehardt and Narayanan [42] describe a scalable, modular, fully instrumented PET (*OpenWPM*) comprising a collection of tools that can be used to collect a variety of data for later analysis. OpenWPM is used for two studies: first, to measure the prevalence of third-party tracking in general across the most popular sites on the web (what the authors call a “Web Census”); and second, to study fingerprinting on the web.

Fingerprinting is a stateless technique for tracking users, measuring properties of users’ devices that together can be used to identify a user across different sites. For example, a script might ask for the list of fonts available to the user’s web browser, and measure the bounding boxes in which they are drawn.

The study updates earlier measurements of known fingerprinting techniques, and also identifies and measures new fingerprinting methods. Six fingerprinting techniques are measured, based respectively on canvas drawing, canvas fonts, local IP discovery via webRTC, audio signal processing, and fine-grained battery status. All but the use of audio processing were previously known.

**4.2.1 The Web Census: results from a crawl of 1 million websites.** Several studies have been published using OpenWPM. The “Web Census” study measured tracking practices across the web, recording detailed information regarding a variety of fingerprinting techniques, and evaluated the effectiveness of a variety of tools in blocking tracking.

The list of websites to visit was determined in advance from the Alexa top 1 million sites.<sup>61</sup> HTTP request/response pairs, Javascript calls, and Javascript files are recorded for each site visited. Trackers were defined as any resource blocked by a “consumer privacy

<sup>59</sup>For example, consider in 5 the blue line for an estimated potential reach of 30 from 70 to 75 uploaded phone numbers, followed by the orange line for an estimated potential reach of 40 for 75 to 80 uploaded phone numbers.

<sup>60</sup>Note that FB cannot claim to be protecting the PII of its users with the transformations it performs—as noted, user PII is already in the hands of the merchant who uses custom audiences, so there is nothing to protect, beyond whether the user is active on FB.

<sup>61</sup>Of the 1 million sites in the list, 82,749 failed to load correctly.

tool—in particular, this includes third parties whose URL matches against either of two “tracking-protection lists,” EasyList and EasyPrivacy.

The crawl found 81,000 third parties present on more than two first party sites, with Google, Facebook, Twitter, and AdNexus each present on more than 100,000 first party sites. Google, in particular, is found to be present on approximately 850,000 first party sites, of which 750,000 are tracking contexts. The next most prevalent organization, Facebook, is present on 350,000 to 400,000 sites, of which between 250,000 and 300,000 are in a tracking context.

*Limitations of data collection methods of the Web Census.* The techniques used by the Web Census for visiting sites were also the foundation for fingerprinting analysis. The use of the Alexa rankings to provide websites for research purposes is not uncommon; however, it bears mentioning that these rankings reflect the browsing habits of a group that is likely to be unrepresentative: users of the Alexa toolbar.<sup>62</sup> These rankings have been found inconsistent with other web analytics companies, including comScore and Nielsen.<sup>63</sup> To better explore the “long tail” of trackers with low prominence that is discussed in the work,<sup>64</sup> an interesting follow on study would explore a list of 1 million websites found by crawling networks of sites connected by hyperlinks.

The OpenWPM HTTP instrumentation does not store HTTP Archive (HAR) files contributors [22]. This is a limitation, since HAR files contain the full response contents that can be used for analysis, rather than only HTTP headers.

In addition to the limitations in the strategy for choosing sites to visit, the nature of the interaction one on a site was minimal: only the homepage was visited. This limitation is acknowledged by the authors,<sup>65</sup> but it is not noted that some of the most valuable contexts for tracking a user arise only on further navigation within a site. For example, tracking users who place an item in a shopping basket but navigate away before making the purchase are an important target for “abandoned cart” follow up emails.<sup>66</sup>

Finally, the definition of ‘tracker’ used by the study appeals to the lists of trackers used by ad blocker tools, including EasyList and EasyPrivacy.<sup>67</sup> This presents a problem for the external validity of the study: in effect, relying on lists makes the list maintainers the only ones capable of arguing that the study applies beyond the particular third parties on the list, to trackers in general.

The limitations of the Web Census should not be overemphasized: the study is an important step toward increased transparency on the web that does not rely on industry cooperation or self report. Such broad sweeps are useful in law enforcement, so that practices evolving to become illegal can be caught expeditiously.<sup>68</sup>

<sup>62</sup> Alexa claims that their ranking system reflects sources of data beyond toolbar users, but this depends on site operators running Javascript on their sites.

<sup>63</sup> See <https://www.techcrunch.com/2007/11/25/alexa-makes-believe-internet>.

<sup>64</sup> “Prominence” is a new metric proposed by the work. See section 4.2.2.

<sup>65</sup> Section 4, “Limitations.”

<sup>66</sup> See for example <https://mailchimp.com/help/create-an-abandoned-cart-email/>, and <https://www.shopify.com/blog/abandoned-cart-emails>.

<sup>67</sup> See <https://easylist-downloads.adblockplus.org/easyprivacy.txt>

<sup>68</sup> Compare the “funeral sweeps” performed by the Federal Trade Commission to police the Funeral Rule. See <https://www.ftc.gov/news-events/press-releases/1996/09/ftc-announces-results-another-funeral-home-sweep>, <https://www.ftc.gov/news-events/press-releases/1996/03/delaware-area-funeral-home-sweep-results>, [http://www.ofdaonline.org/aws/OFDA/pt/sd/news\\_article/11738/\\_PARENT/layout\\_details/false](http://www.ofdaonline.org/aws/OFDA/pt/sd/news_article/11738/_PARENT/layout_details/false).

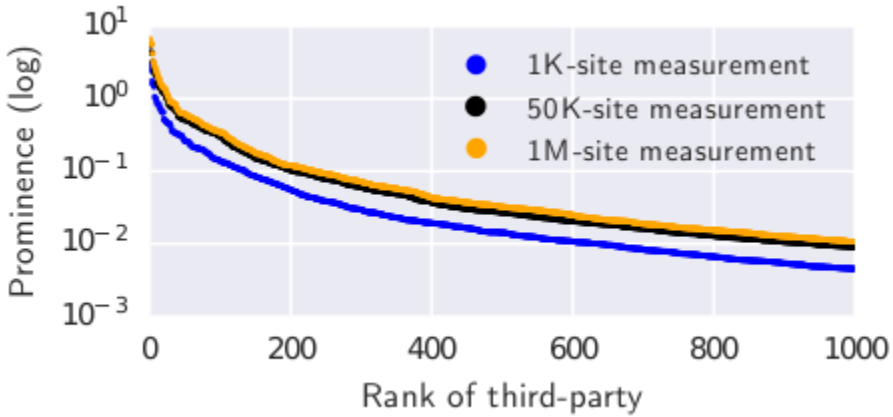


Fig. 6. Frequency with which a user will encounter a third party, according to the Prominence metric, as a function of third-party rank. Note the log scale. This rank-prominence curve is proposed as a basis for summary comparison of the tracking ecosystem over time. [42, 9]

**4.2.2 Quantifying tracking.** A metric is presented that measures the frequency with which a user will encounter a given third party  $t$ . The metric is described schematically as including not only the number of first party sites, but their popularity rank, represented as  $\text{rank}(s)$ :

$$\text{Prominence}(t) = \sum_{\text{edge}(s,t)=1} \text{VisitFrequency}(\text{rank}(s))$$

where  $\text{edge}(s,t)$  indicates whether third party  $t$  is present on site  $s$ , and  $\text{VisitFrequency}$  describes the frequency of visits to  $s$ . The inclusion of site rank penalizes obscure sites. Citing research that  $\text{VisitFrequency}$  should follow a power law distribution, they propose as an instantiation of the metric

$$\text{Prominence}(t) = \sum_{\text{edge}(s,t)=1} \frac{1}{\text{rank}(s)}.$$

A claimed advantage of this metric over a simple count of first party sites is its ability to robustly represent changes in the tracking ecosystem over time via a *rank-prominence curve* (see Figure 6).

