Data Protection

Responsible Data Science
DS-UA 202 and DS-GA 1017

Instructors: Julia Stoyanovich and George Wood

This reader contains links to online materials and excerpts from selected articles on data protection. For convenience, the readings are organized by course week. Please note that some excerpts end in the middle of a section. Where that is the case, the partial section is not required reading.

Weeks 7 & 8: Anonymity and Privacy

Robust De-anonymization of Large Sparse Datasets

Arvind Narayanan and Vitaly Shmatikov

The University of Texas at Austin

Abstract

We present a new class of statistical de-anonymization attacks against high-dimensional micro-data, such as individual preferences, recommendations, transaction records and so on. Our techniques are robust to perturbation in the data and tolerate some mistakes in the adversary's background knowledge.

We apply our de-anonymization methodology to the Netflix Prize dataset, which contains anonymous movie ratings of 500,000 subscribers of Netflix, the world's largest online movie rental service. We demonstrate that an adversary who knows only a little bit about an individual subscriber can easily identify this subscriber's record in the dataset. Using the Internet Movie Database as the source of background knowledge, we successfully identified the Netflix records of known users, uncovering their apparent political preferences and other potentially sensitive information.

1   Introduction

Datasets containing *micro-data*, that is, information about specific individuals, are increasingly becoming public in response to "open government" laws and to support data mining research. Some datasets include legally protected information such as health histories; others contain individual preferences and transactions, which many people may view as private or sensitive.

Privacy risks of publishing micro-data are well-known. Even if identifiers such as names and Social Security numbers have been removed, the adversary can use background knowledge and cross-correlation with other databases to re-identify individual data records. Famous attacks include de-anonymization of a Massachusetts hospital discharge database by joining it with a public voter database [25] and privacy breaches caused by (ostensibly anonymized) AOL search data [16].

Micro-data are characterized by high dimensionality and sparsity. Each record contains many attributes (*i.e.*, columns in a database schema), which can be viewed as dimensions. Sparsity means that for the average record, there are no "similar" records in the multi-dimensional space defined by the attributes. This sparsity is empirically well-established [7, 4, 19] and related to the "fat tail" phenomenon: individual transaction and preference records tend to include statistically rare attributes.

Our contributions. Our first contribution is a formal model for privacy breaches in anonymized micro-data (section 3). We present two definitions, one based on the probability of successful de-anonymization, the other on the amount of information recovered about the target. Unlike previous work [25], we do not assume *a priori* that the adversary's knowledge is limited to a fixed set of "quasi-identifier" attributes. Our model thus encompasses a much broader class of de-anonymization attacks than simple cross-database correlation.

Our second contribution is a very general class of de-anonymization algorithms, demonstrating the fundamental limits of privacy in public micro-data (section 4). Under very mild assumptions about the distribution from which the records are drawn, the adversary with a small amount of background knowledge about an individual can use it to identify, with high probability, this individual's record in the anonymized dataset and to learn all anonymously released information about him or her, including sensitive attributes. For *sparse* datasets, such as most real-world datasets of individual transactions, preferences, and recommendations, very little background knowledge is needed (as few as 5-10 attributes in our case study). Our de-anonymization algorithm is *robust* to the imprecision of the adversary's background knowledge and to perturbation that may have been applied to the data prior to release. It works even if only a *subset* of the original dataset has been published.

Our third contribution is a practical analysis of the Netflix Prize dataset, containing anonymized movie ratings of 500,000 Netflix subscribers (section 5). Netflix—the world's largest online DVD rental

service—published this dataset to support the Netflix Prize data mining contest. We demonstrate that an adversary who knows a little bit about some subscriber can easily identify her record if it is present in the dataset, or, at the very least, identify a small set of records which include the subscriber's record. The adversary's background knowledge need not be precise, *e.g.*, the dates may only be known to the adversary with a 14-day error, the ratings may be known only approximately, and some of the ratings and dates may even be completely wrong. Because our algorithm is robust, if it uniquely identifies a record in the published dataset, with high probability this identification is not a false positive.

2   Related work

Unlike statistical databases [1, 3, 5], micro-data include actual records of individuals even after anonymization. A popular approach to micro-data privacy is k-anonymity [27, 9]. The data publisher decides in advance which of the attributes may be available to the adversary (these are called "quasi-identifiers"), and which are the sensitive attributes to be protected. k-anonymization ensures that each quasi-identifier tuple occurs in at least k records in the anonymized database. This does not guarantee any privacy, because the values of sensitive attributes associated with a given quasi-identifier may not be sufficiently diverse [20, 21] or the adversary may know more than just the quasi-identifiers [20]. Furthermore, k-anonymization completely fails on high-dimensional datasets [2], such as the Netflix Prize dataset and most real-world datasets of individual recommendations and purchases.

The de-anonymization algorithm presented in this paper does not assume that the attributes are divided *a priori* into quasi-identifiers and sensitive attributes. Examples include anonymized transaction records (if the adversary knows a few of the individual's purchases, can he learn *all* of her purchases?), recommendations and ratings (if the adversary knows a few movies that the individual watched, can he learn *all* movies she watched?), Web browsing and search histories, and so on. In such datasets, it is hard to tell in advance which attributes might be available to the adversary; the adversary's background knowledge may even vary from individual to individual. Unlike [25, 22, 14], our algorithm is *robust*. It works even if the published records have been perturbed, if only a subset of the original dataset has been published, and if there are mistakes in the adversary's background knowledge.

Our definition of privacy breach is somewhat similar to that of Chawla *et al.* [8]. We discuss the differences in section 3. There is theoretical evidence that for any (sanitized) database with meaningful utility, there is *always* some auxiliary or background information that results in a privacy breach [11]. In this paper, we aim to quantify the amount of auxiliary information required and its relationship to the percentage of records which would experience a significant privacy loss.

We are aware of only one previous paper that considered privacy of movie ratings. In collaboration with the MovieLens recommendation service, Frankowski *et al.* correlated public mentions of movies in the MovieLens discussion forum with the users' movie rating histories in the internal MovieLens dataset [14]. The algorithm uses the entire public record as the background knowledge (29 ratings per user, on average), and is not robust if this knowledge is imprecise, *e.g.*, if the user publicly mentioned movies which he did not rate.

While our algorithm follows the same basic scoring paradigm as [14], our scoring function is more complex and our selection criterion is nontrivial and an important innovation in its own right. Furthermore, our case study is based solely on public data and does *not* involve cross-correlating internal Netflix datasets (to which we do not have access) with public forums. It requires much less background knowledge (2-8 ratings per user), which need not be precise. Furthermore, our analysis has privacy implications for 500,000 Netflix subscribers whose records have been published; by contrast, the largest public MovieLens datasets contains only 6,000 records.

3   Model

Database. Define database $\mathcal{D}$ to be an $N \times M$ matrix where each row is a record associated with some individual, and the columns are attributes. We are interested in databases containing individual preferences or transactions. The number of columns thus reflects the total number of items in the space we are considering, ranging from a few thousand for movies to millions for (say) the `amazon.com` catalog.

Each attribute (column) can be thought of as a dimension, and each individual record as a point in the multidimensional attribute space. To keep our analysis general, we will not fix the space X from which attributes are drawn. They may be boolean (*e.g.*, has this book been rated?), integer (*e.g.*, the book's rating on a 1-10 scale), date, or a tuple such as a (rating, date) pair.

A typical reason to publish anonymized micro-data is "collaborative filtering," *i.e.*, predicting a consumer's future choices from his past behavior using the knowledge

of what similar consumers did. Technically, the goal is to predict the value of some attributes using a combination of other attributes. This is used in shopping recommender systems, aggressive caching in Web browsers, and other applications [28].

Sparsity and similarity. Preference databases with thousands of attributes are necessarily *sparse*, *i.e.*, each individual record contains values only for a small fraction of attributes. For example, the shopping history of even the most profligate Amazon shopper contains only a tiny fraction of all available items. We call these attributes *non-null*; the set of non-null attributes is the *support* of a record (denoted $\mathsf{supp}(r)$). Null attributes are denoted $\bot$. The support of a column is defined analogously. Even though points corresponding to database records are very sparse in the attribute space, each record may have dozens or hundreds of non-null attributes, making the database truly high-dimensional.

The distribution of per-attribute support sizes is typically heavy- or *long-tailed*, roughly following the power law [7, 4]. This means that although the supports of the columns corresponding to "unpopular" items are small, these items are so numerous that they make up the bulk of the non-null entries in the database. Thus, any attempt to approximate the database by projecting it down to the most common columns is bound to failure.[1]

Unlike "quasi-identifiers" [27, 9], there are no attributes that can be used directly for de-anonymization. In a large database, for any except the rarest attributes, there are hundreds of records with the same value of this attribute. Therefore, it is *not* a quasi-identifier. At the same time, knowledge that a particular individual has a certain attribute value does reveal *some* information, since attribute values and even the mere fact that a given attribute is non-null vary from record to record.

The similarity measure Sim is a function that maps a pair of attributes (or more generally, a pair of records) to the interval $[0, 1]$. It captures the intuitive notion of two values being "similar." Typically, Sim on attributes will behave like an indicator function. For example, in our analysis of the Netflix Prize dataset, Sim outputs 1 on a pair of movies rated by different subscribers if and only if both the ratings and the dates are within a certain threshold of each other; it outputs 0 otherwise.

To define Sim over two records r_1, r_2, we "generalize" the cosine similarity measure:

$$\mathsf{Sim}(r_1, r_2) = \frac{\sum \mathsf{Sim}(r_{1i}, r_{2i})}{|\mathsf{supp}(r_1) \cup \mathsf{supp}(r_2)|}$$

---

[1]The same effect causes k-anonymization to fail on high-dimensional databases [2].
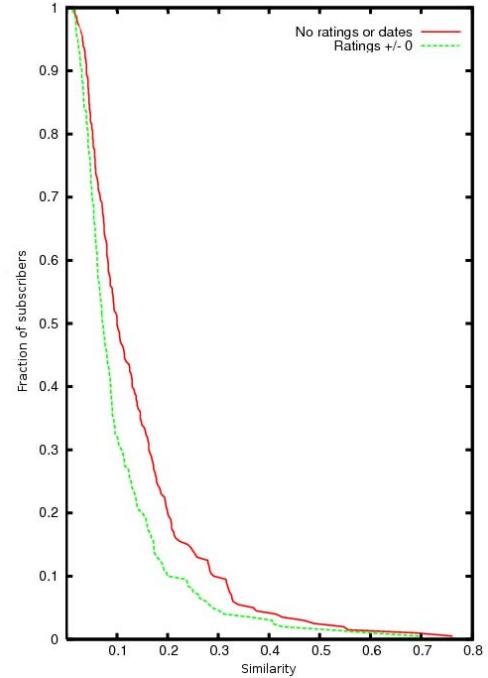
Figure 1. X-axis (x) is the similarity to the "neighbor" with the highest similarity score; Y-axis is the fraction of subscribers whose nearest-neighbor similarity is at least x.

Definition 1 (Sparsity) *A database D is (ϵ, δ)-sparse w.r.t. the similarity measure* Sim *if*

$$\Pr_r[\mathsf{Sim}(r, r') > \epsilon \,\forall r' \neq r] \leq \delta$$

As a real-world example, in fig. 1 we show that the Netflix Prize dataset is overwhelmingly sparse. For the vast majority of records, there isn't a *single* record with similarity score over 0.5 in the entire 500,000-record dataset, even if we consider only the sets of movies rated without taking into account numerical ratings or dates.

Sanitization and sampling. Database sanitization methods include generalization and suppression [26, 9], as well as perturbation. The data publisher may only release a (possibly non-uniform) sample of the database. Our algorithm is designed to work against data that have been both anonymized and sanitized.

If the database is published for collaborative filtering or similar data mining purposes (as in the case of the Netflix Prize dataset), the "error" introduced by sanitization *cannot* be large, otherwise data utility will be

lost. We make this precise in our analysis. Our definition of privacy breach allows the adversary to identify not just his target record, but *any* record as long as it is sufficiently similar (via Sim) to the target and can thus be used to determine its attributes with high probability.

From the viewpoint of our de-anonymization algorithm, there is no difference between the perturbation of the published records and the imprecision of the adversary's knowledge about his target. In either case, there is a small discrepancy between the attribute value(s) in the anonymous record and the same value(s) as known to the adversary. In the rest of the paper, we treat perturbation simply as imprecision of the adversary's knowledge. The algorithm is designed to be robust to the latter.

Adversary model. We sample record r randomly from database D and give *auxiliary information* or background knowledge related to r to the adversary. It is restricted to a subset of (possibly imprecise, perturbed, or simply incorrect) values of r's attributes, modeled as an arbitrary probabilistic function Aux: $X^M \rightarrow X^M$. The attributes given to the adversary may be chosen uniformly from the support of r, or according to some other rule.[2] Given this auxiliary information and an anonymized sample $\hat{D}$ of D, the adversary's goal is to reconstruct attribute values of the entire record r. Note that there is no artificial distinction between quasi-identifiers and sensitive attributes.

If the published records are sanitized by adding random noise Z_S, and the noise used in generating Aux is Z_A, then the adversary's task is equivalent to the scenario where the data are not perturbed but noise $Z_S + Z_A$ is used in generating Aux. This makes perturbation equivalent to imprecision of Aux.

Privacy breach: formal definitions. What does it mean to de-anonymize a record r? The naive answer is to find the "right" anonymized record in the public sample $\hat{D}$. This is hard to capture formally, however, because it requires assumptions about the data publishing process (*e.g.*, what if $\hat{D}$ contains two copies of every original record?). Fundamentally, the adversary's objective is is to learn as much as he can about r's attributes that he doesn't already know. We give two different (but related) formal definitions, because there are two distinct scenarios for privacy breaches in large databases.

The first scenario is automated large-scale de-anonymization. For every record r about which he has some information, the adversary must produce a single

"prediction" for all attributes of r. An example is the attack that inspired k-anonymity [25]: taking the demographic data from a voter database as auxiliary information, the adversary joins it with the anonymized hospital discharge database and uses the resulting combination to determine the values of medical attributes for each person who appears in both databases.

Definition 2 *A database D can be (θ, ω)-deanonymized w.r.t. auxiliary information Aux if there exists an algorithm A which, on inputs D and Aux(r) where $r \leftarrow D$ outputs r' such that*

$$\Pr[\textit{Sim}(r, r') \geq \theta] \geq \omega$$

Definition 2 can be interpreted as an *amplification of background knowledge*: the adversary starts with aux = Aux(r) which is close to r on a small subset of attributes, and uses this to compute r' which is close to r on the entire set of attributes. This captures the **adversary's ability to gain information about his target record**. As long he finds *some* record which is guaranteed to be very similar to the target record, *i.e.*, contains the same or similar attribute values, privacy breach has occurred.

If operating on a sample $\hat{D}$, the de-anonymization algorithm must also detect whether the target record is part of the sample, or has not been released at all. In the following, the probability is taken over the randomness of the sampling of r from $\hat{D}$, Aux and A itself.

Definition 3 (De-anonymization) *An arbitrary subset $\hat{D}$ of a database D can be (θ, ω)-deanonymized w.r.t. auxiliary information Aux if there exists an algorithm A which, on inputs $\hat{D}$ and Aux(r) where $r \leftarrow D$*

- *If $r \in \hat{D}$, outputs r' s.t. $\Pr[\textit{Sim}(r, r') \geq \theta] \geq \omega$*

- *if $r \notin \hat{D}$, outputs $\perp$ with probability at least ω*

The same error threshold $(1 - \omega)$ is used for both false positives and false negatives because the parameters of the algorithm can be adjusted so that both rates are equal; this is the "equal error rate."

In the second privacy breach scenario, the adversary produces a set or "lineup" of candidate records that include his target record r, either because there is not enough auxiliary information to identify r in the lineup or because he expects to perform additional analysis to complete de-anonymization. This is similar to communication anonymity in mix networks [24].

The *number* of candidate records is not a good metric, because some of the records may be much likelier candidates than others. Instead, we consider the probability distribution over the candidate records, and use

---

[2]For example, in the Netflix Prize case study we also pick uniformly from among the attributes whose supports are below a certain threshold, *e.g.*, movies that are outside the most popular 100 or 500 movies.

as the metric the conditional *entropy* of r given aux. In the absence of an "oracle" to identify the target record r in the lineup, the entropy of the distribution itself can be used as a metric [24, 10]. If the adversary has such an "oracle" (this is a technical device used to measure the adversary's success; in the real world, the adversary may not have an oracle telling him whether de-anonymization succeeded), then privacy breach can be quantified as follows: *how many bits of additional information does the adversary need in order to output a record which is similar to his target record?*

Thus, suppose that after executing the de-anonymization algorithm, the adversary outputs records $r'_1, \ldots r'_k$ and the corresponding probabilities $p_1, \ldots p_k$. The latter can be viewed as an *entropy encoding* of the candidate records. According to Shannon's source coding theorem, the optimal code length for record r'_i is $(-\log p_i)$. We denote by $H_S(\Pi, x)$ this Shannon entropy of a record x w.r.t. a probability distribution Π. In the following, the expectation is taken over the coin tosses of A, the sampling of r and Aux.

Definition 4 (Entropic de-anonymization) *A database D can be (θ, H)-deanonymized w.r.t. auxiliary information* Aux *if there exists an algorithm A which, on inputs D and* Aux(r) *where $r \leftarrow D$ outputs a set of candidate records D' and probability distribution Π such that*

$$E[\textit{min}_{r' \in D', \textit{Sim}(r,r') \geq \theta} H_S(\Pi, r')] \leq H$$

This definition measures the minimum Shannon entropy of the candidate set of records which are similar to the target record. As we will show, in sparse databases this set is likely to contain a single record, thus taking the minimum is but a syntactic requirement.

When the minimum is taken over an empty set, we define it to be $H_0 = \log_2 N$, the *a priori* entropy of the target record. This models outputting a random record from the entire database when the adversary cannot compute a lineup of plausible candidates. Formally, the adversary's algorithm A can be converted into an algorithm A', which outputs the mean of two distributions: one is the output of A, the other is the uniform distribution over D. Observe that for A', the minimum is always taken over a non-empty set, and the expectation for A' differs from that for A by at most 1 bit.

Chawla *et al.* [8] give a definition of privacy breach via *isolation* which is similar to ours, but requires a metric on attributes, whereas our general similarity measure does not naturally lead to a metric (there is no feasible way to derive a distance function from it that satisfies the triangle inequality). This appears to be essential for achieving robustness to completely erroneous attributes in the adversary's auxiliary information.

4 De-anonymization algorithm

We start by describing an algorithm template or meta-algorithm. The inputs are a sample $\hat{D}$ of database D and auxiliary information aux $= Aux(r), r \leftarrow D$. The output is either a record $r' \in \hat{D}$, or a set of candidate records and a probability distribution over those records (following Definitions 3 and 4, respectively).

The three main components of the algorithm are the scoring function, matching criterion, and record selection. The **scoring function** Score assigns a numerical score to each record in $\hat{D}$ based on how well it matches the adversary's auxiliary information Aux. The **matching criterion** is the algorithm applied by the adversary to the set of scores to determine if there is a match. Finally, **record selection** selects one "best-guess" record or a probability distribution, if needed.

1. Compute Score(aux, r') for each $r' \in \hat{D}$.

2. Apply the matching criterion to the resulting set of scores and compute the matching set; if the matching set is empty, output $\perp$ and exit.

3. If a "best guess" is required (de-anonymization according to Defs. 2 and 3), output $r' \in \hat{D}$ with the highest score. If a probability distribution over candidate records is required (de-anonymization according to Def. 4), compute and output some non-decreasing distribution based on the scores.

Algorithm Scoreboard. The following simple instantiation of the above template is sufficiently tractable to be formally analyzed in the rest of this section.

- Score(aux, r') $= \min_{i \in \text{supp(aux)}}$ Sim(aux$_i$, r'_i), *i.e.*, the score of a candidate record is determined by the least similar attribute between it and the adversary's auxiliary information.

- The matching set $D' = \{r' \in \hat{D} : $ Score(aux, r') $> \alpha\}$ for some fixed constant α. The matching criterion is that D' be nonempty.

- Probability distribution is uniform on D'.

Algorithm Scoreboard-RH. Algorithm Scoreboard is not sufficiently robust for some applications; in particular, it fails if any of the attributes in the adversary's auxiliary information are completely incorrect.

The following algorithm incorporates several heuristics which have proved useful in practical analysis (see section 5). First, the scoring function gives higher weight to statistically rare attributes. Intuitively, if the auxiliary information tells the adversary that his target has a certain rare attribute, this helps de-anonymization much more than the knowledge of a common attribute (*e.g.*, it is more useful to know that the target has purchased "The Dedalus Book of French Horror" than the fact that she purchased a Harry Potter book).

Second, to improve robustness, the matching criterion requires that the top score be significantly above the second-best score. This measures how much the identified record "stands out" from other candidate records.

- $\mathsf{Score}(\mathsf{aux}, r') = \sum_{i \in \mathsf{supp}(\mathsf{aux})} \mathsf{wt}(i) \mathsf{Sim}(\mathsf{aux}_i, r'_i)$ where $\mathsf{wt}(i) = \frac{1}{\log |\mathsf{supp}(i)|}$. [3]

- If a "best guess" is required, compute $\mathsf{max} = \mathsf{max}(S), \mathsf{max}_2 = \mathsf{max}_2(S)$ and $\sigma = \sigma(S)$ where $S = \{\mathsf{Score}(\mathsf{aux}, r') : r' \in \hat{D}\}$, *i.e.*, the highest and second-highest scores and the standard deviation of the scores. If $\frac{\mathsf{max} - \mathsf{max}_2}{\sigma} < \phi$, where ϕ is a fixed parameter called the *eccentricity*, then there is no match; otherwise, the matching set consists of the record with the highest score.[4]

- If entropic de-anonymization is required, output distribution $\Pi(r') = c \cdot e^{\frac{\mathsf{Score}(\mathsf{aux}, r')}{\sigma}}$ for each r', where c is a constant that makes the distribution sum up to 1. This weighs each matching record in inverse proportion to the likelihood that the match in question is a statistical fluke.

Note that there are two ways in which this algorithm can fail to find the correct record. First, an incorrect record may be assigned the highest score. Second, the correct record may not have a score which is significantly higher than the second-highest score.

4.1 Analysis: general case

We now quantify the amount of auxiliary information needed to de-anonymize an arbitrary dataset using Algorithm $\mathsf{Scoreboard}$. The smaller the required information (*i.e.*, the fewer attribute values the adversary needs to know about his target), the easier the attack.

We start with the worst-case analysis and calculate how much auxiliary information is needed without any

assumptions about the distribution from which the data are drawn. In section 4.2, we will show that much less auxiliary information is needed to de-anonymize records drawn from *sparse* distributions (real-world transaction and recommendation datasets are all sparse).

Let aux be the auxiliary information about some record r; aux consists of m (non-null) attribute values, which are close to the corresponding values of attributes in r, that is, $|\mathsf{aux}| = m$ and $\mathsf{Sim}(\mathsf{aux}_i, r_i) \geq 1 - \epsilon \ \forall i \in \mathsf{supp}(\mathsf{aux})$, where aux_i (respectively, r_i) is the ith attribute of aux (respectively, r).

Theorem 1 *Let* $0 < \epsilon, \delta < 1$ *and let* D *be the database. Let* Aux *be such that* $\mathsf{aux} = \mathsf{Aux}(r)$ *consists of at least* $m \geq \frac{\log N - \log \epsilon}{-\log(1-\delta)}$ *randomly selected attribute values of the target record* r, *where* $\forall i \in \mathsf{supp}(\mathsf{aux})$, $\mathsf{Sim}(\mathsf{aux}_i, r_i) \geq 1 - \epsilon$. *Then* D *can be* $(1 - \epsilon - \delta, 1 - \epsilon)$*-deanonymized w.r.t.* Aux.

Proof. Use Algorithm $\mathsf{Scoreboard}$ with $\alpha = 1 - \epsilon$ to compute the set of all records in $\hat{D}$ that match aux, then output a record r' at random from the matching set. It is sufficient to prove that this randomly chosen r' must be very similar to the target record r. (This satisfies our definition of a privacy breach because it gives the adversary almost everything he may want to learn about r.)

Record r' is a *false match* if $\mathsf{Sim}(r, r') \leq 1 - \epsilon - \delta$ (*i.e.*, the likelihood that r' is similar to the target r is below the threshold). We first show that, with high probability, there are no false matches in the matching set.

Lemma 1 *If* r' *is a false match, then* $\Pr_{i \in \mathsf{supp}(r)}[\mathsf{Sim}(r_i, r'_i) \geq 1 - \epsilon] < 1 - \delta$

Lemma 1 holds, because the contrary implies $\mathsf{Sim}(r, r') \geq (1 - \epsilon)(1 - \delta) \geq (1 - \epsilon - \delta)$, contradicting the assumption that r' is a false match. Therefore, the probability that the false match r' belongs to the matching set is at most $(1 - \delta)^m$. By a union bound, the probability that the matching set contains even a single false match is at most $N(1 - \delta)^m$. If $m = \frac{\log \frac{N}{\epsilon}}{\log \frac{1}{1-\delta}}$, then the probability that the matching set contains any false matches is no more than ϵ.

Therefore, with probability $1 - \epsilon$, there are no false matches. Thus for every record r' in the matching set, $\mathsf{Sim}(r, r') \geq 1 - \epsilon - \delta$, *i.e.*, any r' must be similar to the true record r. To complete the proof, observe that the matching set contains at least one record, r itself.

When δ is small, $m = \frac{\log N - \log \epsilon}{\delta}$. This depends logarithmically on ϵ and linearly on δ: the chance that the algorithm fails completely is very small even if attribute-wise accuracy is not very high. Also note that the matching set need not be small. Even if the algorithm returns

---

[3] Without loss of generality, we assume $\forall i \ |\mathsf{supp}(i)| > 0$.

[4] Increasing ϕ increases the false negative rate, *i.e.*, the chance of erroneously dismissing a correct match, and decreases the false positive rate; ϕ may be chosen so that the two rates are equal.

many records, with high probability they are *all* similar to the target record r, and thus any one of them can be used to learn the unknown attributes of r.

4.2 Analysis: sparse datasets

Most real-world datasets containing individual transactions, preferences, and so on are *sparse*. Sparsity increases the probability that de-anonymization succeeds, decreases the amount of auxiliary information needed, and improves robustness to both perturbation in the data and mistakes in the auxiliary information.

Our assumptions about data sparsity are very mild. We only assume $(1 - \epsilon - \delta, \dots)$ sparsity, *i.e.*, we assume that the average record does not have *extremely* similar peers in the dataset (real-world records tend not to have even *approximately* similar peers—see fig. 1).

Theorem 2 *Let ϵ, δ, and* aux *be as in Theorem 1. If the database D is $(1 - \epsilon - \delta, \epsilon)$-sparse, then D can be $(1, 1 - \epsilon)$-deanonymized.* □

The proof is essentially the same as for Theorem 1, but in this case *any $r' \neq r$ from the matching set must be a false match*. Because with probability $1 - \epsilon$, Scoreboard outputs no false matches, the matching set consists of exactly one record: the true target record r.

De-anonymization in the sense of Definition 4 requires even less auxiliary information. Recall that in this kind of privacy breach, the adversary outputs a "lineup" of k suspect records, one of which is the true record. This k-deanonymization is equivalent to $(1, \frac{1}{k})$-deanonymization in our framework.

Theorem 3 *Let D be $(1 - \epsilon - \delta, \epsilon)$-sparse and* aux *be as in Theorem 1 with $m = \frac{\log \frac{N}{k-1}}{\log \frac{1}{1-\delta}}$. Then*

- *D can be $(1, \frac{1}{k})$-deanonymized.*

- *D can be $(1, \log k)$-deanonymized (entropically).*

By the same argument as in the proof of Theorem 1, if the adversary knows $m = \frac{\log \frac{N}{k-1}}{\log \frac{1}{1-\delta}}$ attributes, then the expected number of false matches in the matching set is at most $k-1$. Let X be the random variable representing this number. A random record from the matching set is a false match with probability of at least $\frac{1}{X}$. Since $\frac{1}{x}$ is a convex function, apply Jensen's inequality [18] to obtain $E[\frac{1}{X}] \geq \frac{1}{E(X)} \geq \frac{1}{k}$.

Similarly, if the adversary outputs the uniform distribution over the matching set, its entropy is $\log X$. Since $\log x$ is a concave function, by Jensen's inequality $E[\log X] \leq \log E(X) \leq \log k$.

Neither claim follows directly from the other. □

4.3 De-anonymization from a sample

We now consider the scenario in which the released database $\hat{D} \subsetneq D$ is a sample of the original database D, *i.e.*, only some of the anonymized records are available to the adversary. This is the case, for example, for the Netflix Prize dataset (the subject of our case study in section 5), where the publicly available anonymized sample contains less than $\frac{1}{10}$ of the original data.

In this scenario, even though the original database D contains the adversary's target record r, this record may not appear in $\hat{D}$ even in anonymized form. The adversary can still apply Scoreboard, but the matching set may be empty, in which case the adversary outputs $\perp$ (indicating that de-anonymization fails). If the matching set is not empty, he proceeds as before: picks a random record r' and learn the attributes of r on the basis of r'. We now demonstrate the equivalent of Theorem 1: de-anonymization succeeds as long as r is in the public sample; otherwise, the adversary can detect, with high probability, that r is not in the public sample.

Theorem 4 *Let ϵ, δ, D, and* aux *be as in Theorem 1, and $\hat{D} \subset D$. Then $\hat{D}$ can be $(1 - \epsilon - \delta, 1 - \epsilon)$-deanonymized w.r.t.* aux. □

The bound on the probability of a false match given in the proof of Theorem 1 still holds, and the adversary is guaranteed at least one match as long as his target record r is in $\hat{D}$. Therefore, if $r \notin \hat{D}$, the adversary outputs $\perp$ with probability at least $1 - \epsilon$. If $r \in \hat{D}$, then again the adversary succeeds with probability at least $1 - \epsilon$.

Theorems 2 and 3 do not translate directly. For each record in the public sample $\hat{D}$, there could be any number of similar records in $D \setminus \hat{D}$, the part of the database that is not available to the adversary.

Fortunately, if D is sparse, then theorems 2 and 3 still hold, and de-anonymization succeeds with a very small amount of auxiliary information. We now show that if the random sample $\hat{D}$ is sparse, then the entire database D must also be sparse. Therefore, the adversary can simply apply the de-anonymization algorithm to the sample. If he finds the target record r, then with high probability this is not a false positive.

Theorem 5 *If database D is not (ϵ, δ)-sparse, then a random $\frac{1}{\lambda}$-subset $\hat{D}$ is not $(\epsilon, \frac{\delta\gamma}{\lambda})$-sparse with probability at least $1 - \gamma$.* □

For each $r \in \hat{D}$, the "nearest neighbor" r' of r in D has a probability $\frac{1}{\lambda}$ of being included in $\hat{D}$. Therefore, the expected probability that the similarity with the

nearest neighbor is at least $1 - \epsilon$ is at least $\frac{\delta}{\lambda}$. (Here the expectation is over the set of all possible samples and the probability is over the choice of the record in $\hat{D}$.) Applying Markov's inequality, the probability, taken over the choice $\hat{D}$, that $\hat{D}$ is sparse, *i.e.*, that the similarity with the nearest neighbor is $\frac{\delta\gamma}{\lambda}$, is no more than γ. $\qquad\square$

The above bound is quite pessimistic. Intuitively, for any "reasonable" dataset, the sparsity of a random sample will be about the same as that of the original dataset.

Theorem 5 can be interpreted as follows. Consider the adversary who has access to a sparse sample $\hat{D}$, but not the entire database D. Theorem 5 says that either a very-low-probability event has occurred, or D itself is sparse. Note that it is meaningless to try to bound the probability that D is sparse because we do not have a probability distribution on how D itself is created.

Intuitively, this says that unless the sample is specially tailored, sparsity of the sample implies sparsity of the entire database. The alternative is that the similarity between a random record in the sample and its nearest neighbor is very different from the corresponding distribution in the full database. In practice, most, if not all anonymized datasets are published to support research on data mining and collaborative filtering. Tailoring the published sample in such a way that its nearest-neighbor similarity is radically different from that of the original data would completely destroy utility of the sample for learning new collaborative filters, which are often based on the set of nearest neighbors. Therefore, in real-world anonymous data publishing scenarios—including, for example, the Netflix Prize dataset—sparsity of the sample should imply sparsity of the original dataset.

5 Case study: Netflix Prize dataset

On October 2, 2006, Netflix, the world's largest online DVD rental service, announced the $1-million Netflix Prize for improving their movie recommendation service [15]. To aid contestants, Netflix publicly released a dataset containing $100, 480, 507$ movie ratings, created by $480, 189$ Netflix subscribers between December 1999 and December 2005.

Among the Frequently Asked Questions about the Netflix Prize [23], there is the following question: "Is there any customer information in the dataset that should be kept private?" The answer is as follows:

> "No, all customer identifying information has been removed; all that remains are ratings and dates. This follows our privacy policy [. . .] Even if, for example, you knew all your own

ratings and their dates you probably couldn't identify them reliably in the data because only a small sample was included (less than one-tenth of our complete dataset) and that data was subject to perturbation. Of course, since you know all your own ratings that really isn't a privacy problem is it?"

Removing identifying information is not sufficient for anonymity. An adversary may have auxiliary information about a subscriber's movie preferences: the titles of a few of the movies that this subscriber watched, whether she liked them or not, maybe even approximate dates when she watched them. We emphasize that even if it is hard to collect such information for a large number of subscribers, targeted de-anonymization—for example, a boss using the Netflix Prize dataset to find an employee's entire movie viewing history after a casual conversation—still presents a serious threat to privacy.

We investigate the following question: *How much does the adversary need to know about a Netflix subscriber in order to identify her record if it is present in the dataset, and thus learn her complete movie viewing history?* Formally, we study the relationship between the size of aux and $(1, \omega)$- and $(1, H)$-deanonymization.

Does privacy of Netflix ratings matter? The issue is *not* "Does the average Netflix subscriber care about the privacy of his movie viewing history?," but "Are there *any* Netflix subscribers whose privacy can be compromised by analyzing the Netflix Prize dataset?" As shown by our experiments below, it is possible to learn sensitive *non-public* information about a person from his or her movie viewing history. We assert that even if the vast majority of Netflix subscribers did not care about the privacy of their movie ratings (which is not obvious by any means), our analysis would still indicate serious privacy issues with the Netflix Prize dataset.

Moreover, the linkage between an individual and her movie viewing history has implications for her *future* privacy. In network security, "forward secrecy" is important: even if the attacker manages to compromise a session key, this should not help him much in compromising the keys of future sessions. Similarly, one may state the "forward privacy" property: if someone's privacy is breached (*e.g.*, her anonymous online records have been linked to her real identity), future privacy breaches should not become easier. Consider a Netflix subscriber Alice whose entire movie viewing history has been revealed. Even if in the future Alice creates a brand-new virtual identity (call her Ecila), Ecila will *never* be able to disclose any non-trivial information about the movies that she had rated within Netflix

because any such information can be traced back to her real identity via the Netflix Prize dataset. In general, once any piece of data has been linked to a person's *real* identity, any association between this data and a *virtual* identity breaks anonymity of the latter.

Finally, the Video Privacy Protection Act of 1988 [13] lays down strong provisions against disclosure of personally identifiable rental records of "prerecorded video cassette tapes or similar audio visual material." While the Netflix Prize dataset does not *explicitly* include personally identifiable information, the issue of whether the implicit disclosure demonstrated by our analysis runs afoul of the law or not is a legal question to be considered.

How did Netflix release and sanitize the data? Figs. 2 and 3 plot the number of ratings X against the number of subscribers in the released dataset who have at least X ratings. The tail is surprisingly thick: thousands of subscribers have rated more than a thousand movies. Netflix claims that the subscribers in the released dataset have been "randomly chosen." Whatever the selection algorithm was, it was not uniformly random. Common sense suggests that with uniform subscriber selection, the curve would be monotonically decreasing (as most people rate very few movies or none at all), and that there would be no sharp discontinuities.

We conjecture that some fraction of subscribers with more than 20 ratings were sampled, and the points on the graph to the left of $X = 20$ are the result of some movies being deleted after sampling.

We requested the rating history as presented on the Netflix website from some of our acquaintances, and based on this data (which is effectively drawn from Netflix's *original*, non-anonymous dataset, since we know the names associated with these records), located two of them in the Netflix Prize dataset. Netflix's claim that the data were perturbed does not appear to be borne out. One of the subscribers had 1 of 306 ratings altered, and the other had 5 of 229 altered. (These are upper bounds, because the subscribers may have changed their ratings after Netflix took the 2005 snapshot that was released.) In any case, the level of noise is far too small to affect our de-anonymization algorithm, which has been specifically designed to withstand this kind of imprecision. We have no way of determining how many dates were altered and how many ratings were deleted, but we conjecture that very little perturbation has been applied.

It is important that the Netflix Prize dataset has been released to support development of better recommendation algorithms. A significant perturbation of individual attributes would have affected cross-attribute corre-

lations and significantly decreased the dataset's utility for creating new recommendation algorithms, defeating the entire purpose of the Netflix Prize competition.

Note that the Netflix Prize dataset clearly has not been k-anonymized for any value of $k > 1$.
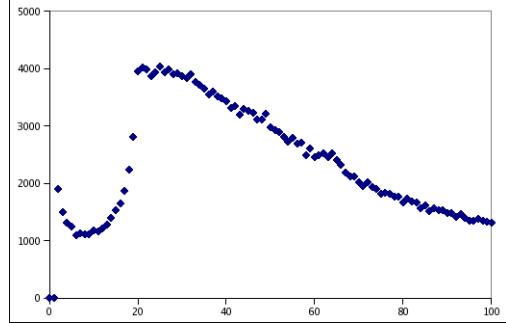

Figure 2. For each $X \leq 100$, the number of subscribers with X ratings in the released dataset.

De-anonymizing the Netflix Prize dataset. We apply Algorithm Scoreboard-RH from section 4. The similarity measure Sim on attributes is a threshold function: Sim returns 1 if and only if the two attribute values are within a certain threshold of each other. For movie ratings, which in the case of Netflix are on the 1-5 scale, we consider the thresholds of 0 (corresponding to exact match) and 1, and for the rating dates, 3 days, 14 days, or ∞. The latter means that the adversary has no information about the date when the movie was rated.

Some of the attribute values known to the attacker may be completely wrong. We say that aux of a record
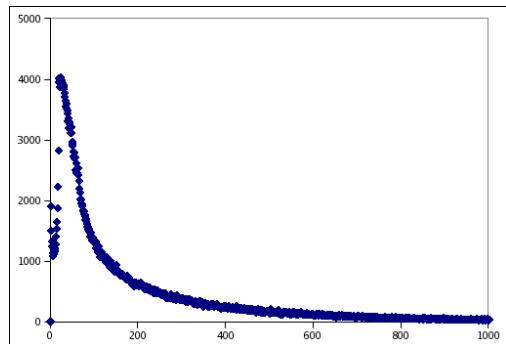

Figure 3. For each $X \leq 1000$, the number of subscribers with X ratings in the released dataset.
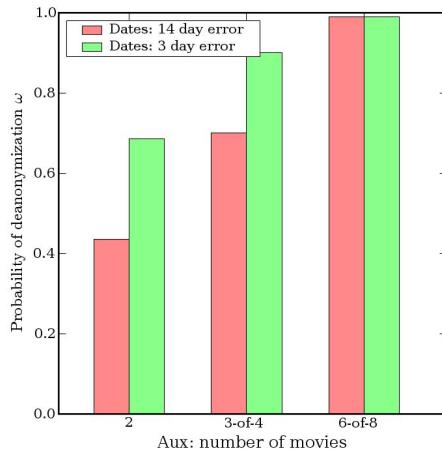
Figure 4. Adversary knows exact ratings and approximate dates.