*Limitations of Prominence.* The risk presented by a website with third-party trackers is related to the probability that the site will be visited by a large number of people. However, the same risk is also related to the nature of the tracking—tricking users into installing a malware toolbar will provide high-fidelity tracking, for example. Since a principled instrument for measuring the severity of individual trackers would be difficult,<sup>69</sup> a simple alternative is to include the number of trackers combined on a single site:

$$\text{Prominence}'(t) = \sum_{\text{edge}(s,t)=1} \frac{1}{\text{rank}(s)\text{tracker\_count}(s)}$$

<sup>69</sup>This is related to the difficulty in defining what constitutes a tracker, which the authors acknowledge in section 4, “What makes a tracker?”

This captures not only the number and popularity of the sites in which a tracker appears, but also the severity of the tracking, since many trackers combined provide tracking that reveals more, and is more difficult to avoid.<sup>70</sup>

**4.2.3 A study of fingerprinting.** Fingerprinting techniques are measured across the approximately 917K sites encountered in the study, using Javascript instrumentation to “... monitor access to all built-in interfaces and objects we suspect may be used for fingerprinting.” (11) This approach makes an end-run around attempts to thwart static analysis via code obfuscation,<sup>71</sup> which is reportedly commonly used by trackers: “We find that fingerprinting and tracking scripts are frequently minified or obfuscated, hence our dynamic approach. With our detection methodology, we intercept and record access to specific Javascript objects, which is not affected by minification or obfuscation of the source code” (18).

*Canvas fingerprinting.* The Canvas interface allows higher level web-applications to accomplish the low-level tasks of drawing realtime graphics. The role of low-level elements makes the performance of this interface different across devices, and therefore useful for fingerprinting: “Differences in font rendering, smoothing, anti-aliasing, as well as other device features cause devices to draw the image differently. This allows the resulting pixels to be used as part of a device fingerprint.”

Canvas fingerprinting was found on 1.6% of sites visited, nearly all (98.2%) from third-party scripts. This is a larger number of sites than had been found in a similar study performed two years prior, and the set of trackers utilizing canvas fingerprinting also changed to be dominated by less prominent trackers. The vast majority of uses of canvas fingerprinting came from sites using the technique to track for (at least, ostensibly) fraud detection. The authors remark that this is “in line with the ad industry’s self-regulatory norm regarding acceptable uses of fingerprinting.”<sup>72</sup>

*Canvas font fingerprinting.* The simple version of canvas font fingerprinting requires only an enumeration of fonts via the Canvas interface. Such enumeration was detected in only a small percentage of sites (<1%), including a slight bias toward more highly ranked sites (2.5%), and nearly all came from a single third party, mathtag.com.

Detection was accomplished by instrumenting the `measureText` method, which returns metrics on HTML elements with text inside them. Cases were categorized as fingerprinting if the font property was reset and measured at least 50 times on the same string.

*WebRTC-based fingerprinting.* WebRTC is a Javascript framework allowing realtime communication in the browser. For example, Jitsi allows self-hosted video conferencing over the web using WebRTC.<sup>73</sup> To provide low-latency communication, peers using WebRTC communicate information about their local network environment without user permission

<sup>70</sup>Avoidance, and the net cost to the consumer of being tracked, are two parts of the accepted analysis given by the Federal Trade Commission of practices actionable under the unfairness standard in Section 5 of the FTC Act. See <https://www.ftc.gov/public-statements/1980/12/ftc-policy-statement-unfairness>.

<sup>71</sup>Although it “has proven hard to pin down exactly what [code] obfuscation is” it can be informally understood as the transformation of a program into a form that is still executable but from which it is hard to extract information. See Collberg and Nagra [19, Ch. 4].

<sup>72</sup>This remark is somewhat undercut by the partnership fraud detection providers with companies serving behavioral ads. For example, the vast majority of uses of canvas fingerprinting were from doubleverify.com, which has partnered with the marketing company InMobi for the purpose of serving native ads in video while reducing fraud.

<sup>73</sup>A popular and free hosted Jitsi service is <https://meet.jit.si>.



(unlike, for example, video or audio access through WebRTC). Local and public IP address information can be used as an identifier for tracking purposes.

To detect this attack, `RTCPeerConnection` is instrumented to detect access to an event handler, `onicecandidate`. The authors then manually tested if the scripts used the IPs for tracking by checking “if the code is located in a script that contains other known fingerprinting techniques” (12).

Again, only a tiny number of sites used WebRTC in the way described, with only 10 scripts reused frequently (83% of usage).

*AudioContext fingerprinting.* By identifying scripts that perform several fingerprinting techniques, the authors were able to identify a previously unreported technique that uses the `AudioContext` interface. The idea behind the technique is at a high level similar to canvas fingerprinting: the `AudioContext` interface mediates interaction with a low-level process via Javascript, just as the `Canvas` interface mediates low-level functionality needed for drawing. Low-level processing varies between devices (including browsers), but are stable within a particular device, so the precise output can be used as part of an identifier.

For concreteness, one attack is performed by passing a prepared audio signal to an `OscillatorNode` object to generate a waveform from the signal (the two attacks seen generated sine and triangle waves, respectively). The waveform is compressed and stored in a buffer as a sequence of numeric values, and the sum of the values is hashed to generate a fingerprint.<sup>74</sup>

*Battery API fingerprinting.* The battery level or charging status of a host device has a large number of states held for sufficient duration to be useful as an identifier. One use of this technique that is mentioned as possible, but is never attested in a particular case, is to overcome privacy protections that are employed while a user is on a site. This would make the Battery API a niche technique to overcome a particular case of increased defenses. Only two scripts were found to be using this technique.

The authors do not describe their method for detecting this attack, though they point to other work that provides greater detail.

*Limitations of the fingerprinting study.* For questions about fingerprinting that are statistical in nature, missing a few websites with especially strong protections against bots may be acceptable. For the purposes of enforcement against illegal practices, the success of sites in avoiding measurement is an important limitation, as such sites are likely to be the most sophisticated users of tracking. A separate study of bot countermeasures would be a useful to address this gap.

The fingerprinting study relies heavily on manual filtering to detect fingerprinting methods. In the description of detection methods for WebRTC, for example, calls to a particular suspicious API lead to a manual check of whether other fingerprinting techniques are present in the same script as the call. This approach is in conflict with a point they make elsewhere against static analysis of fingerprinting scripts (and motivating their dynamic approach), that “[fingerprinting] script content may be obfuscated to the point where manual inspection is difficult and the purpose of the script unclear” (14).<sup>75</sup>

Again, the most sophisticated users of fingerprinting—who are among the most valuable to measure—obfuscation is likely to be employed, resulting in misses by the study. It is also worth adding that manual approaches to measurement do not translate well to tools usable

<sup>74</sup>A more complicated attack is also described, using a Fast Fourier Transform of the signal as well. See Englehardt and Narayanan [42, Appendix section 12].

<sup>75</sup>This point is made to argue against blacklists as an approach to blocking fingerprinting trackers.

as countermeasures to fingerprinting. Given the long tail of little-known third parties acting as trackers, who are less likely to be “responsive to the scrutiny resulting from privacy studies” (2), this shortcoming should not be overlooked.

## 5 USABILITY

Threats to usability can be threats to privacy. Faced with regulatory requirements to provide meaningful notice and substantial consent when collecting personal information,<sup>76</sup> some services make use of crafted user-interfaces (UI) to bias the distribution of responses in toward increased sharing. This practice has become common enough to have acquired a name: “dark patterns.”

From the perspective of the privacy arms race, dark patterns highlight the difficulty with reliance on legal or policy solutions. Dark patterns are in effect a gaming of the paradigm for taking consent; however, they do not fool consumers. Several “name-and-shame” websites have sprung up to collect complaints about dark patterns.<sup>77</sup> Such sites are a variety of APET, increasing pressure on companies by increasing transparency about their practices. Other potential responses include technologies for automating UI interactions, for example, providing a one-click method for deleting a Google account, or setting the most restrictive privacy preference settings in Facebook. Another approach seeks to remove specific problematic UI elements, such as nags or “alerts” to provide more information.<sup>78</sup>

### 5.1 Dark patterns

‘Dark patterns’ is an informal term, with existing empirical study almost nonexistent;<sup>79</sup> however, examples labeled as “dark patterns” have attracted consumer complaints for several years, and recently proposed legislation has targeted dark patterns for federal enforcement.<sup>80</sup> Dark patterns are, most broadly, high-level descriptions of user interface (UI)<sup>81</sup> designs

<sup>76</sup>See <https://gdpr.eu/gdpr-consent-requirements/>.

<sup>77</sup>See <https://darkpatterns.org>, and <https://www.reddit.com/r/asshotedesign/>.

<sup>78</sup>See DeleteMe, available at <https://joindeleteme.com> and JustDeleteMe, available at <https://backgroundchecks.org/justdeleteme/>. A related extension is available for Chrome at <https://chrome.google.com/webstore/detail/justdeleteme/hfpofkfbabpbmchmiekfnlgaedbgcf>. See also AccountKiller, available at <https://www.accountkiller.com/en>. For an example of a preference-setting APET, see “Privacy Settings,” which changes default configuration settings that are not fully available in the browser’s preferences UI to limit sharing and improve resistance to tracking via fingerprinting. See <https://add0n.com/privacy-settings.html>.

<sup>79</sup>In total, one web measurement study and two user studies have been conducted. The measurement study and one of the user studies are unpublished at the time of this writing. For the web measurement study, see Mathur et al. [57]. For the user studies, see Conti and Sobiesk [21], Luguri and Strahilevitz [54]. For a sampling of the popular literature on dark patterns, see <https://www.consumerreports.org/privacy/how-to-spot-manipulative-dark-patterns-online/>, <https://gizmodo.com/senators-introduce-bill-to-stop-dark-patterns-huge-plat-1833929276>, <https://techcrunch.com/2018/07/01/wtf-is-dark-pattern-design/>, <https://www.theverge.com/2013/8/29/4640308/dark-patterns-inside-the-interfaces-designed-to-trick-you>.

<sup>80</sup>See generally Sen. Mark Warren and Sen. Deborah Fischer’s “DETOUR Act,” proposed in April 2019, and the companion whitepaper from Sen. Warren, ‘Potential Policy Proposals for Regulation of Social Media and Technology Firms,’ which was released in the summer of 2019. Warren and his staff have collaborated with Tristan Harris, co-founder of the Center for Human Technology, on the DETOUR Act since 2017. See ‘Deceptive Design and Dark Patterns: What are they? What do they do? How do we stop them?’, available at <https://www.youtube.com/watch?v=OjmMdfeliq4>, at 1:12:56 (in which Harris mentions the length of the collaboration at a panel discussion convened by Sen. Warren) and at 0:38:00 (in which Warren mentions his hope that the DETOUR Act be integrated into a broader federal privacy bill).

<sup>81</sup>A user interface is the arrangement and functionality of elements used to interact with a computer system—for example, the buttons, pictures, and text that allow the user to read the news on a webpage; or

Dark Pattern	Description
False warning	A warning about a problem that does not exist. May be used to discourage a behavior by making it appear to be dangerous, such as adding a warning icon.
Roach Motel	A design making it easy to do something, but difficult to undo it. Often used to establish a subscription relationship that is difficult to cancel.
Social proof	Inclusion of graphic or textual elements falsely indicating popularity, such as a counter with a number of fake people who have purchased an item.
Urgency	Requires a decision to be made “before the opportunity is lost;” for example, under false time pressure or with visible offers from other customers who do not exist.

Table 1. Introductory examples of a dark pattern.

for computer systems that influence people to do something they otherwise wouldn’t—in particular, one common use of dark patterns is to influence people into surrendering personal information that they would otherwise keep confidential.<sup>82</sup> Attempts to isolate a more specific working definition of dark patterns has taken the form of competing taxonomies in the literature Chatellier [12], Conti and Sobiesk [21], Mathur et al. [57], Moser et al. [58] (see Table 1 for several introductory examples).

This section discusses two recent attempts to formalize and study dark patterns; one is a naturalistic study measuring the prevalence of a particular subclass of dark patterns on the web Mathur et al. [57], while the other is an experimental study of the effectiveness of a sequence of dark patterns on user choice Luguri and Strahilevitz [54].

#### 5.1.1 Measuring textual dark patterns on the web at scale.

*Data collection*. Mathur et al. [57] develops a semi-automated approach to identification of textual dark patterns on shopping websites. An automated web browser instance (“crawler”) attempts to make a purchase from a set of shopping websites, and extracts textual “segments” from the page.

English language shopping websites are identified from the Alexa Top Sites list using a website classification service, and an off-the-shelf language detection python package.<sup>83</sup>

Once English shopping sites are identified, a search is conducted for product pages. A supervised classifier is constructed for this purpose. To generate a training set, a depth-first search is performed.<sup>84</sup> The resulting set of pages are manually labelled as “product” or

the timing and appearance of advertisements, signups, and prompts for payment information that allow a user to make a purchase at a point-of-sale terminal. See Wilson [82]

<sup>82</sup>This descriptive use of ‘pattern’ in ‘dark patterns’ is in contrast to the more standard prescriptive usage, on which designers who repeatedly encounter a set of related problems come to identify best practices for their solution, which they call a “design pattern.” The prescriptive use originated with Architecture; see Alexander [2].

<sup>83</sup>Webshrinker and polyglot, respectively. The crawler used is built from the python Selenium package.

<sup>84</sup>We interpret the use of “naive depth-first crawler” as slightly unusual, in that the crawler does not exhaust the space of reachable URLs from a given site, but rather explores one random path until it can proceed not farther. Our interpretation is consistent with a dataset of “several thousand URLs” from 100 websites. See section 4.2.1.

“non-product,” and used to train a Logistic Regression classifier on syntactic features of URLs, including “the length of a URL, the length of its path, the number of forward slashes and hyphens in its path, and whether its path contained the words ‘product’ or ‘category’.”

A crawl is then performed on the English shopping sites, guided by the likelihood score that a URL was a product page.<sup>85</sup> Classifier performance is evaluated by randomly manually review. The reported false negative rate for the sample was 0.14, and the false positive rate was 0. The final false negative rate was likely higher, as only 11,286 shopping websites of the original 19,455 were used (approximately 58%), with 53,180 pages identified as product pages.

*Classifier design.* Textual segments are used to produce clusters that “organize the segments in a manner that would be conducive to scanning, making it easier for an expert analyst to sift through the clusters for possible dark patterns.”

Prior to clustering, two steps of preprocessing are performed on the data, removing stop words and duplicates, and retaining only tokens appearing in at least 100 segments. A feature extraction step follows, in which bag of words features are generated from each segment.<sup>86</sup>

Hierarchical clustering is performed using a “canned” discriminative classifier, Hierarchical DBSCAN. The output of the classifier is used to extract a flat partition of the best textual segment clusters—a hierarchical description of the segments is not used directly.

Next clusters are presented to human reviewers, who identified “... clusters that represented specific types of user interfaces (e.g., login choices, cart totals), website characteristics (e.g., stock notifications), and product options (e.g., small/medium/large) that generally appear on shopping websites.” This represented a scan of 10,277 clusters for the reviewer. This initial step is followed by another manual scan, in which dark patterns were identified based on “a shared understanding of possible dark patterns” using “the literature on dark patterns and impulse buying, and media coverage of high-pressure sales and marketing tactics.” Finally, websites with identified dynamic dark patterns were monitored every four hours for five days.

A fuller description of what was tried before settling on the semi-automated approach presented would provide valuable information about the future prospects for a fully-automated detector of novel dark patterns. From the perspective of the privacy arms race, automated detection of dark patterns increases the risk of suffering costs for their use; if effective, the risk of lost customers may be high enough to undermine the use of dark patterns for increasing conversion of (fewer) visitors into buyers.

*Simulating user interaction.* The dark patterns of interest to Mathur et al. [57] are sometimes accessible only by reaching the checkout stage of interaction with a shopping website. To do so, several simple sorts of website interaction are automated, including selecting items to purchase, and clicking on “cart and checkout buttons” to move the visitor through the checkout process.<sup>87</sup> The success of the crawler is misleadingly presented; Mathur et al. [57] report success for 66 of 100 randomly sampled product pages, but later report only 11K sites out of 56K successfully visited.

<sup>85</sup>The search algorithm is left unspecified beyond that a priority queue ordered by product page likelihood was used.

<sup>86</sup>No justification is given for this number, and no description of how it was reached (for example, whether other numbers were tried).