Figure 5. Same parameters as Fig. 4, but the adversary must also detect when the target record is not in the sample.

r consists of m movies out of m' if $|\mathsf{aux}| = m'$, r_i is non-null for each aux_i, and $\sum_i \mathsf{Sim}(\mathsf{aux}_i, r_i) \geq m$. We instantiate the scoring function as follows:

$$\mathsf{Score}(\mathsf{aux}, r') = \sum_{i \in \mathsf{supp}(\mathsf{aux})} \mathsf{wt}(i)(e^{\frac{\rho_i - \rho_i'}{\rho_0}} + e^{\frac{d_i - d_i'}{d_0}})$$

where $\mathsf{wt}(i) = \frac{1}{\log |\mathsf{supp}(i)|}$ ($|\mathsf{supp}(i)|$ is the number of subscribers who have rated movie i), ρ_i and d_i are the rating and date, respectively, of movie i in the auxiliary information, and ρ_i' and d_i' are the rating and date in the candidate record r'.[5] As explained in section 4, this scoring function was chosen to favor statistically unlikely matches and thus minimize accidental false positives. The parameters ρ_0 and d_0 are 1.5 and 30 days, respectively. These were chosen heuristically, as they gave the best results in our experiments,[6] and used throughout, regardless of the amount of noise in Aux. The eccentricity parameter was set to $\phi = 1.5$, *i.e.*, the algorithm declares there is no match if and only if the difference between the highest and the second highest scores is no more than 1.5 times the standard deviation. (A constant value of ϕ does not always give the equal error rate, but it is a close enough approximation.)

---

[5]$\mathsf{wt}(i)$ is undefined when $|\mathsf{supp}(i)| = 0$, but this is not a concern since every movie is rated by at least 4 subscribers.

[6]It may seem that tuning the parameters to the specific dataset may have unfairly improved our results, but an actual adversary would have performed the same tuning. We do not claim that these numerical parameters should be used for other instances of our algorithm; they must be derived by trial and error for each target dataset.

Didn't Netflix publish only a sample of the data? Because Netflix published less than $\frac{1}{10}$ of its 2005 database, we need to be concerned about the possibility that when our algorithm finds a record matching aux in the published sample, this may be a false match and the real record has not been released at all.

Algorithm $\mathsf{Scoreboard}\text{-}\mathsf{RH}$ is specifically designed to detect when the record corresponding to aux is *not* in the sample. We ran the following experiment. First, we gave aux from a random record to the algorithm and ran it on the dataset. Then we *removed* this record from the dataset and re-ran the algorithm. In the former case, the algorithm should find the record; in the latter, declare that it is not in the dataset. As shown in Fig. 5, the algorithm succeeds with high probability in both cases.

It is possible, although *extremely* unlikely, that the original Netflix dataset is not as sparse as the published sample, *i.e.*, it contains clusters of records which are close to each other, but only one representative of each cluster has been released in the Prize dataset. A dataset with such a structure would be exceptionally unusual and theoretically problematic (see Theorem 4).

If the adversary has less auxiliary information than shown in Fig. 5, false positives cannot be ruled out *a priori*, but there is a lot of extra information in the dataset that can be used to eliminate them. For example, if the start date and total number of movies in a record are part of the auxiliary information (*e.g.*, the adversary knows approximately when his target first joined Netflix), they
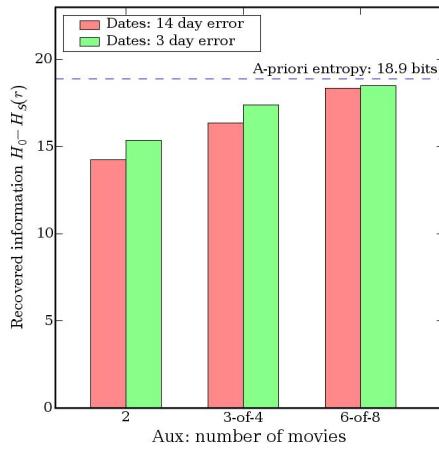
Figure 6. Entropic de-anonymization: same parameters as in Fig. 4.

can be used to eliminate candidate records.

Results of de-anonymization. We carried out the experiments summarized in the following table:

Fig	Ratings	Dates	Type	Aux selection
4	Exact	$\pm 3/\pm 14$	Best-guess	Uniform
5	Exact	$\pm 3/\pm 14$	Best-guess	Uniform
6	Exact	$\pm 3/\pm 14$	Entropic	Uniform
8	Exact	No info.	Best-guess	Not 100/500
9	± 1	± 14	Best-guess	Uniform
10	± 1	± 14	Best-guess	Uniform
11	Exact	No info.	Entropic	Not 100/500
12	± 1	± 14	Best-guess	Uniform

Our conclusion is that very little auxiliary information is needed for de-anonymize an average subscriber record from the Netflix Prize dataset. With 8 movie ratings (of which 2 may be completely wrong) and dates that may have a 14-day error, 99% of records can be uniquely identified in the dataset. For 68%, *two* ratings and dates (with a 3-day error) are sufficient (Fig. 4). Even for the other 32%, the number of possible candidates is brought down dramatically. In terms of entropy, the additional information required for complete de-anonymization is around 3 bits in the latter case (with no auxiliary information, this number is 19 bits). When the adversary knows 6 movies correctly and 2 incorrectly, the extra information he needs for complete de-anonymization is a fraction of a bit (Fig. 6).

Even without any dates, a substantial privacy breach occurs, especially when the auxiliary information consists of movies that are not blockbusters. In Fig. 7, we
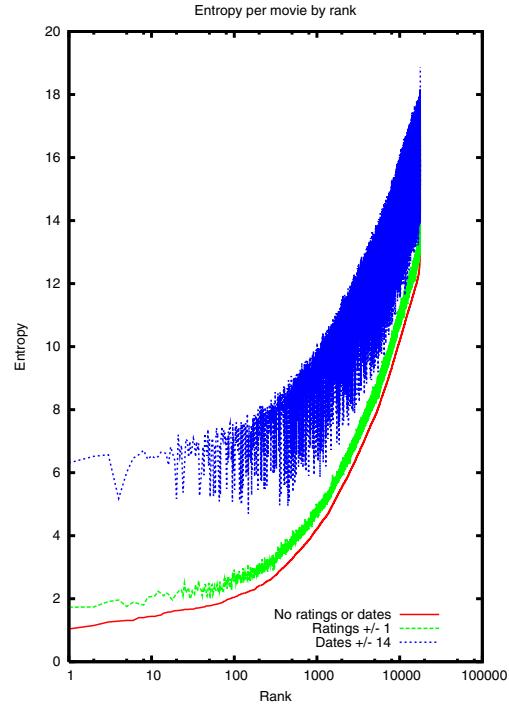

Figure 7. Entropy of movie by rank

demonstrate how much information the adversary gains about his target just from the knowledge that the target watched a particular movie as a function of the rank of the movie.[7] Because there are correlations between the lists of subscribers who watched various movies, we cannot simply multiply the information gain per movie by the number of movies. Therefore, Fig. 7 cannot be used to infer how many movies the adversary needs to know for successful de-anonymization.

As shown in Fig. 8, two movies are no longer sufficient for de-anonymization, but 84% of subscribers present in the dataset can be uniquely identified if the adversary knows 6 out of 8 moves outside the top 500. To show that this is not a significant limitation, consider that most subscribers rate fairly rare movies:

Not in X most rated	% of subscribers who rated . . .		
	≥ 1 movie	≥ 5	≥ 10
$X = 100$	100%	97%	93%
$X = 500$	99%	90%	80%
$X = 1000$	97%	83%	70%

Fig. 9 shows that the effect of relative popularity of movies known to the adversary is not dramatic.

In Fig. 10, we add even more noise to the auxiliary

---

[7]We measure the rank of a movie by the number of subscribers who have rated it.
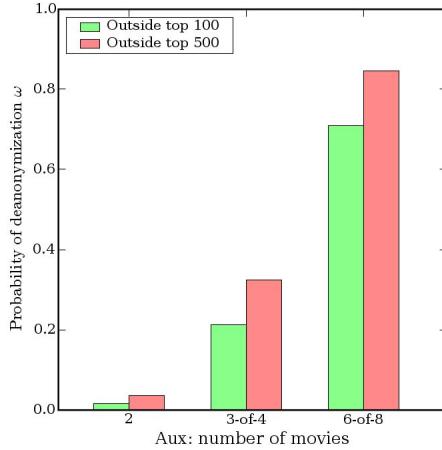
Figure 8. Adversary knows exact ratings but does not know dates at all.

Figure 9. Effect of knowing less popular movies rated by victim. Adversary knows approximate ratings (±1) and dates (14-day error).

information, allowing mistakes about *which* movies the target watched, and not just their ratings and dates.

Fig. 11 shows that even when the adversary's probability to correctly learn the attributes of the target record is low, he gains a lot of information about the target record. Even in the worst scenario, the additional information needed to to complete the de-anonymization has been reduced to less than half of its original value.

Fig. 12 shows why even partial de-anonymization can be very dangerous. There are many things the adversary might know about his target that are not captured by our formal model, such as the approximate number of movies rated, the date when they joined Netflix and so on. Once a candidate set of records is available, further automated analysis or human inspection might be sufficient to complete the de-anonymization. Fig. 12 shows that in some cases, knowing the number of movies the target has rated (even with a 50% error!) can more than double the probability of complete de-anonymization.

Obtaining the auxiliary information. Given how little auxiliary information is needed to de-anonymize the average subscriber record from the Netflix Prize dataset, a determined adversary who targets a specific individual may not find it difficult to obtain such information, especially since it need not be precise. We emphasize that massive collection of data on thousands of subscribers is not the only or even the most important threat. A water-cooler conversation with an office colleague about her cinematographic likes and dislikes may yield enough information, especially if at least a few of the movies

mentioned are outside the top 100 most rated Netflix movies. This information can also be gleaned from personal blogs, Google searches, and so on.

One possible source of a large number of personal movie ratings is the Internet Movie Database (IMDb) [17]. We expect that for Netflix subscribers who use IMDb, there is a strong correlation between their private Netflix ratings and their public IMDb ratings.[8] Our attack does not require that all movies rated by the subscriber in the Netflix system be also rated in IMDb, or vice versa. In many cases, even a handful of movies that are rated by a subscriber in both services would be sufficient to identify his or her record in the Netflix Prize dataset (if present among the released records) with enough statistical confidence to rule out the possibility of a false match except for a negligible probability.

Due to the restrictions on crawling IMDb imposed by IMDb's terms of service (of course, a real adversary may not comply with these restrictions), we worked with a very small sample of around 50 IMDb users. Our results should thus be viewed as a proof of concept. They do not imply anything about the percentage of IMDb users who can be identified in the Netflix Prize dataset.

The auxiliary information obtained from IMDb is quite noisy. First, a significant fraction of the movies rated on IMDb are not in Netflix, and vice versa, *e.g.*,

---

[8]We are *not* claiming that a large fraction of Netflix subscribers use IMDb, or that many IMDb users use Netflix.
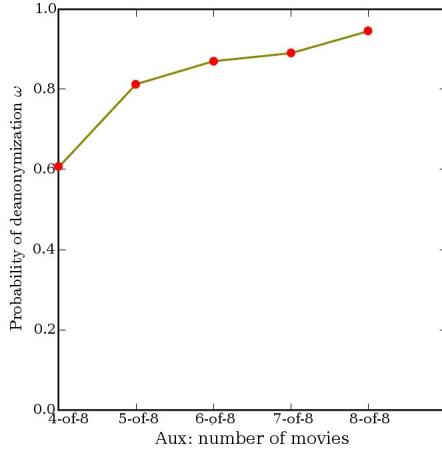
Figure 10. Effect of increasing error in Aux.
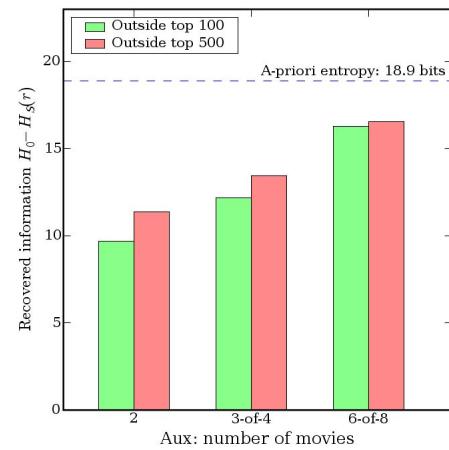
Figure 11. Entropic de-anonymization: same parameters as in Fig. 6.

movies that have not been released in the US. Second, some of the ratings on IMDb are missing (*i.e.*, the user entered only a comment, not a numerical rating). Such data are still useful for de-anonymization because an average user has rated only a tiny fraction of all movies, so the mere fact that a person has watched a given movie tremendously reduces the number of anonymous Netflix records that could possibly belong to that user. Finally, IMDb users among Netflix subscribers fall into a continuum of categories with respect to rating dates, separated by two extremes: some meticulously rate movies on both IMDb and Netflix at the same time, and others rate them whenever they have free time (which means the dates may not be correlated at all). Somewhat offsetting these disadvantages is the fact that we can use all of the user's ratings publicly available on IMDb.

Because we have no "oracle" to tell us whether the record our algorithm has found in the Netflix Prize dataset based on the ratings of some IMDb user indeed belongs to that user, we need to guarantee a very low false positive rate. Given our small sample of IMDb users, our algorithm identified the records of two users in the Netflix Prize dataset with eccentricities of around 28 and 15, respectively. These are exceptionally strong matches, which are highly unlikely to be false positives: the records in questions are **28 standard deviations** (respectively, 15 standard deviations) away from the second-best candidate. Interestingly, the first user was de-anonymized mainly from the ratings and the second mainly from the dates. For nearly all the other IMDb

users we tested, the eccentricity was no more than 2.

Let us summarize what our algorithm achieves. Given a user's *public* IMDb ratings, which the user posted voluntarily to reveal *some* of his (or her; but we'll use the male pronoun without loss of generality) movie likes and dislikes, we discover *all* ratings that he entered *privately* into the Netflix system. Why would someone who rates movies on IMDb—often under his or her real name—care about privacy of his Netflix ratings? Consider the information that we have been able to deduce by locating one of these users' entire movie viewing history in the Netflix Prize dataset and that *cannot* be deduced from his public IMDb ratings.

First, his political orientation may be revealed by his strong opinions about "Power and Terror: Noam Chomsky in Our Times" and "Fahrenheit 9/11," and his religious views by his ratings on "Jesus of Nazareth" and "The Gospel of John." Even though one should not make inferences solely from someone's movie preferences, in many workplaces and social settings opinions about movies with predominantly gay themes such as "Bent" and "Queer as folk" (both present and rated in this person's Netflix record) would be considered sensitive. In any case, it should be for the individual and not for Netflix to decide whether to reveal them publicly.

6 Conclusions

We have presented a de-anonymization methodology for sparse micro-data, and demonstrated its prac-
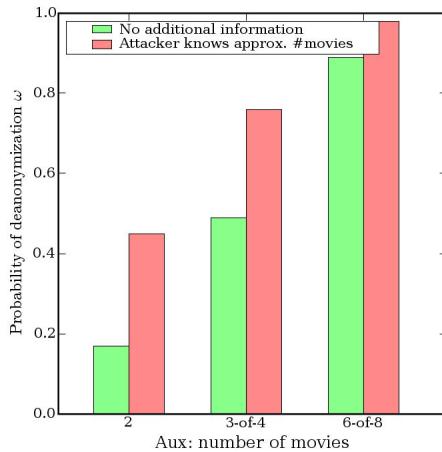
Figure 12. Effect of knowing approximate number of movies rated by victim ($\pm 50\%$). Adversary knows approximate ratings (± 1) and dates (14-day error).

tical applicability by showing how to de-anonymize movie viewing records released in the Netflix Prize dataset. Our de-anonymization algorithm Scoreboard-RH works under very general assumptions about the distribution from which the data are drawn, and is robust to data perturbation and mistakes in the adversary's knowledge. Therefore, we expect that it can be successfully used against any dataset containing anonymous multi-dimensional records such as individual transactions, preferences, and so on.

We conjecture that the amount of perturbation that must be applied to the data to defeat our algorithm will completely destroy their utility for collaborative filtering. Sanitization techniques from the k-anonymity literature such as generalization and suppression [27, 9, 20] do not provide meaningful privacy guarantees, and in any case fail on high-dimensional data. Furthermore, for most records simply knowing *which* columns are non-null reveals as much information as knowing the specific values of these columns. Therefore, any technique such as generalization and suppression which leaves sensitive attributes untouched does not help.

Other possible countermeasures include interactive mechanisms for privacy-protecting data mining such as [5, 12], as well as more recent non-interactive techniques [6]. Both support only limited classes of computations such as statistical queries and learning halfspaces. By contrast, in scenarios such as the Netflix Prize,

the purpose of the data release is precisely to foster computations on the data that have not even been foreseen at the time of release [9], and are vastly more sophisticated than the computations that we know how to perform in a privacy-preserving manner.

An intriguing possibility was suggested by Matthew Wright via personal communication: to release the records without the column identifiers (*i.e.*, movie names in the case of the Netflix Prize dataset). It is not clear how much worse the current data mining algorithms would perform under this restriction. Furthermore, this does not appear to make de-anonymization impossible, but merely harder. Nevertheless, it is an interesting countermeasure to investigate.

References

[1] N. Adam and J. Worthmann. Security-control methods for statistical databases: A comparative study. *ACM Computing Surveys*, 21(4), 1989.

[2] C. Aggarwal. On k-anonymity and the curse of dimensionality. In *VLDB*, 2005.

[3] R. Agrawal and R. Srikant. Privacy-preserving data mining. In *SIGMOD*, 2000.

[4] C. Anderson. *The Long Tail: Why the Future of Business Is Selling Less of More*. Hyperion, 2006.

[5] A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: The SuLQ framework. In *PODS*, 2005.

[6] A. Blum, K. Ligett, and A. Roth. A learning theory approach to non-interactive database privacy. In *STOC*, 2008.

[7] E. Brynjolfsson, Y. Hu, and M. Smith. Consumer surplus in the digital economy. *Management Science*, 49(11), 2003.

---

[9]As of February 2008, the current best algorithm in the Netflix Prize competition is a combination of 107 different techniques.

[8] S. Chawla, C. Dwork, F. McSherry, A. Smith, and H. Wee. Towards privacy in public databases. In *TCC*, 2005.

[9] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati. k-anonymity. *Secure Data Management in Decentralized Systems*, 2007.

[10] C. Díaz, S. Seys, J. Claessens, and B. Preneel. Towards measuring anonymity. In *PET*, 2003.

[11] C. Dwork. Differential privacy. In *ICALP*, 2006.

[12] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, 2006.

[13] Electronic Privacy Information Center. The Video Privacy Protection Act (VPPA). http://epic.org/privacy/vppa/, 2002.

[14] D. Frankowski, D. Cosley, S. Sen, L. Terveen, and J. Riedl. You are what you say: privacy risks of public mentions. In *SIGIR*, 2006.

[15] K. Hafner. And if you liked the movie, a Netflix contest may reward you handsomely. New York Times, Oct 2 2006.

[16] S. Hansell. AOL removes search data on vast group of web users. New York Times, Aug 8 2006.

[17] IMDb. The Internet Movie Database. http://www.imdb.com/, 2007.

[18] J. L. W. V. Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, 30(1), 1906.

[19] J. Leskovec, L. Adamic, and B. Huberman. The dynamics of viral marketing. In *EC*, 2006.

[20] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In *ICDE*, 2006.

[21] A. Machanavajjhala, D. Martin, D. Kifer, J. Gehrke, and J. Halpern. Worst case background knowledge. In *ICDE*, 2007.

[22] B. Malin and L. Sweeney. How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. *J. of Biomedical Informatics*, 37(3), 2004.

[23] Netflix. Netflix Prize: FAQ. http://www.netflixprize.com/faq, Downloaded on Oct 17 2006.

[24] A. Serjantov and G. Danezis. Towards an information theoretic metric for anonymity. In *PET*, 2003.

[25] L. Sweeney. Weaving technology and policy together to maintain confidentiality. *J. of Law, Medicine and Ethics*, 25(2–3), 1997.

[26] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International J. of Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5), 2002.

[27] L. Sweeney. k-anonymity: A model for protecting privacy. *International J. of Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557–570, 2002.

[28] J. Thornton. Collaborative filtering research papers. http://jamesthornton.com/cf/, 2006.

A   Glossary of terms

Symbol	Meaning
D	Database
$\hat{D}$	Released sample
N	Number of rows
M	Number of columns
m	Size of aux
X	Domain of attributes
$\perp$	Null attribute
supp$(.)$	Set of non-null attributes in a row/column
Sim	Similarity measure
Aux	Auxiliary information sampler
aux	Auxiliary information
Score	Scoring function
ϵ	Sparsity threshold
δ	Sparsity probability
θ	Closeness of de-anonymized record
ω	Probability that de-anonymization succeeds
r, r'	Record
Π	P.d.f over records
H_S	Shannon entropy
H	De-anonymization entropy
ϕ	Eccentricity

now
the essence of knowledge

The Algorithmic Foundations
of Differential Privacy

Cynthia Dwork
Microsoft Research, USA
dwork@microsoft.com

Aaron Roth
University of Pennsylvania, USA
aaroth@cis.upenn.edu

1

The Promise of Differential Privacy

"Differential privacy" describes a promise, made by a data holder, or *curator*, to a data subject: "You will not be affected, adversely or otherwise, by allowing your data to be used in any study or analysis, no matter what other studies, data sets, or information sources, are available." At their best, differentially private database mechanisms can make confidential data widely available for accurate data analysis, without resorting to data clean rooms, data usage agreements, data protection plans, or restricted views. Nonetheless, data utility will eventually be consumed: the Fundamental Law of Information Recovery states that overly accurate answers to too many questions will destroy privacy in a spectacular way.[1] The goal of algorithmic research on differential privacy is to postpone this inevitability as long as possible.

Differential privacy addresses the paradox of learning nothing about an individual while learning useful information about a population. A medical database may teach us that smoking causes cancer, affecting an insurance company's view of a smoker's long-term medical costs. Has the smoker been harmed by the analysis? Perhaps — his insurance

---

[1]This result, proved in Section 8.1, applies to *all* techniques for privacy-preserving data analysis, and not just to differential privacy.

premiums may rise, if the insurer knows he smokes. He may also be helped — learning of his health risks, he enters a smoking cessation program. Has the smoker's privacy been compromised? It is certainly the case that more is known about him after the study than was known before, but was his information "leaked"? Differential privacy will take the view that it was not, with the rationale that the impact on the smoker is the same *independent of whether or not he was in the study.* It is the *conclusions reached* in the study that affect the smoker, not his presence or absence in the data set.

Differential privacy ensures that the same conclusions, for example, smoking causes cancer, will be reached, independent of whether any individual opts into or opts out of the data set. Specifically, it ensures that any sequence of outputs (responses to queries) is "essentially" equally likely to occur, independent of the presence or absence of any individual. Here, the probabilities are taken over random choices made by the privacy mechanism (something controlled by the data curator), and the term "essentially" is captured by a parameter, ε. A smaller ε will yield better privacy (and less accurate responses).

Differential privacy is a *definition,* not an algorithm. For a given computational task T and a given value of ε there will be many differentially private algorithms for achieving T in an ε-differentially private manner. Some will have better accuracy than others. When ε is small, finding a highly accurate ε-differentially private algorithm for T can be difficult, much as finding a numerically stable algorithm for a specific computational task can require effort.

1.1   Privacy-preserving data analysis

Differential privacy is a definition of privacy tailored to the problem of privacy-preserving data analysis. We briefly address some concerns with other approaches to this problem.

Data Cannot be Fully Anonymized and Remain Useful. Generally speaking, the richer the data, the more interesting and useful it is. This has led to notions of "anonymization" and "removal of personally identifiable information," where the hope is that portions of the

data records can be suppressed and the remainder published and used for analysis. However, the richness of the data enables "naming" an individual by a sometimes surprising collection of fields, or attributes, such as the combination of zip code, date of birth, and sex, or even the names of three movies and the approximate dates on which an individual watched these movies. This "naming" capability can be used in a *linkage attack* to match "anonymized" records with non-anonymized records in a different dataset. Thus, the medical records of the governor of Massachussetts were identified by matching anonymized medical encounter data with (publicly available) voter registration records, and Netflix subscribers whose viewing histories were contained in a collection of anonymized movie records published by Netflix as training data for a competition on recommendation were identified by linkage with the Internet Movie Database (IMDb).

Differential privacy neutralizes linkage attacks: since being differentially private is a property of the data access mechanism, and is unrelated to the presence or absence of auxiliary information available to the adversary, access to the IMDb would no more permit a linkage attack to someone whose history is in the Netflix training set than to someone not in the training set.

Re-Identification of "Anonymized" Records is Not the Only Risk. Re-identification of "anonymized" data records is clearly undesirable, not only because of the re-identification *per se*, which certainly reveals membership in the data set, but also because the record may contain compromising information that, were it tied to an individual, could cause harm. A collection of medical encounter records from a specific urgent care center on a given date may list only a small number of distinct complaints or diagnoses. The additional information that a neighbor visited the facility on the date in question gives a fairly narrow range of possible diagnoses for the neighbor's condition. The fact that it may not be possible to match a specific record to the neighbor provides minimal privacy protection to the neighbor.

Queries Over Large Sets are Not Protective. Questions about specific individuals cannot be safely answered with accuracy, and indeed one

might wish to reject them out of hand (were it computationally feasible to recognize them). Forcing queries to be over large sets is not a panacea, as shown by the following *differencing attack*. Suppose it is known that Mr. X is in a certain medical database. Taken together, the answers to the two large queries "How many people in the database have the sickle cell trait?" and "How many people, not named X, in the database have the sickle cell trait?" yield the sickle cell status of Mr. X.

Query Auditing Is Problematic. One might be tempted to *audit* the sequence of queries and responses, with the goal of interdicting any response if, in light of the history, answering the current query would compromise privacy. For example, the auditor may be on the lookout for pairs of queries that would constitute a differencing attack. There are two difficulties with this approach. First, it is possible that *refusing* to answer a query is itself disclosive. Second, query auditing can be computationally infeasible; indeed if the query language is sufficiently rich there may not even exist an algorithmic procedure for deciding if a pair of queries constitutes a differencing attack.

Summary Statistics are Not "Safe." In some sense, the failure of summary statistics as a privacy solution concept is immediate from the differencing attack just described. Other problems with summary statistics include a variety of *reconstruction attacks* against a database in which each individual has a "secret bit" to be protected. The utility goal may be to permit, for example, questions of the form "How many people satisfying property P have secret bit value 1?" The goal of the adversary, on the other hand, is to significantly increase his chance of guessing the secret bits of individuals. The reconstruction attacks described in Section 8.1 show the difficulty of protecting against even a *linear* number of queries of this type: unless sufficient inaccuracy is introduced almost all the secret bits can be reconstructed.

A striking illustration of the risks of releasing summary statistics is in an application of a statistical technique, originally intended for confirming or refuting the presence of an individual's DNA in a forensic mix, to ruling an individual in or out of a genome-wide association study. According to a Web site of the Human Genome Project, "Single nucleotide polymorphisms, or SNPs (pronounced "snips"), are DNA

sequence variations that occur when a single nucleotide (A,T,C, or G) in the genome sequence is altered. For example a SNP might change the DNA sequence AAGGCTAA to ATGGCTAA." In this case we say there are two alleles: A and T. For such a SNP we can ask, given a particular reference population, what are the frequencies of each of the two possible alleles? Given the allele frequencies for SNPs in the reference population, we can examine how these frequencies may differ for a subpopulation that has a particular disease (the "case" group), looking for alleles that are associated with the disease. For this reason, genome-wide association studies may contain the allele frequencies of the case group for large numbers of SNPs. By definition, these allele frequencies are only aggregated statistics, and the (erroneous) assumption has been that, by virtue of this aggregation, they preserve privacy. However, given the genomic data of an individual, it is theoretically possible to determine if the individual is in the case group (and, therefore, has the disease). In response, the National Institutes of Health and Wellcome Trust terminated public access to aggregate frequency data from the studies they fund.

This is a challenging problem even for differential privacy, due to the large number — hundreds of thousands or even one million — of measurements involved and the relatively small number of individuals in any case group.

"Ordinary" Facts are Not "OK." Revealing "ordinary" facts, such as purchasing bread, may be problematic if a data subject is followed over time. For example, consider Mr. T, who regularly buys bread, year after year, until suddenly switching to rarely buying bread. An analyst might conclude Mr. T most likely has been diagnosed with Type 2 diabetes. The analyst might be correct, or might be incorrect; either way Mr. T is harmed.

"Just a Few." In some cases a particular technique may in fact provide privacy protection for "typical" members of a data set, or more generally, "most" members. In such cases one often hears the argument that the technique is adequate, as it compromises the privacy of "just a few" participants. Setting aside the concern that outliers may be precisely those people for whom privacy is most important, the "just a few"

philosophy is not intrinsically without merit: there is a social judgment, a weighing of costs and benefits, to be made. A well-articulated definition of privacy consistent with the "just a few" philosophy has yet to be developed; however, for a single data set, "just a few" privacy can be achieved by randomly selecting a subset of rows and releasing them in their entirety (Lemma 4.3, Section 4). Sampling bounds describing the quality of statistical analysis that can be carried out on random sub-samples govern the number of rows to be released. Differential privacy provides an alternative when the "just a few" philosophy is rejected.

1.2   Bibliographic notes

Sweeney [81] linked voter registration records to "anonymized" medical encounter data; Narayanan and Shmatikov carried out a linkage attack against anonymized ranking data published by Netflix [65]. The work on presence in a forensic mix is due to Homer et al. [46]. The first reconstruction attacks were due to Dinur and Nissim [18].

What does it mean to preserve privacy?

BY CYNTHIA DWORK

A Firm Foundation for Private Data Analysis

IN THE INFORMATION realm, loss of privacy is usually associated with failure to control access to information, to control the flow of information, or to control the purposes for which information is employed. Differential privacy arose in a context in which ensuring privacy is a challenge even if all these control problems are solved: privacy-preserving statistical analysis of data.

The problem of *statistical disclosure control*—revealing accurate statistics about a set of respondents while preserving the privacy of individuals—has a venerable history, with an extensive literature spanning statistics, theoretical computer science, security, databases, and cryptography (see, for example, the excellent survey of Adam and Wortmann,[1] the discussion of related work in Blum et al.,[2] and the *Journal of Official Statistics* dedicated to confidentiality and disclosure control).

This long history is a testament to the importance of the problem. Statistical databases can be of enormous social value; they are used for apportioning resources, evaluating medical therapies, understanding the spread of disease, improving economic utility, and informing us about ourselves as a species.

The data may be obtained in diverse ways. Some data, such as census, tax, and other sorts of official data, is compelled; other data is collected opportunistically, for example, from traffic on the Internet, transactions on Amazon, and search engine query logs; other data is provided altruistically, by respondents who hope that sharing their information will help others to avoid a specific misfortune, or more generally, to increase the public good. Altruistic data donors are typically promised their individual data will be kept confidential—in short, they are promised "privacy." Similarly, medical data and legally compelled data, such as census data and tax return data, have legal privacy

» **key insights**

■ **In analyzing private data, only by focusing on rigorous privacy guarantees can we convert the cycle of "propose-break-propose again" into a path of progress.**

■ **A natural approach to defining privacy is to require that accessing the database teaches the analyst nothing about any individual. But this is problematic: the whole point of a statistical database is to teach general truths, for example, that smoking causes cancer. Learning this fact teaches the data analyst something about the likelihood with which certain individuals, not necessarily in the database, will develop cancer. We therefore need a definition that separates the utility of the database (learning that smoking causes cancer) from the increased risk of harm due to joining the database. This is the intuition behind *differential privacy*.**

■ **This can be achieved, often with low distortion. The key idea is to randomize responses so as to effectively hide the presence or absence of the data of any individual *over the course of the lifetime of the database*.**

IMAGE BY BRIAN GREENBERG

mandates. In my view, ethics demand that opportunistically obtained data should be treated no differently, especially when there is no reasonable alternative to engaging in the actions that generate the data in question.

The problems remain: even if data encryption, key management, access control, and the motives of the data curator are all unimpeachable, what does it mean to preserve privacy, and how can it be accomplished?

"How" Is Hard
Let us consider a few common suggestions and some of the difficulties they can encounter.

Large Query Sets. One frequent suggestion is to disallow queries about a specific individual or small set of individuals. A well-known differencing argument demonstrates the inadequacy of the suggestion. Suppose it is known that Mr. X is in a certain medical database. Taken together, the answers to the two large queries "How many people in the database have the sickle cell trait?" and "How many people, not named X, in the database have the sickle cell trait?" yield the sickle cell status of Mr. X. The example also shows that encrypting the data, another frequent suggestion (oddly), would be of no help at all. The privacy compromise arises from correct operation of the database.

In *query auditing*, each query to the database is evaluated in the context of the query history to determine if a response would be disclosive; if so, then the query is refused. For example, query auditing might be used to interdict the pair of queries about sickle cell trait just described. This approach is problematic for several reasons, among them that query monitoring is computationally infeasible[16] and that the refusal to respond to a query may itself be disclosive.[15]

We think of a database as a collection of *rows*, with each row containing the data of a different respondent. In *subsampling* a subset of the rows is chosen at random and released. Statistics can then be computed on the subsample and, if the subsample is sufficiently large, these may be representative of the dataset as a whole. If the size of the subsample is very small compared to the size of the dataset, this approach has the property that every respondent is unlikely to appear in the subsample. However, this is clearly insufficient: Suppose appearing in a subsample has terrible consequences. Then every time subsampling occurs *some* individual suffers horribly.

In *input perturbation*, either the data or the queries are modified before a response is generated. This broad category encompasses a generalization of subsampling, in which the curator first chooses, based on a secret, random, function of the query, a subsample from the database, and then returns the result obtained by applying the query to the subsample.[4] A nice feature of this approach is that repeating the same query yields the same answer, while semantically equivalent but syntactially different queries are made on essentially unrelated subsamples. However, an outlier may only be protected by the unlikelihood of being in the subsample.

In what is traditionally called *randomized response*, the data itself is randomized once and for all and statistics are computed from the noisy responses, taking into account the distribution on the perturbation.[23] The term "randomized response" comes from the practice of having the respondents to a survey flip a coin and, based on the outcome, answering an invasive yes/no question or answering a more emotionally neutral one. In the computer science literature the choice governed by the coin flip is usually between honestly reporting one's value and responding randomly, typically by flipping a second coin and reporting the outcome. Randomized response was devised for the setting in which the individuals do not trust the curator, so we can think of the randomized responses as simply being published. Privacy comes from the uncertainty of how to interpret a reported value. The approach becomes untenable for complex data.

Adding random noise to the output has promise, and we will return to it later; here we point out that if done naïvely this approach will fail. To see this, suppose the noise has mean zero and that fresh randomness is used in generating every response. In this case, if the same query is asked repeatedly, then the responses can be averaged, and the true answer will eventually emerge. This is disastrous: an adversarial analyst could exploit this to carry out the difference attack described above. The approach cannot be "fixed" by recording each query and providing the same response each time a query is re-issued. There are several reasons

for this. For example, syntactically different queries may be semantically equivalent, and if the query language is sufficiently rich, then the equivalence problem itself is undecidable, so the curator cannot even test for this.

Problems with noise addition arise even when successive queries are completely unrelated to previous queries.[5] Let us assume for simplicity that the database consists of a single—but very sensitive—bit per person, so we can think of the database as an n-bit Boolean vector $d = (d_1, \ldots, d_n)$. This is an abstraction of a setting in which the database rows are quite complex, for example, they may be medical records, but the attacker is interested in one specific field, such as HIV status. The abstracted attack consists of issuing a string of queries, each described by a subset S of the database rows. The query is asking how many 1's are in the selected rows. Representing the query as the n-bit characteristic vector $\mathbf{S}$ of the set S, with 1's in all the positions corresponding to rows in S and 0's everywhere else; the true answer to the query is the inner product $A(S) = \sum_{i=1}^{n} d_i \mathbf{S}_i$. Suppose the privacy mechanism responds with $A(S) + $ random noise. How much noise is needed in order to preserve privacy?

Since we have not yet defined privacy, let us consider the easier problem of avoiding blatant "non-privacy," defined as follows: the system is blatantly non-private if an adversary can construct a candidate database that agrees with the real database D in, say, 99% of the entries. An easy consequence of the following theorem is that a privacy mechanism adding noise with magnitude always bounded by, say, $n/401$ is blatantly non-private against an adversary that can ask all 2^n possible queries.[5] There is nothing special about 401; any number exceeding 400 would work.

THEOREM 1. *Let $\mathcal{M}$ be a mechanism that adds noise bounded by E. Then there exists an adversary that can reconstruct the database to within $4E$ positions.*[5]

Blatant non-privacy with $E = n/401$ follows immediately from the theorem, as the reconstruction will be accurate in all but at most $4E = n \cdot \frac{4}{401} < n/100$ positions.

PROOF. Let d be the true database. The adversary can attack in two phases:

1. **Estimate the number of 1's in all possible sets:** Query $\mathcal{M}$ on all subsets $S \subseteq [n]$.
2. **Rule out "distant" databases:** For every candidate database $c \in \{0, 1\}^n$, if, for any $S \subseteq [n]$, $\left| \sum_{i \in S} c_i - \mathcal{M}(S) \right| > E$, then rule out c. If c is not ruled out, then output c and halt.

Since $\mathcal{M}(S)$ never errs by more than E, the real database will not be ruled out, so this simple (but inefficient!) algorithm will output *some* database; let us call it c. We will argue that the number of positions in which c and d differ is at most $4 \cdot E$.

Let I_0 be the indices in which $d_i = 0$, that is, $I_0 = \{i \mid d_i = 0\}$. Similarly, define $I_1 = \{i \mid d_i = 1\}$. Since c was not ruled out, $\left| \mathcal{M}(I_0) - \sum_{i \in I_0} c_i \right| \leq E$. However, by assumption $\left| \mathcal{M}(I_0) - \sum_{i \in I_0} d_i \right| \leq E$. It follows from the triangle inequality that c and d differ in at most $2E$ positions in I_0; the same argument shows that they differ in at most $2E$ positions in I_1. Thus, c and d agree on all but at most $4E$ positions. $\square$

What if we consider more realistic bounds on the number of queries? We think of $\sqrt{n}$ as an interesting threshold on noise, for the following reason: If the database contains n people drawn uniformly at random from a population of size $N \gg n$, and the fraction of the population satisfying a given condition is p, then we expect the number of rows in the database satisfying p to be roughly $np \pm \Theta\left(\sqrt{n}\right)$, by the properties of the hypergeometric distribution. That is, the sampling error is on the order of $\sqrt{n}$. We would like that the noise introduced for privacy is smaller than the sampling error, ideally $o\left(\sqrt{n}\right)$. Unfortunately, noise of magnitude $o\left(\sqrt{n}\right)$ is blatantly non-private against a series of $n \log^2 n$ randomly generated queries,[5] no matter the distribution on the noise. Several strengthenings of this pioneering result are now known. For example, if the entries in S are chosen independently according to a standard normal distribution, then blatant non-privacy continues to hold even against an adversary asking only $\Theta(n)$ questions, and even if more than a fifth of the responses have arbitrarily

wild noise magnitudes, provided the other responses have noise magnitude $o(\sqrt{n})$.[8]

These are not just interesting mathematical exercises. We have been focusing on *interactive* privacy mechanisms, distinguished by the involvement of the curator in answering each query. In the *noninteractive* setting the curator publishes some information of arbitrary form, and the data is not used further. Research statisticians like to "look at the data," and we have frequently been asked for a method of generating a "noisy table" that will permit highly accurate answers to be derived for computations that are not specified at the outset. The noise bounds say this is impossible: No such table can safely provide very accurate answers to too many weighted subset sum questions; otherwise the table could be used in a simulation of the interactive mechanism, and an attack could be mounted against the table. Thus, even if the analyst only requires the responses to a small number of unspecified queries, the fact that the table can be exploited to gain answers to other queries is problematic.