<sup>87</sup>An additional part of simulated user interaction involved avoiding bot detection by websites, or content delivery networks that serve website traffic. This part of the project is not discussed in Mathur et al. [57]; it was mentioned by Arvind Narayanan to the author.

The poor performance of the tool in user simulation precludes its use to audit individual sites for bad privacy practices. An adversarial site can thwart a very brittle measurement tool, or even behave differently when it is being measured. More generally, from the arms race perspective, broad-spectrum blocking of measurement bots is likely to be effective against measurement bots, since they are not engineered for interaction with any particular site (unlike, say, a scraper for LinkedIn), but are designed to interact with many sites. The upshot of this is that broad-spectrum blocking makes conditions for privacy less transparent on the web.

*Result analysis.* 1,818 instances of dark patterns were found on 1,254 websites of the total 11K crawled. This is a lower bound, measuring only particularly simple textual dark patterns, and failing to measure large portions of the data set.

Dark patterns are found to be more likely to appear on popular websites, suggesting a lack of cost in the form of reputational harm even when large numbers of visitors are confronted with a dark pattern.

In addition, an ecosystem of third parties were found to be providing dark patterns on behalf of many shopping sites.<sup>88</sup> 22 third parties of this type were found on a vast majority of websites with dark patterns (1,066). The existence of a market to meet demand for dark patterns is a testament to their effectiveness in converting visitors to buyers.<sup>89</sup> Whether the development of dark patterns driven by competition in the market will result in more aggressive dark patterns is not clear; if so, a complementary market for circumventing dark patterns would be a clear manifestation of the privacy arms race in the subfield of usability for shopping websites.

*5.1.2 Exploring the effect of dark patterns on user choice.* Luguri and Strahilevitz [54] notes that “many have assumed that dark patterns are efficacious. Why else would large, well-capitalized companies that are known to engage in A-B testing be rolling them out?” (5) The study they conduct aims to produce evidence supporting this assumption, and finds a strong effect for at least one particular sequence of dark pattern manipulations: more than double the percentage of consumers were signed up for a dubious identity theft protection service when exposed to a sequence of mild dark patterns, while the percentage nearly quadrupled for a stronger sequence of dark patterns.

The experiment design was a “bait-and-switch scenario” in which a survey on privacy attitudes would be presented, and used to deceive participants into believing that the strong interest in privacy that they presented was used to justify signing them up for an expensive identity-protection service, for which they could opt out.

The opportunity to opt-out was the locus for the dark pattern manipulation: a control group received no dark pattern, while two others received dark patterns described as “mild” and “aggressive:”

In the mild dark patterns condition, subjects could either click “Accept and continue (recommended)” or “Other options,” and the button that accepted the program was selected by default. [...] If subjects selected “Other options,” they were directed to the next screen, which asked them to choose between “I do not want to protect my data or credit history” or “After reviewing my options, I would like to protect my privacy and receive data protection and credit history

<sup>88</sup>HAR files were instrumental in the detection of third party entities, underscoring their usefulness were they to be included in the Web Census discussed in 4.2.1.

<sup>89</sup>The effectiveness of dark patterns is discussed further in the sequel, section 5.1.2.

monitoring.” [...] Next, if subjects did not accept the program, they were asked to tell us why they declined the valuable protection. Several non-compelling options were listed, including “My credit rating is already bad,” “Even though 16.7 million Americans were victimized by identity theft last year, I do not believe it could happen to me or my family,” “I’m already paying for identity theft and credit monitoring services,” and “I’ve got nothing to hide so if hackers gain access to my data I won’t be harmed.” They also could choose “Other” and type in their reason, or choose “On second thought, please sign me up for 6 months of free credit history monitoring and data protection services.” (19-20)

In the aggressive dark pattern condition, the first two screen matched the mild condition, but

Participants attempting to decline the identity theft protection were then told that since they indicated they did not want to protect their data, we would like to give them more information so they could make an informed choice. We asked them to read a paragraph of information about what identity theft is. Participants could either choose “Accept data protection plan and continue” or “I would like to read more information.” They were forced to remain on the page for at least ten seconds before being able to advance, and they were shown a countdown timer during this period. (20)

*Results.* Acceptance rates for the identity theft program were lowest in the control condition, at 11.3%. Rates more than doubled in the mild condition, to 25.8%, and nearly doubled again for the aggressive condition to 41.9%. The largest change in acceptance rates came with the initial screen, constituting approximately 75% of participants in the mild condition, and 65% in the aggressive condition, with acceptance rates falling for every additional dark pattern afterward. Rather than being worn down in their defenses by increasingly burdensome requirements to resist acceptance, it seems that those who continued were an increasingly resolute group (the final dark pattern, which forced a user to decline the service by acceding that the decision was being made for bad reasons, had almost no effect). The acceptance rates were further found to be robust to even large increases in the cost of the service (up to triple the amount had no effect).

After the study, a questionnaire was administered. The difference in negative affect between “mild” dark patterns and control were not significant, while the “aggressive” dark patterns produced a significant increase in negative affect. The authors note that “[t]hese results suggest that if companies go too far and present customers with a slew of blatant dark patterns designed to nudge them, they might experience backlash and the loss of good will.” This effect is familiar to marketers and social psychologists, who have given it the name “reactance.”<sup>90</sup>

Combining the two results, “mild” dark patterns in the UI of a website had a large effect on acceptance of a dubious identity theft protection plan, but did not provoke significant reactance relative to the control condition without dark patterns.

*Limitations of the study.* The control condition in the study is described as including no dark pattern, but a “neutral choice architecture” is problematic to define. The control condition should be described in much greater detail, and include pictures of the design, to allow fair assessment.

<sup>90</sup>See Steindl et al. [71].

For the treatment conditions, given the general downward trend in acceptance rates it is surprising that the authors allow both that “[d]ark patterns that were used later in the manipulation are less likely to work by the very fact that people who were most susceptible to dark patterns were no longer in the sample” and that the study “demonstrates the substantial cumulative power that different kinds of dark patterns can have.” (23) A cumulative effect would suggest a growing percentage of the remaining participants would accept the plan, but that is not what is reported to have happened.

The nature of the questionnaire administered after the survey and protection plan offer is left unspecified. Given the effects of different modes of administration, this oversight undermines the interpretation given to the questionnaire results.<sup>91</sup> For example, in-person verbal administration of a questionnaire by a Professor would be more likely to produce demand characteristics, as compared to written administration. This would explain the similar levels of negative affect in the control and “mild” treatment conditions, where the latter is within expected norms for persuading a student by presenting new information, making a student participant more likely to respond favorably to the sort of didactic manipulations performed. Moreover, it would account for the most effective dark pattern, in which the acceptance button was changed to read “Accept and continue (recommended)” rather than “Accept.”

More generally, text is important to the dark patterns used in the study. For example, users were forced to choose options with text that read “After reviewing my options, I would like to protect my privacy and receive data protection and credit history monitoring,” or, to decline, “Even though 16.7 million Americans were victimized by identity theft last year, I do not believe it could happen to me or my family,” “I’m already paying for identity theft and credit monitoring services,” and “I’ve got nothing to hide so if hackers gain access to my data I won’t be harmed.” In the context of a research study, such text may be persuasive, supporting an interpretation on which the participants were presented with an argument that led them to change their minds. Whether or this sort of text *should* be persuasive is an additional question; however, an argument should be considered distinct from a UI design decision that tricks or manipulates a person.

## 6 CONCLUSION: STRATEGIC RESPONSE TO PRIVACY ATTACK

Tools for systematic, coordinated response to privacy attack are a distinctive variety of PET, prompted by an escalating arms race for personal information on the Internet. This survey begins the task of organizing the landscape of such “strategic” PETs, revisiting older taxonomies of “hard” or “soft,” and of privacy as “confidentiality,” “control,” or “practice.” A new perspective reshuffles the relative prominence of some of these notions: what had been called “soft” technologies in earlier context appear to be more central to strategic PETs than “hard” technologies, for example, since minimization of disclosure for the data subject is largely supplanted by transformation of incentives for the data controller.

Review of the theory of anonymity reveals a relatively mature foundation, with consensus on terminology and a broad array of measures for quantifying anonymity. These foundations are important, as strategically impacting a data controller in response to attack demands precision to mount a proportional response.

From discussion of projects focused on transparency and usability there emerges a new privacy property not previously visible: human-bot indistinguishability, or “replicant-anonymity.” Replicant-anonymity underlies the automation required for many strategic PETs—much as

<sup>91</sup>See Tourangeau et al. [75].

technologies for building anonymous and confidential channels underlies many older “hard” PETs.

This article takes a first step in a much larger project of making the case for strategic PETs. Future expansion of this initial work must consider obfuscation in greater depth, including not only practical technologies but theoretical results, such as exploring the tradeoff between data quality and data quantity: when is a decrease in quality produced by a strategic PET enough to offset gains in data quantity achieved by privacy attack? Questions like these will need greater attention if the field is to progress.

The Internet has changed immensely since its earliest days, not least by the powerful influence of public and private actors with incentive to make ever greater use of personal data. With new abiding conditions comes a need for new vision: strategic PETs are tools for future privacy, and for a say in a future Internet as well.

## APPENDIX

### Privacy policy changes and natural experiments exploring privacy harms

Much of the survey traffics in the notion of a “privacy attack,” as though this term were entirely clear. It is not. In particular, subjective harms are a subtle and difficult legal question. The following analysis of recent work by James Cooper considers the experimental methodology used to explore the question of subjective harm in the context of Google’s then-new practice of sharing data between all its 60+ services.<sup>92</sup>

In January 2012, Google announced a change to its privacy policy, allowing data about a user to be consolidated across the more than 60 services that it offered; the change was to be instituted in March of the same year.<sup>93</sup> The Federal Trade Commission (*FTC, Commission*) took no action in response to this announcement, prompting the Electronic Privacy Information Center (*EPIC*) to file a complaint and motion compelling the FTC to act.<sup>94</sup> EPIC said that the FTC’s failure to act to stop the change in data collection practices prior to March 1 would cause “irreparable injury” to users of Google’s services, whose information would be “[...] misappropriated, combined, and disclosed by Google.”<sup>95</sup> The nature of some of the injury caused by Google’s consolidated collection is the focus of a recent article by James C. Cooper [23].

More generally, Cooper [23] considers the unwelcome mental states that data collection can create in a collection subject,<sup>96</sup> which he calls ‘subjective harms,’ and which he distinguishes

<sup>92</sup>Note to reviewers: This section has been lightly updated from work completed for an Information Privacy Seminar. It is not included to fulfill any requirement (the article is already too long), but I include it as an addendum to introduce an analysis and evaluation of empirical research originating from a law enforcement and policy perspective.

<sup>93</sup>Center [11]. These included a search engine, work and productivity services, email services, an operating system, a calendar service, instant messaging and chat, social networking services, maps and related services (Google Latitude), online payment processing, video sharing, photo organizing and editing, advertising services, and analytics services, including cross-domain tracking. For empirical evidence of Google’s disproportionate role in online tracking, see section 4.2.1).

<sup>94</sup>Specifically, the motion was for a temporary restraining order and preliminary injunction to enforce an earlier consent order that the FTC had issued against Google for engaging in unfair and deceptive trade practices by using user information from the Gmail service to “seed” a new social networking service Google had created, called “Google Buzz.” For the complaint see <http://epic.org/privacy/ftc/google/EPIC-Complaint-Final.pdf>; for the motion, see <http://epic.org/privacy/ftc/google/TRO-Motion-final.pdf>.

<sup>95</sup>Id. at 20.

<sup>96</sup>The focus of the study is more narrowly focused on the Google case in particular, in which data is collected in a commercial setting rather than as part of a scientific research project or state action.



from ‘objective harms.’<sup>97</sup> Subjective harms are arrayed around the perception of unwanted observation, and are manifest in feelings of discomfort, anxiety, or embarrassment.<sup>98</sup> Subjective privacy harms restrict autonomy via a well-worn connection between privacy and the development of personal identity.<sup>99</sup>

Certain actions observation present a threat of dignitary harm, and may be purposely avoided in the presence of an observer;<sup>100</sup> the comprehensiveness of changes in behavior made in response to observation is the extent to which a person has not acted freely.<sup>101</sup> Cooper’s focus on these sorts of dignitary harm in his account of reduced autonomy situates his work in the context of common law privacy torts, contrary to much contemporary privacy scholarship.<sup>102</sup>

There are two related motivations animating Cooper in his article: the first is a desire to contribute to increased analytic rigor in characterizing subjective privacy harms to consumers caused by unwanted data collection:

one of the most vexing problems in privacy policy is identifying consumer harm from unwanted observation; because it is highly subjective and is likely to vary greatly throughout the population, it doesn’t lend itself to easy measurement. (4)

The pay-off of rigorous understanding is a rational treatment of the trade-offs between privacy and the useful data flows they are presumed to impede, perhaps with optimization methods from economics. In other articles, Cooper has lamented the dearth of such rigorous reasoning in case law and policy debates on privacy, as compared to other areas, in particular antitrust.<sup>103</sup> As a result, the balance of opinion in “what masquerades as privacy policy analysis” has tipped too far in favor of privacy, without enough thought spared for the cost: “Imagine,” we are earnestly implored, “the innovation squandered without even the scantest justification.”<sup>104</sup> A second motivation for Cooper is a specific application of a proposed quantification of consumer harm to the case of Google’s 2012 change in collection practices, to evaluate the merits of claims of “irreparable injury” forecast by EPIC if Google’s data collection practices were allowed to change. For this purpose Cooper presents a naturalistic study in which reduced autonomy is measured after Google’s new privacy policy took effect by quantifying the reduction in sensitive searches performed before and after the change.

The theory is simple. After March 1, 2012, Google combined user information across platforms [...]. Some may want to avoid this intrusion and forego

<sup>97</sup>This subjective/objective contrast among harms is borrowed from Ryan Calo Calo [10].

<sup>98</sup>Cooper, *supra* note 4, at 4-5.

<sup>99</sup>Cohen [18].

<sup>100</sup>A clear example of such response to sensitive action is found in the statistical literature on response bias in sensitive surveys, where a question such as “have you cheated on you husband or wife?” leads to an increase in the average rate at which respondents lie in order to choose the more socially-acceptable “no” answer than “yes.” Tourangeau et al. [75], Warner [79].

<sup>101</sup>The case in which a person is comfortable being watched is interesting as well, though slightly off the path here. One example is of people who accidentally “break the ice” in doing something embarrassing in front of a camera, and afterwards become disinhibited under observation, purposely performing the same action. See Oulasvirta et al. [62].

<sup>102</sup>Bambauer [7, 108, note 10]. The alternative school of thought on the question of privacy trade-offs might be represented by work that treats privacy as a commodity, as in Abowd and Schmutte [1]; Richards [67].

<sup>103</sup>“The WhatsApp Privacy Policy Change: No Cause for Alarm,” accessed March 23, 2017, <https://www.forbes.com/sites/jamesccooper1/2016/09/07/the-whatsapp-privacy-policy-change-no-cause-for-alarm/#11ee10796992>.

<sup>104</sup>In this he echoes several others, including Bambauer [6], Wittes and Liu [83].

using Google to perform sensitive searches. In this manner, a reduction in sensitive search is an indirect measure of reduced autonomy. (2-3)

There were no control and treatment groups in the study; instead variables for number of searches for “sensitive terms” in a specific time window were compared with variables for “neutral terms.” Sensitive terms are selectively borrowed from another study without explanation.<sup>105</sup> Reduction in sensitive search is not directly measured; instead Google Trends data on search behavior is used.<sup>106</sup> The outcome of the study was, in effect, a null result: only for a narrow window was any effect on sensitive search found, and this was small, and soon vanished.

For each of the two main legs of Cooper’s approach to the Google 2012 case, there is a problem to be faced: one for the analysis of subjective harm as reduction in autonomy, and the other in the measurement of reduction of autonomy with reduction in sensitive search.

The problem with taking reduction in autonomy to be evidence for harm in the Google 2012 case is that often no subjective harm is allowed to develop by companies like Google, who nevertheless are the source of objective harms. Reduction in autonomy such as is caused by filter bubbles, price discrimination, discriminatory decision-making, or manipulative targeted advertising is compartmentalized by data collectors, who carefully study how to collect data surreptitiously and avoid revealing that they are “creepily” informed and thereby provoking “negative reactance”—privacy defensive behaviors.<sup>107</sup> This being the case, to pronounce consumers unharmed given lack of evidence of unwanted feelings about Google’s new data collection practices seems unjustified.