In the case of "Internet scale" datasets, obtaining responses to, say, $n \geq 10^8$ queries is infeasible. What happens if the curator permits only a sublinear number of questions? This inquiry led to the first algorithmic results in differential privacy, in which it was shown how to maintain privacy against a sublinear number of *counting* queries, that is, queries of the form "How many rows in the database satisfy property P?" by adding noise of order $o(\sqrt{n})$—less than the sampling error —to each answer.[12] The cumbersome privacy guarantee, which focused on the question of what an adversary can learn about a row in the database, is now known to imply a natural and still very powerful relaxation of differential privacy, defined here.

"What" Is Hard

Newspaper horror stories about "anonymized" and "de-identified" data typically refer to noninteractive approaches in which certain kinds of information in each data record have been suppressed or altered. A famous example is AOL's release of a set of "anonymized" search query

> It has taken several years to fully appreciate the importance of taking auxiliary information into account in privacy-preserving data release.

logs. People search for many "obviously" disclosive things, such as their full names (vanity searches), their own social security numbers (to see if their numbers are publicly available on the Web, possibly with a goal of assessing the threat of identity theft), and even the combination of mother's maiden name and social security number. AOL carefully redacted such obviously disclosive "personally identifiable information," and each user id was replaced by a random string. However, search histories can be very idiosyncratic, and a *New York Times* reporter correctly traced such an "anonymized" search history to a specific resident of Georgia.

In a *linkage attack*, released data are linked to other databases or other sources of information. We use the term *auxiliary information* to capture information about the respondents *other* than that which is obtained through the (interactive or noninteractive) statistical database. Any priors, beliefs, or information from newspapers, labor statistics, and so on, all fall into this category.

In a notable demonstration of the power of auxiliary information, medical records of the governor of Massachusetts were identified by linking voter registration records to "anonymized" Massachusetts Group Insurance Commission (GIC) medical encounter data, which retained the birthdate, sex, and zip code of the patient.[22]

Despite this exemplary work, it has taken several years to fully appreciate the importance of taking auxiliary information into account in privacy-preserving data release. Sources and uses of auxiliary information are endlessly varied. As a final example, it has been proposed to modify search query logs by mapping *all* terms, not just the user ids, to random strings. In *token-based hashing* each query is tokenized, and then an uninvertible hash function is applied to each token. The intuition is that the hashes completely obscure the terms in the query. However, using a statistical analysis of the hashed log and *any* (unhashed) query log, for example, the released AOL log discussed above, the anonymization can be severely compromised, showing that token-based hashing is unsuitable for anonymization.[17]

As we will see next, there are deep reasons for the fact that auxiliary information plays such a prominent role in these examples.

Dalenius's Desideratum

In 1977 the statistician Tore Dalenius articulated an "*ad omnia*" (as opposed to *ad hoc*) privacy goal for statistical databases: Anything that can be learned about a respondent from the statistical database should be learnable without access to the database. Although informal, this feels like the "right" direction. The breadth of the goal captures all the common intuitions for privacy. In addition, the definition only holds the database accountable for whatever "extra" is learned about an individual, beyond that which can be learned from other sources. In particular, an extrovert who posts personal information on the Web may destroy his or her own privacy, and the database should not be held accountable.

Formalized, Dalenius' goal is strikingly similar to the gold standard for security of a cryptosystem against a passive eavesdropper, defined five years later. *Semantic security* captures the intuition that the encryption of a message reveals no information about the message. This is formalized by comparing the ability of a computationally efficient adversary, having access to both the ciphertext and any auxiliary information, to output (anything about) the plaintext, to the ability of a computationally efficient party having access *only* to the auxiliary information (and not the ciphertext), to achieve the same goal.[13] Abilities are measured by probabilities of success, where the probability space is over the random choices made in choosing the encryption keys, the ciphertexts, and by the adversaries. Clearly, if this difference is very, very tiny, then in a rigorous sense the ciphertext leaks (almost) no information about the plaintext.

The formal definition of semantic security is a pillar of modern cryptography. It is therefore natural to ask whether a similar property, such as Dalenius' goal, can be achieved for statistical databases. But there is an essential difference in the two problems. Unlike the eavesdropper on a conversation, the statistical database attacker is also a user, that is, a legitimate

> The formal definition of semantic security is a pillar of modern cryptography. It is therefore natural to ask whether a similar property, such as Dalenius' goal, can be achieved for statistical databases.

consumer of the information provided by the statistical database (not to mention the fact that he or she may also be a respondent in the database).

Many papers in the literature attempt to formalize Dalenius' goal (in some cases unknowingly) by requiring that the adversary's prior and posterior beliefs about an individual (that is, before and after having access to the statistical database) should not be "too different," or that access to the statistical database should not change the adversary's views about any individual "too much." The difficulty with this approach is that if the statistical database teaches us anything at all, then it *should* change our beliefs about individuals. For example, suppose the adversary's (incorrect) prior view is that everyone has two left feet. Access to the statistical database teaches that almost everyone has one left foot and one right foot. The adversary now has a very different view of whether or not any given respondent has two left feet. But has privacy been compromised?

The last hopes for Dalenius' goal evaporate in light of the following parable, which again involves auxiliary information. Suppose we have a statistical database that teaches average heights of population subgroups, and suppose further that it is infeasible to learn this information (perhaps for financial reasons) in any other way (say, by conducting a new study). Finally, suppose that one's true height is considered sensitive. Given the auxiliary information "Turing is two inches taller than the average Lithuanian woman," access to the statistical database teaches Turing's height. In contrast, anyone without access to the database, knowing only the auxiliary information, learns much less about Turing's height.

A rigorous impossibility result generalizes this argument, extending to essentially any notion of privacy compromise, *assuming the statistical database is useful*. The heart of the attack uses extracted randomness from the statistical database as a one-time pad for conveying the privacy compromise to the adversary/user.[6, 9]

Turing did not have to be a member of the database for the attack described earlier to be prosecuted against him. More generally, the things that

statistical databases are designed to teach can, sometimes indirectly, cause damage to an individual, even if this individual is not in the database.

In practice, statistical databases are (typically) created to provide some anticipated social gain; they teach us something we could not (easily) learn without the database. Together with the attack against Turing, and the fact that he did not have to be a member of the database for the attack to work, this suggests a new privacy goal: Minimize the increased risk to an individual incurred by joining (or leaving) the database. That is, we move from comparing an adversary's prior and posterior views of an individual to comparing the risk to an individual when included in, versus when not included in, the database. This makes sense. A privacy guarantee that limits risk incurred by joining encourages participation in the dataset, increasing social utility. This is the starting point on our path to *differential privacy*.

Differential Privacy

Differential privacy will ensure that the ability of an adversary to inflict harm (or good, for that matter)—of any sort, to any set of people—should be essentially the same, independent of whether any individual opts in to, or opts out of, the dataset. We will do this indirectly, simultaneously addressing all possible forms of harm and good, by focusing on the probability of any given output of a privacy mechanism and how this probability can change with the addition or deletion of any row. Thus, we will concentrate on pairs of databases (D, D') differing only in one row, meaning one is a subset of the other and the larger database contains just one additional row. Finally, to handle worst-case pairs of databases, our probabilities will be over the random choices made by the privacy mechanism.

DEFINITION 1. *A randomized function* $\mathcal{K}$ *gives* ε-*differential privacy if for all datasets* D *and* D' *differing on at most one row, and all* $S \subseteq Range(\mathcal{K})$,

$$\Pr[\mathcal{K}(D) \in S] \leq \exp(\varepsilon)$$
$$\times \Pr[\mathcal{K}(D') \in S], \quad (1)$$

where the probability space in each case is over the coin flips of $\mathcal{K}$.

The multiplicative nature of the guarantee implies that an output whose probability is zero on a given database must also have probability zero on any neighboring database, and hence, by repeated application of the definition, on any other database. Thus, Definition 1 trivially rules out the subsample-and-release paradigm discussed: For an individual x not in the dataset, the probability that x's data is sampled and released is obviously zero; the multiplicative nature of the guarantee ensures that the same is true for an individual whose data *is* in the dataset.

Any mechanism satisfying this definition addresses all concerns that any participant might have about the leakage of his or her personal information, regardless of any auxiliary information known to an adversary: Even if the participant removed his or her data from the dataset, no outputs (and thus consequences of outputs) would become significantly more or less likely. For example, if the database were to be consulted by an insurance provider before deciding whether or not to insure a given individual, then the presence or absence of *any* individual's data in the database will not significantly affect his or her chance of receiving coverage.

Definition 1 extends naturally to group privacy. Repeated application of the definition bounds the ratios of probabilities of outputs when a collection C of participants opts in or opts out, by a factor of $e^{|C|\varepsilon}$. Of course, the point of the statistical database is to disclose aggregate information about large groups (while simultaneously protecting individuals), so we should expect privacy bounds to disintegrate with increasing group size.

The parameter ε is public, and its selection is a social question. We tend to think of ε as, say, 0.01, 0.1, or in some cases, ln 2 or ln 3.

Sometimes, for example, in the census, an individual's participation is known, so hiding presence or absence makes no sense; instead we wish to hide the values in an individual's row. Thus, we can (and sometimes do) extend "differing in at most one row" to mean having symmetric difference at most 1 to capture both possibilities. However, we will continue to use the original definition.

Returning to randomized response, we see that it yields ε-differential privacy for a value of ε that depends on the universe from which the rows are chosen and the probability with which a random, rather than non-random, value is contributed by the respondent. As an example, suppose each row consists of a single bit, and that the respondent's instructions are to first flip an unbiased coin to determine whether he or she will answer randomly or truthfully. If heads (respond randomly), then the respondent is to flip a second unbiased coin and report the outcome; if tails, the respondent answers truthfully. Fix $b \in \{0, 1\}$. If the true value of the input is b, then b is output with probability 3/4. On the other hand, if the true value of the input is $1 - b$, then b is output with probability 1/4. The ratio is 3, yielding (ln 3)-differential privacy.

Suppose n respondents each employ randomized response independently, but using coins of known, fixed, bias. Then, given the randomized data, by the properties of the binomial distribution the analyst can approximate the true answer to the question "How many respondents have value b?" to within an expected error on the order of $\Theta\left(\sqrt{n}\right)$. As we will see, it is possible to do much better—obtaining *constant* expected error, independent of n.

Generalizing in a different direction, suppose each row now has two bits, each one randomized independently, as described earlier. While each bit remains (ln 3)-differentially private, their logical-AND enjoys less privacy. That is, consider a privacy mechanism in which each bit is protected by this exact method of randomized response, and consider the query: "What is the logical-AND of the bits in the row of respondent i (after randomization)?" If we consider the two extremes, one in which respondent i has data 11 and the other in which respondent i has data 00, we see that in the first case the probability of output 1 is 9/16, while in the second case the probability is 1/16. Thus, this mechanism is at best (ln 9)-differentially private, not ln 3. Again, it is possible to do much better, even while releasing the entire 4-element histogram, also known as a *contingency table*, with only constant expected error in each cell.

Achieving Differential Privacy

Achieving differential privacy revolves around hiding the presence or absence of a single individual. Consider the query "How many rows in the database satisfy property P?" The presence or absence of a single row can affect the answer by at most 1. Thus, a differentially private mechanism for a query of this type can be designed by first computing the true answer and then adding random noise according to a distribution with the following property:

$$\forall z, z' \; s.t. \; |z - z'| = 1 : \Pr[z] \leq e^{\varepsilon} \Pr[z']. \quad (2)$$

To see why this is desirable, consider any feasible response r. For any m, if m is the true answer and the response is r then the random noise must have value $r - m$; similarly, if $m - 1$ is the true answer and the response is r, then the random noise must have value $r - m + 1$. In order for the response r to be generated in a differentially private fashion, it suffices for

$$e^{-\varepsilon} \leq \frac{\Pr[\text{noise} = r - m]}{\Pr[\text{noise} = r - m + 1]} \leq e^{\varepsilon}.$$

In general we are interested in vector-valued queries; for example, the data may be points in $\mathbf{R}^d$ and we wish to carry out an analysis that clusters the points and reports the location of the largest cluster.

DEFINITION 2. *For* $f : \mathcal{D} \rightarrow \mathbf{R}^d$, *the* L_1 *sensitivity of* f *is*[7]

$$\Delta f = \max_{D, D'} \| f(D) - f(D') \|_1$$

$$= \max_{D, D'} \sum_{i=1}^{d} | f(D)_i - f(D')_i | \quad (3)$$

for all D, D' *differing in at most one row.*

In particular, when $d = 1$ the sensitivity of f is the maximum difference in the values that the function f may take on a pair of databases that differ in only one row. This is the difference our noise must be designed to hide. For now, let us focus on the case $d = 1$.
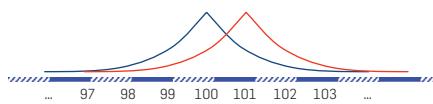
The Laplace distribution with parameter b, denoted Lap(b), has density function $P(z|b) = \frac{1}{2b} \exp(-|z|/b)$; its variance is $2b^2$. Taking $b = 1/\varepsilon$ we have that the density at z is proportional to $e^{-\varepsilon |z|}$. This distribution has highest density at 0 (good for accuracy), and for any z, z' such that $|z - z'| \leq 1$ the density at z is at most e^{ε} times the density at z', satisfying the condition in Equation 2. It is also symmetric about 0, and this is important. We cannot, for example, have a distribution that only yields non-negative noise. Otherwise the only databases on which a counting query could return a response of 0 would be databases in which no row satisfies the query. Letting D be such a database, and letting $D' = D \cup \{r\}$ for some row r satisfying the query, the pair D, D' would violate ε-differential privacy. Finally, the distribution gets flatter as ε decreases. This is correct: smaller ε means better privacy, so the noise density should be less "peaked" at 0 and change more gradually as the magnitude of the noise increases.

There is nothing special about the cases $d = 1$, $\Delta f = 1$:

THEOREM 2. *For* $f : \mathcal{D} \rightarrow \mathbf{R}^d$, *the mechanism* $\mathcal{K}$ *that adds independently generated noise with distribution Lap* $(\Delta f/\varepsilon)$ *to each of the d output terms enjoys* ε-differential privacy.[7]

Before proving the theorem, we illustrate the situation for the case of a counting query ($\Delta f = 1$) when $\varepsilon = \ln 2$ and the true answer to the query is 100. The distribution on the outputs (in gray) is centered at 100. The distribution on outputs when the true answer is 101 is shown in orange.

... 97 98 99 100 101 102 103 ...

PROOF. (Theorem 2) The proof is a simple generalization of the reasoning we used to illustrate the case of a single counting query.

Consider any subset $S \subseteq \text{Range}(\mathcal{K})$, and let D, D' be any pair of databases differing in at most one row. When the database is D, the probability density at any $r \in S$ is proportional to $\exp(-\| f(D) - r \|_1 (\varepsilon/\Delta f))$. Similarly, when the database is D', the probability density at any $r \in \text{Range}(K)$ is proportional to $\exp(-\| f(D') - r \|_1 (\varepsilon/\Delta f))$.

We have

$$\frac{e^{-\| f(D) - r \|(\varepsilon/\Delta f)}}{e^{-\| f(D') - r \|(\varepsilon/\Delta f)}} = \frac{e^{\| f(D') - r \|(\varepsilon/\Delta f)}}{e^{\| f(D) - r \|(\varepsilon/\Delta f)}}$$

$$= e^{(\| f(D') - r \| - \| f(D) - r \|)(\varepsilon/\Delta f)}$$

$$\leq e^{(\| f(D') - f(D) \|)(\varepsilon/\Delta f)}$$

where the inequality follows from the triangle inequality. By definition of sensitivity, $\| f(D') - f(D) \|_1 \leq \Delta f$, and so the ratio is bounded by $\exp(\varepsilon)$. Integrating over S yields ε-differential privacy. $\square$

Given any query sequence $f_1, \ldots, f_m$, ε-differential privacy can be achieved by running $\mathcal{K}$ with noise distribution Lap$\left(\sum_{i=1}^{m} \Delta f_i / \varepsilon \right)$ on *each* query, even if the queries are chosen adaptively, with each successive query depending on the answers to the previous queries. In other words, by allowing the quality of each answer to deteriorate in a controlled way with the sum of the sensitivities of the queries, we can maintain ε-differential privacy.

With this in mind, let us return to some of the suggestions we considered earlier. Recall that using the specific randomized response strategy described above, for a single Boolean attribute, yielded error $\Theta(\sqrt{n})$ on databases of size n and (ln 3)-differential privacy. In contrast, using Theorem 2 with the same value of ε, noting that $\Delta f = 1$ yields a variance of $2(1/\ln 3)^2$, or an expected error of $\sqrt{2}/\ln 3$. More generally, to obtain ε-differential privacy we get an expected error of $\sqrt{2}/\varepsilon$. Thus, our expected error magnitude is constant, independent of n.

What about two queries? The sensitivity of a sequence of two counting queries is 2. Applying the theorem with $\Delta f/\varepsilon = 2/\varepsilon$, adding independently generated noise distributed as Lap$(2/\varepsilon)$ to each true answer yields ε-differential privacy. The variance is $2(2/\varepsilon)^2$, or standard deviation $2\sqrt{2}/\varepsilon$. Thus, for any desired ε we can achieve ε-differential privacy by increasing the expected magnitude of the errors as a function of the total sensitivity of the two-query sequence. This holds equally for:

- Two instances of the *same query*, addressing the repeated query problem
- One count for each of two different bit positions, for example, when each row consists of two bits
- A pair of queries of the form: "How many rows satisfy property P?" and "How many rows satisfy property Q?" (where possibly $P = Q$)
- An arbitrary pair of queries

However, the theorem also shows we can sometimes do better. The logical-AND count we discussed earlier, even though it involves two different bits in each row, still only has sensitivity 1: The number of 2-bit rows whose entries are both 1 can change by at most 1 with the addition or deletion of a single row. Thus, this more complicated query can be answered in an ε-differentially private fashion using noise distributed as $Lap(1/\varepsilon)$; we do not need to use the distribution $Lap(2/\varepsilon)$.

Histogram Queries. The power of Theorem 2 really becomes clear when considering *histogram queries*, defined as follows. If we think of the rows of the database as elements in a universe X, then a histogram query is a partitioning of X into an arbitrary number of disjoint regions $X_1, X_2, \ldots, X_d$. The implicit question posed by the query is: "For $i = 1, 2, \ldots, d$, how many points in the database are contained in X_i?" For example, the database may contain the annual income for each respondent, and the query is a partitioning of incomes into ranges: $\{[0, 50K), [50K, 100K), \ldots, \geq 500K\}$. In this case $d = 11$, and the question is asking, for each of the d ranges, how many respondents in the database have annual income in the given range. This looks like d separate counting queries, but the entire query actually has sensitivity $\Delta f = 1$. To see this, note that if we remove one row from the database, then only one cell in the histogram changes, and that cell only changes by 1; similarly for adding a single row. So Theorem 2 says that ε-differential privacy can be maintained by perturbing each cell with an independent random draw from $Lap(1/\varepsilon)$. Returning to our example of 2-bit rows, we can pose the 4-ary histogram query requesting, for each pair of literals $v_1 v_2$, the number of rows with value $v_1 v_2$, adding noise of order $1/\varepsilon$ to each of the four cells.

When Noise Makes No Sense. There are times when the addition of noise for achieving privacy makes no sense. For example, the function f might map databases to strings, strategies, or trees, or it might be choosing the "best" among some specific, not necessarily continuous, set of real-valued objects. The problem of optimizing the

> **There are times when the addition of noise for achieving privacy makes no sense.**

output of such a function while preserving ε-differential privacy requires additional technology.

Assume the curator holds a database D and the goal is to produce an object y. The *exponential mechanism*[19] works as follows. We assume the existence of a *utility function* $u(D, y)$ that measures the quality of an output y, given that the database is D. For example, the data may be a set of labeled points in $\mathbf{R}^d$ and the output y might be a d-ary vector describing a $(d - 1)$-dimensional hyperplane that attempts to classify the points, so that those labeled with $+1$ have non-negative inner product with y and those labeled with -1 have negative inner product. In this case the utility would be the number of points correctly classified, so that higher utility corresponds to a better classifier. The exponential mechanism outputs y with probability proportional to $\exp(u(D,y)\varepsilon/\Delta u)$ and ensures ε-differential privacy. Here Δu is the sensitivity of the utility function bounding, for all databases (D, D') differing in only one row and potential outputs y, the difference $|u(D, y) - u(D',y)|$. In our example, $\Delta u = 1$. The mechanism assigns most mass to the best classifier, and the mass assigned to any other drops off exponentially in the decline in its utility for the current dataset—hence the name "exponential mechanism."

When Sensitivity Is Hard to Analyze. The Laplace and exponential mechanisms provide a differentially private interface through which the analyst can access the data. Such an interface can be useful even when it is difficult to determine the sensitivity of the desired function or query sequence; it can also be used to run an iterative algorithm, composed of easily analyzed steps, for as many iterations as a given privacy budget permits. This is a powerful observation; for example, using only noisy sum queries, it is possible to carry out many standard data mining tasks, such as singular value decompositions, finding an ID3 decision tree, clustering, learning association rules, and learning anything learnable in the statistical queries learning model, frequently with good accuracy, in a privacy-preserving fashion.[2] This approach has been generalized to yield a publicly available codebase for

writing programs that ensure differential privacy.[18]

k-Means Clustering. As an example of "private programming,"[2] consider *k*-means clustering, described first in its usual, non-private form. The input consists of points $p_1, \ldots, p_n$ in the *d*-dimensional unit cube $[0, 1]^d$. Initial candidate means $\mu_1, \ldots, \mu_k$ are chosen randomly from the cube and updated as follows:

1. Partition the samples $\{p_i\}$ into *k* sets $S_1, \ldots, S_k$, associating each p_i with the nearest μ_j.
2. For $1 \leq j \leq k$, set $\mu'_j = \sum_{i \in S_j} p_i / |S_j|$, the mean of the samples associated with μ_j.

This update rule is typically iterated until some convergence criterion has been reached, or a fixed number of iterations have been applied.

Although computing the nearest mean of any one sample (Step 1) would breach privacy, we observe that to compute an average among an unknown set of points it is enough to compute their sum and divide by their number. Thus, the computation only needs to expose the approximate cardinalities of the S_j, not the sets themselves. Happily, the *k* candidate means implicitly define a histogram query, since they partition the space $[0, 1]^d$ according to their Voronoi cells, and so the vector $(|S_1|, \ldots, |S_k|)$ can be released with very low noise in each coordinate. This gives us a differentially private approximation to the denominators in Step 2. As for the numerators, the sum of a subset of the p_i has sensitivity at most *d*, since the points come from the bounded region $[0,1]^d$. Even better, the sensitivity of the *d*-ary function that returns, for each of the *k* Voronoi cells, the *d*-ary sum of the points in the cell is at most *d*: Adding or deleting a single *d*-ary point can affect at most one sum, and that sum can change by at most 1 in each of the *d* dimensions. Thus, using a query sequence with total sensitivity at most $d + 1$, the analyst can compute a new set of candidate means by dividing, for each μ_j, the approximate sum of the points in S_j by the approximation to the cardinality $|S_j|$.

If we run the algorithm for a fixed number *N* of iterations we can use the noise distribution $\mathrm{Lap}((d + 1)N/\varepsilon)$ to obtain ε-differential privacy. If we do not know the number of iterations in advance we can increase the noise parameter as the computation proceeds. There are many ways to do this. For example, we can answer in the first iteration with parameter $(d + 1)(\varepsilon/2)$, in the next with parameter $(d + 1)(\varepsilon/4)$, and so on, each time using up half of the remaining "privacy budget."

Generating Synthetic Data
The idea of creating a synthetic dataset whose statistics closely mirror those of the original dataset, but which preserves privacy of individuals, was proposed in the statistics community no later than 1993.[21] The lower bounds on noise discussed at the end of Section on "How Is Hard" imply that no such dataset can safely provide very accurate answers to too many weighted subset sum questions, motivating the interactive approach to private data analysis discussed herein. Intuitively, the advantage of the interactive approach is that only the questions actually asked receive responses.

Against this backdrop, the non-interactive case was revisited from a learning theory perspective, challenging the interpretation of the noise lower bounds as a limit on the number of queries that can be answered privately.[3] This work, described next, has excited interest in interactive and non-interactive solutions yielding noise in the range $\left[\omega\left(\sqrt{n}\right), o(n)\right]$.

Let *X* be a universe of data items and let $\mathcal{C}$ be a *concept class* consisting of functions $c : X \to \{0,1\}$. We say $x \in X$ *satisfies* a concept $c \in \mathcal{C}$ if and only if $c(x) = 1$. A concept class can be extremely general; for example, it might consist of all rectangles in the plane, or all Boolean circuits containing a given number of gates.

Given a sufficiently large database $D \in X^n$, it is possible to privately generate a synthetic database that maintains approximately correct fractional counts for *all* concepts in $\mathcal{C}$ (there may be infinitely many!). That is, letting *S* denote the synthetic database produced, with high probability over the choices made by the privacy mechanism, for every concept $c \in \mathcal{C}$, the fraction of elements in *S* that satisfy *c* is approximately the same as the fraction of elements in *D* that satisfy *c*.

The minimal size of the input database depends on the quality of the approximation, the logarithm of the cardinality of the universe *X*, the privacy parameter ε, and the *Vapnick–Chervonenkis dimension* of the concept class $\mathcal{C}$ (for finite $|C|$ this is at most $\log_2 |\mathcal{C}|$). The synthetic dataset, chosen by the exponential mechanism, will be a set of $m = O(\mathrm{VCdim}(\mathcal{C})/\gamma^2)$, elements in X (γ governs the maximum permissible inaccuracy in the fractional count). Letting *D* denote the input dataset and $\hat{D}$ a candidate synthetic dataset, the utility function for the exponential mechanism is given by

$$u(D, \hat{D}) = -\max_{h \in \mathcal{C}} \left| h(D) - \frac{n}{m} h(\hat{D}) \right|.$$

Pan-Privacy
Data collected by a curator for a given purpose may be subject to "mission creep" and legal compulsion, such as a subpoena. Of course, we could analyze data and then throw it away, but can we do something even stronger, never storing the data in the first place? Can we strengthen our notion of privacy to capture the "never store" requirement?

These questions suggest an investigation of differentially private streaming algorithms with small state—much too small to store the data. However, nothing in the definition of a streaming algorithm, even one with very small state, precludes storing a few individual data points. Indeed, popular techniques from the streaming literature, such as Count-Min Sketch and subsampling, do precisely this. In such a situation, a subpoena or other intrusion into the local state will breach privacy.

A *pan-private* algorithm is private "inside and out," remaining differentially private even if its internal state becomes visible to an adversary.[10] To understand the pan-privacy guarantee, consider click stream data. This data is generated by individuals, and an individual may appear many times in the stream. Pan-privacy requires that any two streams differing only in the information of a single individual should produce very similar distributions on the *internal states* of the algorithm *and on its outputs*, even though the data of an individual are

interleaved arbitrarily with other data in the stream.

As an example, consider the problem of *density estimation*. Assuming, for simplicity, that the data stream is just a sequence of IP addresses in a certain range, we wish to know what fraction of the set of IP addresses in the range actually appears in the stream. A solution inspired by randomized response can be designed using the following technique.[10]

Define two probability distributions, D_0 and D_1, on the set $\{0, 1\}$. D_0 assigns equal mass to zero and to one. D_1 has a slight bias toward 1; specifically, 1 has mass $1/2 + \varepsilon/4$, while 0 has mass $1/2 - \varepsilon/4$.

Let X denote the set of all possible IP addresses in the range of interest. The algorithm creates a table, with a 1-bit entry b_x for each $x \in X$, initialized to an independent random draw from D_0. So initially the table is roughly half zeroes and half ones.

In an atomic step, the algorithm receives an element from the stream, changes state, and discards the element. When processing $x \in X$, the algorithm makes a fresh random draw from D_1, and stores the result in b_x. This is done no matter how many times x may have appeared in the past. Thus, for any x appearing at least once, b_x will be distributed according to D_1. However, if x never appears, then the entry for x is the bit drawn according to D_0 during the initialization of the table.

As with randomized response, the density in X of the items in the stream can be approximated from the number of 1's in the table, taking into account the expected fraction of "false positives" from the initialization phase and the "false negatives" when sampling from D_1. Letting θ denote the fraction of entries in the table with value 1, the output is $4(\theta - 1/2)/\varepsilon + \text{Lap}(1/\varepsilon|X|)$.

Intuitively, the internal state is differentially private because, for each $b \in \{0, 1\}$, $e^{-\varepsilon} \le \Pr_{D_1}[b]/\Pr_{D_0}[b] \le e^{\varepsilon}$; privacy for the output is ensured by the addition of Laplacian noise. Overall, the algorithm is 2ε-differentially pan-private.

Conclusion

The differential privacy frontier is expanding rapidly, and there is insufficient space here to list all the interesting directions currently under investigation by the community. We identify a few of these.

The Geometry of Differential Privacy. Sharper upper and lower bounds on noise required for achieving differential privacy against a sequence of linear queries can be obtained by understanding the geometry of the query sequence.[14] In some cases dependencies among the queries can be exploited by the curator to markedly improve the accuracy of the responses. Generalizing this investigation to the nonlinear and interactive cases would be of significant interest.

Algorithmic Complexity. We have so far ignored questions of computational complexity. Many, but not all, of the techniques described here have efficient implementations. For example, there are instances of the synthetic data generation problem that, under standard cryptographic assumptions, have no polynomial time implementation.[11] It follows that there are cases in which the exponential mechanism has no efficient implementation. When can this powerful tool be implemented efficiently, and how?

An Alternative to Differential Privacy? Is there an alternative, "*ad omnia*," guarantee that composes automatically, and permits even better accuracy than differential privacy? Can cryptography be helpful in this regard?[20]

The work described herein has, for the first time, placed private data analysis on a strong mathematical foundation. The literature connects differential privacy to decision theory, economics, robust statistics, geometry, additive combinatorics, cryptography, complexity theory learning theory, and machine learning. Differential privacy thrives because it is natural, it is not domain-specific, and it enjoys fruitful interplay with other fields. This flexibility gives hope for a principled approach to privacy in cases, like private data analysis, where traditional notions of cryptographic security are inappropriate or impracticable.

References
1. Adam, N.R., Wortmann, J. Security-control methods for statistical databases: A comparative study. *ACM Comput. Surv. 21* (1989), 515–556.
2. Blum, A., Dwork, C., McSherry, F., Nissim, K. Practical privacy: The SuLQ framework. In *Proceedings of the 24th ACM Symposium on Principles of Database Systems* (2005), 128–138.
3. Blum, A., Ligett, K., Roth, A. A learning theory approach to non-interactive database privacy. In *Proceedings of the 40th ACM Symposium on Theory of Computing* (2008), 609–618.
4. Denning, D.E. Secure statistical databases with random sample queries. *ACM Trans. Database Syst. 5* (1980), 291–315.
5. Dinur, I., Nissim, K. Revealing information while preserving privacy. In *Proceedings of the 22nd ACM Symposium on Principles of Database Systems* (2003), 202–210.
6. Dwork, C. Differential privacy. In *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP) (2)* (2006), 1–12.
7. Dwork, C., McSherry, F., Nissim, K., Smith, A. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Theory of Cryptography Conference* (2006), 265–284.
8. Dwork, C., McSherry, F., Talwar, K. The price of privacy and the limits of lp decoding. In *Proceedings of the 39th ACM Symposium on Theory of Computing* (2007), 85–94.
9. Dwork, C., Naor, M. On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. *J. Privacy Confidentiality 2* (2010). Available at: http://repository.cmu.edu/jpc/vol2/iss1/8.
10. Dwork, C., Naor, M., Pitassi, T., Rothblum, G., Yekhanin, S. Pan-private streaming algorithms. In *Proceedings of the 1st Symposium on Innovations in Computer Science* (2010).
11. Dwork, C., Naor, M., Reingold, O., Rothblum, G., Vadhan, S. When and how can privacy-preserving data release be done efficiently? In *Proceedings of the 41st ACM Symposium on Theory of Computing* (2009), 381–390.
12. Dwork, C., Nissim, K. Privacy-preserving datamining on vertically partitioned databases. In *Advances in Cryptology—CRYPTO'04* (2004), 528–544.
13. Goldwasser, S., Micali, S. Probabilistic encryption. *JCSS 28* (1984), 270–299.
14. Hardt, M., Talwar, K. On the geometry of differential privacy, (2009). In *Proceedings of the 42nd ACM Symposium on Theory of Computing* (2010), 705–714.
15. Kenthapadi K., Mishra, N., Nissim, K. Simulatable auditing. In *Proceedings of the 24th ACM Symposium on Principles of Database Systems* (2005), 118–127.
16. Kleinberg, J., Papadimitriou, C., Raghavan, P. Auditing boolean attributes. In *Proceedings of the 19th ACM Symposium on Principles of Database Systems* (2000), 86–91.
17. Kumar, R., Novak, J., Pang, B., Tomkins, A. On anonymizing query logs via token-based hashing. In *Proceedings of the WWW 2007* (2007), 629–638.
18. McSherry, F. Privacy integrated queries (codebase). Available on Microsoft Research downloads website. See also *Proceedings of SIGMOD* (2009), 19–30.
19. McSherry, F., Talwar, K. Mechanism design via differential privacy. In *Proceedings of the 48th Annual Symposium on Foundations of Computer Science* (2007).
20. Mironov, I., Pandey, O., Reingold, O., Vadhan, S. Computational differential privacy. In *Advances in Cryptology—CRYPTO'09* (2009), 126–142.
21. Rubin, D. Discussion: Statistical disclosure limitation. *J. Official Statist. 9* (1993), 462–468.
22. Sweeney, L. Weaving technology and policy together to maintain confidentiality. *J. Law Med. Ethics 25* (1997), 98–110.
23. Warner, S. Randomized response: a survey technique for eliminating evasive answer bias. *JASA* (1965), 63–69.

Cynthia Dwork (dwork@microsoft.com) is a principal researcher at Microsoft Research, Silicon Valley Campus, Mountain View, CA.

Understanding
DATABASE RECONSTRUCTION ATTACKS
on Public Data

THESE ATTACKS ON STATISTICAL DATABASES ARE NO LONGER A THEORETICAL DANGER.

SIMSON GARFINKEL
JOHN M. ABOWD, AND
CHRISTIAN MARTINDALE
U.S. CENSUS BUREAU

In 2020 the U.S. Census Bureau will conduct the Constitutionally mandated decennial Census of Population and Housing. Because a census involves collecting large amounts of private data under the promise of confidentiality, traditionally statistics are published only at high levels of aggregation. Published statistical tables are vulnerable to DRAs (*database reconstruction attacks*), in which the underlying microdata is recovered merely by finding a set of microdata that is consistent with the published statistical tabulations. A DRA can be performed by using the tables to create a set of mathematical constraints and then solving the resulting set of simultaneous equations. This article shows how such an attack can be addressed by adding noise to the published tabulations, so that the reconstruction no longer results in the original data. This has implications for the 2020 Census.

The goal of the census is to count every person once,

and only once, and in the correct place. The results are used to fulfill the Constitutional requirement to apportion the seats in the U.S. House of Representatives among the states according to their respective numbers.

In addition to this primary purpose of the decennial census, the U.S. Congress has mandated many other uses for the data. For example, the U.S. Department of Justice uses block-by-block counts by race for enforcing the Voting Rights Act. More generally, the results of the decennial census, combined with other data, are used to help distribute more than $675 billion in federal funds to states and local organizations.

Beyond collecting and distributing data on the American people, the Census Bureau is also charged with protecting the privacy and confidentiality of survey responses. All census publications must uphold the confidentiality standard specified by Title 13, Section 9 of the U.S. Code, which states that Census Bureau publications are prohibited from identifying "the data furnished by any particular establishment or individual." This section prohibits the Census Bureau from publishing respondents' names, addresses, or any other information that might identify a specific person or establishment.

Upholding this confidentiality requirement frequently poses a challenge, because many statistics can inadvertently provide information in a way that can be attributed to a particular entity. For example, if a statistical agency *accurately* reports that there are two persons living on a block and that the average age of the block's residents is 35, that would constitute an improper disclosure of personal information, because

one of the residents could look up the data, subtract their contribution, and infer the age of the other.

Of course, this is an extremely simple example. Statistical agencies have understood the risk of such unintended disclosure for decades and have developed a variety of techniques to protect data confidentiality while still publishing useful statistics. These techniques include *cell suppression*, which prohibits publishing statistical summaries from small groups of respondents; *top-coding*, in which ages higher than a certain limit are coded as that limit before statistics are computed; *noise-injection*, in which random values are added to some attributes; and *swapping*, in which some of the attributes of records representing different individuals or families are swapped. Together, these techniques are called SDL (statistical disclosure limitation).

Computer scientists started exploring the issue of statistical privacy in the 1970s with the increased availability of interactive query systems. The goal was to build a system that would allow users to make queries that would produce summary statistics without revealing information about individual records. Three approaches emerged: auditing database queries, so that users would be prevented from issuing queries that zeroed in on data from specific individuals; adding noise to the data stored within the database; and adding noise to query results.[1] Of these three, the approach of adding noise proved to be the easiest, because the complexity of auditing queries increased exponentially over time—and, in fact, was eventually shown to be NP (nondeterministic polynomial)-hard.[8] Although these results were all couched in the

language of interactive query systems, they apply equally well to the activities of statistical agencies, with the *database* being the set of confidential survey responses, and the *queries* being the schedule of statistical tables that the agency intends to publish.

In 2003, Irit Dinur and Kobbi Nissim showed that it isn't even necessary for an attacker to construct queries on a database carefully to reveal its underlying confidential data.[4] Even a surprisingly small number of random queries can reveal confidential data, because the results of the queries can be combined and then used to "reconstruct" the underlying confidential data. Adding noise to either the database or to the results of the queries decreases the accuracy of the reconstruction, but it also decreases the accuracy of the queries. The challenge is to add sufficient noise in such a way that each individual's privacy is protected, but not so much noise that the utility of the database is ruined.

Subsequent publications[3,6] refined the idea of adding noise to published tables to protect the privacy of the individuals in the data set. Then in 2006, Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith proposed a formal framework for understanding these results. Their paper, "Calibrating Noise to Sensitivity in Private Data Analysis,"[5] introduced the concept of *differential privacy*. They provided a mathematical definition of the privacy loss that persons suffer as a result of a data publication, and they proposed a mechanism for determining how much noise needs to be added for any given level of privacy protection. (The authors were awarded the Test of Time award at the Theory of Cryptography Conference in 2016

and the Godel Prize in 2017.)

The 2020 census is expected to count roughly 330 million people living on roughly 8.5 million blocks, with some inhabited blocks having as few as a single person and other blocks having thousands. With this level of scale and diversity, it is difficult to visualize how such a data release might be susceptible to database reconstruction. We now know, however, that reconstruction would in fact pose a significant threat to the confidentiality of the 2020 microdata that underlies unprotected statistical tables if privacy-protecting measures are not implemented. To help understand the importance of adopting formal privacy methods, this article presents a database reconstruction of a much smaller statistical publication: a hypothetical block containing seven people distributed over two households. (The 2010 U.S. Census contained 1,539,183 census blocks in the 50 states and the District of Columbia with between one and seven residents. The data can be downloaded from https://www2.census.gov/census_2010/01-Redistricting_File--PL_94-171/.)