But even if it is allowed that subjective harm is the best way to understand the harms of Google’s new policy, its measurement by reduction in sensitive search is the wrong choice. Information from over 60 services was newly combined by Google; a broad and richly detailed picture could be obtained of a user of even a minority of these services. The central threat from such a change is of data linkage [15]; that is, the combination of sets of attributes that are separately insufficient to identify the entity that they describe, but that together become sufficient for that purpose. Insensitive search items are as much a threat as sensitive ones when combined with attributes from several years of email, document, messaging, travel, and phone history. In light of this, it is plausible to hypothesize that many users may have decided to leave Google altogether rather than carefully curate their searches for respectability. To the extent that we assume that such users had performed a random

<sup>105</sup>Cooper claimed to have chosen the twenty terms with the highest embarrassment scores, but he did not. See note 10 at 10, Appendix Table A1 at 33, and ? , table 14 at 35 [hereinafter the Snowden Effect Study]marthews2017government. Here are some top-20 terms that were excluded (rank is in parentheses): 1) ‘cutting’ (2) 2) ‘white pride’ (5) 3) ‘viagra’ (10) 4) ‘alcoholics anonymous’, ‘weed’ and ‘gender reassignment’ (11) For comparison, ‘KKK’, which was included, was 11th.

<sup>106</sup>The methodology behind Google Trends is mostly opaque. Google Trend values are the result of an analysis to determine how many searches were done as a percentage of all Google searches. The results are therefore relative measures: the interest data for a keyword is divided by the highest point of interest for the date range requested. Some of what is said explicitly about its procedure for generating Google Trends data raises more questions; for example, in The Snowden Effect Study, it is reported that “Google [...] says it excludes duplicate searches and searches made by a few people.”<sup>1</sup> Especially for sensitive searches, the privacy conscious use means to mix their IPs with a large amount of other traffic: VPNs and Tor, for example. This would lead to under reported trends data, as a large amount of traffic for different users would appear to be “searches made by a few people.”

<sup>107</sup>Deirdre Murphy made a similar point during the University of Arizona “Conversations on Privacy” series, available at <https://sidiprojects.us/mediawiki-1.23.2/images/0/0f/Corporations.m4a>. See also Goldfarb and Tucker [48], White et al. [80].

sampling of insensitive search terms prior to leaving, the effect of this important group would wash out in Google Trends, and remain invisible to Cooper's analysis.<sup>108</sup>

Cooper designed a naturalistic study attempting to measure the subjective harms suffered by consumers after Google's 2012 change of privacy policy, in the mold of a recent study of the "Snowden effect." Unlike the Snowden Effect study, however, Cooper's study showed no effect on sensitive search beyond a narrow window.<sup>109</sup> Despite its roots in an improving desire for sober, careful analysis, Cooper's study has severe methodological problems, such as an unexplained selective choice of sensitive terms for a treatment group and use of opaquely-generated Google Trends data. In addition, more fundamental theoretical problems were found to derive from an inaccurate understanding of the objective harms of Google's new data collection practices, and how they relate to the subjective harms.

## ZERO-KNOWLEDGE PROOF EXAMPLE

ZKP allows a prover, Peggy, to demonstrate with high probability that she knows some information, but not reveal the information in the process.<sup>110</sup> Suppose Peggy wishes to demonstrate her identity to a verifier, Victor, online. We will follow a protocol built on the ideas of Rabin encryption.

Peggy sets things up by picking a large  $n$  that is hard to compute square roots on.<sup>111</sup> She chooses a secret  $s$  in  $(\mathbb{Z}/n\mathbb{Z})^\times$ ,<sup>112</sup> and publishes  $t \equiv s^2 \pmod{n}$  and the modulus  $n$ , which together are Peggy's identifier. She must show Victor that she knows the secret value of the square root of  $t$  modulo  $n$ . She proves this in rounds, with repeated rounds making it increasingly improbable that Peggy does not know the information she is trying to prove she knows:

- (1) Peggy picks a random integer  $z$  from  $(\mathbb{Z}/n\mathbb{Z})^\times$ , and sends Victor  $z^2$  modulo  $n$ .
- (2) Victor randomly chooses either to ask for  $z$ , or  $zs$  modulo  $n$ , and Peggy answers.
- (3) Victor verifies Peggy's answer is consistent with what has previously been revealed. If Victor asked for  $z$ , then he checks  $z^2 \equiv r^2 \pmod{n}$ ; if he asked for  $sz$ , he checks instead  $r^2 \equiv s^2 z^2 \equiv t z^2 \pmod{n}$ .

After Peggy passes enough rounds, Victor has proved that she possesses the secret, but knows no more about it than before the protocol began.

## REFERENCES

- [1] Abowd, J. M., Schmutte, I. M., 2015. Revisiting the Economics of Privacy: Population Statistics and Confidentiality Protection as Public Goods.  
URL [http://digitalcommons.ilr.cornell.edu/ldi/22/?utm\\_source=digitalcommons.ilr.cornell.edu%2Fldi%2F22&utm\\_medium=PDF&utm\\_campaign=PDFCoverPages](http://digitalcommons.ilr.cornell.edu/ldi/22/?utm_source=digitalcommons.ilr.cornell.edu%2Fldi%2F22&utm_medium=PDF&utm_campaign=PDFCoverPages)
- [2] Alexander, C., Sep. 2018. A Pattern Language: Towns, Buildings, Construction. Oxford University Press, google-Books-ID: FTpxDwAAQBAJ.
- [3] Anderson, M., Feil, T., Jan. 2005. A First Course in Abstract Algebra: Rings, Groups and Fields, Second Edition. CRC Press, google-Books-ID: MYjLBQAAQBAJ.

<sup>108</sup>This assumes that Google Trends works as advertised; the actual algorithm for producing Google Trends data is not publicly available, which is a separate problem with its use.

<sup>109</sup>Because the study was a null result, the interpretation that "those who were uncomfortable with the Google's new policy of combining data were able to leave", Calo [10, 3], note is entirely unsupported by evidence.

<sup>110</sup>The following discussion follows a description of ZKP with ideas that were first used for encryption by Rabin [64]. The presentation that follows is adapted from discussions with Jeremy Boomer for an Elementary Number Theory course.

<sup>111</sup>In the Rabin cryptosystem, two large primes are multiplied to find such an  $n$ .

<sup>112</sup>See 'multiplicative group' [3].

- [4] Anderson, R., 1996. The eternity service. In: Proceedings of PRAGOCRYPT. Vol. 96. pp. 242–252.
- [5] Anderson, R., 2008. Security engineering. John Wiley & Sons.
- [6] Bambauer, J. R., Mar. 2011. Tragedy of the Data Commons. SSRN Scholarly Paper ID 1789749, Social Science Research Network, Rochester, NY.  
URL <https://papers.ssrn.com/abstract=1789749>
- [7] Bambauer, J. Y., 2012. The new intrusion. *Notre Dame L. Rev.* 88, 205.  
URL [http://heinonline.org/hol/cgi-bin/get\\_pdf.cgi?handle=hein.journals/tndl88&section=8](http://heinonline.org/hol/cgi-bin/get_pdf.cgi?handle=hein.journals/tndl88&section=8)
- [8] Barlow, J. P., 1996. A Declaration of the Independence of Cyberspace. *The Humanist* 56 (3), 18.
- [9] Bennett, C. J., 1995. Implementing Privacy Codes of Practice (plus 8830). Canadian Standards Association.
- [10] Calo, M. R., 2011. The Boundaries of Privacy Harm. *Indiana Law Journal* 86, 1131–1617.  
URL <https://litigation-essentials.lexisnexis.com/webcd/app?action=DocumentDisplay&crawlid=1&doctype=cite&docid=86+Ind.+L.J.+1131&srctype=smi&srcid=3B15&key=155ce426986283545306b5f422ea9e9b>
- [11] Center, E. P. I., ??? EPIC - EPIC v. FTC (Enforcement of the Google Consent Order).  
URL <https://epic.org/privacy/ftc/google/consent-order.html>
- [12] Chatellier, R., ??? Shaping Choices in the Digital World - From dark patterns to data protection: the influence of ux/ui design on user empowerment. Shaping Choices in the Digital World.  
URL [https://www.academia.edu/38856784/Shaping\\_Choices\\_in\\_the\\_Digital\\_World\\_-\\_From\\_dark\\_patterns\\_to\\_data\\_protection\\_the\\_influence\\_of\\_ux\\_ui\\_design\\_on\\_user\\_empowerment](https://www.academia.edu/38856784/Shaping_Choices_in_the_Digital_World_-_From_dark_patterns_to_data_protection_the_influence_of_ux_ui_design_on_user_empowerment)
- [13] Chaum, D., 1981. Untraceable Electronic Mail, Return Addresses, and Digital Pseudonyms. *Communications*.
- [14] Chen, T., Chaabane, A., Tournoux, P. U., Kaafar, M.-A., Boreli, R., 2013. How Much Is Too Much? Leveraging Ads Audience Estimation to Evaluate Public Profile Uniqueness. In: Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J. M., Mattern, F., Mitchell, J. C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M. Y., Weikum, G., De Cristofaro, E., Wright, M. (Eds.), *Privacy Enhancing Technologies*. Vol. 7981. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 225–244.  
URL [http://link.springer.com/10.1007/978-3-642-39077-7\\_12](http://link.springer.com/10.1007/978-3-642-39077-7_12)
- [15] Christen, P., 2012. Data Matching. Springer Berlin Heidelberg, Berlin, Heidelberg.  
URL <http://link.springer.com/10.1007/978-3-642-31164-2>
- [16] Clarke, I., Sandberg, O., Wiley, B., Hong, T. W., 2001. Freenet: A distributed anonymous information storage and retrieval system. In: *Designing privacy enhancing technologies*. Springer, pp. 46–66.
- [17] Clauß, S., Schiffner, S., 2006. Structuring anonymity metrics. In: *Proceedings of the second ACM workshop on Digital identity management - DIM '06*. ACM Press, Alexandria, Virginia, USA, p. 55.  
URL <http://portal.acm.org/citation.cfm?doid=1179529.1179539>
- [18] Cohen, J. E., 2000. Examined lives: Informational privacy and the subject as object. *Stanford Law Review*, 1373–1438.  
URL <http://www.jstor.org/stable/1229517>
- [19] Collberg, C., Nagra, J., 2009. *Surreptitious Software: Obfuscation, Watermarking, and Tamperproofing for Software Protection*. Addison-Wesley Software Security Series. Addison-Wesley.
- [20] Confidentiality and Data Access Committee, 2005. Report on statistical disclosure limitation methodology. Tech. Rep. 22, Office of Management and Budget.
- [21] Conti, G., Sobiesk, E., 2010. Malicious interface design: exploiting the user. In: *Proceedings of the 19th international conference on World wide web*. ACM, pp. 271–280.
- [22] contributors, W., Oct. 2019. HAR (file format). Page Version ID: 921314736.  
URL [https://en.wikipedia.org/w/index.php?title=HAR\\_\(file\\_format\)&oldid=921314736](https://en.wikipedia.org/w/index.php?title=HAR_(file_format)&oldid=921314736)
- [23] Cooper, J. C., 2017. Anonymity, Autonomy, and the Collection of Personal Data: Measuring the Privacy Impact of Google's 2012 Privacy Policy Change.  
URL [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2909148](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2909148)
- [24] Cranor, L. F., Guduru, P., Arjula, M., 2006. User interfaces for privacy agents. *ACM Transactions on Computer-Human Interaction (TOCHI)* 13 (2), 135–178.  
URL <http://dl.acm.org/citation.cfm?id=1165735>
- [25] Czeskis, A., Mah, D., Sandoval, O., Smith, I., Koscher, K., Appelbaum, J., Kohno, T., Schneier, B., 2013. DeadDrop/StrongBox security assessment.
- [26] Danezis, G., Goldberg, I., 2009. Sphinx: A compact and provably secure mix format. In: *Security and Privacy, 2009 30th IEEE Symposium on*. IEEE, pp. 269–282.

- [27] Danezis, G., Gürses, S., 2010. A critical review of 10 years of privacy technology. *Proceedings of Surveillance Cultures: A Global Surveillance Society*.  
URL <http://homes.esat.kuleuven.be/~sguurses/papers/DanezisGuersesSurveillancePets2010.pdf>
- [28] Das, A., Degeling, M., Smullen, D., Sadeh, N., ??? Personal Privacy Assistants for the Internet of Things.
- [29] Davidowitz, S. S., 2017. *Everybody lies: Big Data New Data and what the Internet can tell us about who we really are*. HarperCollins, New-York NY USA.
- [30] De Cristofaro, E., Soriente, C., Tsudik, G., Williams, A., 2012. Hummingbird: Privacy at the time of twitter. In: *2012 IEEE Symposium on Security and Privacy*. IEEE, pp. 285–299.
- [31] Diaz, C., Danezis, G., Syverson, P., Aug. 2010. Anonymous Communication. In: *Handbook of Financial Cryptography and Security*. CRC Press.
- [32] Diaz, C., Gürses, S., 2012. Understanding the landscape of privacy technologies. Extended abstract of invited talk in proceedings of the Information Security Summit, 58–63.  
URL <http://cosic.esat.kuleuven.be/publications/article-2215.pdf>
- [33] Diffie, W., Hellman, M., 1976. New directions in cryptography. *IEEE transactions on Information Theory* 22 (6), 644–654.
- [34] Dingledine, R., Freedman, M. J., Molnar, D., 2001. The free haven project: Distributed anonymous storage service. In: *Designing Privacy Enhancing Technologies*. Springer, pp. 67–95.
- [35] Duncan, G., Lambert, D., 1989. The risk of disclosure for microdata. *Journal of Business & Economic Statistics* 7 (2), 207–217.
- [36] Dwork, C., 2006. Differential privacy. *Automata, Languages and Programming*, Pt 2, 4052: 1–12, 2006. Bugliesi, M Prennel, B Sassone, V Wegener. In: *I 33rd International Colloquium on Automata, Languages and Programming JUL*. pp. 10–14.
- [37] Dwork, C., 2011. Differential Privacy. In: van Tilborg, H. C. A., Jajodia, S. (Eds.), *Encyclopedia of Cryptography and Security*. Springer US, Boston, MA, pp. 338–340.  
URL [https://doi.org/10.1007/978-1-4419-5906-5\\_752](https://doi.org/10.1007/978-1-4419-5906-5_752)
- [38] Dwork, C., 2011. A firm foundation for private data analysis. *Communications of the ACM* 54 (1), 86–95.
- [39] Dwork, C., McSherry, F., Nissim, K., Smith, A., 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In: Halevi, S., Rabin, T. (Eds.), *Theory of Cryptography*. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 265–284.
- [40] Dwork, C., Roth, A., others, 2014. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science* 9 (3-4), 211–407.  
URL <http://isites.harvard.edu/fs/docs/icb.topic1465468.files/The%20Algorithmic%20Foundations%20of%20Differential%20Privacy.pdf>
- [41] Eckersley, P., 2010. How unique is your web browser? In: *Privacy Enhancing Technologies*. Springer, pp. 1–18.  
URL [http://link.springer.com/chapter/10.1007/978-3-642-14527-8\\_1](http://link.springer.com/chapter/10.1007/978-3-642-14527-8_1)
- [42] Englehardt, S., Narayanan, A., 2016. Online tracking: A 1-million-site measurement and analysis. In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, pp. 1388–1401.
- [43] Evfimievski, A., Gehrke, J., Srikant, R., 2003. Limiting privacy breaches in privacy preserving data mining. In: *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, pp. 211–222.
- [44] Fung, B., Wang, K., Chen, R., Yu, P. S., 2010. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys (CSUR)* 42 (4), 14.  
URL <http://dl.acm.org/citation.cfm?id=1749605>
- [45] Goldberg, I., 2003. Privacy-enhancing technologies for the Internet, II: Five years later. *Privacy Enhancing Technologies*, Proceedings 2482, 1–12.
- [46] Goldberg, I., 2007. Privacy-enhancing technologies for the internet III: ten years later. In: *Digital Privacy*. Auerbach Publications, pp. 25–40.
- [47] Goldberg, I., Wagner, D., Brewer, E., 1997. Privacy-enhancing technologies for the Internet. In: *Proceedings IEEE COMPCON 97. Digest of Papers*. IEEE, pp. 103–109.
- [48] Goldfarb, A., Tucker, C., 2011. Online display advertising: Targeting and obtrusiveness. *Marketing Science* 30 (3), 389–404.  
URL <http://pubsonline.informs.org/doi/abs/10.1287/mksc.1100.0583>