Even a relatively small number of constraints results in an exact solution for the blocks' inhabitants. Differential privacy can protect the published data by creating uncertainty. Although readers may think that the reconstruction of a block with just seven people is an insignificant risk for the country as a whole, this attack can be performed for virtually every block in the United States using the data provided in the 2010 census. The final section of this article discusses the implications of this for the 2020 decennial census.

AN EXAMPLE DATABASE RECONSTRUCTION ATTACK

To present the attack, let's consider the census of a fictional geographic frame (for example, a suburban block), conducted by a fictional statistical agency. For every block, the agency collects each resident's age, sex, and race, and publishes a variety of statistics. To simplify the example, this fictional world has only two races—black or African American, and white—and two sexes—female and male.

The statistical agency is prohibited from publishing the raw microdata and instead publishes a tabular report. Table 1 shows fictional statistical data for a fictional block

TABLE 1: **FICTIONAL STATISTICAL DATA FOR A FICTIONAL BLOCK**

STATISTIC	GROUP	AGE		
		COUNT	MEDIAN	MEAN
1A	total population	7	30	38
2A	female	4	30	33.5
2B	male	3	30	44
2C	black or African American	4	51	48.5
2D	white	3	24	24
3A	single adults	(D)	(D)	(D)
3B	married adults	4	51	54
4A	black or African American female	3	36	36.7
4B	black or African American male	(D)	(D)	(D)
4C	white male	(D)	(D)	(D)
4D	white female	(D)	(D)	(D)
5A	persons under 5 years	(D)	(D)	(D)
5B	persons under 18 years	(D)	(D)	(D)
5C	persons 64 years or over	(D)	(D)	(D)

Note: Married persons must be 15 or over

published by the fictional statistics agency. The "statistic" column is for identification purposes only.

Notice that a substantial amount of information in table 1 has been suppressed—marked with a (D). In this case, the statistical agency's disclosure-avoidance rules prohibit it from publishing statistics based on one or two people. This suppression rule is sometimes called "the rule of three," because cells in the report sourced from fewer than three people are suppressed. In addition, complementary suppression has been applied to prevent subtraction attacks on the small cells.

Encoding the constraints

The database can be reconstructed by treating the attributes of the persons living on the block as a collection of variables. A set of constraints is then extracted from the published table. The database reconstruction finds a set of attributes that are consistent with the constraints. If the statistics are highly constraining, then there will be a single possible reconstruction, and the reconstructed microdata will necessarily be the same as the underlying microdata used to create the original statistical publication. Note that there must be at least one solution because the table is known to be tabulated from a real database.

For example, statistic 2B states that three males live in the geography. This fictional statistical agency has previously published technical specifications that its computers internally represent each person's age as an integer. The oldest verified age of any human being was 122.[14] If we allow for unreported supercentenarians and consider 125 to the oldest possible age of a human

being, there are only a finite number of possible age combinations, specifically:

$$\binom{125}{3} = \frac{125 \times 124 \times 123}{3 \times 2 \times 1} = 317{,}750$$

Within the 317,750 possible age combinations, however, there are only 30 combinations that satisfy the constraints of having a median of 30 and a mean of 44 (see Table 1). (Notice that the table does not depend on the oldest possible age, so long as it is 101 or over.) Applying the constraints imposed by the published statistical tables, the possible combinations of ages for the three males can be reduced from 317,750 to 30. Table 2 shows the 30 possible ages for which the median is 30 and the mean is 44

TABLE 2: **POSSIBLE AGES FOR A MEDIAN OF 30 AND MEAN OF 44**

A	B	C	A	B	C	A	B	C
1	30	101	11	30	91	21	30	81
2	30	100	12	30	90	22	30	80
3	30	99	13	30	89	23	30	79
4	30	98	14	30	88	24	30	78
5	30	97	15	30	87	25	30	77
6	30	96	16	30	86	26	30	76
7	30	95	17	30	85	27	30	75
8	30	94	18	30	84	28	30	74
9	30	93	19	30	83	29	30	73
10	30	92	20	30	82	30	30	72

To mount a full reconstruction attack, an attacker extracts all of these constraints and then creates a single mathematical model embodying all constraints. An automated solver can then find an assignment of the variables that satisfies these constraints.

To continue with the example, statistic 1A establishes the universe of the constraint system. Because the block contains seven people, and each has four attributes (age, sex, race, and marital status), that creates 28 variables, representing those four attributes for each person. These variables are A1... A7 (age), S1... S7 (sex), R1... R7 (race), and M1... M7 (marital status), as shown in table 3. The table

TABLE 3: **VARIABLES ASSOCIATED WITH THE RECONSTRUCTION ATTACK.**

PERSON	AGE	SEX	RACE	MARITAL STATUS
1	A1	S1	R1	M1
2	A2	S2	R2	M2
3	A3	S3	R3	M3
4	A4	S4	R4	M4
5	A5	S5	R5	M5
6	A6	S6	R6	M6
7	A7	S7	R7	M7
KEY				
female		0		
male		1		
black or African American			0	
white			1	
single				0
married				1

shows the variables associated with the DRA. The coding
for the categorical attributes is presented in the key.

Because the mean age is 38, we know that:

$$A1 + A2 + A3 + A4 + A5 + A6 + A7 = 7 \times 38$$

The language Sugar[13] is used to encode the constraints
in a form that can be processed by a SAT (satisfiability)
solver. Sugar represents constraints as s-expressions.[11] For
example, the age combination equation can be represented
as:

```
; First define the integer variables,
; with the range 0..125
(int A1 0 125)
(int A2 0 125)
(int A3 0 125)
(int A4 0 125)
(int A5 0 125)
(int A6 0 125)
(int A7 0 125)


; Statistic 1A: Mean age is 38
(= (+ A1 A2 A3 A4 A5 A6 A7)
    (* 7 38)
)
```

Once the constraints in the statistical table are turned
into s-expressions, Sugar solves them with a brute-force
algorithm. Essentially, Sugar explores every possible
combination of the variables, until a combination is

found that satisfies the constraints. Using a variety of heuristics, SAT solvers are able to rapidly eliminate many combinations of variable assignments.

Despite their heuristic complexity, SAT solvers can process only those systems that have Boolean variables, so Sugar transforms the s-expressions into a much larger set of Boolean constraints. For example, each age variable is encoded using unary notation as 126 Boolean variables. Using this notation, the decimal value 0 is encoded as 126 false Boolean variables, the decimal value 1 is encoded as 1 true and 125 false values, and so on. Although this conversion is not space efficient, it is fast, provided that the integers have a limited range.

To encode the median age constraint, the median of a group of numbers is precisely defined as the value of the middle number when the numbers are arranged in sorted order (for the case in which there is an odd number of numbers). Until now, persons 1 through 7 have not been distinguished in any way: the number labels are purely arbitrary. To make it easier to describe the median constraints, we can assert that the labels must be assigned in order of age. This is done by introducing five constraints, which has the side effect of eliminating duplicate answers that have simply swapped records, an approach called *breaking symmetry*:[12]

```
(<=  A1  A2)
(<=  A2  A3)
(<=  A3  A4)
(<=  A4  A5)
(<=  A6  A7)
```

Having asserted that the labels are in chronological order, we can constrain the age of the person in the middle to be the median:

```
(= A4 30)
```

This code fragment assures that the output is sorted by age. This technique also does a good job of eliminating duplicate answers that have swapped records.

Sugar has an "`if`" function that allows encoding constraints for a subset of the population. Recall that statistic 2B contains three constraints: there are three males, their median age is 30, and their average age is 44. The value 0 represents a female, and 1 represents a male:

```
#define FEMALE 0
#define MALE 1
```

Using the variable Sn to represent the sex of person n, we then have the constraint:

```
S1 + S2 + S3 + S4 + S5 + S6 + S7 = 3
```

This can be represented as:

```
(= (+ S1 S2 S3 S4 S5 S6 S7) 3)
```

Now, using the `if` function, it is straightforward to create a constraint for the mean age 44 of male persons:

```
(= (+ (if (= S1 MALE) A1 0) ; average male age
       (if (= S2 MALE) A2 0)
       (if (= S3 MALE) A3 0)
       (if (= S4 MALE) A4 0)
       (if (= S5 MALE) A5 0)
       (if (= S6 MALE) A6 0)
       (if (= S7 MALE) A7 0)
       )
   (* 3 44))
```

Table 1 translates into 164 individual s-expressions extending over 457 lines. Sugar then translates this into a single Boolean formula consisting of 6,755 variables arranged in 252,575 clauses. This format is called the CNF (conjunctive normal form) because it consists of many clauses that are combined using the Boolean AND operation.

Interestingly, we can even create constraints for the suppressed data. Statistic 3A is suppressed, so we know that there are 0, 1, or 2 single adults, assuming that no complementary suppression was required. Let Mn represent the marital status of person n:

```
#define SINGLE 0
#define MARRIED 1

(int SINGLE_ADULT_COUNT 0 2)
(= (+ (if (and (= M1 SINGLE) (> A1 17)) 1 0)
       (if (and (= M2 SINGLE) (> A2 17)) 1 0)
       (if (and (= M3 SINGLE) (> A3 17)) 1 0)
       (if (and (= M4 SINGLE) (> A4 17)) 1 0)
```

```
      (if (and (= M5 SINGLE) (> A5 17)) 1 0)
      (if (and (= M6 SINGLE) (> A6 17)) 1 0)
      (if (and (= M7 SINGLE) (> A7 17)) 1 0))
 SINGLE_ADULT_COUNT)

(>= SINGLE_ADULT_COUNT 0)
(<= SINGLE_ADULT_COUNT 2)
```

Translating the constraints into CNF allows them to be solved using any solver that can solve NP-complete program, such as a SAT solver, an SMT (satisfiability module theories) solver, or MIP (mixed integer programming) solver. There are many such solvers, and most take input in the so-called DIMACS file format, which is a standardized form for representing CNF equations. The DIMACS format (named for the Center for Discrete Mathematics and Theoretical Computer Science) was popularized by a series of annual SAT solver competitions. One of the results of these competitions was a tremendous speed-up of SAT solvers over the past two decades. Many solvers can now solve CNF systems with millions of variables and clauses in just a few minutes, although some problems do take much longer. Marijn Heule and Oliver Kullmann discussed the rapid advancement and use of SAT solvers in their 2017 article, "The Science of Brute Force."[7]

The open-source PicoSAT[2] solver is able to find a solution to the CNF problem detailed here in roughly two seconds on a 2013 MacBook Pro with a 2.8-GHz Intel i7 processor and 16 GB of RAM (although the program is not limited by RAM), while the open-source Glucose SAT solver can solve the problem in under 0.1 seconds on the same

computer. The stark difference between the two solvers shows the speed-up possible with an improved solving algorithm.

Exploring the solution universe

PicoSAT creates a satisfying assignment for the 6,755 Boolean variables. After the solver runs, Sugar can translate these assignments back into integer values of the constructed variables. (SMT and MIP solvers can represent the constraints at a higher level of abstraction, but for our purposes a SAT solver is sufficient.)

There exists a solution universe of all the possible solutions to this set of constraints. If the solution universe contains a single possible solution, then the published statistics completely reveal the underlying confidential data—provided that noise was not added to either the microdata or the tabulations as a disclosure-avoidance mechanism. If there are multiple satisfying solutions, then any element (person) in common among all of the solutions is revealed. If the equations have no solution, either the set of published statistics is inconsistent with the fictional statistical agency's claim that it is tabulated from a real confidential database or an error was made in that tabulation. This doesn't mean that a high-quality reconstruction is not possible. Instead of using the published statistics as a set of constraints, they can be used as inputs to a multidimensional objective function: the system can then be solved using another kind of solver called an optimizer.

Normally SAT, SMT, and MIP solvers will stop when they find a single satisfying solution. One of the advantages

of PicoSAT is that it can produce the solution universe of all possible solutions to the CNF problem. In this case, however, there is a single satisfying assignment that produces the statistics in table 1. That assignment is seen in Table 4.

Table 1 provides some redundant constraints on the solution universe: some of the constraints can be dropped while preserving a unique solution. For example, dropping statistic 2A, 2B, 2C, or 2D still yields a single solution, but dropping 2A *and* 2B increases the solution universe to eight satisfying solutions. All of these solutions contain the reconstructed microdata records 8FBS, 36FBM, 66FBM, and 84MBM. This means that even if statistics 2A and 2B are suppressed, we can still infer that these four microdata records must be present.

Statistical agencies have long used suppression in an attempt to provide privacy to those whose attributes are present in the microdata, although the statistics that they typically drop are those that are based on a small number

TABLE 4: **A SINGLE SATISFYING ASSIGNMENT**

AGE	SEX	RACE	MARITAL STATUS		SOLUTION #1
8	F	B	S		8FBS
18	M	W	S		18MWS
24	F	W	S	=	24FWS
30	M	W	M		30MWM
36	F	B	M		36FBM
66	F	B	M		66FBM
84	M	B	M		84MBM

of persons. How effective is this approach?

In table 1, statistic 4A is an obvious candidate for suppression—especially given that statistics 4B, 4C, and 4D have already been suppressed to avoid an inappropriate statistical disclosure.

Removing the constraints for statistic 4A increases the number of solutions from one to two, shown in table 5.

DEFENDING AGAINST A DRA

There are three approaches for defending against a database reconstruction attack. The first is to publish less statistical data—this is the approach taken by legacy disclosure-avoidance techniques (cell suppression, top-coding, and generalization). The second and third approaches involve adding noise, or randomness. Noise can be added to the statistical data being tabulated or to the results after tabulation. Each approach is considered here.

Option 1: Publish less data

Although it might seem that publishing less statistical data is a reasonable defense against the DRA, this choice

TABLE 5: **SOLUTIONS WITHOUT STATISTIC 4A**

SOLUTION #1	SOLUTION #2
8FBS	2FBS
18MWS	12MWS
24FWS	24FWM
30MWM	30MBM
36FBM	36FWS
66FBM	72FBM
84MBM	90MBM

may severely limit the number of tabulations that can be published. A related problem is that, with even a moderately small population, it may be computationally infeasible to determine when the published statistics still identify a sizable fraction of individuals in the population.

Option 2: Apply noise before tabulation

This approach is called *input noise injection*. For example, each respondent's age might be randomly altered by a small amount. Input noise injection doesn't prevent finding a set of microdata that is consistent with the published statistics, what we call *database reconstruction*, but it limits the value of the reconstructed microdata, since what is reconstructed is the microdata *after* the noise has been added.

Swapping, the disclosure-avoidance approach used in the 2010 census, is a kind of input noise injection. In swapping, some of the attributes are exchanged, or *swapped*, between records. The advantage of swapping is that it has no impact on some kinds of statistics: if people are swapped only within a county, then any tabulation at the county level will be unaffected by swapping. The disadvantage of swapping is that it can have significant impact on statistics at lower levels of geography, and values that are not swapped are unprotected.

Option 3: Apply noise to the published statistics

This approach is called *output noise injection*. Whereas input noise injection applies noise to the microdata directly, output noise injection applies output to the statistical publications. Output noise injection complicates database

reconstruction by eliminating naïve approaches based on the straightforward application of SAT solvers. Also, even if a set of microdata is constructed that is mostly consistent with the published statistics, this microdata will be somewhat different from the original microdata that was collected. The more noise that was added to the tabulation, the more the microdata will be different.

When noise is added to either the input data (option 2) or the tabulation results (option 3), with all records having equal probability of being altered, it is possible to mathematically describe the resulting privacy protection. This is the basis of differential privacy.

Implications for the 2020 Census

The Census Bureau has announced that it is adopting a noise-injection mechanism based on differential privacy to provide privacy protection for the underlying microdata collected as part of the 2020 census. Following is the motivation for that decision.

The protection mechanism developed for the 2010 census was based on a swapping.[15] The swapping technique was not designed to protect the underlying data against a DRA. Indeed, it is the Census Bureau's policy that both the swapped and the unswapped microdata are considered confidential.

The 2010 census found a total population of 308,745,538. These people occupied 10,620,683 habitable blocks. Each person was located in a residential housing unit or institutional housing arrangement (what the Census Bureau calls "group quarters"). For each person, the Census Bureau tabulated the person's location, as well as

sex, age, race, and ethnicity, and the person's relationship to the head of the household—that is, six attributes per person, for a total of approximately 1.5 billion attributes. Using this data, the Census Bureau published approximately 7.7 billion linearly independent statistics, including 2.7 billion in the PL94-171 redistricting file, 2.8 billion in the balance of summary file 1, 2 billion in summary file 2, and 31 million records in a public-use microdata sample. This results in approximately 25 statistics per person. Given these numbers and the example in this article, it is clear that there is a theoretical possibility that the national-level census could be reconstructed, although tools such as Sugar and PicoSAT are probably not powerful enough to do so.

To protect the privacy of census respondents, the Census Bureau is developing a privacy-protection system based on differential privacy. This system will ensure that every statistic and the corresponding microdata receive some amount of privacy protection, while providing that the resulting statistics are sufficiently accurate for their intended purpose.

This article has explained the motivation for the decision to use differential privacy. Without a privacy-protection system based on noise injection, it would be possible to reconstruct accurate microdata using only the published statistics. By using differential privacy, we can add the minimum amount of noise necessary to achieve the Census Bureau's privacy requirements. A future article will explain how that system works.

RELATED WORK

In 2003 Irit Dinur and Kobbi Nissim[4] showed that the amount of noise that needs to be added to a database to prevent a reconstruction of the underlying data is on the order of $\Omega\ (\sqrt{n})$ where n is the number of bits in the database. In practice, many statistical agencies do not add this much noise when they release statistical tables. (In our example, each record contains 11 bits of data, so the confidential database has 77 bits of information. Each statistic in Table 3 can be modeled as a four-bit of count, a seven-bit of median, and a seven-bit of mean, for a total of 18 bits; Table 3 releases 126 bits of information.) Dinur and Nissim's primary finding is that many statistical agencies leave themselves open to the risk of database reconstruction. This article demonstrates one way to conduct that attack.

Statistical tables create the possibility of

SAT and SAT Solvers

The Boolean SAT problem was the first to be proven NP-complete.[9] This problem asks, for a given Boolean formula, whether replacing each variable with either true or false can make the formula evaluate to true. Modern SAT solvers work well and reasonably quickly in a variety of SAT problem instances and up to reasonably large instance sizes.

Many modern SAT solvers use a heuristic technique called CDCL (conflict-driven clause learning).[10] Briefly, a CDCL algorithm:

1. Assigns a value to a variable arbitrarily.

2. Uses this assignment to determine values for the other variables in the formula (a process known as unit propagation).

3. If a conflict is found, backtracks to the clause that made the conflict occur and undoes variable assignments made after that point.

4. Adds the negation of the conflict-causing clause as a new clause to the master formula and resumes from step 1.

This process is fast at solving SAT problems because adding conflicts as new clauses has

the potential to avoid wasteful "repeated backtracks." Additionally, CDCL and its predecessor algorithm, DPLL (Davis–Putnam–Logemann–Loveland), are both provably complete algorithms: they will always return either a solution or "Unsatisfiable" if given enough time and memory. Another advantage is that CDCL solvers reuse past work when producing the universe of all possible solutions.

A wide variety of SAT solvers are available to the public for minimal or no cost. Although a SAT solver requires the user to translate the problem into Boolean formulae before use, programs such as Naoyuki Tamura's Sugar facilitate this process by translating user-input mathematical and English constraints into Boolean formulae automatically.

database reconstruction because they form a set of constraints for which there is ultimately only one exact solution when the published table is correctly tabulated from a real confidential database. Restricting the number or specific types of queries—for example, by suppressing results from a small number of respondents—is often insufficient to prevent access to indirectly identifying information, because the system's refusal to answer a "dangerous" query itself provides the attacker with information.

CONCLUSION

With the dramatic improvement in both computer speeds and the efficiency of SAT and other NP-hard solvers in the last decade, DRAs on statistical databases are no longer just a theoretical danger. The vast quantity of data products published by statistical agencies each year may give a determined attacker more than enough information to reconstruct some or all of a target database and breach the privacy of millions of people. Traditional disclosure-avoidance techniques are not designed to protect against this kind of attack.

Sugar Input

Sugar input is given in a standard CSP (constraint satisfaction problem) file format. A constraint must be given on a single line of the file, but here we separate most constraints into multiple lines for readability. Constraint equations are separated by comments describing the statistics they encode.

Input for the model in this article is available online at https://queue.acm.org/appendices/Garfinkel_SugarInput.txt.

Faced with the threat of database reconstruction, statistical agencies have two choices: they can either publish dramatically less information or use some kind of noise injection. Agencies can use differential privacy to determine the minimum amount of noise necessary to add, and the most efficient way to add that noise, in order to achieve their privacy protection goals.

Acknowledgments

References

1. Adam, N.R., Worthmann, J.C. 1989. Security-control methods for statistical databases: A comparative study. *ACM Computing Surveys* 21(4), 515–556); http://doi.acm.org/10.1145/76894.76895.
2. Biere, A. 2008. PicoSAT essentials. *Journal on Satisfiability, Boolean Modeling and Computation* 4, 75-97; https://pdfs.semanticscholar.org/7ea4cdd0003234f9e98ff5a080d9191c398e26c2.pdf.
3. Blum, A., Dwork, C., McSherry, F., Nissim, K. 2005.

Practical privacy: the SuLQ framework. In *Proceedings of the 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 128–138; https://dl.acm.org/citation.cfm?id=1065184.

4. Dinur, I., Nissim, K. 2003. Revealing information while preserving privacy. In *Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Principles of Database Systems*, 202–210; https://dl.acm.org/citation.cfm?id=773173.

5. Dwork, C., McSherry, F., Nissim, K., Smith, A. 2006. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, 265–284. Berlin, Heidelberg: Springer-Verlag; http://dx.doi.org/10.1007/11681878_14.

6. Dwork, C., Nissim, K. 2004. Privacy-preserving datamining on vertically partitioned databases. In *Proceedings of the 24th International Cryptology Conference* 3152, 528–544. Santa Barbara, California: Springer Verlag; https://www.microsoft.com/en-us/research/publication/privacy-preserving-datamining-on-vertically-partitioned-databases/.

7. Heule, M.J.H., Kullmann, O. 2017. The science of brute force. *Communications of the ACM* 60(8), 70–79; http://doi.acm.org/10.1145/3107239.

8. Kleinberg, J., Papadimitriou, C., Raghavan, P. 2000. Auditing Boolean attributes. In *Proceedings of the 19th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 86–91; http://doi.acm.org/10.1145/335168.335210.

9. Kong, S., Malec, D. 2007. Cook-Levin theorem. Lecture, University of Wisconsin.

10. Marques-Silva, J., Lynce, I., Malik, S. 2009. Conflict-driven clause learning SAT solvers. In *Handbook of Satisfiability*, 131–153. Amsterdam, The Netherlands: IOS Press.

11. McCarthy, J. 1960. Recursive functions of symbolic expressions and their computation by machine, part I. *Communications of the ACM* 3(4), 184–195; https://dl.acm.org/citation.cfm?id=367199.

12. Metin, H., Baarir, S., Colange, M., Kordon, F. 2018. CDCLSym: Introducing effective symmetry breaking in SAT solving. In *Tools and Algorithms for the Construction and Analysis of Systems*, ed. D. Beyer and M. Huisman, 99–114. Springer International Publishing; https://link.springer.com/chapter/10.1007/978-3-319-89960-2_6.

13. Tamura, N., Taga, A., Kitagawa, S., Banbara, M. 2009. Compiling finite linear CSP into SAT. *Contraints* 14(2), 254–272; https://dl.acm.org/citation.cfm?id=1527316; http://bach.istc.kobe-u.ac.jp/sugar/.

14. Whitney, C.R. 1997. Jeanne Calment, world's elder, dies at 122. *New York Times* (August 5); https://nyti.ms/2kM4oFb.

15. Zayatz, L., Lucero, J., Massell, P., Ramanayake, A. 2009. Disclosure avoidance for Census 2010 and American Community Survey five-year tabular data products. Statistical Research Division, U.S. Census Bureau; https://www.census.gov/srd/CDAR/rrs2009-10_ACS_5yr.pdf.

Simson L. Garfinkel *is the Senior Computer Scientist for Confidentiality and Data Access at the U.S. Census Bureau*

and the Chair of the Census Bureau's Disclosure Review Board.

John M. Abowd *is the Chief Scientist and Associate Director for Research and Methodology at the U.S. Census Bureau, where he serves on leave from his position as the Edmund Ezra Day Professor of Economics, professor of information science, and member of the Department of Statistical Sciences at Cornell University.*

Christian Martindale *is a senior at Duke University. He intends to pursue a career in management consulting or law.*

Related articles

➡️ Go Static or Go Home
In the end, dynamic systems are simply less secure.
Paul Vixie
https://queue.acm.org/detail.cfm?ref=rss&id=2721993

➡️ Privacy, Anonymity, and Big Data in the Social Sciences
Quality social science research and the privacy of human subjects requires trust.
Jon P. Daries, et al.
https://queue.acm.org/detail.cfm?id=2661641

➡️ Research for Practice:
Private Online Communication;
Highlights in Systems Verification
Expert-curated Guides to the
Best of CS Research
Albert Kwon, James Wilcox
https://queue.acm.org/detail.cfm?id=3149411

WILLIAM DUKE/PHOTO BY DAXIAO PRODUCTIONS/SHUTTERSTOCK

Can a set of equations keep U.S. census data private?

By **Jeffrey Mervis** | Jan. 4, 2019 , 2:50 PM

The U.S. Census Bureau is making waves among social scientists with what it calls a "sea change" in how it plans to safeguard the confidentiality of data it releases from the decennial census.

The agency announced in September 2018 that it will apply a mathematical concept called differential privacy to its release of 2020 census data after conducting experiments that suggest current approaches can't assure confidentiality. But critics of the new policy believe the Census Bureau is moving too quickly to fix a system that isn't broken. They also fear the changes will degrade the quality of the information used by thousands of researchers, businesses, and government agencies.

The move has implications that extend far beyond the research community. Proponents of differential privacy say a fierce, ongoing legal battle over plans to add a citizenship question to the 2020 census has only underscored the need to assure people that the government will protect their privacy.

A noisy conflict

The Census Bureau's job is to collect, analyze, and disseminate useful information about the U.S. population. And there's a lot of it: The agency generated some 7.8 billion statistics about the 308 million people counted in the 2010 census, for example.

At the same time, the bureau is prohibited by law from releasing any information for which "the data furnished by any particular establishment or individual … can be identified."

Once upon a time, meeting that requirement meant simply removing the names and addresses of respondents. Over the past several decades, however, census officials have developed a bag of statistical tricks aimed at providing additional protection without undermining the quality of the data.

Got a tip?   How to contact the news team

Related Jobs

Scientist, Surgical Technician
Pfizer
Groton, Connecticut

Senior Scientist, B Cell Repertoires and Antibody Optimization
Pfizer
Cambridge, Massachusetts

Scientist (non-PhD), Immunosuppression
Pfizer
La Jolla, California

MORE JOBS ▶

Latest News

Trending | Most Read

1. **France grossly underestimated radioactive fallout from atom bomb tests, study finds**

2. **Martian rover sends back 'overwhelming' video, audio from the Red Planet**

3. **How big is the average penis?**

4. **Abortion opponents protest COVID-19 vaccines' use of fetal cells**

5. **What is research misconduct? European countries can't agree**

Sifter

Scientists use 'x-ray vision' to read a letter sealed in 1697

BY SOFIA MOUTINHO | MAR. 2, 2021

Such perturbations, also known as injecting noise, are meant to foil attempts to reidentify individuals by combining census data with other publicly available information, such as credit reports, voter registration rolls, and property records. But preventing reidentification has grown more challenging with the advent of ever-more-powerful computational tools capable of stripping away privacy.

Census officials now believe those ad hoc methods are no longer good enough to satisfy the law. "The problem is real, and it has moved from a concern to an issue," says John Thompson, who stepped down as census director in June 2017, and who recently retired as head of the Council of Professional Associations on Federal Statistics in Arlington, Virginia. "In Census Bureau lingo, that means it's no longer simply a risk, but rather something you have to deal with."

The agency's decision to adopt differential privacy was spurred, in part, by recent work on what is known as the "database reconstruction theorem." The theorem shows that, given access to a sufficiently large amount of information, someone can reconstruct underlying databases and, in theory, identify individuals.

"Database reconstruction theorem is the death knell for traditional [data] publication systems from confidential sources," says John Abowd, chief scientist and associate director for research at the Census Bureau, located in Suitland, Maryland. "It exposes a vulnerability that we were not designing our systems to address," says Abowd, who has spearheaded the agency's efforts to adopt differential privacy.

But some users of census data strongly disagree. Steven Ruggles, a population historian at the University of Minnesota in Minneapolis, is leading the charge against the new policy.

Ruggles says traditional methods have successfully prevented any identity disclosures and, thus, there's no urgency to do more. If the Census Bureau is hell-bent on imposing differential privacy, he adds, officials should work with the community to iron out the kinks before applying it to the 2020 census and its smaller cousin, the American Community Survey.

"Differential privacy goes above and beyond what is necessary to keep data safe under census law and precedent," says Ruggles, who also manages a university-based social research institute that disseminates census data. "This is not the time to impose arbitrary and burdensome new rules that will sharply restrict or eliminate access to the nation's core data sources."

"My central concern about differential privacy is that it's a blunt instrument," he adds. "If you want to provide the same level of protection against reidentification that current methods do, you're going to have to do a lot more damage to the data than is done now."

Ways to protect confidentiality

Protecting confidentiality has been a priority for the Census Bureau for most—but not all—of its existence. After the first U.S. census was conducted in 1790, officials posted the results so that residents could correct errors. But in 1850, the interior secretary decreed that the returns would be kept confidential. They were "not to be used in any way to the gratification of curiosity and census officials," or "the exposure of any man's business or pursuits," notes an official history of the census published in 1900. In 1954 the agency's confidentiality mandate was codified in Title 13 of the U.S. Code.

Publicly available census data come in two flavors. One type, called small-area data, provides the basic characteristics of residents—age, sex, and race/ethnicity—down to the census block level. A census block, often the size of a city block, is the smallest geographic area for which data are reported. There were some 11 million blocks in 2010, of which 6.3 million were inhabited.

The second is called microdata, which are the full records collected by the Census Bureau on individuals—including, for example, the size of the household and the relationships between the residents. When microdata are reported, they are lumped together by areas containing at least

100,000 people.

Together, these census products provide fodder for thousands of researchers. Census data are also the basis for surveys by other government agencies and the private sector that shape decisions ranging from locating new factories or shopping malls to building new roads and schools.

The Census Bureau has used a variety of methods to preserve the confidentiality of these data as it moved from print to magnetic tape to digital distribution. Officials can, for instance, mask the responses of outliers—such as the income of a billionaire. They can also be less precise, for example, by reporting ages within 5-year ranges rather than a single year. Another technique involves swapping information with a respondent possessing many similar characteristics who lives in a different block.

How much noise to inject depends on many factors. However, census officials have never disclosed details of their formula or said how often a particular method is used. They fear that such information could help someone to reverse engineer the process.

A mathematical approach

Differential privacy, first described in 2006, isn't a substitute for swapping and other ways to perturb the data. Rather, it allows someone—in this case, the Census Bureau—to measure the likelihood that enough information will "leak" from a public data set to open the door to reconstruction.

"Any time you release a statistic, you're leaking something," explains Jerry Reiter, a professor of statistics at Duke University in Durham, North Carolina, who has worked on differential privacy as a consultant with the Census Bureau. "The only way to absolutely ensure confidentiality is to release no data. So the question is, how much risk is OK? Differential privacy allows you to put a boundary" on that risk.

A database can be considered differentially protected if the information it yields about someone doesn't depend on whether that person is part of the database. Differential privacy was originally designed to apply to situations in which outsiders make a series of queries to extract information from a database. In that scenario, each query consumes a little bit of what the experts call a "privacy budget." After that budget is exhausted, queries are halted in order to prevent database reconstruction.

In the case of census data, however, the agency has already decided what information it will release, and the number of queries is unlimited. So its challenge is to calculate how much the data must be perturbed to prevent reconstruction.

Abowd says the privacy budget "can be set at wherever the agency thinks is appropriate." A low budget increases privacy with a corresponding loss of accuracy, whereas a high budget reveals more information with less protection. The mathematical parameter is called epsilon; Reiter likens setting epsilon to "turning a knob." And epsilon can be fine-tuned: Data deemed especially sensitive can receive more protection.

The epsilon can be made public, along with the supporting equations on how it was calculated. In contrast, Abowd says, traditional approaches to limiting disclosure are "fundamentally dishonest" from a scientific perspective because of their underlying uncertainty. "At the moment," he says, the public doesn't "know the global disclosure risk. … That's because the agency doesn't tell you everything it did to the data before releasing it."

A simulated attack

A professor of labor economics at Cornell University, Abowd first learned that traditional procedures to limit disclosure were vulnerable—and that algorithms existed to quantify the risk— at a 2005 conference on privacy attended mainly by cryptographers and computer scientists. "We were speaking different languages, and there was no Rosetta Stone," he says.

He took on the challenge of finding common ground. In 2008, building on a long relationship with the Census Bureau, he and a team at Cornell created the first application of differential privacy to a census product. It is a web-based tool, called OnTheMap, that shows where people work and live.

Abowd took leave from Cornell to join the Census Bureau in June 2016, and one of his first moves was to test the vulnerability of the 2010 census data to an outside attack. The goal was to see how well a census team could reconstruct individual records from the thousands of tables the agency had published—and then try to identify those individuals.

The three-step process required substantial computing power. First, the researchers reconstructed records for individuals—say, a 55-year-old Hispanic woman—by mining the aggregated census tables. Then, they tried to match the reconstructed individuals to even more detailed census block records (that still lacked names or addresses); they found "putative matches" about half the time.

Finally, they compared the putative matches to commercially available credit databases in hopes of attaching a name to a particular record. Even if they could, however, the team didn't know whether they had actually found the right person.

Abowd won't say what proportion of the putative matches appeared to be correct. (He says a forthcoming paper will contain the ratio, which he calls "the amount of uncertainty an attacker would have once they claim to have reidentified a person from the public data.") Although one of Abowd's recent papers notes that "the risk of re-identification is small," he believes the experiment proved reidentification "can be done." And that, he says, "is a strong motivation for moving to differential privacy."

Too far, too fast?

Such arguments haven't convinced Ruggles and other social scientists opposed to applying differential privacy on the 2020 census. They are circulating manuscripts that question the significance of the census reconstruction exercise and that call on the agency to delay and change its plan.

Last month they had their first public opportunity to express their opposition during a meeting at census headquarters of the Federal Economic Statistics Advisory Committee (FESAC), which advises the Census Bureau and two other major federal statistical agencies. Abowd and Ruggles went toe to toe during a panel discussion on differential privacy, and council members had a chance to quiz them.

One point of disagreement is the interpretation of federal law. Title 13 requires the agency to mask only the identity of individuals, critics argue, not their characteristics. If identifying characteristics is illegal, Ruggles writes in a recent paper, then "virtually all Census Bureau microdata and small-area products currently fail to meet that standard."

Abowd reads the law differently. "Steve has gotten it wrong," he says flatly. "The statute says that what is prohibited is releasing the data in an identifiable way."

At the meeting, several members of the advisory committee peppered Abowd with questions about the significance of being able to reconstruct 50% of microdata files. That percentage is rather low, they argue. In any event, they say, reconstruction is a far cry from reidentification, which is what the law prohibits. They also wondered why anyone would go to the trouble of messing with census data when there are other, better ways to obtain scads of personal information that can be used to identify individuals.

"I'm not surprised that someone has reconstructed the fact that there are 45-year-old white men living in a particular block," said Colm O'Muircheartaigh, a professor of public policy at the University of Chicago in Illinois and a member of FESAC. "But that kind of information is neither very interesting or useful."

Identifying individuals based on household data might be more valuable, he said. "But I imagine

it would be much harder to reconstruct a household," O'Muircheartaigh said. "And even if we could, reconstructing a typical American household—say, two adults and two children—would hardly be a killer identification."

Census data also don't age well because of high mobility rates, he added. "These are static data," he said. "Even if you knew that such and such a person lived somewhere in 2010, how valuable would that be in 2014 or 2018?"

Some meeting attendees also accused Abowd of failing to address the practical effects of applying differential privacy. One skeptic was Kirk Wolter, chief statistician for NORC at the University of Chicago, a research institution that does survey work for many federal agencies. He argued that noisier census data would have a major ripple effect, degrading the quality of many other surveys that rely on census data to select their samples. "These surveys provide the information infrastructure for the country," he noted. "And all of them would suffer."

Correcting for those problems will cost money, he predicted, with organizations like NORC having to adjust samples and redesign surveys. And given the tight budgets of most survey research organizations, those could translate into fewer studies—and less information about the country's residents.

Thompson agrees. "Kirk is exactly right," he says. Applying differential privacy means "those surveys will take longer and cost more. And they may be less accurate. But you don't have a choice."

The citizenship elephant

Proponents of adopting differential privacy say there is also another compelling reason to move forward quickly: a controversial decision made last March by Commerce Secretary Wilbur Ross to **add a citizenship question** to the 2020 census.

A slew of local and state officials have joined civil rights groups in suing the federal government in a bid to block the question. They argue that adding the question will lead nonresidents and other vulnerable populations to avoid filling out the census form, **leading to a significant undercount**. And they are worried about privacy, too. Knowing how someone answered the citizenship question, critics say, would allow a government agency to take punitive action against nonresidents.

"Maybe a researcher wouldn't try to do that," says Thompson, a witness for the plaintiffs in one of the suits. "But there are a lot of people who might. And I think that [federal immigration officials] would love to have that information."

Abowd knows the extreme sensitivity of the citizenship question. His emails last year to Ross expressing reservations about adding it to the 2020 census have been publicly revealed by the litigation. And although he tiptoed around the topic during the recent FESAC discussion, it was clear that he was worried about the damage it could wreak on the agency's credibility.

"The entire history of traditional disclosure limitation was aimed at preventing attackers, armed with external data, from using it in combination with the variables on the [census] microdata file to attach a name and address," Abowd said during the roundtable. "With regard to 2010, most of those databases did not have race and ethnicity on them. And none have citizenship, to just bring into the room the variable that we probably should be discussing more explicitly."

Practical issues

Ruggles, meanwhile, has spent a lot of time thinking about the kinds of problems differential privacy might create. His Minnesota institute, for instance, disseminates data from the Census Bureau and 105 other national statistical agencies to 176,000 users. And he fears differential privacy will put a serious crimp in that flow of information.

In the most extreme scenario, he says, the Census Bureau could decide to make 2020 census data available only through its network of 29 secure Federal Statistical Research Data Centers.

That would impose serious hardships on users, Ruggles says, because the centers require users to obtain a security clearance, which often involves lengthy waiting periods. Such rules could also prevent most international scholars from using the centers, he says, as well as graduate students seeking a quick turnaround for a dissertation. In addition, researchers are only cleared if their project is deemed to benefit the agency's mission.

There are also questions of capacity and accessibility. The centers require users to do all their work onsite, so researchers would have to travel, and the centers offer fewer than 300 workstations in total.

Thompson says the Census Bureau needs to address those issues regardless of whether it adopts differential privacy. He agrees with Ruggles that it takes too long to gain access to the research centers, and he thinks the bureau needs to change its definition of what research serves its mission. "I have argued that anyone advancing the science of using data" should be eligible, he says. "We need a 21st-century Census Bureau, and that will take a lot of fixing."

(With regard to access, Abowd says the agency is considering setting up "virtual" centers that would allow a much broader audience to work with the data. But Ruggles is skeptical that such a system would satisfy the bureau's own definition of confidentiality.)

A need to communicate

Abowd has said, "The deployment of differential privacy within the Census Bureau marks a sea change for the way that official statistics are produced and published." And Ruggles agrees. But he says the agency hasn't done enough to equip researchers with the maps and tools needed to navigate the uncharted waters.
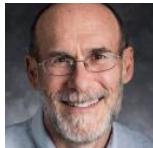
 "It's pretty clear we are going to have a new methodology," Ruggles concedes. "But I think it could be implemented in a better or worse way. I would like them to consider the trade-offs, and not take such an absolutist stand on the risks."

Meanwhile, NORC's Wolter says regardless of whether his concerns are addressed, the bureau must do more outreach—and not just in peer-reviewed journals. "Census badly needs a communications strategy, by real communications specialists," he said. "There are thousands of users [of census data] who won't understand any of this stuff. And they need to know what is going to happen."

Clarification, *17 January 2019, 5:00 p.m.: The first quote from John Abowd in the story has been revised to make it clear that the Census Bureau is now addressing the vulnerability of census data to reidentification.*

Posted in: **Science and Policy**, **Scientific Community**
doi:10.1126/science.aaw5470

Jeffrey Mervis

Jeff tries to explain how government works to readers of *Science*.
✉ **Email Jeffrey**

More from News

DataSynthesizer: Privacy-Preserving Synthetic Datasets

Haoyue Ping
Drexel University, USA
hp354@drexel.edu

Julia Stoyanovich*
Drexel University, USA
stoyanovich@drexel.edu

Bill Howe†
University of Washington, USA
billhowe@cs.washington.edu

ABSTRACT

To facilitate collaboration over sensitive data, we present DataSynthesizer, a tool that takes a sensitive dataset as input and generates a structurally and statistically similar synthetic dataset with strong privacy guarantees. The data owners need not release their data, while potential collaborators can begin developing models and methods with some confidence that their results will work similarly on the real dataset. The distinguishing feature of DataSynthesizer is its usability — the data owner does not have to specify any parameters to start generating and sharing data safely and effectively.

DataSynthesizer consists of three high-level modules — DataDescriber, DataGenerator and ModelInspector. The first, DataDescriber, investigates the data types, correlations and distributions of the attributes in the private dataset, and produces a data summary, adding noise to the distributions to preserve privacy. DataGenerator samples from the summary computed by DataDescriber and outputs synthetic data. ModelInspector shows an intuitive description of the data summary that was computed by DataDescriber, allowing the data owner to evaluate the accuracy of the summarization process and adjust any parameters, if desired.

We describe DataSynthesizer and illustrate its use in an urban science context, where sharing sensitive, legally encumbered data between agencies and with outside collaborators is reported as the primary obstacle to data-driven governance.

The code implementing all parts of this work is publicly available at https://github.com/DataResponsibly/DataSynthesizer.

CCS CONCEPTS

• **Security and privacy → Data anonymization and sanitization**; **Privacy protections**; **Usability in security and privacy**;

KEYWORDS

Data Sharing; Synthetic Data; Differential Privacy

1 INTRODUCTION

Collaborative projects in the social and health sciences increasingly require sharing sensitive, privacy-encumbered data. Social scientists, government agencies, health workers, and non-profits are eager to collaborate with data scientists, but formal data sharing agreements are too slow and expensive to create in ad hoc situations — our colleagues report that 18 months is a typical timeframe to establish such agreements! As a result, many promising collaborations can fail before they even begin. Data scientists require access to the data before they can understand the problem or even determine whether they can help. But data owners cannot share data without significant legal protections in place. Beyond legal concerns, there is a general reluctance to share sensitive data with non-experts before they have "proven themselves," since they do not understand the context in which the data was collected and may be distracted by spurious results.

To bootstrap these collaborations without incurring the cost of formal data sharing agreements, we saw a need to generate datasets that are *structurally and statistically similar* to the real data but that are 1) obviously synthetic to put the data owners at ease, and 2) offer strong privacy guarantees to prevent adversaries from extracting any sensitive information. These two requirements are not redundant: strong privacy guarantees are not always sufficient to convince data owners to release data, and even seemingly random datasets may not prevent subtle privacy attacks. With this approach, data scientists can begin to develop models and methods with synthetic data, but maintain some degree of confidence that their work will remain relevant when applied to the real data once proper data sharing agreements are in place.

We propose a tool named DataSynthesizer to address this problem. Assume that the private dataset contains one table with m attributes and n tuples, and that the values in each attribute are homogeneous, that is, they are all of the same data type. We are interested in producing a synthetic dataset such that summary statistics of all numerical, categorical, string, and datetime attributes are similar to the private dataset. What statistics we preserve depends on the data type, as we will discuss in Section 3.

DataSynthesizer infers the domain of each attribute and derives a description of the distribution of attribute values in the private dataset. This information is saved in a dataset description file, to which we refer as data summary. Then DataSynthesizer is able to generate synthetic datasets of arbitrary size by sampling from the probabilistic model in the dataset description file.

DataSynthesizer can operate in one of three modes: In *correlated attribute mode*, we learn a differentially private Bayesian network capturing the correlation structure between attributes, then draw samples from this model to construct the result dataset. In cases where the correlated attribute mode is too computationally expensive or when there is insufficient data to derive a reasonable model, one can use *independent attribute mode*. In this mode, a histogram is derived for each attribute, noise is added to the histogram to achieve differential privacy, and then samples are drawn for each attribute. Finally, for cases of extremely sensitive data, one can use *random* mode that simply generates type-consistent random values for each attribute.

DataSynthesizer uses principled methods to infer attribute data types, learn statistical properties of the attributes, and protect privacy. While the specific techniques we employ are standard and have been used elsewhere (see Section 4 for a description of relevant work), the primary contribution of our system is in its usability. The system supports three intuitive modes of operation, and requires minimal input from the user.

2 DEMONSTRATION SCENARIO

We now briefly describe the main components of the DataSynthesizer system, and then explain the demonstration scenarios. The high-level architecture of the system is presented in Figure 1. The system has three modules — DataDescriber, DataGenerator and ModelInspector. Each component of the system will be explained in detail in Section 3. A data owner interacts with DataSynthesizer through a Web-based UI that serves as a wrapper for 1) invoking the DataDescriber library to compute a data summary, 2) generating a synthetic dataset from the summary using DataGenerator, and 3) inspecting and comparing datasets and data summaries with ModelInspector.

DataDescriber can be invoked on any CSV file. In particular, we can invoke it on both the input file and the output file, and compare the resulting summaries. Based on this comparison, we will convey an important point to the demonstration attendees: While the input and output datasets themselves are clearly very different, the statistical descriptions of the datasets are very similar. What sort of a comparison is drawn between the input and the output depends on the mode of operation of DataSynthesizer. When the tool in invoked in correlated attribute mode, a comparison of the learned Bayesian networks and of the pair-wise attribute correlations is shown to the user. In independent attribute mode, per-attribute histograms are presented, along with the ranges of attribute values.

During the demonstration, we will showcase the functionality of DataSynthesizer on a variety of datasets. We will prepare several interesting datasets for the interactive session, including an urban homelessness dataset from the University of Washington Data Science Institute (where the relevant tasks include making targeted service recommendations), the criminal sentencing dataset from a ProPublica investigation [1] (where the relevant tasks include analyzing the bias of recidivism models), and several datasets from the UCI Machine Learning Repository [6]. During the demonstration we will encourage users to download additional datasets from portals such as data.seattle.gov and data.ny.gov and to run the tool.
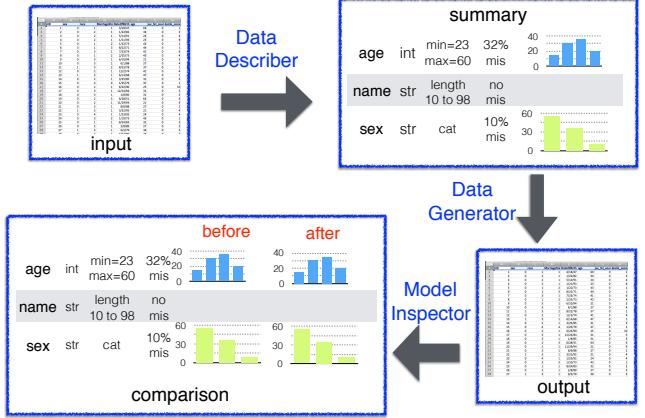

Figure 1: The DataSynthesizer system architecture.

Attendees will play the role of a data owner who is preparing a synthetic dataset for release to a set of new collaborators. After reviewing the raw dataset, they will generate a summary model and inspect the results, verifying that the results are sensible. Then, the attendees will generate a synthetic dataset from this model and inspect individual records, observing that the records are visibly "scrambled". Finally, they will generate a new model from the synthetic dataset and observe that the statistical properties are intact. At this point, we will explain the privacy guarantees that are enforced, and will illustrate them using the before-and-after visualizations.

3 SYSTEM OVERVIEW

DataSynthesizer is an end-to-end system that takes a private dataset as input and generates synthetic datasets. The system is implemented in Python 3. It assumes that the private dataset is presented in CSV format.

The input dataset is first processed by the DataDescriber module. The domains and the estimates of distributions of the attributes are inferred and saved in a dataset description file. For each attribute identified as categorical, DataDescriber computes the frequency distribution of values, represented as a bar chart. DataGenerator then samples from this distribution when deriving the synthetic dataset. For non-categorical numerical and datetime attributes, DataDescriber derives an equi-width histogram to represent the distribution. DataGenerator draws samples from this histogram during data generation using uniform sampling. For non-categorical string attributes, their minimum and maximum lengths are recorded. DataDescriber generates random strings within the length range during data generation stage.

We now describe each of the modules of DataSynthesizer in turn.

3.1 DataDescriber

3.1.1 Inferring data types and domains. The domain of an attribute is the set of its legal values. The data type is an important ingredient of the attribute domain. DataSynthesizer supports four data types. The system allows users to explicitly specify attribute data types. If an attribute data type is not specified by the user,

Table 1: Data types supported by DataSynthesizer

Data Type	Example
integer	id, age, ...
float	score, rating, ...
string	name, gender, ...
datetime	birthday, event time, ...

it is inferred by the DataDescriber. For each attribute, DataDescriber first detects whether it is numerical, and if so — whether it is an *integer* or a *float*. If the attribute is non-numerical, DataDescriber attempts to parse it as *datetime*. Any attribute that is neither numerical nor *datetime* is considered a *string*.

The data type of an attribute restricts its domain, which may be limited further if the attribute is categorical, where only specific values are legal. For example, in a dataset of students, the attribute *degree* may only take on values *BS*, *BA*, *MS*, or *PhD*. The domain of *degree* is {*BS, BA, MS, PhD*}. Note that *integer*, *float* and *datetime* attributes can also be categorical. In general, an attribute is considered to be categorical if a limited number of distinct values for that attribute is observed in the input dataset. DataDescriber has a parameter *categorical threshold* that defaults to 10 and represents the maximum number of distinct values for a categorical attribute.

The *categorical threshold*, like any threshold, may be challenging, or even impossible, to set in a way that reflects user preferences for all attributes in the input. Some attributes may appear categorical, but the user may prefer to treat them as numerical instead. For example, age of elementary school children may take on only a handful of distinct values but the user may nonetheless wish to generate data for this attribute from a continuous range. The opposite situation may also arise — an attribute may take on 200 or so distinct values, as is the case with country names, and so would not be considered categorical under any reasonable threshold. Still, the user may prefer to treat this attribute as categorical, so that a valid country name, rather than a random string, is generated for this attribute in the synthesized dataset. For this reason, DataSynthesizer allows users to specify a data type, and state whether an attribute is categorical, over-riding defaults on a per-attribute basis.

Note that the actual datatype of a categorical attribute is immaterial in terms of the statistical properties and the privacy guarantees in synthetic data generation. For example, an attribute such as *sex* may be encoded as M/F, as 0/1 or using a Boolean flag (e.g., True for male and False for female). In all cases, the tool will compute the frequencies of each attribute value in the input, and will subsequently sample values from the resulting data summary.

There might be missing values in the input dataset, and these are important to represent in the summary. DataDescriber calculates the missing rate for each attribute — the number of observed missing values divided by n, the size of the dataset.

3.1.2 Differential privacy. Differential privacy is a family of techniques that guarantee that the output of an algorithm is statistically indistinguishable on a pair of *neighboring* databases; that is, a pair of databases that differ by only one tuple. This concept is formalized with the following definition.

Algorithm 1 GreedyBayes(D, A, k)

Require: Dataset D, set of attributes A, maximum number of parents k
1: Initialize $\mathcal{N} = \emptyset$ and $V = \emptyset$.
2: Randomly select an attribute X_1 from A.
3: Add $(X_1, \emptyset)$ to $\mathcal{N}$; add X_1 to V.
4: **for** $i = 2, .., |A|$ **do**
5:     Initialize $\Omega = \emptyset$
6:     $p = min(k, |V|)$
7:     **for** each $X \in A \backslash V$ and each $\Pi \in \binom{V}{p}$ **do**
8:         Add (X, Π) to Ω
9:     **end for**
10:    Compute mutual information based on D for all pairs in Ω.
11:    Select (X_i, Π_i) from Ω with maximal mutual information.
12:    Add (X_i, Π_i) to $\mathcal{N}$.
13: **end for**
14: **return** $\mathcal{N}$

(ϵ-Differential Privacy [3]) Let ϵ be a positive number. Let $\mathcal{A}$ be a randomized algorithm taking a dataset as input. Let D be a dataset. For any D' that differs from D on at most one tuple, for any legal output O of $\mathcal{A}$, $Pr(\mathcal{A}(D) = O) \leq e^{\epsilon} \times Pr(\mathcal{A}(D') = O)$.

When ϵ approaches 0, $Pr(\mathcal{A}(D) = O) = Pr(\mathcal{A}(D') = O)$. That is, the presence or absence of a single individual in the input to the algorithm will be undetectable when one looks at the output.

3.1.3 Independent attribute mode. When invoked in independent attribute mode, DataDescriber performs frequency-based estimation of the unconditioned probability distributions of the attributes. The distribution is captured by bar charts for categorical attributes, and by histograms for numerical attributes. Precision of the histograms can be refined by a parameter named *histogram size*, which represents the number of bins and is set to 20 by default.

DataDescriber implements a differentially private mechanism, adding controlled noise into the learned distributions. The noise is from a Laplace distribution with location 0 and scale $\frac{1}{n\epsilon}$, where n is the size of the input, denoted $Lap(\frac{1}{n\epsilon})$, setting $\epsilon = 0.1$ by default. DataDescriber also adds Laplace noise to the missing rate. When Laplace noise is added to histogram frequencies, the value may become negative. In that case the value is reset to 0 [8].

3.1.4 Correlated attribute mode. Attribute values are often correlated, e.g., *age* and *income* of a person. When invoked in correlated attribute mode, DataDescriber uses the GreedyBayes algorithm to construct Bayesian networks (BN) to model correlated attributes [8].

Algorithm 1 constructs a BN $\mathcal{N}$ from input dataset D, attributes A, and the maximum number of BN node parents k, which defaults to 4. In this algorithm, V is the set of visited attributes, and Π is a subset of V that will become parents of node X if added to $\mathcal{N}$. Which attributes Π are selected as parents of X is determined greedily, by maximizing mutual information (X, Π). Algorithm 1 learns the structure of the BN with privacy guarantees when the mutual information computation is differentially private [8].

The Bayesian networks constructed in Algorithm 1 gives the sampling order for generating attribute values. The distribution from which a dependent attribute is sampled is called a conditioned
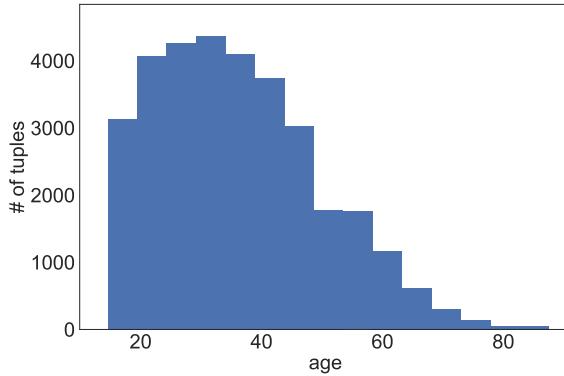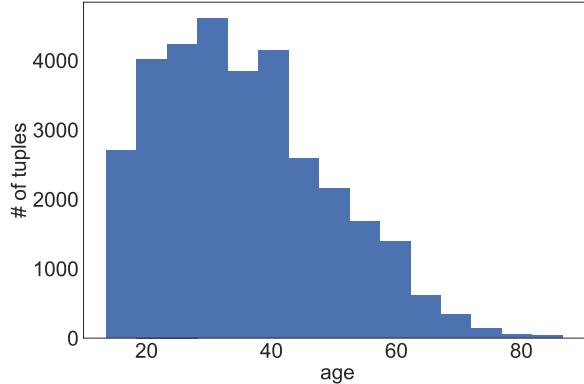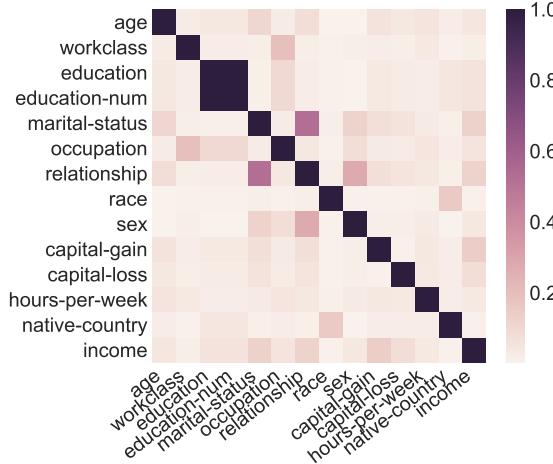
Figure 2: Histogram on *age*: Adult Income [6].

Figure 3: Histogram on *age*: synthetic.

Figure 4: Pair-wise correlations: Adult Income [6].
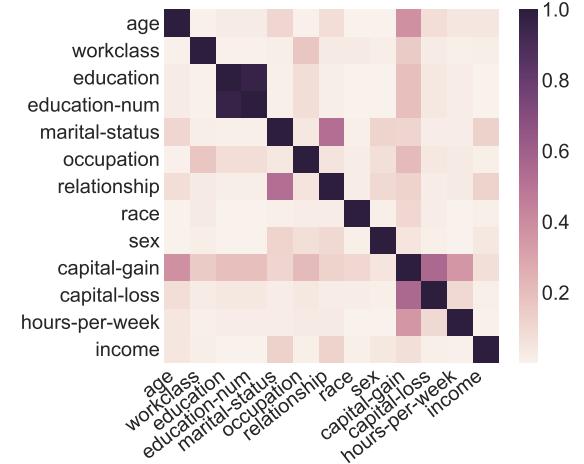
Figure 5: Pair-wise correlations: synthetic.
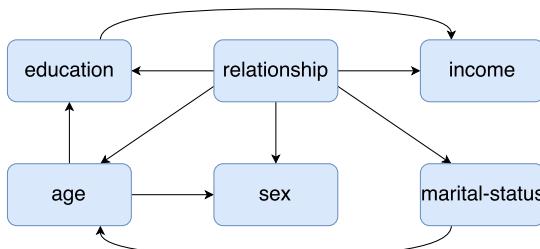
Figure 6: Bayesian network: Adult Income [6].
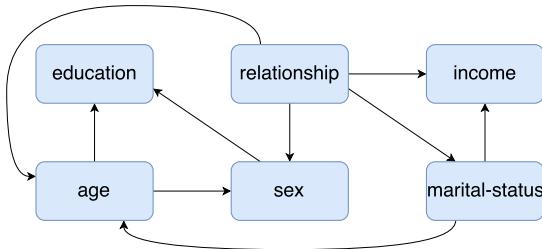
Figure 7: Bayesian network: synthetic.

distribution. When constructing noisy conditioned distributions, $Lap(\frac{4(d-k)}{n \cdot \epsilon})$ is injected to preserve privacy. Here, d is the number of attributes, k is the maximum number of parents of a BN node, and n is the number of tuples in the input dataset. We construct conditional distributions according to Algorithm 1 of [8].

The parents of a dependent attribute can be categorical or numerical, whose distributions are modeled by bar charts and histograms, respectively. The conditions for this dependent attribute are the legal values of the categorical parents and the intervals of the numerical parents. Here, the intervals are formed in the same way as the unconditioned distributions of the parent attributes. For example, the *age* attribute has intervals {[10, 20),[20, 30),[30, 40)} in its unconditioned distribution. Assume *education* only depends on *age*. Its conditioned distributions will be under the same intervals, i.e., age ∈ [10, 20), age ∈ [20, 30) and age ∈ [30, 40) respectively.

Algorithm 2 DataGenerator($n, \mathcal{M}, \mathcal{S}, A_U, s$)

Require: number of tuples n to generate, mode $\mathcal{M}$, dataset description $\mathcal{S}$, uniform attributes A_U, seed s
1: Set seed = s for pseudo-random number generator.
2: **if** $\mathcal{M}$ is independent attribute mode **then**
3:   Read all attributes A from $\mathcal{S}$.
4:   **for** $X \in A$ **do**
5:     **if** $X \in A_U$ **then**
6:       Read the domain of X from $\mathcal{S}$.
7:       Sample n values uniformly from its domain.
8:     **else**
9:       Read the distribution of X from $\mathcal{S}$.
10:      Sample n values from its distribution.
11:    **end if**
12:   **end for**
13: **else if** $\mathcal{M}$ is correlated attribute mode **then**
14:   Read Bayesian network $\mathcal{N}$ from $\mathcal{S}$.
15:   Sample root attribute from an unconditional distribution.
16:   Sample remaining attributes from conditional distributions.
17: **end if**
18: **return** Sampled dataset

3.2 DataGenerator

Given a dataset description file generated by DataDescriber, Data-Generator samples synthetic data from the distributions in this file. The size of the output dataset is specified by the user, and defaults to n, the size of the input dataset.

Algorithm 2 describes the data generation process. When invoked in random mode, DataGenerator generates type-consistent random values for each attribute. When invoked in independent attribute mode, DataGenerator draws samples from bar charts or histograms using uniform sampling. Finally, when invoked in correlated attribute mode, DataDescriber samples attribute values in appropriate order from the Bayesian network.

An important property of differential privacy is that its performance degrades with repeated queries. To prevent leaking private information through adversaries repeatedly sending data generation requests, the system administrator can assign a unique random seed for each person who requires a synthetic dataset. To support this, DataGenerator provides per-user seed functionality.

3.3 Model Inspector

ModelInspector provides several built-in functions to inspect the similarity between the private input dataset and the output synthetic dataset. The data owner can quickly test whether the tuples in the synthetic dataset are detectable by inspecting and comparing the first 5 and last 5 tuples in both datasets.

The synthetic dataset should have similar distributions of attribute values as the input dataset. In independent attribute mode, ModelInspector allows users to visually compare attribute distributions in the form of bar charts or histograms. Figures 2 and 3 present such summaries for the attribute *age* from the Adult Income dataset [6]. The system also computes the correlation coefficient

of an appropriate kind, depending on attribute type, and the KL-divergence value, which measure how much the "after" probability distribution diverges from the "before" probability distribution for a given attribute. In correlated attribute mode, ModelInspector presents "before" and "after" pairwise mutual information matrices (Figures 4 and 5) and describes Bayesian networks (such as those shown graphically in Figures 6 and 7), enabling an at-a-glance comparison of the statistical properties of the datasets.

4 RELATED WORK

In our work on DataSynthesizer we leverage recent advances in practical differential privacy [4] and privacy-preserving generation of synthetic datasets [7, 8]. In particular, we make use of the privacy-preserving learning of the structure and conditional probabilities of a Bayesian network in PrivBayes [8], and are inspired in our implementation by the work on DPBench [4].

Data sharing systems, including SQLShare [5] and DataHub [2], aim to facilitate collaborative data analysis, but do not incorporate privacy preserving features or purport to manage sensitive data. We see these systems efforts as a potential delivery vector for DataSynthesizer capabilities.

5 TAKE-AWAY MESSAGES

We have presented a demonstration of DataSynthesizer, a privacy-preserving synthetic data generator designed to facilitate collaborations between domain-expert data owners and external data scientists. The cost of establishing formal data sharing agreements limits the impact of these ad hoc collaborations in government, social sciences, health, or other areas where data is heavily encumbered by privacy rules. Given a dataset, DataSynthesizer can derive a structurally and statistically similar dataset, at a configurable level of statistical fidelity, while ensuring strong privacy guarantees. DataSynthesizer is designed with usability in mind. The system supports three intuitive modes of operation, and requires minimal input from the user.

We see DataSynthesizer being used in a variety of application contexts, both as a stand-alone library and as a component of more comprehensive data sharing platforms. As part of ongoing work, we are studying how best to deliver these features to data owners, and determining how additional requirements can be met. DataSynthesizer is open source, and is available for download at https://github.com/DataResponsibly/DataSynthesizer.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. *ProPublica* (23 May 2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
[2] Anant P. Bhardwaj and others. 2015. DataHub: Collaborative Data Science & Dataset Version Management at Scale. In *CIDR*.
[3] Cynthia Dwork and others. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *TCC*.
[4] Michael Hay and others. 2016. Principled Evaluation of Differentially Private Algorithms using DPBench. In *SIGMOD*.
[5] Shrainik Jain and others. 2016. SQLShare: Results from a Multi-Year SQL-as-a-Service Experiment. In *SIGMOD*. ACM, New York, NY, USA.
[6] M. Lichman. 2013. UCI Machine Learning Repository. (2013). http://archive.ics.uci.edu/ml
[7] Wentian Lu and others. 2014. Generating private synthetic databases for untrusted system evaluation. In *ICDE*.
[8] Jun Zhang and others. 2014. PrivBayes: private data release via Bayesian networks. In *SIGMOD*.

WINNING THE NIST CONTEST: A SCALABLE AND GENERAL APPROACH TO DIFFERENTIALLY PRIVATE SYNTHETIC DATA

RYAN MCKENNA, GEROME MIKLAU, AND DANIEL SHELDON

College of Information & Computer Sciences, The University of Massachusets, Amherst, MA 10002
e-mail address: rmckenna@cs.umass.edu

College of Information & Computer Sciences, The University of Massachusets, Amherst, MA 10002
e-mail address: miklau@cs.umass.edu

College of Information & Computer Sciences, The University of Massachusets, Amherst, MA 10002
e-mail address: sheldon@cs.umass.edu

ABSTRACT. We propose a general approach for differentially private synthetic data generation, that consists of three steps: (1) **select** a collection of low-dimensional marginals, (2) **measure** those marginals with a noise addition mechanism, and (3) **generate** synthetic data that preserves the measured marginals well. Central to this approach is `Private-PGM` [42], a post-processing method that is used to estimate a high-dimensional data distribution from noisy measurements of its marginals. We present two mechanisms, `NIST-MST` and `MST`, that are instances of this general approach. `NIST-MST` was the winning mechanism in the 2018 NIST differential privacy synthetic data competition, and `MST` is a new mechanism that can work in more general settings, while still performing comparably to `NIST-MST`. We believe our general approach should be of broad interest, and can be adopted in future mechanisms for synthetic data generation.

1. INTRODUCTION

Data sharing within the modern enterprise is extremely constrained by privacy concerns. Privacy-preserving synthetic data is an appealing solution: it allows existing analytics workflows and machine learning methods to be used while the original data remains protected. But recent research has shown that unless a formal privacy standard is adopted, synthetic data can violate privacy in subtle ways [18,25]. Differential privacy offers such a formalism, and the problem of differentially private synthetic data generation has therefore received considerable research attention in recent years [3, 6, 9, 13, 14, 26, 31, 32, 39, 40, 52–55, 59, 60, 66, 68, 70, 71].

In 2018, the National Institute of Standards and Technology (NIST) highlighted the importance of this problem by organizing the *Differential Privacy Synthetic Data Competition* [56]. This competition was the first of its kind for the privacy research community, and it encouraged privacy researchers and practitioners to develop novel practical mechanisms for this task. The competition consisted of three rounds of increasing complexity. In this paper we describe `NIST-MST`, the winning entry in the third and final round of the competition. Our algorithm is an instance of a general template for differentially private synthetic data generation that we believe will simplify design of future mechanisms for synthetic data.

Our approach to differentially private synthetic data generation consists of three high-level steps, as show in Figure 1: (1) query selection, (2) query measurement and (3) synthetic
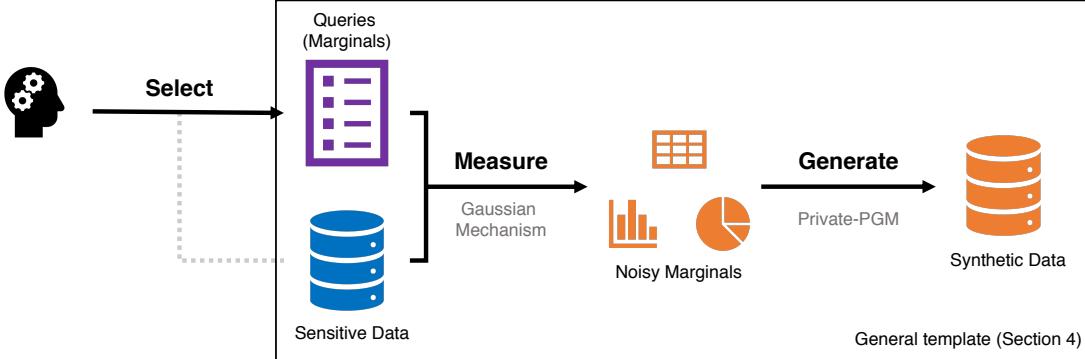
Figure 1. A general template for differentially private synthetic data generation. First, a collection of marginal queries is **selected**, either manually (e.g., by a domain expert) or automatically by an algorithm. Second, the Gaussian mechanism is used to **measure** those marginals while preserving differential privacy. Finally, `Private-PGM` is used to post-process the noisy marginals and **generate** a synthetic dataset that respects them.

data generation. For step (1), there are various ways to approach query selection; a domain expert familiar with the data and its use cases can specify the set of queries, or they can be automatically determined by an algorithm. The selected queries are important because they will ultimately determine the statistics for which the synthetic data preserves accuracy. For step (2), after the queries are fixed, they are measured privately with a noise-addition mechanism, in our case, with the Gaussian mechanism. In step (3), the noisy measurements are processed through `Private-PGM` [42], a post-processing method that can estimate a high-dimensional data distribution from noisy measurements and generate synthetic data.

This approach is similar in spirit to the widely studied select-measure-reconstruct paradigm for linear query answering under differential privacy [4, 16, 17, 29, 35–38, 38, 41, 46–48, 57, 58, 61–65, 67, 69]. However, the output is now synthetic data, rather than query answers. In addition, most existing methods from this paradigm suffer from the curse of dimensionality, and have trouble scaling to high-dimensional domains. Our approach is simple and modular but there are three main technical challenges to using it in practice. These are (1) identifying what statistics to measure about the dataset, (2) generating synthetic data that effectively preserves the measured statistics, and (3) overcoming the challenges of high-dimensional domains. Fortunately, `Private-PGM` solves problem (2) and (3) above, as long as the measured statistics only depend on the data through its low-dimensional marginals. This allows the mechanism designer to focus on problem (1), and frees them from the burden of figuring out *how* to generate synthetic data with differential privacy, allowing them instead to focus on *what* statistics to measure, based on what they want the synthetic data to preserve. Thus, we believe this approach to differentially privacy synthetic data, using `Private-PGM`, will be broadly applicable.

In Figure 1, there is a dashed gray line connecting the select step with the sensitive data. This indicates that query selection may or may not depend on the sensitive data, but if it does, it must be via a differentially private mechanism. The rules of the NIST competition permitted the use of a public provisional dataset which the `NIST-MST` algorithm uses for query selection. The effectiveness of query selection relies on the similarity of the public data to the private data being synthesized. Since high quality provisional data may not

always be available, we propose a variant of the algorithm, called `MST`, that does not require public data and instead uses a portion of the privacy budget to select measurements. The novelty of this algorithm is that it uses the data (privately) to select marginals to measure that support efficient synthesis in step (3). This extension leads to an algorithm that can be applied in a wider variety of settings. We show experimentally that, without the advantage of provisional data, it nevertheless performs comparably to `NIST-MST`.

This paper is organized as follows. In Section 2, we set up notation, state assumptions, and summarize relevant background in differential privacy. In Section 3, we summarize the important aspects of the competition. In Section 4, we describe the general template from Figure 1 in greater detail. In Section 5, we present `NIST-MST`, the winning mechanism from the competition, by building on the general template. In Section 6, we present `MST`, a novel mechanism inspired by `NIST-MST` that does not rely on the existence of public provisional data. We conclude with a simple experimental evaluation and discussion of results.

2. Background

The algorithms described in this paper take as input a dataset, assumed to be a single table, and generate a synthetic dataset satisfying (ϵ, δ)-differential privacy. Below we provide the relevant background and notation on datasets, marginals, and differential privacy.

2.1. **Data.** The input is a dataset D consisting of m records, each containing potentially sensitive information about one individual. Each record has d attributes $\mathcal{A} = \{A_1, \ldots, A_d\}$, and the domain of possible values for an attribute A_i is denoted by Ω_i. We assume Ω_i is finite and has size $|\Omega_i| = n_i$. The full domain of possible values is thus $\Omega = \Omega_1 \times \cdots \times \Omega_d$ which has size $\prod_i n_i = n$. We use $\mathcal{D}$ to denote the set of all possible datasets, which is equal to $\mathcal{D} = \cup_{m=0}^{\infty} \Omega^m$.

2.2. **Marginals.** A marginal is a key statistic that captures low-dimensional structure in a high-dimensional data distribution. We will explain (in Section 3) that the evaluation metrics of the contest can be defined in terms of marginals computed on the dataset. In addition, our algorithms will privately measure selected marginals and use the resulting noisy measurements to construct synthetic data. More precisely, a marginal, for a set of attributes C, is a table that counts the number of occurrences of each combination of possible values for attributes C.

Definition 1 (Marginal). *Let $C \subseteq \mathcal{A}$ be a subset of attributes, $\Omega_C = \prod_{i \in C} \Omega_i$, and $n_C = |\Omega_C|$. The marginal on C is a vector $\mu \in \mathbb{R}^{n_C}$, indexed by domain elements $t \in \Omega_C$, such that each entry is a count, i.e., $\mu_t = \sum_{x \in D} \mathbb{1}[x_C = t]$. We let $M_C : \mathcal{D} \to \mathbb{R}^{n_C}$ denote the function that computes the marginal on C, i.e., $\mu = M_C(D)$.*

2.3. **Differential privacy.** Differential privacy [20, 21] protects individuals by bounding the impact any one individual can have on the output of an algorithm. This is formalized using the notion of neighboring datasets. Two datasets $D, D' \in \mathcal{D}$ are neighbors (denoted $D \sim D'$) if D' can be obtained from D by adding or removing a single record.

Definition 2 (Differential Privacy [20]). *A randomized mechanism $\mathcal{M} : \mathcal{D} \rightarrow \mathcal{R}$ satisfies (ϵ, δ)-differential privacy (DP) if for any neighboring datasets $D \sim D' \in \mathcal{D}$, and any subset of possible outputs $S \subseteq \mathcal{R}$,*

$$\Pr[\mathcal{M}(D) \in S] \leq \exp(\epsilon) \Pr[\mathcal{M}(D') \in S] + \delta.$$

This definition requires that, on any two neighboring input databases, the difference in the output distributions of the randomized algorithm $\mathcal{M}$ is bounded by e^ϵ, except with a small failure probability δ. This failure probability δ is usually assumed to be cryptographically small; in the contest it was set to $\delta \approx 2 \cdot 10^{-12}$. The algorithms in this paper achieve differential privacy by repeated application of the Gaussian mechanism and the Exponential Mechanism, defined below:

Definition 3 (Gaussian Mechanism). *Let $f : \mathcal{D} \rightarrow \mathbb{R}^p$ be a vector-valued function of the input data. The Gaussian Mechanism adds i.i.d. Gaussian noise with scale σ to $f(D)$:*

$$\mathcal{M}(D) = f(D) + \mathcal{N}(0, \sigma^2 \mathbf{I}).$$

Definition 4 (Exponential Mechanism). *Let $q : \mathcal{D} \times \mathcal{R} \rightarrow \mathbb{R}$ be quality score function and ϵ be a parameter. Then the exponential mechanism outputs a candidate $r \in \mathcal{R}$ according to the following distribution:*

$$\Pr[\mathcal{M}(D) = r] \propto \exp\left(\epsilon \cdot q(D, r)\right)$$

To accurately analyze the privacy of multiple invocations of the Gaussian/Exponential mechanisms (i.e., to derive the (ϵ, δ) parameters) we use the tools of Rényi Differential Privacy (RDP), a variant of differential privacy so named because it uses the Rényi divergence in the bound on a mechanism's output distributions for neighboring inputs.

Definition 5 (Rényi Differential Privacy [45]). *A randomized mechanism $\mathcal{M} : \mathcal{D} \rightarrow \mathcal{R}$ satisfies (α, γ)-Rényi differential privacy (RDP) for $\alpha \geq 1$ and $\gamma \geq 0$, if for any neighboring datasets $D \sim D' \in \mathcal{D}$, we have:*

$$D_\alpha(\mathcal{M}(D) \,||\, \mathcal{M}(D')) \leq \gamma,$$

where $D_\alpha(\cdot \,||\, \cdot)$ is the Rényi divergence of order α between two probability distributions.

To analyze the privacy of the mechanisms above under Rényi-DP, we define the sensitivity of a vector-valued query as follows:

Definition 6 (Sensitivity). *Let $f : \mathcal{D} \rightarrow \mathbb{R}^p$ be a vector-valued function of the input data. The L_2 sensitivity of f is $\Delta_f = \max_{D \sim D'} \|f(D) - f(D')\|_2$.*

It is easy to verify that the L_2 sensitivity of any marginal function M_C is 1, regardless of the attributes in C. This is because one individual can only contribute a count of 1 to a single cell of the output vector. A single invocation of the Gaussian Mechanism satisfies Rényi-DP with parameters determined by the noise scale σ and the sensitivity Δ_f of the function. Similarly, a single invocation of the Exponential Mechanism also satisfies Rényi-DP with parameters determined by ϵ and Δ_q.

Proposition 1 (Rényi-DP of the Gaussian Mechanism [22, 45]). *The Gaussian Mechanism applied to the function* $f : \mathcal{D} \to \mathbb{R}^p$ *satisfies* $\left(\alpha, \alpha \frac{\Delta_f^2}{2\sigma^2}\right)$*-RDP for all* $\alpha \geq 1$.

Proposition 2 (Rényi-DP of the Exponential Mechanism [12, 44]). *The Exponential Mechanism applied to the quality score function* $q : \mathcal{D} \times \mathcal{R} \to \mathbb{R}$ *satisfies* $(2\epsilon\Delta, 0)$*-DP and* $(\alpha, \alpha \frac{(2\epsilon\Delta)^2}{8})$*-RDP for all* $\alpha \geq 1$, *where* $\Delta = \max_{r \in \mathcal{R}} \Delta_{q(\cdot, r)}$ *is the maximum sensitivity of* q.

Note that any mechanism that is $(\alpha, \alpha\rho)$-RDP for all $\alpha \geq 1$ is also ρ-zCDP [11] and vice-versa. We rely on the following propositions to reason about multiple adaptive invocations of RDP mechanisms, and the translation between Rényi-DP and (ϵ, δ)-DP.

Proposition 3 (Adaptive Composition of RDP Mechanisms [45]). *Let* $\mathcal{M}_1 : \mathcal{D} \to \mathcal{R}_1$ *be* (α, γ_1)*-RDP and* $\mathcal{M}_2 : \mathcal{D} \times \mathcal{R}_1 \to \mathcal{R}_2$ *be* (α, γ_2)*-RDP. Then the mechanism* $\mathcal{M} = \mathcal{M}_2(D, \mathcal{M}_1(D))$ *is* $(\alpha, \gamma_1 + \gamma_2)$*-RDP.*

Proposition 4 (RDP to DP [45]). *If a mechanism* $\mathcal{M}$ *satisfies* (α, γ)*-Rényi differential privacy, it also satisfies* $\left(\gamma + \frac{\log(1/\delta)}{\alpha - 1}, \delta\right)$*-differential privacy for all* $\delta \in (0, 1]$.