- [49] Golle, P., 2006. Revisiting the uniqueness of simple demographics in the US population. In: *Proceedings of the 5th ACM workshop on Privacy in electronic society - WPES '06*. ACM Press, Alexandria, Virginia, USA, p. 77.  
URL <http://portal.acm.org/citation.cfm?doid=1179601.1179615>
- [50] Gray, M., 2018. Understanding and Improving Privacy 'Audits' Under FTC Orders. *SSRN Electronic Journal*.  
URL <https://www.ssrn.com/abstract=3165143>
- [51] Gürses, S., 2014. Can you engineer privacy? *Communications of the ACM* 57 (8), 20–23.  
URL <http://dl.acm.org/citation.cfm?id=2633029>
- [52] Kearns, M., Roth, A., Oct. 2019. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford University Press, google-Books-ID: z5OzDwAAQBAJ.
- [53] Le Métayer, D., 2016. Whom to Trust? Using Technology to Enforce Privacy. In: *Enforcing Privacy*. Springer, pp. 395–437.
- [54] Luguri, J., Strahilevitz, L., Aug. 2019. Shining a Light on Dark Patterns. *SSRN Scholarly Paper ID 3431205*, Social Science Research Network, Rochester, NY.  
URL <https://papers.ssrn.com/abstract=3431205>
- [55] MacKay, D. J., 2003. *Information theory, inference, and learning algorithms*. Vol. 7. Citeseer.
- [56] Markoff, J., Apr. 2005. What the Dormouse Said: How the Sixties Counterculture Shaped the Personal Computer Industry. Penguin.
- [57] Mathur, A., Acar, G., Friedman, M. J., Lucherini, E., Mayer, J., Chetty, M., Narayanan, A., 2019. Dark Patterns at Scale: Findings from a Crawl of 11k Shopping Websites. *Proceedings of the ACM on Human-Computer Interaction* 3 (CSCW), 81.
- [58] Moser, C., Schoenebeck, S. Y., Resnick, P., 2019. Impulse Buying: Design Practices and Consumer Needs. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, p. 242.
- [59] Narayanan, A., Shmatikov, V., 2008. Robust de-anonymization of large sparse datasets. In: *Security and Privacy, 2008. SP 2008. IEEE Symposium on*. IEEE, pp. 111–125.
- [60] Nelson, T. H., 1987. *Computer lib*, rev. ed. Edition. Tempus Books of Microsoft Press, Redmond, Wash.
- [61] OHRP, Oct. 1997. Chapter III: Basic IRB Review. In: *IRB Guidebook*.  
URL [http://wayback.archive-it.org/org-745/20150930182812/http://www.hhs.gov/ohrp/archive/irb/irb\\_chapter3.htm#e4](http://wayback.archive-it.org/org-745/20150930182812/http://www.hhs.gov/ohrp/archive/irb/irb_chapter3.htm#e4)
- [62] Oulasvirta, A., Pihlajamaa, A., Perkiö, J., Ray, D., Vähäkangas, T., Hasu, T., Vainio, N., Myllymäki, P., 2012. Long-term effects of ubiquitous surveillance in the home. In: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, pp. 41–50.  
URL <http://dl.acm.org/citation.cfm?id=2370224>
- [63] Pfitzmann, A., Kohntopp, M., 2001. Anonymity, unobservability, and pseudonymity—a proposal for terminology. In: *Designing privacy enhancing technologies*. Springer, pp. 1–9.
- [64] Rabin, M. O., 1979. Digitalized signatures and public-key functions as intractable as factorization. *Tech. rep.*, Massachusetts Inst of Tech Cambridge Lab for Computer Science.
- [65] Ratha, N. K., Connell, J. H., Bolle, R. M., 2001. Enhancing security and privacy in biometrics-based authentication systems. *IBM systems Journal* 40 (3), 614–634.
- [66] Reiter, J. P., 2005. Estimating risks of identification disclosure in microdata. *Journal of the American Statistical Association* 100 (472), 1103–1112.
- [67] Richards, N. M., 2011. *The Limits of Tort Privacy*.  
URL [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=1862264](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1862264)
- [68] Rocher, L., Hendrickx, J. M., De Montjoye, Y.-A., 2019. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature communications* 10 (1), 1–9.
- [69] Shannon, C. E., 1949. *The Mathematical Theory of Communication*. University of Illinois Press.
- [70] Shostack, A., Feb. 2014. *Threat Modeling: Designing for Security*. John Wiley & Sons, google-Books-ID: YiHcAgAAQBAJ.
- [71] Steindl, C., Jonas, E., Sittenthaler, S., Traut-Mattausch, E., Greenberg, J., Oct. 2015. Understanding Psychological Reactance: New Developments and Findings. *Zeitschrift für Psychologie* 223 (4), 205–214.  
URL <https://econtent.hogrefe.com/doi/10.1027/2151-2604/a000222>
- [72] Sweeney, L., 2000. Uniqueness of simple demographics in the US population. *Tech. rep.*, Technical report, Carnegie Mellon University.
- [73] Sweeney, L., 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10 (05), 557–570.

URL <http://www.worldscientific.com/doi/abs/10.1142/S0218488502001648>

- [74] Takeuchi, T. T., 2010. Constructing a bivariate distribution function with given marginals and correlation: application to the galaxy luminosity function. *Monthly Notices of the Royal Astronomical Society* 406 (3), 1830–1840.
- [75] Tourangeau, R., Rips, L., Rasinski, K., 2000. *The Psychology of Survey Response*. Cambridge University Press.  
URL <https://books.google.com/books?id=bjVYdyXXT3oC>
- [76] VanderSloot, B., Amann, J., Bernhard, M., Durumeric, Z., Bailey, M., Halderman, J. A., 2016. Towards a complete view of the certificate ecosystem. In: *Proceedings of the 2016 Internet Measurement Conference*. ACM, pp. 543–549.
- [77] Venkatadri, G., Lucherini, E., Sapiezynski, P., Mislove, A., 2019. Investigating sources of PII used in Facebook’s targeted advertising. *Proceedings on Privacy Enhancing Technologies* 2019 (1), 227–244.
- [78] Wagner, I., Eckhoff, D., 2018. Technical privacy metrics: a systematic survey. *ACM Computing Surveys (CSUR)* 51 (3), 57.
- [79] Warner, S. L., Mar. 1965. Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association* 60 (309), 63.  
URL <http://www.jstor.org/stable/2283137?origin=crossref>
- [80] White, T. B., Zahay, D. L., Thorbjørnsen, H., Shavitt, S., 2008. Getting too personal: Reactance to highly personalized email solicitations. *Marketing Letters* 19 (1), 39–50.  
URL <http://link.springer.com/article/10.1007/s11002-007-9027-9>
- [81] Wilcox-O’Hearn, Z., Warner, B., 2008. Tahoe: the least-authority filesystem. In: *Proceedings of the 4th ACM international workshop on Storage security and survivability*. ACM, pp. 21–26.
- [82] Wilson, C., Sep. 2009. *User Experience Re-Mastered: Your Guide to Getting the Right Design*. Morgan Kaufmann, google-Books-ID: URgdIJSLrkC.
- [83] Wittes, B., Liu, J. C., 2015. The privacy paradox: The privacy benefits of privacy threats. Center for Technology Innovation at Brookings, 1–21.  
URL [https://www.brookings.edu/wp-content/uploads/2016/06/Wittes-and-Liu\\_Privacy-paradox\\_v10.pdf](https://www.brookings.edu/wp-content/uploads/2016/06/Wittes-and-Liu_Privacy-paradox_v10.pdf)