3. Competition setup

In this section we will summarize the format of the competition and the different components of the challenge problem. The competition consisted of three rounds of increasing complexity, but our focus is on the third round, which built on the previous two rounds.

3.1. **Competition Format.** Competitors were given approximately one month to design their differentially private synthetic data mechanism. The competition organizers provided contestants with a precise problem specification along with a "competitor pack", which included a provisional dataset to test and develop mechanisms, a file that contained domain information for each attribute in the dataset, a script to evaluate the quality of the synthetic data according to their custom scoring criteria, and a baseline mechanism. Each of these components will be described in detail in the subsequent sections. During the one-month competition period, competitors could submit the synthetic data produced by their mechanism to be scored and attain a spot on the provisional leaderboard. This was a good way to gauge how well other competitors were doing, although it was not an authoritative source as the submissions had not been vetted to ensure they satisfied differential privacy (e.g., someone could submit the true data and get a perfect score on the provisional leaderboard).

At the end of the competition period, competitors had to submit their source code along with a document describing the solution and a proof of privacy. A team of experts unknown to the competitors checked the final submitted algorithms to ensure that they satisfied (ϵ, δ)-differential privacy. This was done by checking the written description of the algorithm as well as the source code, to ensure there were no privacy violations or mistakes. After the mechanism was verified to be differentially private, its utility was evaluated on the final dataset (different from the provisional dataset) and performance results were added to a final leaderboard that would determine the ranking of solutions.

In the subsequent subsections, we will specify the details of each component of the problem — i.e., the dataset, the domain, and the evaluation criteria.

3.2. **Dataset.** Algorithms were designed for and evaluated on data from the 1940 U.S. decennial census. The provisional dataset contained data for one state (Colorado) and the final holdout data was for a different state (unknown at the time of the competition). We remark that the provisional dataset was treated as public information. Therefore, any analysis and insights derived from it were not considered to violate privacy. However, solutions that used information from the provisional dataset too aggressively risked over-fitting, and scoring poorly on the final dataset. The provisional dataset contained 98 attributes and about 661 thousand records. All attributes were discrete, taking on values from the domain $\Omega_i = \{0, \ldots, n_i - 1\}$. The value of n_i for each attribute i was provided in a separate specs file. The values of n_i ranged from 2 (for binary attributes like `SEX`) to $10,000,000$ (for numerical attributes like `INCWAGE`). The total number of possible database rows (i.e., the full domain size) was about $|\Omega| = 5 \times 10^{205}$. We provide a full breakdown of the domain in Table 4.

3.3. **Evaluation metrics.** The utility of the synthetic data was measured by how well it preserved key statistics in the ground truth data with respect to three main criteria, enumerated below. We state below the statistics that need to be preserved to score well, but not the exact formula for calculating score. For the precise information, please refer to the official challenge problem statement [56]. Note that the scores for each evaluation metric were normalized to the same range, and averaged across the three metrics (with equal weights). Algorithms were evaluated at three privacy levels, with $\epsilon = 0.3, 1.0, 8.0$ and $\delta \approx 2 \cdot 10^{-12}$, and these scores were averaged to obtain the final score. Computational efficiency was not taken into consideration; several of the solutions (including `NIST-MST`) required up to 10 minutes or more to run.

(1) **3-way Marginals.** The synthetic data was evaluated by comparing its marginals with the marginals of the true data for 100 random triples of attributes, unknown to competitors at submission time. Therefore, a synthetic dataset $\tilde{D}$ scores well on this metric (in expectation) if $M_C(D) \approx M_C(\tilde{D})$ *for all* triples C. There are a total of $\binom{98}{3} = 152096$ possible triples, so this evaluation criteria requires the synthetic data to preserve a large number of marginals to consistently score well.

(2) **High-order conjunctions.** The synthetic data was evaluated by looking at how well it preserved high-order conjunctions. Probabilistically, a high-order conjunction for a set of attributes $C \subseteq \mathcal{A}$ assumes the form $\Pr[\bigwedge_{i \in C}[t_i \in S_i] \mid t \sim D]$, where $S_i \subseteq \Omega_i$ is a subset of the domain for attribute i. This quantity can be expressed in terms of the marginal $\mu = M_C(D)$ via $\sum_{t \in S} \mu_t$, where S is the Cartesian product of S_i's, i.e., $S = \prod_{i \in C} S_i$ and t is a tuple restricted to the attributes in the set C. Synthetic data was evaluated on 300 random high-order conjunctions, where C is generated with a simple random sample of the attributes $\mathcal{A}$ with selection probability 0.1, and S_i is a random subset of Ω_i. There are a total of $2^{98} \approx 10^{29}$ possible choices for C, and the expected size of C is $0.1 \cdot 98 \approx 10$. Even without accounting for the variability in S_i, it is clear that the number of statistics that need to be preserved is enormous.

(3) **Income inequality and gender wage gap.** The synthetic data was evaluated by how well it preserved statistics relating to income inequality and gender wage gap, broken down by city. While the precise details of this metric can be found in the official problem statement, to score well, it suffices for the synthetic data to be accurate with respect to the marginal on (`SEX`, `CITY`, `INCWAGE`). Unlike metrics (1) and (2), above, this metric is relatively easy to score well on, because it just requires preserving one marginal well.

```
1   from private_pgm import FactoredInference
2   from scipy.sparse import identity
3   from numpy.random import normal
4
5   data = load_NIST()
6   queries = [("SEX","LABFORCE"), ("LABFORCE","SCHOOL")]
7   measurement_log = []
8   for c in queries:
9       M_c = data.project(c).datavector()
10      y_c = M_c + normal(loc=0, scale=50, size=M_c.size)
11      measurement_log.append( (identity(M_c.size), y_c, 50, c) )
12  engine = FactoredInference(data.domain)
13  model = engine.estimate(measurements)
14  synth = model.synthetic_data()
```

Figure 2. A demonstration of how to generate synthetic data with `Private-PGM` using real `Python` code. In this case, the selected marginals are (`SEX,LABFORCE`) and (`LABFORCE,SCHOOL`). In Lines 9-12, these marginals are measured with Gaussian noise to protect privacy. In Lines 14-16, `Private-PGM` takes these noisy measurements as input, estimates a model, and generates synthetic data. The `Private-PGM` library provides a straightforward interface that allows users to quickly write end-to-end code to generate synthetic data; different statistics can be preserved by changing Line 6.

4. Overview of Measurement and Inference with Private-PGM

In this section, we elaborate on the general template for a mechanism outlined in Figure 1. Recall there are three high-level steps:

(1) **Select.** Select a collection of marginals to measure.
(2) **Measure.** Use the Gaussian mechanism to measure each marginal in the collection.
(3) **Generate.** Use `Private-PGM` to estimate a data distribution from the noisy measurements and generate synthetic data that preserves the measured marginals well.

In this section, we describe the latter two steps, which form the core of the mechanism. In the next section, we will describe the full mechanism `NIST-MST` in detail, including the *select* step and many other details relating specifically to the NIST contest and dataset. Figure 2 shows how simple and modular this framework for synthetic data generation is. The open source `Private-PGM` library[1] provides a simple interface to the key routines so that an end-to-end synthetic data generation mechanism can be written with very little code, allowing the modeler to focus on tailoring the procedure to the workload and domain. Under the hood, `Private-PGM` has thousands of lines of code, but it exposes a simple interface that is easy to use. In this example, there are only two selected marginal queries: (`SEX,LABFORCE`) and (`LABFORCE,SCHOOL`), but the code can be readily modified (Line 6) to accommodate other marginal queries. In the rest of the section, we describe this general approach in more detail, and give some insight into the steps described in this code snippet.

4.1. **Measuring Marginals with the Gaussian Mechanism.** Algorithm 1 shows the method for measuring marginals. Given a collection of attribute subsets $\mathcal{C}$, it computes the marginal for each $C \in \mathcal{C}$ and adds i.i.d. Gaussian noise to preserve privacy. It also accepts a weight w_C for each attribute subset, which represents the relative importance of

---

[1]https://github.com/ryan112358/private-pgm/

Algorithm 1: Measure Marginals

Input: D (sensitive dataset), $\mathcal{C}$ (a collection of attribute subsets), w_C (weights for each $C \in \mathcal{C}$), σ (noise scale)

Output: log (a list of noisy measurements together with metadata)

(1) Normalize weights, $w_C \leftarrow w_C / \sqrt{\sum_C w_C^2}$.

(2) For $C \in \mathcal{C}$:

(3)   Calculate noisy marginal, $\tilde{\mu} = w_C M_C(D) + \mathcal{N}(0, \sigma^2 I)$

(4)   Append 4-tuple $(w_C I, \tilde{\mu}, \sigma, C)$ to measurement log

---

that marginal. It collects all of these noisy measurements into a *measurement log*, which will be passed to `Private-PGM` for post-processing. The measurement log records, for each marginal defined by a subset of attributes, the noisy marginal query answers together with information about the weight assigned to the marginal and the magnitude of noise used to measure it. It is easy to verify the privacy properties of Algorithm 1, as it is a direct application of the Gaussian mechanism on a sensitivity-1 quantity.[2]

Theorem 1. *Algorithm 1 satisfies $(\alpha, \frac{\alpha}{2\sigma^2})$-RDP for all $\alpha \geq 1$.*

4.2. `Private-PGM`: **Inference and Synthetic Data Generation.** `Private-PGM` is a general-purpose post-processing tool to infer a data distribution given noisy measurements [42]. It is compatible with measurements from a wide variety of mechanisms for discrete data, and can often improve utility at no cost to privacy. Because it infers a representation of a full data distribution, it produces query answers that are *consistent* with one another, even if the noisy measurements are inconsistent. It uses a compact representation of the data distribution to avoid exponential complexity in many cases, though the size of the representation will depend on the measurements, as we describe below.

The high-level idea of `Private-PGM` is to solve an optimization problem to find a data distribution that would produce measurements close to the ones that were observed. It applies to cases when private measurements depend on the data through marginals. For example, suppose the measurements are of the form

$$y_C = Q_C M_C(D) + \xi$$

for all attributes sets C in some collection $\mathcal{C}$, where $Q_C \in \mathbb{R}^{p_C \times n_C}$ is a linear transformation applied to the marginal prior to release and $\xi \in \mathbb{R}^{p_C}$ is zero-centered noise (e.g., Laplace or Gaussian) with known standard deviation. The measurements taken in Algorithm 1 represent the common case where Q_C is just the identity matrix, so that we observe the noisy marginals directly. However, the ability to measure arbitrary linear transformations of marginals is a nice feature that is useful for some types of measurements that occur in practice.[3] Examples of this include hierarchical measurements for answering range queries [29, 35, 47], and optimized measurements for answering general linear query workloads [36, 41].

Given these measurements, `Private-PGM` infers a data distribution P that best explains the measurements by solving the optimization problem

$$\underset{P}{\text{argmin}} \sum_{C \in \mathcal{C}} \|Q_C M_C(P) - y_C\|_2^2. \tag{4.1}$$

---

[2]The weights are explicitly normalized so that the collection of marginals has sensitivity 1.

[3]In fact, Q_C can be replaced with an arbitrary non-linear differentiable transformation, and `Private-PGM` will accept that as input as well.

The objective of this optimization problem is the negative log-likelihood of the noisy measurements under the Gaussian release mechanism, so Equation (4.1) can be seen as a maximum likelihood estimator. We have abused notation by allowing M_C to operate on a data distribution rather than a dataset; the correct interpretation is to substitute the probability vector P for the contingency table representation D of the dataset for computing the marginal. An obvious issue with the optimization problem in Equation (4.1) is that the dimension of the decision variable P is equal to the domain size n, which is exponential in the number of attributes, so we cannot usually solve this problem directly. The key observation of `Private-PGM` is

Fact 1. *Equation* (4.1) *has an optimum of the form* P_θ, *where* P_θ *is a graphical model with one factor for each set* $C \in \mathcal{C}$ *of attributes for which the mechanism measured a marginal.*

This allows us to solve the much lower-dimensional optimization problem

$$\underset{\theta}{\mathrm{argmin}} \sum_{C \in \mathcal{C}} \|Q_C M_C(P_\theta) - y_C\|_2^2 \,, \tag{4.2}$$

with no loss in solution quality. The decision variable θ is the parameter vector of the graphical model, and has dimension equal to the total length of the set of measured marginals. A simple proximal algorithm is given in [42] that solves this optimization problem using only repeated calls to a routine to perform marginal inference in a discrete graphical model — i.e., computing $M_C(P_\theta)$ for all $C \in \mathcal{C}$ and various different θ. The procedure is efficient whenever marginal inference in the graphical model is efficient. Belief propagation is the standard way to perform marginal inference in practice, as it efficiently computes $M_C(P_\theta)$ directly in terms of θ without ever explicitly materializing the full joint distribution P_θ [33].

Remark 1 (Scalability of `Private-PGM`). `Private-PGM` is able to scale to very high-dimensional domains. The main factors that influence it's scalability are (1) the total size of the parameter vector θ and (2) the structure of the set $\mathcal{C}$. The size of θ is the same as the size of all of the relevant marginals combined, and that must not be too large. The size of each marginal depends directly on $|\Omega_i|$, the number of possible values for each attribute. Furthermore, the set $\mathcal{C}$ is important because it corresponds to the structure of the graphical model, and belief propagation is most efficient for tree-structured models. The scalability of belief propagation and `Private-PGM` for non tree-structured models depends on a quantity known as the tree width, which is a measure of how "tree-like" the model is [33].

Remark 2 (Lack of modeling assumptions). It is easy to misconstrue the meaning of the graphical model representation. The inferred distribution P_θ will satisfy conditional independence properties dictated by the structure of the model. However, *no approximation is made when solving the optimization problem in Equation* (4.1), and *the independence properties do not arise from assumptions made by the modeler about the structure of the data distribution.* By Fact 1, there is an optimum to Equation (4.1) that is a graphical model, and hence satisfies these conditional independence properties. Moreover, the graphical model solution P_θ can be shown to have maximum entropy among all optima of Equation (4.1) [42].

Once P_θ is estimated, `Private-PGM` can be used for multiple purposes: reducing error on measured marginals, estimating unmeasured marginals, and even generating synthetic tabular data. These use cases are explained in greater detail below:

Reducing error on measured marginals. First, `Private-PGM` improves utility by combining all sources of measured information into a single cohesive estimate for the data distribution. When the measurements are inconsistent with each other, `Private-PGM` resolves these inconsistencies in a principled manner, reducing variance and boosting utility. One achieves this by using $\bar{y}_C = Q_C M_C(P_\theta)$ in place of the noisy observation y_C. The estimated marginal $\bar{y}_C$ will typically have smaller variance than y_C and will often have lower overall error as well, therefore offering immediate utility improvements at no cost to privacy. Example 1 demonstrates this idea more concretely in a toy setting.

Example 1 (Boosting utility on measured marginals). *We draw* 1000 *tuples from the actual contest dataset and measure two of their marginals, (SEX,LABFORCE) and (LABFORCE,SCHOOL), using the Gaussian mechanism with $\sigma = 50$. Tables (a–c) below show the true marginals, the noisy marginals, and the marginals estimated by* **Private-PGM**. *One can easily verify that the noisy marginals are not consistent: the (SEX,LABFORCE) marginal implies the total number of people with LABFORCE=N is $124.549 + 318.029 = 442.578$, while the (LABFORCE,SCHOOL) marginal implies the same that number is $287.215 + 171.134 = 458.349$. These are two different estimates for the same quantity, which is a consistency problem. In contrast, the* **Private-PGM** *estimated marginals are consistent: they both agree that the total number is 436.873. Additionally,* **Private-PGM** *better estimates the true marginals than the noisy marginals do: the L_1 distances are about 213 and 272 for* **Private-PGM**, *while they are about 251 and 295 for the noisy marginals, which is a significant boost in utility.*

SEX	LABFORCE	count
M	—	156
M	N	65
M	Y	316
F	—	158
F	N	282
F	Y	23

LABFORCE	SCHOOL	count
—	N	159
—	Y	155
N	N	288
N	Y	59
Y	N	336
Y	Y	3

(a) True marginals

SEX	LABFORCE	count
M	—	132.428
M	N	124.549
M	Y	244.365
F	—	173.633
F	N	318.029
F	Y	−21.358

LABFORCE	SCHOOL	count
—	N	116.021
—	Y	186.826
N	N	287.215
N	Y	171.134
Y	N	278.498
Y	Y	−46.497

(b) Noisy marginals

SEX	LABFORCE	count
M	—	124.829
M	N	121.696
M	Y	254.636
F	—	166.034
F	N	315.177
F	Y	0

LABFORCE	SCHOOL	count
—	N	110.029
—	Y	180.834
N	N	276.477
N	Y	160.396
Y	N	254.636
Y	Y	0

(c) `Private-PGM` marginals

Estimating unmeasured marginals. Second, `Private-PGM` can be used to answer new queries that were never measured directly by using P_θ in place of the true data D. This allows us to estimate new marginals without spending the privacy budget, saving a precious resource. Example 2 demonstrates this idea in a toy setting.

Example 2 (Estimating new marginals). *Building on Example 1, recall that we measured the marginals on (SEX,LABFORCE) and (LABFORCE,SCHOOL). We can use* **Private-PGM** *to estimate the marginal on (SEX,SCHOOL), even though we never measured it and it can not be directly inferred from the other marginals that were measured. As shown below in Table (b), the provided estimate is reasonable, given that we never measured it, and we added significant noise to the marginals we did measure. We reiterate that we obtained this estimate "for free", without spending additional privacy budget. Additionally,* **Private-PGM** *can estimate*

the marginal on (SEX,LABFORCE,SCHOOL), which is shown in Table (d). This is pretty close to the true 3-way marginal, shown in Table (c). In fact, the normalized L_1 error is only 0.135. While there may be other equally good estimates for the 3-way marginal (according to the loss function in Equation (4.1)), the estimate provided by Private-PGM has maximum entropy among all of them. In this case, Private-PGM was fairly accurate because SEX and SCHOOL are (approximately) conditionally independent given LABFORCE in the true data.

SEX	SCHOOL	count
M	N	423
M	Y	114
F	N	360
F	Y	103

(a) True 2-way marginal

SEX	SCHOOL	count
M	N	378.873
M	Y	122.289
F	N	262.269
F	Y	218.942

(b) Estimated 2-way marginal

SEX	LABFORCE	SCHOOL	count
M	—	N	74
M	—	Y	82
M	N	N	36
M	N	Y	29
M	Y	N	313
M	Y	Y	3
F	—	N	85
F	—	Y	73
F	N	N	252
F	N	Y	30
F	Y	N	23
F	Y	Y	0

(c) True 3-way marginal

SEX	LABFORCE	SCHOOL	count
M	—	N	47.221
M	—	Y	77.608
M	N	N	77.016
M	N	Y	44.68
M	Y	N	254.636
M	Y	Y	0.000
F	—	N	62.808
F	—	Y	103.226
F	N	N	199.461
F	N	Y	115.716
F	Y	N	0.000
F	Y	Y	0.000

(d) Estimated 3-way marginal

Generating synthetic data. Third, Private-PGM can be used to generate synthetic data $\bar{D}$ in tabular format. $\bar{D}$ can be used in place of P_θ and will generally give similar results. They will not give exactly the same results because $\bar{D}$ has integer-valued marginals while P_θ has real-valued marginals, so some additional rounding error is unavoidable. One can obtain the synthetic data in multiple ways; a simple and natural approach would be to sample records from P_θ to form a synthetic dataset. This naive approach would introduce sampling error which is undesirable. Private-PGM uses an alternative approach to reduce error from additional sources of randomness. The details of this procedure are available in the open-source implementation of Private-PGM, and are summarized in the supplementary material. In Example 3, we give an intuitive idea of how this procedure works, and illustrate why it is preferable to the sampling approach.

Example 3 (Generating synthetic data)**.** *Building on Examples 1 and 2, we calculate the LABFORCE marginal from the Private-PGM model in Table (a) below, which has fractional counts. We also use Private-PGM to generate synthetic data and show the same marginal in Table (b), which has integer counts. These two marginals almost exactly match, because Private-PGM tries to preserve the model marginals as closely as possible when generating synthetic data. However, synthetic data obtained by i.i.d sampling will not match the model marginals as closely due to the randomness in sampling, as shown in Table (c).*

LABFORCE	count
—	290.863
N	436.873
Y	254.636

(a) Private-PGM
model marginal

LABFORCE	count
—	291
N	437
Y	254

(b) Private-PGM
synthetic data marginal

LABFORCE	count
—	262
N	468
Y	252

(c) Sampled
synthetic data marginal

Algorithm 2: `NIST-MST`		
(1) **Calibrate Noise**	Derive noise scale σ from target privacy parameters (ϵ, δ)	Equation (5.1)
(2) **Encode Domain**	Use public information to encode attribute domains	
(3) **Transform Data**	Transform data using insights from provisional data	Algorithm 3
(4) **Compress Domain**	Use data to reduce domain:	
Measure	Measure all one-way marginals	Algorithm 1
Compress	Remove domain elements failing threshold test	Algorithm 4
(5) **Select Marginals**	Select a subset of 2- and 3-way marginals	Algorithm 5
(6) **Measure Marginals**	Measure selected marginals	Algorithm 1
(7) **Synthesize data**	Synthesize records using `Private-PGM`:	
Estimate	Estimate distribution from Step 4 and 6 measurements	Equation (4.2)
Generate	Generate synthetic records	Algorithm 8
(8) **Reverse**	Reverse the transformation made in Step 3	Algorithm 9

5. Algorithm Description

In this section we describe `NIST-MST`, which takes the basic mechanism template outlined in the previous section, and applies it to the setting of the NIST competition. `NIST-MST` simply invokes the Gaussian mechanism to measure a carefully chosen subset of 1, 2, and 3-way marginals. Then the resulting noisy measurements are post-processed using `Private-PGM` to obtain synthetic data that is most consistent with those marginals. `NIST-MST` does not follow the template from the previous section exactly, as there are two rounds of measurements, and an additional domain compression step developed specifically to deal with some of the challenges around the dataset used in the competition. The high-level steps of `NIST-MST` are stated in Algorithm 2, and a detailed description of each step will be provided in this section, with motivations and intuitions for the various design choices.

Step 1: Calibrate Noise. In this step, σ is calibrated to ensure the whole algorithm satisfies (ϵ, δ)-differential privacy. Note that only steps (4) and (6) in Algorithm 2 use the sensitive data, and these are both invocations of Algorithm 1, which is $(\alpha, \frac{\alpha}{2\sigma^2})$-RDP (Theorem 1). The data transformations made in steps (3) and (4) do not affect the privacy analysis of Algorithm 1 since one individual can still only affect each marginal by at most one. Hence `NIST-MST` is $(\alpha, \frac{\alpha}{\sigma^2})$-RDP by two-fold adaptive composition (Proposition 3). Moreover, by Proposition 4, `NIST-MST` is $\left(\frac{\alpha}{\sigma^2} + \frac{\log(1/\delta)}{\alpha-1}, \delta\right)$-DP for all $\alpha \geq 1$.

For a fixed α, it is easy to determine σ by solving the equation $\frac{\alpha}{\sigma^2} + \frac{\log(1/\delta)}{\alpha-1} = \epsilon$ for σ. The best value of σ can be obtained by minimizing over all α. In the contest, this computation was done by invoking the `moments accountant` [2], which minimizes over $\alpha = 1, \ldots, 512$. However, this minimization can actually be done in closed form [72], leading to the following equation for σ:

$$\sigma = \frac{\sqrt{\log(1/\delta)} + \sqrt{\log(1/\delta) + \epsilon}}{\epsilon} \tag{5.1}$$

Theorem 2 (Privacy of `NIST-MST`). *Algorithm 2 is (ϵ, δ)-differentially private.*

Proof. From the analysis above, we know that `NIST-MST` is $\left(\frac{\alpha}{\sigma^2} + \frac{\log(1/\delta)}{\alpha-1}, \delta\right)$-DP for all $\alpha \geq 1$. By plugging in $\alpha = 1 + \sigma\sqrt{\log(1/\delta)}$ and $\sigma = \frac{\sqrt{\log(1/\delta)} + \sqrt{\log(1/\delta) + \epsilon}}{\epsilon}$ and simplifying, we see that $\frac{\alpha}{\sigma^2} + \frac{\log(1/\delta)}{\alpha-1} = \epsilon$, and hence `NIST-MST` is (ϵ, δ)-DP as desired. The algebraic

1. SPLIT	2/2	21. NCHLT5	7/7	41. SIZEPL	31/19	61. SUPDIST	631/631	81. ENUMDIST	3021/3021
2. SLREC	3/2	22. RACE	7/7	42. EMPSTATD	35/15	62. METAREA	657/334	82. CITYPOP	3225/3225
3. SEX	3/2	23. WKSWORK2	7/7	43. WKSWORK1	53/53	63. PRESGL	816/816	83. URBPOP	3225/3225
4. SCHOOL	3/2	24. VET1940	9/4	44. SEA	54/54	64. BPL	901/163	84. METAREAD	6561/378
5. URBAN	3/3	25. UCLASSWK	9/8	45. OCCSCORE	81/81	65. MBPL	901/164	85. MTONGUED	9602/489
6. FARM	3/3	26. VETPER	9/8	46. AGEMARR	90/89	66. FBPL	901/165	86. MIGMET5	10000/379
7. OWNERSHP	3/3	27. HISPRULE	9/9	47. MIGRATE5D	91/15	67. IND1950	998/162	87. MIGCOUNTY	10000/385
8. RESPONDT	3/3	28. HRSWORK2	9/9	48. MTONGUE	97/92	68. MIGSEA5	998/510	88. ERSCOR50	10000/1002
9. SPANNAME	3/3	29. CLASSWKR	10/4	49. SEI	97/97	69. OCC	999/231	89. EDSCOR50	10000/1002
10. LABFORCE	3/3	30. INCNONWG	10/4	50. CLASSWKRD	99/18	70. EDUCD	1000/44	90. NPBOSS50	10000/1002
11. VETWWI	3/3	31. SAMEPLAC	10/4	51. HRSWORK1	99/99	71. HIGRADED	1000/69	91. CITY	10000/1164
12. SSENROLL	3/3	32. VETSTAT	10/4	52. VETSTATD	100/10	72. GQTYPED	1000/92	92. MIGCITY5	10000/1164
13. METRO	4/4	33. VETCHILD	10/5	53. GQFUNDS	100/13	73. IND	1000/136	93. RENT	10000/10000
14. EMPSTAT	4/4	34. MIGTYPE5	10/6	54. EDUC	100/13	74. UIND	1000/136	94. MBPLD	90021/537
15. HISPAN	5/5	35. SAMESEA5	10/6	55. AGEMONTH	100/15	75. MIGPLAC5	1000/199	95. FBPLD	90021/539
16. CITIZEN	5/5	36. MIGRATE5	10/7	56. HIGRADE	100/25	76. UOCC	1000/231	96. BPLD	90022/536
17. WARD	6/6	37. GQTYPE	10/10	57. CHBORN	100/62	77. UOCC95	1000/279	97. VALUEH	10000000/5003
18. NATIVITY	6/6	38. MARRNO	10/10	58. AGE	109/109	78. OCC1950	1000/283	98. INCWAGE$_A$	—/52
19. MARST	7/6	39. OWNERSHPD	21/8	59. HISPAND	481/55	79. DURUNEMP	1000/1000	99. INCWAGE$_B$	—/8
20. GQ	7/7	40. FAMSIZE	22/22	60. RACED	621/238	80. COUNTY	1251/385	INCWAGE	10000000/—

Table 4. Domain information for the census dataset used in the third round of the competition. Table specifies attribute names, and number of possible values for that attribute according to (1) the provided specs file and (2) the specs file combined with IPUMS documentation.

manipulation is routine but messy; it can easily be verified with `sympy` (see Figure 4 in the supplement). ☐

Remark 3 (Noise Calibration). It is well known that calibrating σ via an RDP analysis does not give the smallest possible value required to achieve (ϵ, δ)-DP, and an analytic calibration gives strictly better results [7], at least for a *single* invocation of the Gaussian mechanism. However, at the time of the competition, adaptive composition of *two* Gaussian mechanisms was needed, and RDP was chosen because of its clean and well-understood guarantees. If using the analytic Gaussian mechanism, advanced composition would be necessary to reason about the privacy of two-fold adaptive composition [21]. Since these are somewhat loose bounds, the benefit of the analytic calibration would be lost. However, since the time of the competition, much progress has been made on understanding the behavior of the Gaussian mechanism under composition and it is now known that the analytic Gaussian mechanism can be used to calibrate noise for multiple (adaptive) invocations of the Gaussian mechanism [19, 51]. This would give a smaller value of σ than the one shown in Equation (5.1), typically offering an improvement of 10 to 20 percent.

Step 2: Encode Domain. Before running `NIST-MST`, it is necessary to know the data domain Ω. The supplied competitor pack came with a "SPECS" file that contained some domain information. Specifically, it supplied a single positive integer n_i for each attribute i, and the domain for that attribute was assumed to be $\Omega_i^{\text{SPECS}} = \{0, \ldots, n_i - 1\}$. However, because the data is derived from a census source, the domain is very thoroughly documented on the Integrated Public Use Microdata Series (IPUMS) website [1]. IPUMS offers a much finer grained view of the data domain, specifying the exact set of possible values for most attributes. We use Ω_i^{IPUMS} to denote the domain of possible values for attribute i according to IPUMS. Example 4 demonstrates the benefit of using the finer grained IPUMS domain information.

Algorithm 3: Transform data

Input: D (sensitive dataset)
Output: D (transformed sensitive dataset)

(1) Replace `VALUEH` attribute in D using transformation:

(2) Split `INCWAGE` attribute in D into two attributes `INCWAGE`$_\text{A}$ and `INCWAGE`$_\text{B}$ using transformation:[4]

$$\texttt{VALUEH} = \begin{cases} \lfloor \frac{\texttt{VALUEH}}{5} \rfloor & \texttt{VALUEH} \leq 25{,}000 \\ 5001 & \texttt{VALUEH} = 9{,}999{,}998 \\ 5002 & \texttt{VALUEH} = 9{,}999{,}999 \\ 5000 & \text{otherwise} \end{cases}$$

$$\texttt{INCWAGE}_\text{A} = \begin{cases} \lfloor \frac{\texttt{INCWAGE}}{100} \rfloor & \texttt{INCWAGE} \leq 5000 \\ 50 & \texttt{INCWAGE} < 9{,}999{,}998 \\ 51 & \texttt{INCWAGE} = 9{,}999{,}998 \end{cases}$$

$$\texttt{INCWAGE}_\text{B} = \begin{cases} 0 & \texttt{INCWAGE} \equiv 0 \mod 100 \\ 1 & \texttt{INCWAGE} \equiv 0 \mod 20 \\ 2 & \texttt{INCWAGE} \equiv 0 \mod 50 \\ 3 & \texttt{INCWAGE} \equiv 0 \mod 25 \\ 4 & \texttt{INCWAGE} \equiv 0 \mod 10 \\ 5 & \texttt{INCWAGE} \equiv 0 \mod 5 \\ 6 & \texttt{INCWAGE} \equiv 0 \mod 2 \\ 7 & \text{otherwise} \end{cases}$$

Example 4 (SPECS vs. IPUMS). *From the specs file and the IPUMS website, we see the domain for the `EDUC` attribute is $\Omega_i^{SPECS} = \{0, 1, \ldots, 99\}$ and $\Omega_i^{IPUMS} = \{0, 1, \ldots, 11, 99\}$. Note that 99 is a special code that typically corresponds to missing data. While both sources agree that 99 is the largest possible value, the IPUMS documentation suggests that values in the range $12, \ldots, 98$ are not possible. Using the finer granularity domain from IPUMS reduces the number of possible values for `EDUC` from 100 to 13. This has two important ramifications. First, it will make `Private-PGM` more efficient in later steps, since the scalability of that tool depends directly on the domain sizes of the attributes. Second, it will prevent `NIST-MST` from inadvertently introducing out-of-domain tuples to the synthetic data which could otherwise occur by adding positive noise to zero counts.*

Often Ω_i^{IPUMS} is a subset of Ω_i^{SPECS}, although this is not always the case. In some cases, IPUMS documents a certain value as being possible that never appeared in the provisional dataset or the supplied specs file. To account for this `NIST-MST` uses the intersection of the two domains, i.e., $\Omega_i = \Omega_i^{\text{SPECS}} \cap \Omega_i^{\text{IPUMS}}$. Table 4 enumerates the attributes in the dataset along with the domain size provided in the specs file, and the compressed domain size derived by `NIST-MST`.

Step 3: Transform Data. In addition to the general domain encoding outlined above, `NIST-MST` gave special attention to two of the attributes with the largest domain: `INCWAGE` and `VALUEH`. Both of these attributes started out with 10 million possible values, and the IPUMS documentation provided limited information on these attributes. Therefore, `NIST-MST` leveraged the provisional dataset to try to identify a domain that captured all or most of the observed values for these attributes. Algorithm 3 shows how these attributes are transformed to reduce the domain size. The intuition behind this pre-processing procedure is to compress the domain, while ensuring the compressed domain still covers all or most of the values observed in the provisional dataset. For example, other than the special codes of 9,999,998 and 9,999,999, 99.2% of records have `VALUEH` that is a multiple of 5 and less

---

[4]Each condition in the piecewise definition of `INCWAGE`$_\text{A}$ and `INCWAGE`$_\text{B}$ should be interpreted as an "else if" statement rather than an "if" statement, as clearly multiple conditions can be true at the same time

---

Algorithm 4: Domain compression

    Input: log (list of noisy measurements), D (sensitive dataset), Ω (domain)
    Output: D (transformed sensitive dataset), Ω (transformed domain)

(1)   For each measurement $(\_\_, \tilde{\mu}, \sigma, \{i\})$ in log

(2)   Replace values for attribute i in dataset D using transformation:

$$t \leftarrow \begin{cases} t & \tilde{\mu}_t \geq 3\sigma \\ \varnothing & \text{otherwise} \end{cases}$$

(3)   Modify domain accordingly, $\Omega_i \leftarrow \{t \mid \tilde{\mu}_t \geq 3\sigma\} \cup \{\varnothing\}$

---

than or equal to 25,000.[5] This allows us to compress the domain of `VALUEH` to a much more manageable size of 5003 while still covering about 99.7% of the observed values.

    `NIST-MST` uses a similar approach to handle `INCWAGE`. Over 99.95% of records in the provisional dataset had an `INCWAGE` value of either 9,999,998 or something in the range $[0, 5000]$. For that reason, it is reasonably safe to truncate values above 5000 without introducing too much bias. This transformation reduces the domain size of `INCWAGE` to 5002, but `NIST-MST` takes things one step further. There are clear periodic patterns in the `INCWAGE` marginal, as the most common values are all multiples of 100. Multiples of 20, 50, and 25 are also common. To exploit this observation, `NIST-MST` splits up `INCWAGE` into two attributes: `INCWAGE`$_{\text{A}}$ and `INCWAGE`$_{\text{B}}$, and never measures `INCWAGE` directly, but only indirectly through these two derived attributes. `INCWAGE`$_{\text{A}}$ is meant to capture the coarse-grained income by discretizing it into width 100 bins, whereas `INCWAGE`$_{\text{B}}$ is meant to capture the periodicity in the last two digits. These two derived attributes have smaller domains of size 52 and 8, respectively. The exact formulas are given in Algorithm 3.

Step 4: Compress Domain. In step (1), `NIST-MST` was able to greatly reduce the domain size by incorporating information from IPUMS. However, even after this domain encoding, some of the attributes in the data remain fairly sparse. For example, only 17.4% percent of counts in the `VALUEH` marginal exceed 100. In this step, we answer all 1-way marginals, i.e., we pass $\mathcal{C} = \{\{i\} \mid i \in \mathcal{A}\}$ into Algorithm 1. Every marginal is assigned an equal weight of 1, with the exception of `INCWAGE`$_{\text{A}}$, which is assigned a weight of 2.

    After obtaining noisy 1-way marginals, Algorithm 4 is called, which searches for domain elements for which the noisy count fell below the threshold of 3σ. These domain elements were merged into a single "other" domain element, denoted $\varnothing$. Later steps of `NIST-MST` operate over the resulting transformed dataset and domain. This step has two main benefits, similar to the ones from the domain encoding step. First, it improves scalability of `Private-PGM` in later steps by reducing the domain size of the attributes. Second, it ensures that the tuples generated by `NIST-MST` (probably) have attribute values that actually occurred in the dataset.

Step 5: Select Measurements. The next step of `NIST-MST` is to identify a collection of 2- and 3-way marginals that will later be measured. This is one of the most crucial components of `NIST-MST`, because the marginals selected in this step will ultimately determine the marginals that will be preserved in the generated synthetic data. It is important to note that this step selects measurements without using the sensitive dataset, although it does rely heavily on the provisional dataset. This algorithm does take the privacy budget ϵ as

---

[5]9,999,998 and 9,999,999 are special codes corresponding to missing values of N/A values.

Algorithm 5: Marginal selection algorithm

Input: $\hat{D}$ (the provisional dataset), ϵ (privacy budget)
Output: $\mathcal{C}$ (a collection of attribute subsets), w (weights for each $C \in \mathcal{C}$)

(1) Construct a complete graph G where vertices are attributes in the dataset, and edge (i,j) is weighted according to the mutual information between attribute i and attribute j in the dataset $\hat{D}$. Add 100 to the edge weights for (SEX,CITY), (SEX,INCWAGE$_A$) and (CITY,INCWAGE$_A$).

(2) Identify the maximum spanning tree (MST) of the graph, and for each edge in the tree, add the attribute pair to $\mathcal{C}$. Also add (SEX,CITY), (SEX,INCWAGE$_A$), (CITY,INCWAGE$_A$), and (SEX,CITY,INCWAGE$_A$) to $\mathcal{C}$ if they are not already included.

(3) For each pair of adjacent edges $(i,j),(i,k)$ in the MST, compute the marginals $M_{ij}(\hat{D})$, $M_{ik}(\hat{D})$, and $M_{ijk}(\hat{D})$. Use Private-PGM to estimate $\tilde{M}_{ijk}$ from M_{ij} and M_{ik}, and record the error in the estimate $E_{ijk} = \left\| M_{ijk} - \tilde{M}_{ijk} \right\|_1$.

(4) For each attribute i, construct a complete graph consisting of nodes that are neighbors of i in the MST, and each edge (j,k) in the new graph is assigned a weight of E_{ijk}. Remove edges whose weight is below a threshold of 0.1, and compute the maximum spanning tree of the resulting graph. For each edge (j,k) in the new MST, add the (j,k) and (i,j,k) marginals to $\mathcal{C}$.

(5) Remove attribute subsets whose marginal is too large, i.e., $C \in \mathcal{C}$ such that $\prod_{i \in C} n_i \geq 10^6$.

(6) Assign weights to selected attribute subsets using formula:

$$w_C \propto \begin{cases} 8 & C = (\texttt{SEX},\texttt{CITY},\texttt{INCWAGE}_A), \epsilon \leq 0.3 \\ 4 & C = (\texttt{SEX},\texttt{CITY},\texttt{INCWAGE}_A), \epsilon \geq 4.0 \\ 6 & C = (\texttt{SEX},\texttt{CITY},\texttt{INCWAGE}_A), 0.3 < \epsilon < 4.0 \\ 2 & C \in \{(\texttt{SEX},\texttt{CITY}),(\texttt{SEX},\texttt{INCWAGE}_A),(\texttt{CITY},\texttt{INCWAGE}_A)\} \\ 1 & \text{otherwise} \end{cases}$$

---

input, but it does not "consume" it — it only uses it to determine weights to assign to each selected marginal.

Algorithm 5 shows how NIST-MST selects 2- and 3-way marginals for measurement. This algorithm is inspired by a similar approach used by two other mechanisms for differentially private synthetic data [14,66]. It combines one principled step, which is to find the maximum spanning tree (MST) on the graph where edge weights correspond to mutual information between two attributes, with some additional heuristics to ensure that certain important attribute pairs are selected, and more heuristics to select some triples while keeping the graph tree-like. A reader familiar with graphical models with recognize the MST step as the famous Chow-Liu algorithm for structure learning in a graphical model [15]. Intuitively, highly correlated marginals should be measured because attributes that are independent can trivially be preserved without direct measurement.

Figure 3 shows the marginals selected by this algorithm in graphical format. Each edge in the graph represents a pair of attributes whose marginal will be measured by NIST-MST, and each triangle in the graph represents a triple of attributes whose marginal will be measured by NIST-MST. The structure of this measurement graph also corresponds to the structure of the graphical model used by Private-PGM. The tree-like structure of the graph will ultimately allow Private-PGM to run efficiently. The green subgraph corresponds to the (SEX,CITY,INCWAGE$_A$) clique, the black edges form a maximum spanning tree of the underlying correlation graph, and the dotted red edges are additional edges that enhance the expressive capacity of the model while retaining the tree-like structure.
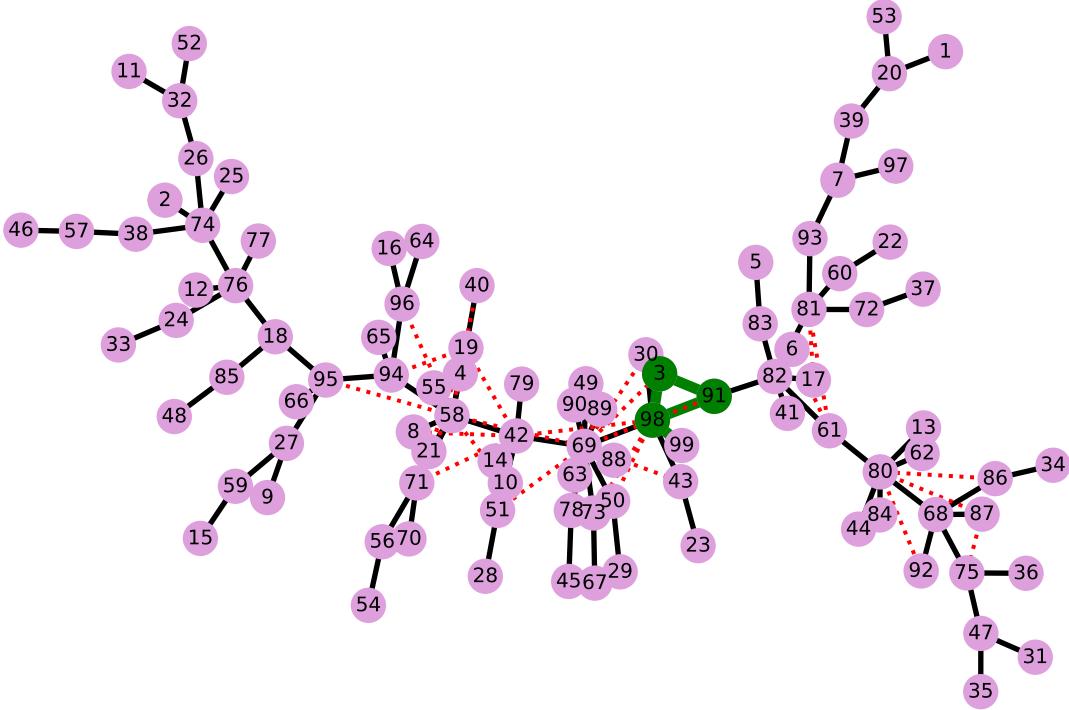
Figure 3. A graphical depiction of the 2- and 3-way marginals selected by `NIST-MST`. Nodes correspond to attributes in the dataset, and edges correspond to marginals selected by `NIST-MST`. Nodes are labeled by the 2-digit code for the attribute given in Table 4. Black edges form a *maximum spanning tree* of the underlying correlation graph. Green nodes and edges correspond to the special (`SEX, CITY, INCWAGE_A`) attributes. Dotted red edges identify the extra marginals chosen to improve expressive capacity of the model while maintaining tractability of `Private-PGM`. All dotted red edges form a triangle, and for each of those `NIST-MST` also included the 3-way marginal corresponding to the three nodes that make up the triangle in the list of selected marginals.

Step 6: Measure Marginals. The next step of `NIST-MST` is to measure the marginals selected in the previous step with Algorithm 1. The result is a collection of noisy measurements contained within a measurement log, and suitable for post-processing with `Private-PGM`.

Step 7: Synthesize data. The next step of `NIST-MST` is to combine the measurement logs from Steps 4 and 6 and pass them to `Private-PGM`, which returns a synthetic dataset whose marginals approximately match those in the measurement log. Because the measurements in Steps 4 and 6 were made on the uncompressed and compressed domains, respectively, the measurements from Step 4 had to be re-expressed over the compressed domain.

Step 8: Reverse Transformation. The final step of `NIST-MST` is to reverse the transformations made in Steps 4 and 3, to bring the data back to the original domain. For Step 4, this requires evenly distributing any instances of ∅ among the original domain elements mapped to it. For Step 3, this requires modifying `VALUEH` and combining `INCWAGE_A` and `INCWAGE_B` back into a single attribute. The details are given in Algorithm 9 of the appendix.

Algorithm 6: Differentially private measurement selection

Input: D (sensitive dataset), log (measurements of 1-way marginals), ρ (privacy parameter), $\mathcal{C}$ (initial set of (i,j) pairs to measure; empty by default)

Output: $\mathcal{C}$ (final set of (i,j) pairs to measure)

(1) Use `Private-PGM` to estimate all 2-way marginals $\bar{M}_{ij}$ from log

(2) Compute L_1 error between estimated 2-way marginal and actual 2-way marginal for all i,j:
$q_{ij}(D) = \left\| M_{ij}(D) - \bar{M}_{ij} \right\|_1$ (this is a sensitivity 1 quantity)

(3) Let $G = (\mathcal{A}, \mathcal{C})$ be the graph where attributes are vertices and edges are pairs of attributes

(4) Let r be the number of connected components in G [6]

(5) Let $\epsilon = \sqrt{\frac{8\rho}{r-1}}$

(6) Repeat $r - 1$ times

(7)   Let S be the set of all attribute pairs (i,j), where i and j are in different connected components of G

(8)   Select attribute pair (i,j) by running the `exponential mechanism` with quality score function q_{ij} on set S and privacy parameter ϵ.

(9)   Add attribute pair (i,j) to $\mathcal{C}$

---

6. Extensions

One limitation with `NIST-MST` is that it is highly tailored to the setting of the NIST competition, and crucially relies on the existence of a public provisional dataset that can be used to select marginals. In more general settings, we will not always have access to a provisional dataset that follows a similar distribution as the sensitive data. For that reason, we propose `MST`, a general purpose mechanism that is inspired by the `NIST-MST` mechanism, but doesn't rely on the existence of provisional data. The basic mechanism is the same as `NIST-MST` outlined in Algorithm 2, with a couple minor exceptions. First, the preprocessing transformations and corresponding reverse transformations are not done — those were specific to the U.S. Census dataset used in the competition and not generally applicable beyond that setting. Second, the measurement selection step, which previously relied on a provisional dataset to select correlated marginals, is replaced by a differentially-private version that uses the sensitive dataset. `MST` devotes $\frac{1}{3}$ of the RDP budget towards measurement selection, and uses the remaining $\frac{2}{3}$ of the RDP budget for measuring the marginals. Privacy of `MST` follows by adaptive composition Proposition 3. For completeness, this calculation is given in the appendix.

The measurement selection algorithm is shown in Algorithm 6. Just like Algorithm 5, this algorithm tries to find a collection of attribute pairs that form a maximum spanning tree of an underlying correlation graph. However, as it uses the sensitive dataset, it must do this in a differentially private way. To achieve this, we first use a low-sensitivity approximation of the mutual information for assigning edge weights. We assume that we already measured all 1-way marginals, so we can get reasonable estimates of 2-way marginals by invoking `Private-PGM`.[7] The edge weights are then computed as the L_1 distance between the true 2-way marginal and the estimated one (a sensitivity 1 quantity). After computing the edge weights, Algorithm 6 can be seen as a differentially private version of Kruskal's algorithm [34] for computing a maximum spanning tree. It consists of $d - 1$ steps (the number of edges in a

---

[6]If $\mathcal{C}$ is empty, this is just the number of attributes d.

[7]In this simple case, `Private-PGM` estimates 2-way marginals under an independence assumption, which could alternatively be achieved by multiplying the (noisy) one-way marginals together.

18

spanning tree), and in each step, it adds a highly weighted edge that connects two different connected components. In Kruskal's algorithm, the highest weighted edge is chosen, but this would not be differentially private. We instead invoke the `exponential mechanism` to select a highly weighted edge in a differentially private way. In principle, we could apply any private selection algorithm here, including report-noisy-max [21] and the recently developed permute-and-flip mechanism [43]. While permute-and-flip is known to dominate the exponential mechanism under ϵ-DP [43], the exponential mechanism enjoys a tighter privacy analysis under Rényi-DP [12].

The result of this algorithm is a collection of attribute pairs that will be measured by `MST`. Algorithm 6 has an optional argument, $\mathcal{C}$, which is an initial set of attribute pairs to measure. If this is supplied, the algorithm will always include those in the result, and then constructs a maximum spanning tree around them. This enables some extra flexibility that may be beneficial in certain settings where some marginals are more important than others, and need to be preserved even if they are not the most highly correlated. For many applications, $\mathcal{C}$ can just be empty. In the context of the competition, this feature is useful because the marginals relating `SEX`, `CITY`, and `INCWAGE_A` are very important (since their accuracy determines $\frac{1}{3}$ of the final score).

Theorem 3 (Privacy of Algorithm 6). *Algorithm 6 is $(\alpha, \alpha\rho)$-RDP for all $\alpha \geq 1$.*

Proof. Step 4a is $(\alpha, \alpha\frac{1}{8}\epsilon^2)$-RDP by Proposition 2. Substituting $\epsilon = \sqrt{\frac{8\rho}{r-1}}$, we see that it is equivalent to $(\alpha, \alpha\frac{\rho}{r-1})$-RDP. It is called $r-1$ times, so the entire mechanism is $(\alpha, \alpha\rho)$-RDP by Proposition 3. $\square$

7. Experiments

In this section we discuss the experimental evaluation carried out by the contest organizers, and how `NIST-MST` compared to the submissions from other teams. We offer our own insights into the numbers and explanations for the differences between mechanisms.

Evaluations were carried out on U.S. Census data for two different states: Arizona and Vermont. These datasets had 293,999 and 211,228 records, respectively. Scores were calculated separately for each of the three evaluation metrics described in Section 3.3. The final score was calculated by averaging the scores for each metric, each value of ϵ, and each of the two datasets. In Table 5 we show the score breakdown by metric (averaged over Arizona and Vermont) for the top five submitted algorithms. We also evaluated our new mechanism (`MST`) and included it as an extra row (highlighted), even though it was not evaluated by the contest organizers at the time of the competition. This was possible because the final evaluation datasets were released after the competition, and the script used to evaluate the synthetic data was provided as part of the competitor pack.

`MST` was described in Section 6. We instantiate it with $\mathcal{C} = \{(\texttt{SEX},\texttt{CITY}), (\texttt{SEX},\texttt{INCWAGE_A}), (\texttt{CITY},\texttt{INCWAGE_A})\}$, and add $(\texttt{SEX},\texttt{CITY},\texttt{INCWAGE_A})$ to the returned result as well. These marginals are weighted using the same formula as `NIST-MST` (see Algorithm 5).

Among the contest submissions, `NIST-MST` consistently performed the best, for most metrics and values of ϵ. Compared to DPSyn, it did a much better job at answering 3-way marginals and high order conjunctions, but performed slightly worse at handling the income inequality metric. Compared to every other mechanism, `NIST-MST` did better on every metric.

ϵ	Team	3-way marginals	High order conjunctions	Income inequality	Overall
0.3	RMcKenna (NIST-MST)	0.12	0.17	0.10	0.13
0.3	MST	**0.11**	**0.16**	0.10	**0.12**
0.3	DPSyn	0.15	0.31	**0.07**	0.18
0.3	PrivBayes	0.19	0.29	0.18	0.22
0.3	Gardn999	0.21	0.32	0.25	0.26
0.3	UCLANESL	0.57	0.72	0.22	0.50
1.0	RMcKenna (NIST-MST)	**0.09**	**0.15**	**0.04**	**0.09**
1.0	MST	**0.09**	**0.15**	0.05	0.10
1.0	DPSyn	0.11	0.23	0.05	0.13
1.0	PrivBayes	0.17	0.26	0.09	0.17
1.0	Gardn999	0.18	0.28	0.22	0.23
1.0	UCLANESL	0.42	0.53	0.28	0.41
8.0	RMcKenna (NIST-MST)	**0.07**	**0.14**	0.04	**0.08**
8.0	MST	0.08	**0.14**	0.05	0.09
8.0	DPSyn	0.09	0.20	**0.02**	0.10
8.0	PrivBayes	0.13	0.23	0.09	0.15
8.0	Gardn999	0.17	0.26	0.24	0.22
8.0	UCLANESL	0.35	0.41	0.25	0.34

Table 5. Evaluation of NIST-MST and other mechanisms from competing teams, broken down by the three scoring metrics: 3-way marginals, high-order conjunctions, and income inequality. MST is also shown for comparison, even though that mechanism was not submitted during the competition. If it was submitted instead of NIST-MST, it would have placed first overall.

Generally speaking, NIST-MST and DPSyn (and to a lesser extent PrivBayes) seemed to be the only mechanisms that scored well on the income inequality metric, which is surprising given that it was the simplest metric and only required preserving one 3-way marginal accurately. This raises an important point: mechanisms that understood the evaluation criteria well, and designed their mechanisms around it, generally performed better than mechanisms that just tried to generate good synthetic data without thinking about how its utility would be evaluated. NIST-MST and DPSyn did a good job of designing their mechanism for the task at hand, which was an important contributing factor for why they outperformed the other solutions.

While NIST-MST relied heavily on the provisional dataset for measurement selection, the more general variant MST still performs well, without explicitly relying on the provisional data. In fact, MST would have won first place if it was submitted instead of NIST-MST at the time of the competition; it was only slightly worse than NIST-MST and still better than the other submissions. The difference in performance between MST and NIST-MST was at most 0.01 for every metric and privacy budget evaluated.

At $\epsilon = 0.3$, MST even achieved a smaller overall score than NIST-MST. NIST-MST measures both 2- and 3-way marginals, while MST only measures 2-way marginals. Since the privacy budget is relatively small, it makes sense that fewer measurements would work better here.

A promising direction for future work is to adaptively select the number of marginals to measure based on the privacy budget and the amount of data available.

For a much more comprehensive evaluation of these mechanisms, we refer the reader to [10], which goes well beyond the metrics used in the competition to evaluate these mechanisms. This work is also a useful resource that summarizes each mechanism at a high level. Their results generally show that `NIST-MST` was the best performing mechanism for many of the additional evaluation metrics not used as part of the official scoring criteria. This suggests that `NIST-MST` produces the most generally useful synthetic data among the submitted mechanisms.

8. Related Work

Differentially private synthetic data has been an active area of research for several years. One of the earliest mechanisms proposed for this task was the "small database mechanism" [21], which instantiates the exponential mechanism over a set of small databases to select one that is statistically similar to the true data. Unfortunately, this mechanism is not able to run in practical settings, as it requires enumerating all possible datasets of a fixed sizes, resulting in a combinatorial explosion even for small dataset sizes (especially if the domain size is large).

In the competition, the top four submissions all followed the same basic template outlined in Figure 1: they selected and measured a collection of marginals, and then used those to estimate synthetic data. Moreover this approach has been applied more generally in the literature [9, 14, 66, 71]. One key difference that sets our approach apart is `Private-PGM`. While [14, 66] do leverage graphical models to generate synthetic data, their approach is limited in how it makes use of the noisy measurements: they more or less treat the noisy marginals as the true marginals (with some lightweight post-processing), whereas `Private-PGM` makes use of all available measurements to resolve inconsistencies and boost utility in a principled manner. An alternative method for resolving inconsistencies and generating synthetic data from noisy marginals is proposed in [71]. This method does not construct an intermediate representation of the data distribution as `Private-PGM` does. As a result, their consistency resolution step only ensures *local consistency* (i.e., that all marginals internally agree on the common marginals), and does not satisfy the stronger notion of *global consistency* (i.e., that there is a data distribution that has all stated marginals), as `Private-PGM` does.

Another popular class of approaches for differentially private synthetic data is based on generative adversarial networks (GANs) [27]. Several differentially private GANs have been proposed for the purpose of generating synthetic data [3, 8, 23, 32, 52–54, 59, 70]. In fact, two teams in the NIST competition adopted a GAN-based approach (UCLANESL in Table 5, and one other team that did not place in the top five), however, their scores were not generally competitive with the other approaches that used the marginal-based framework. GANs are notoriously hard to train in practice [49] and when differential privacy constraints are enforced, it is even more difficult. It typically requires running an algorithm like DP-SGD [2] to train, and if it fails to converge (which is common) the privacy budget used for training is essentially wasted. Setting the right hyper-parameters is also a major challenge for this approach.

Another important and related problem is how to evaluate the quality of synthetic data [5, 10, 30, 50]. Beyond the metrics used in the NIST competition, one alternative is

pMSE, which is a general measure of distributional similarity [50]. Another alternative measure is machine learning efficacy, or how well the synthetic data supports machine learning applications [30]. A number of other measures for evaluating synthetic data can be found in [10]. We believe that there may be no universal answer to this question: it should ultimately depend on the data and its use cases. In general, it would be nice to have a mechanism that can automatically adapt to an analyst-provided *workload*, and generate synthetic data that provides high utility on the queries and tasks in that workload. Several workload-adaptive mechanisms exist, but they are generally restricted to settings where the full high-dimensional histogram can be explicitly materialized in vector form, and are thus unable to scale to high-dimensional domains [21, 24, 28, 36, 41]. When combined with `Private-PGM`, the scalability (and utility) of some of these mechanisms can be improved, however [42].

9. Conclusions

In this paper, we described `NIST-MST`, the winning mechanism from the NIST differential privacy synthetic data competition, and `MST` a new mechanism inspired by `NIST-MST` that works almost as well, without relying on public provisional data. While these mechanisms are state-of-the-art, the problem of differentially private synthetic data is far from solved. Nevertheless, we believe our basic framework centered around `Private-PGM` can serve as a core component of new mechanisms for this task. `Private-PGM` allows the mechanism designer to focus on *what to measure*, rather than *how to post-process* those measurements to get synthetic data while extracting the most utility from them. In fact, in the final round of the recently completed follow-up challenge, the NIST 2020 Temporal Map Challenge, *both* the first and second place teams used `Private-PGM` for post-processing, with each team developing novel techniques for measurement selection.

References

[1] Ipums usa. https://usa.ipums.org/usa/.

[2] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[3] N. C. Abay, Y. Zhou, M. Kantarcioglu, B. Thuraisingham, and L. Sweeney. Privacy preserving synthetic data release using deep learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 510–526. Springer, 2018.

[4] G. Ács, C. Castelluccia, and R. Chen. Differentially private histogram publishing through lossy compression. In *ICDM*, pages 1–10, 2012.

[5] C. Arnold and M. Neunhoeffer. Really useful synthetic data–a framework to evaluate the quality of differentially private synthetic data. *arXiv preprint arXiv:2004.07740*, 2020.

[6] H. J. Asghar, M. Ding, T. Rakotoarivelo, S. Mrabet, and M. A. Kaafar. Differentially private release of high-dimensional datasets using the gaussian copula. *arXiv preprint arXiv:1902.01499*, 2019.

[7] B. Balle and Y.-X. Wang. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International Conference on Machine Learning*, pages 394–403, 2018.

[8] B. K. Beaulieu-Jones, Z. S. Wu, C. Williams, R. Lee, S. P. Bhavnani, J. B. Byrd, and C. S. Greene. Privacy-preserving generative deep neural networks support clinical data sharing. *Circulation: Cardiovascular Quality and Outcomes*, 12(7):e005122, 2019.

[9] V. Bindschaedler, R. Shokri, and C. A. Gunter. Plausible deniability for privacy-preserving data synthesis. *Proceedings of the VLDB Endowment*, 10(5), 2017.

[10] C. M. Bowen and J. Snoke. Comparative study of differentially private synthetic data algorithms from the nist pscr differential privacy synthetic data challenge. *Journal of Privacy and Confidentiality*, 11(1), 2021.

[11] M. Bun and T. Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016.

[12] M. Cesar and R. Rogers. Bounding, concentrating, and truncating: Unifying privacy loss composition for data analytics. In *Algorithmic Learning Theory*, pages 421–457. PMLR, 2021.

[13] A.-S. Charest. How can we analyze differentially-private synthetic datasets? *Journal of Privacy and Confidentiality*, 2(2), 2011.

[14] R. Chen, Q. Xiao, Y. Zhang, and J. Xu. Differentially private high-dimensional data publication via sampling-based inference. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 129–138. ACM, 2015.

[15] C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, 14(3):462–467, 1968.

[16] G. Cormode, C. Procopiuc, D. Srivastava, E. Shen, and T. Yu. Differentially private spatial decompositions. In *Data engineering (ICDE), 2012 IEEE 28th international conference on*, pages 20–31. IEEE, 2012.

[17] B. Ding, M. Winslett, J. Han, and Z. Li. Differentially private data cubes: optimizing noise sources and consistency. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 217–228. ACM, 2011.

[18] I. Dinur and K. Nissim. Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 202–210, 2003.

[19] J. Dong, A. Roth, and W. Su. Gaussian differential privacy. *Journal of the Royal Statistical Society*, 2021.

[20] C. Dwork, F. M. K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284, 2006.

[21] C. Dwork and A. Roth. *The Algorithmic Foundations of Differential Privacy*. Found. and Trends in Theoretical Computer Science, 2014.

[22] V. Feldman, I. Mironov, K. Talwar, and A. Thakurta. Privacy amplification by iteration. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 521–532. IEEE, 2018.

[23] L. Frigerio, A. S. de Oliveira, L. Gomez, and P. Duverger. Differentially private generative adversarial networks for time series, continuous, and discrete open data. In *IFIP International Conference on ICT Systems Security and Privacy Protection*, pages 151–164. Springer, 2019.

[24] M. Gaboardi, E. J. G. Arias, J. Hsu, A. Roth, and Z. S. Wu. Dual query: Practical private query release for high dimensional data. In *International Conference on Machine Learning*, pages 1170–1178. PMLR, 2014.

[25] S. Garfinkel, J. M. Abowd, and C. Martindale. Understanding database reconstruction attacks on public data. *Communications of the ACM*, 62(3):46–53, 2019.

[26] C. Ge, S. Mohapatra, X. He, and I. F. Ilyas. Kamino: Constraint-aware differentially private data synthesis. *Proceedings of the VLDB Endowment*, 14(3), 2020.

[27] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014.

[28] M. Hardt, K. Ligett, and F. Mcsherry. A simple and practical algorithm for differentially private data release. *Advances in Neural Information Processing Systems*, 25:2339–2347, 2012.

[29] M. Hay, V. Rastogi, G. Miklau, and D. Suciu. Boosting the accuracy of differentially private histograms through consistency. *PVLDB*, 3(1-2):1021–1032, 2010.

[30] M. Hittmeir, A. Ekelhart, and R. Mayer. On the utility of synthetic data: an empirical evaluation on machine learning tasks. In *Proceedings of the 14th International Conference on Availability, Reliability and Security*, pages 1–6, 2019.

[31] Z. Huang, R. McKenna, G. Bissias, G. Miklau, M. Hay, and A. Machanavajjhala. Psyndb: accurate and accessible private data generation. *Proceedings of the VLDB Endowment (Demo)*, 12(12):1918–1921, 2019.

[32] J. Jordon, J. Yoon, and M. Van Der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations*, 2018.

[33] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[34] J. B. Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical society*, 7(1):48–50, 1956.

[35] C. Li, M. Hay, G. Miklau, and Y. Wang. A data-and workload-aware algorithm for range queries under differential privacy. *PVLDB*, 7(5):341–352, 2014.

[36] C. Li, M. Hay, V. Rastogi, G. Miklau, and A. McGregor. Optimizing linear counting queries under differential privacy. In *Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 123–134. ACM, 2010.

[37] C. Li and G. Miklau. An adaptive mechanism for accurate query answering under differential privacy. *PVLDB*, 5(6):514–525, 2012.

[38] C. Li, G. Miklau, M. Hay, A. McGregor, and V. Rastogi. The matrix mechanism: optimizing linear counting queries under differential privacy. *The VLDB Journal*, 24(6):757–781, 2015.

[39] H. Li, L. Xiong, and X. Jiang. Differentially private synthesization of multi-dimensional data using copula functions. In *Advances in database technology: proceedings. International Conference on Extending Database Technology*, volume 2014, page 475. NIH Public Access, 2014.

[40] F. Liu. Model-based differentially private data synthesis. *arXiv preprint arXiv:1606.08052*, 2016.

[41] R. McKenna, G. Miklau, M. Hay, and A. Machanavajjhala. Optimizing error of high-dimensional statistical queries under differential privacy. *Proceedings of the VLDB Endowment*, 11(10):1206–1219, 2018.

[42] R. McKenna, D. Sheldon, and G. Miklau. Graphical-model based estimation and inference for differential privacy. In *International Conference on Machine Learning*, pages 4435–4444, 2019.

[43] R. McKenna and D. R. Sheldon. Permute-and-flip: A new mechanism for differentially private selection. *Advances in Neural Information Processing Systems*, 33, 2020.

[44] F. McSherry and K. Talwar. Mechanism design via differential privacy. In *FOCS*, 2007.

[45] I. Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE, 2017.

[46] W. Qardaji, W. Yang, and N. Li. Differentially private grids for geospatial data. In *Intl. Conference on Data Engineering (ICDE)*, pages 757–768. IEEE, 2013.

[47] W. Qardaji, W. Yang, and N. Li. Understanding hierarchical methods for differentially private histograms. *PVLDB*, 6(14):1954–1965, 2013.

[48] W. Qardaji, W. Yang, and N. Li. Priview: practical differentially private release of marginal contingency tables. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 1435–1446. ACM, 2014.

[49] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *NIPS*, 2016.

[50] J. Snoke, G. M. Raab, B. Nowok, C. Dibben, A. Slavkovic, et al. General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society Series A*, 181(3):663–688, 2018.

[51] D. M. Sommer, S. Meiser, and E. Mohammadi. Privacy loss classes: The central limit theorem in differential privacy. *Proceedings on privacy enhancing technologies*, 2019(2):245–269, 2019.

[52] U. Tantipongpipat, C. Waites, D. Boob, A. A. Siva, and R. Cummings. Differentially private mixed-type data generation for unsupervised learning. *arXiv preprint arXiv:1912.03250*, 2019.

[53] A. Torfi, E. A. Fox, and C. K. Reddy. Differentially private synthetic medical data generation using convolutional gans. *arXiv preprint arXiv:2012.11774*, 2020.

[54] R. Torkzadehmahani, P. Kairouz, and B. Paten. Dp-cgan: Differentially private synthetic data and label generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[55] G. Vietri, G. Tian, M. Bun, T. Steinke, and S. Wu. New oracle-efficient algorithms for private synthetic data release. In *International Conference on Machine Learning*, pages 9765–9774. PMLR, 2020.

[56] www.nist.gov. 2018 differential privacy synthetic data challenge. https://www.nist.gov/ctl/pscr/open-innovation-prize-challenges/past-prize-challenges/2018-differential-privacy-synthetic, 2018.

[57] X. Xiao, G. Wang, and J. Gehrke. Differential privacy via wavelet transforms. *IEEE Transactions on Knowledge and Data Engineering*, 23(8):1200–1214, 2011.

[58] Y. Xiao, L. Xiong, L. Fan, S. Goryczka, and H. Li. DPCube: Differentially private histogram release through multidimensional partitioning. *Transactions of Data Privacy*, 7(3), 2014.

[59] L. Xie, K. Lin, S. Wang, F. Wang, and J. Zhou. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*, 2018.

[60] C. Xu, J. Ren, Y. Zhang, Z. Qin, and K. Ren. Dppro: Differentially private high-dimensional data release via random projection. *IEEE Transactions on Information Forensics and Security*, 12(12):3081–3093, 2017.

[61] J. Xu, Z. Zhang, X. Xiao, Y. Yang, and G. Yu. Differentially private histogram publication. In *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*, pages 32–43, 2012.

[62] J. Xu, Z. Zhang, X. Xiao, Y. Yang, G. Yu, and M. Winslett. Differentially private histogram publication. *The VLDB Journal*, pages 1–26, 2013.

[63] G. Yaroslavtsev, G. Cormode, C. M. Procopiuc, and D. Srivastava. Accurate and efficient private release of datacubes and contingency tables. In *ICDE*, 2013.

[64] G. Yuan, Y. Yang, Z. Zhang, and Z. Hao. Convex optimization for linear query processing under approximate differential privacy. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2005–2014. ACM, 2016.

[65] G. Yuan, Z. Zhang, M. Winslett, X. Xiao, Y. Yang, and Z. Hao. Low-rank mechanism: optimizing batch queries under differential privacy. *PVLDB*, 5(11):1352–1363, 2012.

[66] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao. Privbayes: Private data release via bayesian networks. *ACM Transactions on Database Systems (TODS)*, 42(4):1–41, 2017.

[67] J. Zhang, X. Xiao, and X. Xie. Privtree: A differentially private algorithm for hierarchical decompositions. In *SIGMOD*, 2016.

[68] W. Zhang, J. Zhao, F. Wei, and Y. Chen. Differentially private high-dimensional data publication via markov network. *EAI Endorsed Transactions on Security and Safety*, 6(19), 2019.

[69] X. Zhang, R. Chen, J. Xu, X. Meng, and Y. Xie. Towards accurate histogram publication under differential privacy. In *SDM*, 2014.

[70] X. Zhang, S. Ji, and T. Wang. Differentially private releasing via deep generative model (technical report). *arXiv preprint arXiv:1801.01594*, 2018.

[71] Z. Zhang, T. Wang, N. Li, J. Honorio, M. Backes, S. He, J. Chen, and Y. Zhang. Privsyn: Differentially private data synthesis. *arXiv preprint arXiv:2012.15128*, 2020.

[72] J. Zhao, T. Wang, T. Bai, K.-Y. Lam, Z. Xu, S. Shi, X. Ren, X. Yang, Y. Liu, and H. Yu. Reviewing and improving the gaussian mechanism for differential privacy. *arXiv preprint arXiv:1911.12060*, 2019.

Noise Calibration for MST. We begin by calculating our total RDP budget from (ϵ, δ) by invoking Proposition 4. In particular, we find the largest value of ρ such that $(\alpha, \alpha\rho)$-RDP implies (ϵ, δ)-DP by Proposition 4. We accomplish this numerically. We will now divide our RDP budget "ρ" equally among the three steps (4) and (5) and (6) of MST. To achieve $\frac{\rho}{3}$-RDP in steps (4) and (6), it suffices to set $\sigma = \sqrt{\frac{3}{2\rho}}$ by Proposition 1. To achieve $\frac{\rho}{3}$-RDP in step (5), we simply call Algorithm 6 with privacy parameter $\frac{\rho}{3}$. The correctness of this is proven in Theorem 3. The entire algorithm satisfies ρ-RDP by three-fold composition (Proposition 3), which translates to (ϵ, δ)-DP by Proposition 4.

```
In [1]:  from sympy import *

         α,ε,δ,σ = symbols("alpha epsilon delta sigma")
         expr = α/σ**2 + log(1/δ)/(α - 1)
         expr = expr.subs(α, 1 + σ*sqrt(log(1/δ)))
         expr = expr.subs(σ,(sqrt(log(1/δ))+sqrt(log(1/δ)+ε))/ε)
         expr.simplify()

Out[1]:  ε
```

Figure 4. `sympy` code to augment the proof of Theorem 2

Additional Experiment on MST vs. NIST-MST. In Section 7, we evaluated MST in the context of the challenge test suite and evaluation metrics, and found that it performed comparably to NIST-MST. In this section, we compare the quality of the marginals selected from public data (Algorithm 5) and from the sensitive data (Algorithm 6) using various privacy budgets. Specifically, we logged the marginals selected by MST, and compared them to the publicly chosen marginals used in NIST-MST (ignoring the extra 2- and 3-way marginals selected in step 4) using the mutual information criteria described in Algorithm 5. We also show the scores that would be achieved by the best marginals (i.e., the true maximum spanning tree), as well as the scores that would be achieved by a random spanning tree. As shown in the table below, both the publicly chosen marginals and the privately chosen marginals nearly match the score acheived by the best marginals, for both the Arizona and Vermont datasets. The publicly chosen marginals are slightly better at $\epsilon = 0.3$ and $\epsilon = 1.0$, and slightly worse at $\epsilon = 8.0$. Both variants are much better than the random baseline.

Selected Marginals	Arizona	Vermont
Best	73.64	63.32
Public (Algorithm 5)	72.81	62.79
$\epsilon = 0.3$ (Algorithm 6)	70.53	61.15
$\epsilon = 1.0$ (Algorithm 6)	72.27	62.35
$\epsilon = 8.0$ (Algorithm 6)	73.05	63.09
Random	8.55	6.95

Algorithm 7: Synthetic column

Input: μ (vector of fractional counts), n (total number of records to generate)
Output: column (synthetic column of data)

(1) Generate $\lfloor \mu_t \rfloor$ items with value t and add to column for each t in domain

(2) Calculate remainders, $p_t = \mu_t - \lfloor \mu_t \rfloor$

(3) Sample $n - \sum_t \lfloor \mu_t \rfloor$ items (without replacement) from distribution proportional to p_t, and add to column

(4) Shuffle values in column

Algorithm 8: Synthetic data generation

Input: graphical model
Output: dataset (synthetic dataset)

(1) Initialize the set of processed attributes to the empty set

(2) For each attribute i

(3)   Let C be the set of all neighbors of i in the graphical model, intersected with the set of processed attributes

(4)   Group data by C, and for each group in C

(5)   Calculate μ from the graphical model, the vector of fractional counts for every possible value of attribute i, for the given group of other attributes

(6)   Generate synthetic column for this group using Algorithm 7

(7)   Add this partial column to the grouped rows in the dataset

(8)   Add i to the set of processed attributes

Algorithm 9: Reverse transformation of Algorithm 3

Input: D (sensitive dataset)
Output: D (transformed sensitive dataset)

(1) Compute `DIGITS` from `INCWAGE`$_\text{B}$ using Algorithm 10

(2) Replace `VALUEH` and `INCWAGE` attributes in D using transformations:

$$\texttt{VALUEH} = \begin{cases} 5 \cdot \texttt{VALUEH} & \texttt{VALUEH} \leq 5000 \\ 9{,}999{,}998 & \texttt{VALUEH} = 5001 \\ 9{,}999{,}999 & \texttt{VALUEH} = 5002 \end{cases} \qquad \texttt{INCWAGE} = \begin{cases} 100 \cdot \texttt{INCWAGE}_\text{A} + \texttt{DIGITS} & \texttt{INCWAGE}_\text{A} \leq 50 \\ 9{,}999{,}998 & \texttt{INCWAGE}_\text{A} = 51 \end{cases}$$

Algorithm 10: Convert `INCWAGE`$_\text{B}$ to `DIGITS`

Input: k (a value for `INCWAGE`$_\text{B}$)
Output: l (a value for `DIGITS`)

(1) Let $L = \{0, \dots, 99\}$

(2) Let $m = [100, 20, 50, 25, 10, 5, 2, 1]$

(3) For $i = 0, \dots, k$

(4)   Let $L_i = \{l \in L \mid l \equiv 0 \mod m_i\}$

(5)   Let $L = L \setminus L_i$

(6) Sample l uniformly from L_k

Week 9: Ethical Frameworks

Privacy and human behavior in the age of information

Alessandro Acquisti,[1]* Laura Brandimarte,[1] George Loewenstein[2]

This Review summarizes and draws connections between diverse streams of empirical research on privacy behavior. We use three themes to connect insights from social and behavioral sciences: people's uncertainty about the consequences of privacy-related behaviors and their own preferences over those consequences; the context-dependence of people's concern, or lack thereof, about privacy; and the degree to which privacy concerns are malleable—manipulable by commercial and governmental interests. Organizing our discussion by these themes, we offer observations concerning the role of public policy in the protection of privacy in the information age.

I f this is the age of information, then privacy is the issue of our times. Activities that were once private or shared with the few now leave trails of data that expose our interests, traits, beliefs, and intentions. We communicate using e-mails, texts, and social media; find partners on dating sites; learn via online courses; seek responses to mundane and sensitive questions using search engines; read news and books in the cloud; navigate streets with geotracking systems; and celebrate our newborns, and mourn our dead, on social media profiles. Through these and other activities, we reveal information—both knowingly and unwittingly—to one another, to commercial entities, and to our governments. The monitoring of personal information is ubiquitous; its storage is so durable as to render one's past undeletable (1)—a modern digital skeleton in the closet. Accompanying the acceleration in data collection are steady advancements in the ability to aggregate, analyze, and draw sensitive inferences from individuals' data (2).

Both firms and individuals can benefit from the sharing of once hidden data and from the application of increasingly sophisticated analytics to larger and more interconnected databases (3). So too can society as a whole—for instance, when electronic medical records are combined to observe novel drug interactions (4). On the other hand, the potential for personal data to be abused—for economic and social discrimination, hidden influence and manipulation, coercion, or censorship—is alarming. The erosion of privacy can threaten our autonomy, not merely as consumers but as citizens (5). Sharing more personal data does not necessarily always translate into more progress, efficiency, or equality (6).

Because of the seismic nature of these developments, there has been considerable debate about individuals' ability to navigate a rapidly evolving privacy landscape, and about what, if anything, should be done about privacy at a policy level. Some trust people's ability to make self-interested

[1]H. John Heinz III College, Carnegie Mellon University, Pittsburgh, PA, USA. [2]Dietrich College, Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA, USA.
*Corresponding author. E-mail: acquisti@andrew.cmu.edu

decisions about information disclosing and withholding. Those holding this view tend to see regulatory protection of privacy as interfering with the fundamentally benign trajectory of information technologies and the benefits such technologies may unlock (7). Others are concerned about the ability of individuals to manage privacy amid increasingly complex trade-offs. Traditional tools for privacy decision-making such as choice and consent, according to this perspective, no longer provide adequate protection (8). Instead of individual responsibility, regulatory intervention may be needed to balance the interests of the subjects of data against the power of commercial entities and governments holding that data.

Are individuals up to the challenge of navigating privacy in the information age? To address this question, we review diverse streams of empirical privacy research from the social and behavioral sciences. We highlight factors that influence decisions to protect or surrender privacy and how, in turn, privacy protections or violations affect people's behavior. Information technologies have progressively encroached on every aspect of our personal and professional lives. Thus, the problem of control over personal data has become inextricably linked to problems of personal choice, autonomy, and socioeconomic power. Accordingly, this Review focuses on the concept of, and literature around, informational privacy (that is, privacy of personal data) but also touches on other conceptions of privacy, such as anonymity or seclusion. Such notions all ultimately relate to the permeable yet pivotal boundaries between public and private (9).

We use three themes to organize and draw connections between streams of privacy research that, in many cases, have unfolded independently. The first theme is people's uncertainty about the nature of privacy trade-offs, and their own preferences over them. The second is the powerful context-dependence of privacy preferences: The same person can in some situations be oblivious to, but in other situations be acutely concerned about, issues of privacy. The third theme is the malleability of privacy preferences, by which we mean that privacy preferences are subject to

influence by those possessing greater insight into their determinants. Although most individuals are probably unaware of the diverse influences on their concern about privacy, entities whose interests depend on information revelation by others are not. The manipulation of subtle factors that activate or suppress privacy concern can be seen in myriad realms—such as the choice of sharing defaults on social networks, or the provision of greater control on social media—which creates an illusion of safety and encourages greater sharing.

Uncertainty, context-dependence, and malleability are closely connected. Context-dependence is amplified by uncertainty. Because people are often "at sea" when it comes to the consequences of, and their feelings about, privacy, they cast around for cues to guide their behavior. Privacy preferences and behaviors are, in turn, malleable and subject to influence in large part because they are context-dependent and because those with an interest in information divulgence are able to manipulate context to their advantage.

Uncertainty

Individuals manage the boundaries between their private and public spheres in numerous ways: via separateness, reserve, or anonymity (10); by protecting personal information; but also through deception and dissimulation (11). People establish such boundaries for many reasons, including the need for intimacy and psychological respite and the desire for protection from social influence and control (12). Sometimes, these motivations are so visceral and primal that privacy-seeking behavior emerges swiftly and naturally. This is often the case when physical privacy is intruded—such as when a stranger encroaches in one's personal space (13–15) or demonstratively eavesdrops on a conversation. However, at other times (often including when informational privacy is at stake) people experience considerable uncertainty about whether, and to what degree, they should be concerned about privacy.

A first and most obvious source of privacy uncertainty arises from incomplete and asymmetric information. Advancements in information technology have made the collection and usage of personal data often invisible. As a result, individuals rarely have clear knowledge of what information other people, firms, and governments have about them or how that information is used and with what consequences. To the extent that people lack such information, or are aware of their ignorance, they are likely to be uncertain about how much information to share.

Two factors exacerbate the difficulty of ascertaining the potential consequences of privacy behavior. First, whereas some privacy harms are tangible, such as the financial costs associated with identity theft, many others, such as having strangers become aware of one's life history, are intangible. Second, privacy is rarely an unalloyed good; it typically involves trade-offs (16). For example, ensuring the privacy of a consumer's

purchases may protect her from price discrimination but also deny her the potential benefits of targeted offers and advertisements.

Elements that mitigate one or both of these exacerbating factors, by either increasing the tangibility of privacy harms or making trade-offs explicit and simple to understand, will generally affect privacy-related decisions. This is illustrated by one laboratory experiment in which participants were asked to use a specially designed search engine to find online merchants and purchase from them, with their own credit cards, either a set of batteries or a sex toy (17). When the search engine only provided links to the merchants' sites and a comparison of the products' prices from the different sellers, a majority of participants did not pay any attention to the merchants' privacy policies; they purchased from those offering the lowest price. However, when the search engine also provided participants with salient, easily accessible information about the differences in privacy protection afforded by the various merchants, a majority of participants paid a roughly 5% premium to buy products from (and share their credit card information with) more privacy-protecting merchants.

A second source of privacy uncertainty relates to preferences. Even when aware of the consequences of privacy decisions, people are still likely to be uncertain about their own privacy preferences. Research on preference uncertainty (18) shows that individuals often have little sense of how much they like goods, services, or other people. Privacy does not seem to be an exception. This can be illustrated by research in which people were asked sensitive and potentially incriminating questions either point-blank, or followed by credible assurances of confidentiality (19). Although logically such assurances should lead to greater divulgence, they often had the opposite effect because they elevated respondents' privacy concerns, which without assurances would have remained dormant.

The remarkable uncertainty of privacy preferences comes into play in efforts to measure individual and group differences in preference for privacy (20). For example, Westin (21) famously used broad (that is, not contextually specific) privacy questions in surveys to cluster individuals into privacy segments: privacy fundamentalists, pragmatists, and unconcerned. When asked directly, many people fall in the first segment: They profess to care a lot about privacy and express particular concern over losing control of their personal information or others gaining unauthorized access to it (22, 23). However, doubts about the power of attitudinal scales to predict actual privacy behavior arose early in the literature (24). This discrepancy between attitudes and behaviors has become known as the "privacy paradox."

In one early study illustrating the paradox, participants were first classified into categories of privacy concern inspired by Westin's categorization based on their responses to a survey dealing with attitudes toward sharing data (25). Next, they were presented with products

to purchase at a discount with the assistance of an anthropomorphic shopping agent. Few, regardless of the group they were categorized in, exhibited much reluctance to answering the increasingly sensitive questions the agent plied them with.

Why do people who claim to care about privacy often show little concern about it in their daily behavior? One possibility is that the paradox is illusory—that privacy attitudes, which are defined broadly, and intentions and behaviors, which are defined narrowly, should not be expected to be closely related (26, 27). Thus, one might care deeply about privacy in general but, depending on the costs and benefits prevailing in a specific situation, seek or not seek privacy protection (28) .

This explanation for the privacy paradox, however, is not entirely satisfactory for two reasons. The first is that it fails to account for situations in which attitude-behavior dichotomies arise under high correspondence between expressed concerns and behavioral actions. For example, one study compared attitudinal survey answers to actual social media behavior (29). Even within the subset of participants who expressed the highest degree of concern over strangers being able to easily find out their sexual orientation, political views, and partners' names, 48% did in fact publicly reveal their sexual orientation online, 47% revealed their political orientation, and 21% revealed their current partner's name. The second reason is that privacy decision-making is only in part the result of a rational "calculus" of costs and benefits (16, 28); it is also affected by misperceptions of those costs and benefits, as well as social norms, emotions, and heuristics. Any of these factors may affect behavior differently from how they affect attitudes. For instance, present-bias can cause even the privacy-conscious to engage in risky revelations of information, if the immediate gratification from disclosure trumps the delayed, and hence discounted, future consequences (30).

Preference uncertainty is evident not only in studies that compare stated attitudes with behaviors, but also in those that estimate monetary valuations of privacy. "Explicit" investigations ask people to make direct trade-offs, typically between privacy of data and money. For instance, in a study conducted both in Singapore and the United States, students made a series of hypothetical choices about sharing information with websites that differed in protection of personal information and prices for accessing services (31). Using conjoint analysis, the authors concluded that subjects valued protection against errors, improper access, and secondary use of personal information between $30.49 and $44.62. Similar to direct questions about attitudes and intentions, such explicit investigations of privacy valuation spotlight privacy as an issue that respondents should take account of and, as a result, increase the weight they place on privacy in their responses.

Implicit investigations, in contrast, infer valuations of privacy from day-to-day decisions in

which privacy is only one of many considerations and is typically not highlighted. Individuals engage in privacy-related transactions all the time, even when the privacy trade-offs may be intangible or when the exchange of personal data may not be a visible or primary component of a transaction. For instance, completing a query on a search engine is akin to selling personal data (one's preferences and contextual interests) to the engine in exchange for a service (search results). "Revealed preference" economic arguments would then conclude that because technologies for information sharing have been enormously successful, whereas technologies for information protection have not, individuals hold overall low valuations of privacy. However, that is not always the case: Although individuals at times give up personal data for small benefits or discounts, at other times they voluntarily incur substantial costs to protect their privacy. Context, as further discussed in the next section, matters.

In fact, attempts to pinpoint exact valuations that people assign to privacy may be misguided, as suggested by research calling into question the stability, and hence validity, of privacy estimates. In one field experiment inspired by the literature on endowment effects (32), shoppers at a mall were offered gift cards for participating in a non-sensitive survey. The cards could be used online or in stores, just like debit cards. Participants were given either a $10 "anonymous" gift card (transactions done with that card would not be traceable to the subject) or a $12 trackable card (transactions done with that card would be linked to the name of the subject). Initially, half of the participants were given one type of card, and half the other. Then, they were all offered the opportunity to switch. Some shoppers, for example, were given the anonymous $10 card and were asked whether they would accept $2 to "allow my name to be linked to transactions done with the card"; other subjects were asked whether they would accept a card with $2 less value to "prevent my name from being linked to transactions done with the card." Of the subjects who originally held the less valuable but anonymous card, five times as many (52.1%) chose it and kept it over the other card than did those who originally held the more valuable card (9.7%). This suggests that people value privacy more when they have it than when they do not.

The consistency of preferences for privacy is also complicated by the existence of a powerful countervailing motivation: the desire to be public, share, and disclose. Humans are social animals, and information sharing is a central feature of human connection. Social penetration theory (33) suggests that progressively increasing levels of self-disclosure are an essential feature of the natural and desirable evolution of interpersonal relationships from superficial to intimate. Such a progression is only possible when people begin social interactions with a baseline level of privacy. Paradoxically, therefore, privacy provides an essential foundation for intimate disclosure. Similar to privacy, self-disclosure confers numerous objective and subjective benefits, including psychological

and physical health (*34*, *35*). The desire for interaction, socialization, disclosure, and recognition or fame (and, conversely, the fear of anonymous unimportance) are human motives no less fundamental than the need for privacy. The electronic media of the current age provide unprecedented opportunities for acting on them. Through social media, disclosures can build social capital, increase self-esteem (*36*), and fulfill ego needs (*37*). In a series of functional magnetic resonance imaging experiments, self-disclosure was even found to engage neural mechanisms associated with reward; people highly value the ability to share thoughts and feelings with others. Indeed, subjects in one of the experiments were willing to forgo money in order to disclose about themselves (*38*).

Context-dependence

Much evidence suggests that privacy is a universal human need (Box 1) (*39*). However, when people are uncertain about their preferences they often search for cues in their environment to provide guidance. And because cues are a function of context, behavior is as well. Applied to privacy, context-dependence means that individuals can, depending on the situation, exhibit anything ranging from extreme concern to apathy about privacy. Adopting the terminology of Westin, we are all privacy pragmatists, privacy fundamentalists, or privacy unconcerned, depending on time and place (*40*).

The way we construe and negotiate public and private spheres is context-dependent because the boundaries between the two are murky (*41*): The rules people follow for managing privacy vary by situation, are learned over time, and are based on cultural, motivational, and purely situational criteria. For instance, usually we may be more comfortable sharing secrets with friends, but at times we may reveal surprisingly personal information to a stranger on a plane (*42*). The theory of contextual "integrity" posits that social expectations affect our beliefs regarding what is private and what is public, and that such expectations vary with specific contexts (*43*). Thus, seeking privacy in public is not a contradiction; individuals can manage privacy even while sharing information, and even on social media (*44*). For instance, a longitudinal study of actual disclosure behavior of online social network users highlighted that over time, many users increased the amount of personal information revealed to their friends (those connected to them on the network) while simultaneously decreasing the amounts revealed to strangers (those unconnected to them) (Fig. 1) (*45*).

The cues that people use to judge the importance of privacy sometimes result in sensible behavior. For instance, the presence of government regulation has been shown to reduce consumer concern and increase trust; it is a cue that people use to infer the existence of some degree of privacy protection (*46*). In other situations, however, cues can be unrelated, or even negatively related,
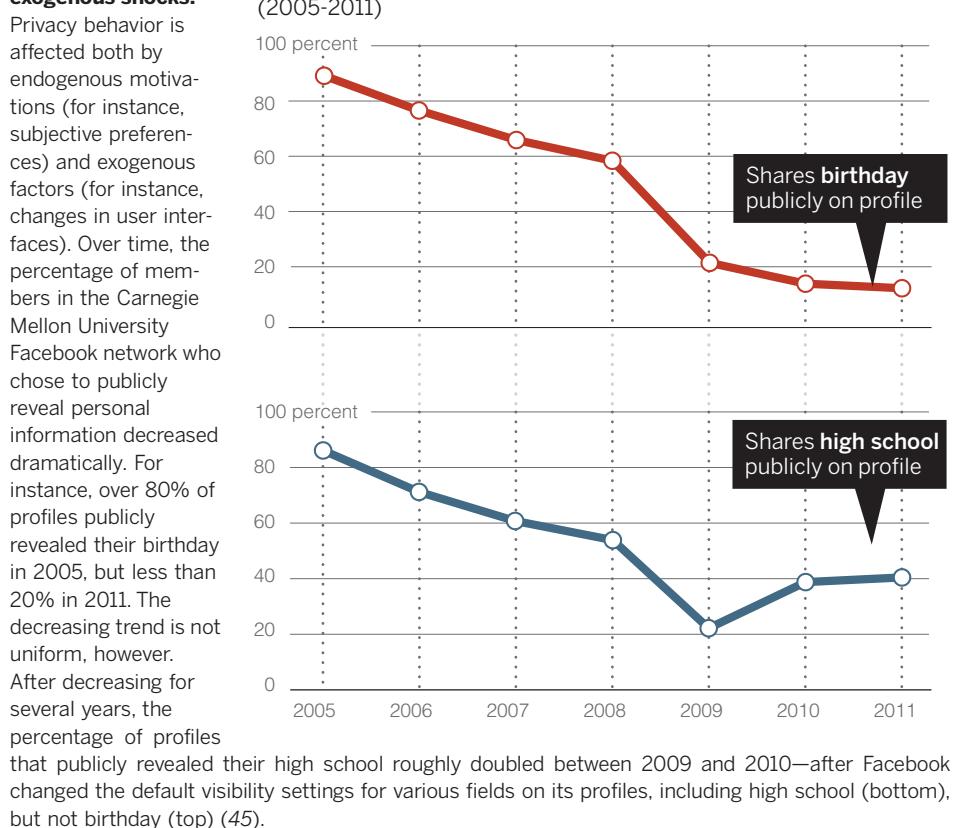
to normative bases of decision-making. For example, in one online experiment (*47*) individuals were more likely to reveal personal and even incriminating information on a website with an unprofessional and casual design with the banner "How Bad R U" than on a site with a formal interface—even though the site with the formal interface was judged by other respondents to be much safer (Fig. 2). Yet in other situations, it is the physical environment that influences privacy concern and associated behavior (*48*), sometimes even unconsciously. For instance, all else being equal, intimacy of self-disclosure is higher in warm, comfortable rooms, with soft lighting, than in cold rooms with bare cement and overhead fluorescent lighting (*49*).

Some of the cues that influence perceptions of privacy are one's culture and the behavior of other people, either through the mechanism of descriptive norms (imitation) or via reciprocity (*50*). Observing other people reveal information increases the likelihood that one will reveal it oneself (*51*). In one study, survey-takers were asked a series of sensitive personal questions regarding their engagement in illegal or ethically questionable behaviors. After answering each question, participants were provided with information, manipulated unbeknownst to them, about the percentage of other participants who in the same survey had admitted to having engaged in a given behavior. Being provided with information that suggested that a majority of survey takers had admitted a certain questionable behavior increased participants' willingness to disclose their engagement in other, also sensitive, behaviors. Other studies have found that the tendency to reciprocate information disclosure is so ingrained that people will reveal more information even to a computer agent that provides information about itself (*52*). Findings such as this may help to explain the escalating amounts of self-disclosure we witness online: If others are doing it, people seem to reason unconsciously, doing so oneself must be desirable or safe.

Other people's behavior affects privacy concerns in other ways, too. Sharing personal information with others makes them "co-owners" of that information (*53*) and, as such, responsible for its protection. Mismanagement of shared information by one or more co-owners causes "turbulence" of the privacy boundaries and, consequently, negative reactions, including anger or mistrust. In a study of undergraduate Facebook users (*54*), for instance, turbulence of privacy boundaries, as a result of having one's profile exposed to unintended audiences, dramatically increased the odds that a user would restrict profile visibility to friends-only.

Likewise, privacy concerns are often a function of past experiences. When something in an environment changes, such as the introduction of a camera or other monitoring devices, privacy concern is likely to be activated. For instance, surveillance can produce discomfort (*55*) and negatively affect worker productivity (*56*). However, privacy concern, like other motivations, is adaptive; people get used to levels of intrusion that do not

Fig. 1. Endogenous privacy behavior and exogenous shocks. Privacy behavior is affected both by endogenous motivations (for instance, subjective preferences) and exogenous factors (for instance, changes in user interfaces). Over time, the percentage of members in the Carnegie Mellon University Facebook network who chose to publicly reveal personal information decreased dramatically. For instance, over 80% of profiles publicly revealed their birthday in 2005, but less than 20% in 2011. The decreasing trend is not uniform, however. After decreasing for several years, the percentage of profiles that publicly revealed their high school roughly doubled between 2009 and 2010—after Facebook changed the default visibility settings for various fields on its profiles, including high school (bottom), but not birthday (top) (*45*).

Disclosure behavior in online social media

Percentage of profiles publicly revealing information over time (2005-2011)

Shares **birthday** publicly on profile

Shares **high school** publicly on profile

change over time. In an experiment conducted in Helsinki (*57*), the installation of sensing and monitoring technology in households led family members initially to change their behavior, particularly in relation to conversations, nudity, and sex. And yet, if they accidentally performed an activity, such as walking naked into the kitchen in front of the sensors, it seemed to have the effect of "breaking the ice"; participants then showed less concern about repeating the behavior. More generally, participants became inured to the presence of the technology over time.

The context-dependence of privacy concern has major implications for the risks associated with modern information and communication technology (*58*). With online interactions, we no longer have a clear sense of the spatial boundaries of our listeners. Who is reading our blog post? Who is looking at our photos online? Adding complexity to privacy decision-making, boundaries between public and private become even less defined in the online world (*59*) where we become social media friends with our coworkers and post pictures to an indistinct flock of followers. With different social groups mixing on the Internet, separating online and offline identities and meeting our and others' expectations regarding privacy becomes more difficult and consequential (*60*).

Malleability and influence

Whereas individuals are often unaware of the diverse factors that determine their concern about privacy in a particular situation, entities whose prosperity depends on information revelation by others are much more sophisticated. With the emergence of the information age, growing institutional and economic interests have developed around disclosure of personal information, from online social networks to behavioral advertising. It is not surprising, therefore, that some entities have an interest in, and have developed expertise in, exploiting behavioral and psychological processes to promote disclosure (*61*). Such efforts play on the malleability of privacy preferences, a term we use to refer to the observation that various, sometimes subtle, factors can be used to activate or suppress privacy concerns, which in turn affect behavior.

Default settings are an important tool used by different entities to affect information disclosure. A large body of research has shown that default settings matter for decisions as important as organ donation and retirement saving (*62*). Sticking to default settings is convenient, and people often interpret default settings as implicit recommendations (*63*). Thus, it is not surprising that default settings for one's profile's visibility on social networks (*64*), or the existence of opt-in or opt-out privacy policies on websites (*65*), affect individuals' privacy behavior (Fig. 3).

In addition to default settings, websites can also use design features that frustrate or even confuse users into disclosing personal information (*66*), a practice that has been referred to as "malicious interface design" (*67*). Another obvious strategy that commercial entities can use to avoid raising privacy concerns is not to "ring alarm bells"

when it comes to data collection. When companies do ring them—for example, by using overly finetuned personalized advertisements—consumers are alerted (*68*) and can respond with negative "reactance" (*69*).

Various so-called "antecedents" (*70*) affect privacy concerns and can be used to influence privacy behavior. For instance, trust in the entity receiving one's personal data soothes concerns. Moreover, because some interventions that are intended to protect privacy can establish trust, con-

cerns can be muted by the very interventions intended to protect privacy. Perversely, 62% of respondents to a survey believed (incorrectly) that the existence of a privacy policy implied that a site could not share their personal information without permission (*40*), which suggests that simply posting a policy that consumers do not read may lead to misplaced feelings of being protected.

Control is another feature that can inculcate trust and produce paradoxical effects. Perhaps because of its lack of controversiality, control has

---
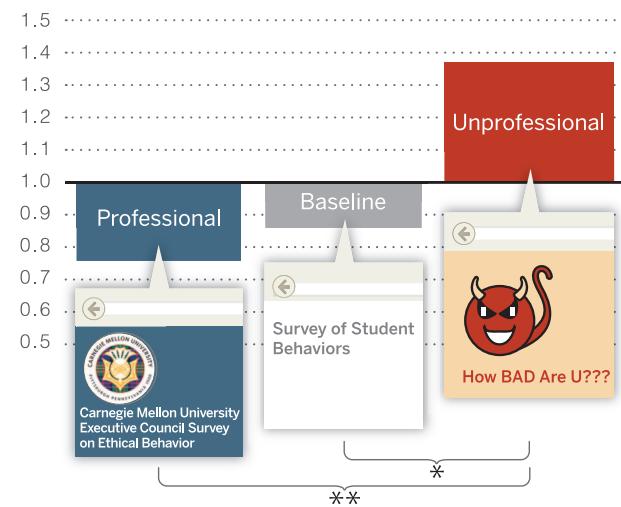
Box 1. Privacy: A modern invention?

Is privacy a modern, bourgeois, and distinctly Western invention? Or are privacy needs a universal feature of human societies? Although access to privacy is certainly affected by socioeconomic factors (*87*) [some have referred to privacy as a "luxury good" (*15*)], and privacy norms greatly differ across cultures (*65, 85*), the need for privacy seems to be a universal human trait. Scholars have uncovered evidence of privacy-seeking behaviors across peoples and cultures separated by time and space: from ancient Rome and Greece (*39, 88*) to preindustrialized Javanese, Balinese, and Tuareg societies (*89, 90*). Privacy, as Altman (*91*) noted, appears to be simultaneously culturally specific and culturally universal. Cues of a common human quest for privacy are also found in the texts of ancient religions: The Quran (49:12) instructs against spying on one another (*92*); the Talmud (Bava Batra 60a) advises home-builders to position windows so that they do not directly face those of one's neighbors (*93*); the Bible (Genesis, 3:7) relates how Adam and Eve discovered their nakedness after eating the fruit of knowledge and covered themselves in shame from the prying eyes of God (*94*) [a discussion of privacy in Confucian and Taoist cultures is available in (*95*)]. Implicit in this heterogeneous selection of historical examples is the observation that there exist multiple notions of privacy. Although contemporary attention focuses on informational privacy, privacy has been also construed as territorial and physical, and linked to concepts as diverse as surveillance, exposure, intrusion, insecurity, appropriation, as well as secrecy, protection, anonymity, dignity, or even freedom [a taxonomy is provided in (*9*)].

Fig. 2. The impact of cues on disclosure behavior. A measure of privacy behavior often used in empirical studies is a subject's willingness to answer personal, sometimes sensitive questions— for instance, by admitting or denying having engaged in questionable behaviors. In an online experiment (*47*), individuals were asked a series of intrusive questions about their behaviors, such as "Have you ever tried to peek at someone else's e-mail without them knowing?" Across conditions, the interface of the questionnaire was manipulated to look more or less professional. The *y* axis captures the mean affirmative admission rates (AARs) to questions that were rated as intrusive (the proportion of questions answered affirmatively) normed, question by ques-

A measure of privacy behavior

Relative admission rates in an experiment testing the impact of different survey interfaces on willingness to answer questions about various sensitive behaviors
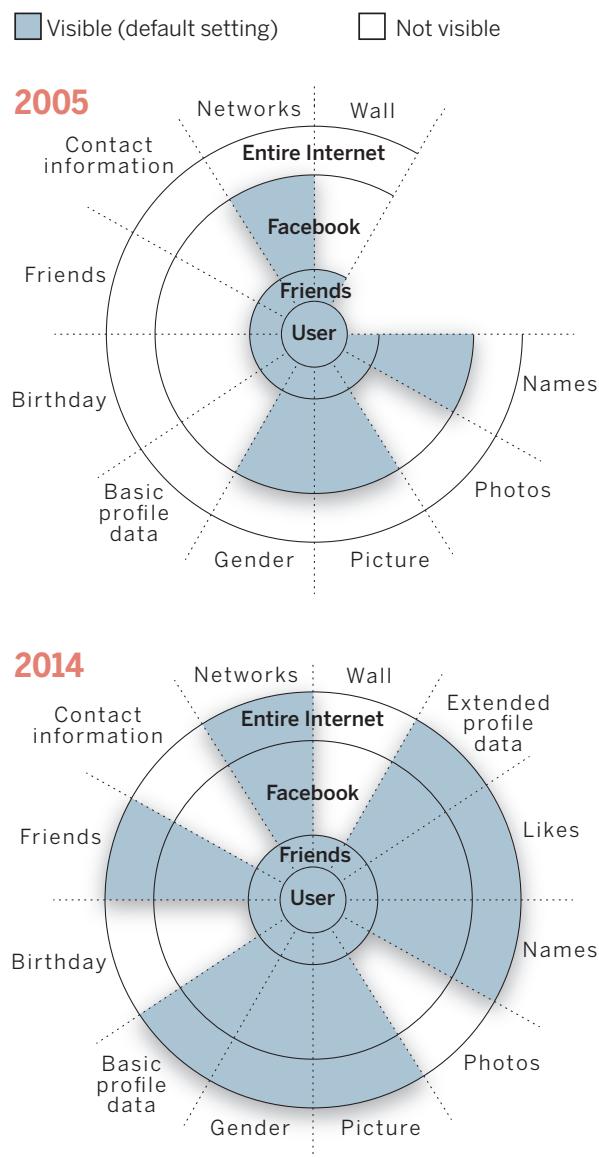
$F_{2,196} = 6.8, P = 0.001$

tion, on the overall average AAR for the question. Subjects revealed more personal and even incriminating information on the website with a more casual design, even though the site with the formal interface was judged by other respondents to be much safer. The study illustrates how cues can influence privacy behavior in a fashion that is unrelated, or even negatively related, to normative bases of decision-making.

Fig. 3. Changes in Facebook default profile visibility settings over time (2005–2014). Over time, Facebook profiles included an increasing amount of fields and, therefore, types of data. In addition, default visibility settings became more revelatory between 2005 (top) and 2014 (bottom), disclosing more personal information to larger audiences, unless the user manually overrode the defaults (fields such as "Likes" and "Extended Profile Data" did not exist in 2005). "Basic profile data" includes hometown, current city, high school, school (status, concentration, secondary concentration), interested in, relationship, workplace, about you, and quotes. Examples of "Extended profile data" include life events such as new job, new school, engagement, expecting a baby, moved, bought a home, and so forth. "Picture" refers to the main profile image. "Photos" refers to the additional images that users might have shared in their account. "Names" refers to the real name, the username, and the user ID. This figure is based on the authors' data and the original visualization created by M. McKeon, available at http://mattmckeon.com/facebook-privacy.

Default visibility settings in social media over time

Legend: ■ Visible (default setting) □ Not visible

2005 — radial chart with axes: Networks, Wall, Names, Photos, Picture, Gender, Basic profile data, Birthday, Friends, Contact information. Rings labeled (center to outer): User, Friends, Facebook, Entire Internet.

2014 — radial chart with axes: Networks, Wall, Extended profile data, Likes, Names, Photos, Picture, Gender, Basic profile data, Birthday, Friends, Contact information. Rings labeled (center to outer): User, Friends, Facebook, Entire Internet.

been one of the capstones of the focus of both industry and policy-makers in attempts to balance privacy needs against the value of sharing. Control over personal information is often perceived as a critical feature of privacy protection (39). In principle, it does provide users with the means to manage access to their personal information. Research, however, shows that control can reduce privacy concern (46), which in turn can have unintended effects. For instance, one study found that participants who were provided with greater explicit control over whether and how much of their personal information researchers could publish ended up sharing more sensitive information with a broader audience—the opposite of the ostensible purpose of providing such control (71).

Similar to the normative perspective on control, increasing the transparency of firms' data practices would seem to be desirable. However, transparency mechanisms can be easily rendered ineffective. Research has highlighted not only that an overwhelming majority of Internet users do not read privacy policies (72), but also that few users would benefit from doing so; nearly half of a sample of online privacy policies were found to be written in language beyond the grasp of most Internet users (73). Indeed, and somewhat amusingly, it has been estimated that the aggregate opportunity cost if U.S. consumers actually read the privacy policies of the sites they visit would be $781 billion/year (74).

Although uncertainty and context-dependence lead naturally to malleability and manipulation, not all malleability is necessarily sinister. Consider monitoring. Although monitoring can cause discomfort and reduce productivity, the feeling of being observed and accountable can induce people to engage in prosocial behaviors or (for better or for worse) adhere to social norms (75). Prosocial behavior can be heightened by monitoring cues as simple as three dots in a stylized face configuration (76). By the same token, the depersonalization induced by computer-mediated interaction (77), either in the form of lack of identifiability or of visual anonymity (78), can have beneficial effects, such as increasing truthful responses to sensitive surveys (79, 80). Whether elevating or suppressing privacy concerns is socially beneficial critically depends, yet again, on context [a meta-analysis of the impact of de-identification on behavior is provided in (81)]. For example, perceptions of anonymity can alternatively lead to dishonest or prosocial behavior. Illusory anonymity induced by darkness caused participants in an experiment (82) to cheat in order to gain more money. This can be interpreted as a form of disinhibition effect (83), by which perceived anonymity licenses people to act in ways that they would otherwise not even consider. In other circumstances, though, anonymity leads to prosocial behavior—for instance, higher willingness to share money in a dictator game, when coupled with priming of religiosity (84).

Conclusions

Norms and behaviors regarding private and public realms greatly differ across cultures (85). Americans, for example, are reputed to be more open about sexual matters than are the Chinese, whereas the latter are more open about financial matters (such as income, cost of home, and possessions). And even within cultures, people differ substantially in how much they care about privacy and what information they treat as private. And as we have sought to highlight in this Review, privacy concerns can vary dramatically for the same individual, and for societies, over time.

If privacy behaviors are culture- and context-dependent, however, the dilemma of what to share and what to keep private is universal across societies and over human history. The task of navigating those boundaries, and the consequences of mismanaging them, have grown increasingly complex and fateful in the information age, to the point that our natural instincts seem not nearly adequate.

In this Review, we used three themes to organize and draw connections between the social and behavioral science literatures on privacy and behavior. We end the Review with a brief discussion of the reviewed literature's relevance to privacy policy.

Uncertainty and context-dependence imply that people cannot always be counted on to navigate the complex trade-offs involving privacy in a self-interested fashion. People are often unaware of the information they are sharing, unaware of how it can be used, and even in the rare situations when they have full knowledge of the consequences of sharing, uncertain about their own preferences. Malleability, in turn, implies that people are easily influenced in what and how much they disclose. Moreover, what they share can be used to influence their emotions, thoughts, and behaviors in many aspects of their lives, as individuals, consumers, and citizens. Although such influence is not always or necessarily malevolent or dangerous, relinquishing control over one's personal data and over one's privacy alters the

balance of power between those holding the data and those who are the subjects of that data.

Insights from the social and behavioral empirical research on privacy reviewed here suggest that policy approaches that rely exclusively on informing or "empowering" the individual are unlikely to provide adequate protection against the risks posed by recent information technologies. Consider transparency and control, two principles conceived as necessary conditions for privacy protection. The research we highlighted shows that they may provide insufficient protections and even backfire when used apart from other principles of privacy protection.

The research reviewed here suggests that if the goal of policy is to adequately protect privacy (as we believe it should be), then we need policies that protect individuals with minimal requirement of informed and rational decision-making—policies that include a baseline framework of protection, such as the principles embedded in the so-called fair information practices (86). People need assistance and even protection to aid in navigating what is otherwise a very uneven playing field. As highlighted by our discussion, a goal of public policy should be to achieve a more even equity of power between individuals, consumers, and citizens on the one hand and, on the other, the data holders such as governments and corporations that currently have the upper hand. To be effective, privacy policy should protect real people—who are naïve, uncertain, and vulnerable—and should be sufficiently flexible to evolve with the emerging unpredictable complexities of the information age.

REFERENCES AND NOTES

1. V. Mayer-Schönberger, *Delete: The Virtue of Forgetting In the Digital Age* (Princeton Univ. Press, Princeton, 2011).
2. L. Sweeney, *Int. J. Uncert. Fuzziness Knowl. Based Syst.* **10**, 557–570 (2002).
3. A. McAfee, E. Brynjolfsson, *Harv. Bus. Rev.* **90**, 60–66, 68, 128 (2012).
4. N. P. Tatonetti, P. P. Ye, R. Daneshjou, R. B. Altman, *Sci. Transl. Med.* **4**, 125ra31 (2012).
5. J. E. Cohen, *Stanford Law Rev.* **52**, 1373–1438 (2000).
6. K. Crawford, K. Miltner, M. L. Gray, *Int. J. Commun.* **8**, 1663–1672 (2014).
7. R. A. Posner, *Am. Econ. Rev.* **71**, 405–409 (1981).
8. D. J. Solove, *Harv. Law Rev.* **126**, 1880–1903 (2013).
9. D. J. Solove, *Univ. Penn. L. Rev.* **154**, 477–564 (2006).
10. F. Schoeman, Ed., *Philosophical dimensions of privacy—An anthology* (Cambridge Univ. Press, New York, 1984).
11. B. M. DePaulo, C. Wetzel, R. Weylin Sternglanz, M. J. W. Wilson, *J. Soc. Issues* **59**, 391–410 (2003).
12. S. T. Margulis, *J. Soc. Issues* **59**, 243–261 (2003).
13. E. Goffman, *Relations in Public: Microstudies of the Public Order* (Harper & Row, New York, 1971).
14. E. Sundstrom, I. Altman, *Hum. Ecol.* **4**, 47–67 (1976).
15. B. Schwartz, *Am. J. Sociol.* **73**, 741–752 (1968).
16. R. S. Laufer, M. Wolfe, *J. Soc. Issues* **33**, 22–42 (1977).
17. J. Y. Tsai, S. Egelman, L. Cranor, A. Acquisti, *Inf. Syst. Res.* **22**, 254–268 (2011).
18. P. Slovic, *Am. Psychol.* **50**, 364–371 (1995).
19. E. Singer, H. Hippler, N. Schwarz, *Int. J. Public Opin. Res.* **4**, 256–268 (1992).
20. V. P. Skotko, D. Langmeyer, *Sociometry* **40**, 178–182 (1977).
21. A. Westin, Harris Louis & Associates, Harris-Equifax Consumer Privacy Survey (Tech. rep. 1991).
22. M. J. Culnan, P. K. Armstrong, *Organ. Sci.* **10**, 104–115 (1999).
23. H. J. Smith, S. J. Milberg, S. J. Burke, *Manage. Inf. Syst. Q.* **20**, 167–196 (1996).
24. B. Lubin, R. L. Harrison, *Psychol. Rep.* **15**, 77–78 (1964).
25. S. Spiekermann, J. Grossklags, B. Berendt, *E-Privacy in 2nd Generation E-Commerce: Privacy Preferences versus Actual Behavior* (Third ACM Conference on Electronic Commerce, Tampa, 2001), pp. 38–47.
26. P. A. Norberg, D. R. Horne, D. A. Horne, *J. Consum. Aff.* **41**, 100–126 (2007).
27. I. Ajzen, M. Fishbein, *Psychol. Bull.* **84**, 888–918 (1977).
28. P. H. Klopfer, D. I. Rubenstein, *J. Soc. Issues* **33**, 52–65 (1977).
29. A. Acquisti, R. Gross, in *Privacy Enhancing Technologies*, G. Danezis, P. Golle Eds. (Springer, New York, 2006), pp. 36–58.
30. A. Acquisti, *Privacy in Electronic Commerce and the Economics of Immediate Gratification* (Fifth ACM Conference on Electronic Commerce, New York, 2004), pp. 21–29.
31. I. Hann, K. Hui, S. T. Lee, I. P. L. Png, *J. Manage. Inf. Syst.* **24**, 13–42 (2007).
32. A. Acquisti, L. K. John, G. Loewenstein, *J. Legal Stud.* **42**, 249–274 (2013).
33. I. Altman, D. Taylor, *Social Penetration: The Development of Interpersonal Relationships* (Holt, Rinehart & Winston, New York, 1973).
34. J. Frattaroli, *Psychol. Bull.* **132**, 823–865 (2006).
35. J. W. Pennebaker, *Behav. Res. Ther.* **31**, 539–548 (1993).
36. C. Steinfield, N. B. Ellison, C. Lampe, *J. Appl. Dev. Psychol.* **29**, 434–445 (2008).
37. C. L. Toma, J. T. Hancock, *Pers. Soc. Psychol. Bull.* **39**, 321–331 (2013).
38. D. I. Tamir, J. P. Mitchell, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 8038–8043 (2012).
39. A. Westin, *Privacy and Freedom* (Athenäum, New York, 1967).
40. C. J. Hoofnagle, J. M. Urban, *Wake Forest Law Rev.* **49**, 261–321 (2014).
41. G. Marx, *Ethics Inf. Technol.* **3**, 157–169 (2001).
42. J. W. Thibaut, H. H. Kelley, *The Social Psychology Of Groups* (Wiley, Oxford, 1959).
43. H. Nissenbaum, *Privacy in Context: Technology, Policy, and the Integrity of Social Life* (Stanford Univ. Press, Redwood City, 2009).
44. d. boyd, *It's Complicated: The Social Lives of Networked Teens* (Yale Univ. Press, New Haven, 2014).
45. F. Stutzman, R. Gross, A. Acquisti, *J. Priv. Confidential.* **4**, 7–41 (2013).
46. H. Xu, H. H. Teo, B. C. Tan, R. Agarwal, *J. Manage. Inf. Syst.* **26**, 135–174 (2009).
47. L. K. John, A. Acquisti, G. Loewenstein, *J. Consum. Res.* **37**, 858–873 (2011).
48. I. Altman, *The Environment and Social Behavior: Privacy, Personal Space, Territory, and Crowding* (Cole, Monterey, 1975).
49. A. L. Chaikin, V. J. Derlega, S. J. Miller, *J. Couns. Psychol.* **23**, 479–481 (1976).
50. V. J. Derlega, A. L. Chaikin, *J. Soc. Issues* **33**, 102–115 (1977).
51. A. Acquisti, L. K. John, G. Loewenstein, *J. Mark. Res.* **49**, 160–174 (2012).
52. Y. Moon, *J. Consum. Res.* **26**, 323–339 (2000).
53. S. Petronio, *Boundaries of Privacy: Dialectics of Disclosure* (SUNY Press, Albany, 2002).
54. F. Stutzman, J. Kramer-Duffield, *Friends Only: Examining a Privacy-Enhancing Behavior in Facebook* (SIGCHI Conference on Human Factors in Computing Systems, ACM, Atlanta, 2010), pp. 1553–1562.
55. T. Honess, E. Charman, *Closed Circuit Television in Public Places: Its Acceptability and Perceived Effectiveness* (Police Research Group, London, 1992).
56. M. Gagné, E. L. Deci, *J. Organ. Behav.* **26**, 331–362 (2005).
57. A. Oulasvirta et al., *Long-Term Effects of Ubiquitous Surveillance in the Home* (ACM Conference on Ubiquitous Computing, Pittsburgh, 2012), pp. 41–50.
58. L. Palen, P. Dourish, *Unpacking "Privacy" For A Networked World* (SIGCHI Conference on Human Factors in Computing Systems, ACM, Fort Lauderdale, 2003), pp. 129–136.
59. Z. Tufekci, *Bull. Sci. Technol. Soc.* **28**, 20–36 (2008).
60. J. A. Bargh, K. Y. A. McKenna, G. M. Fitzsimons, *J. Soc. Issues* **58**, 33–48 (2002).
61. R. Calo, *Geo. Wash. L. Rev.* **82**, 995–1304 (2014).
62. E. J. Johnson, D. Goldstein, *Science* **302**, 1338–1339 (2003).
63. C. R. McKenzie, M. J. Liersch, S. R. Finkelstein, *Psychol. Sci.* **17**, 414–420 (2006).
64. R. Gross, A. Acquisti, *Information Revelation and Privacy in Online Social Networks* (ACM Workshop–Privacy in the Electronic Society, New York, 2005), pp. 71–80.
65. E. J. Johnson, S. Bellman, G. L. Lohse, *Mark. Lett.* **13**, 5–15 (2002).
66. W. Hartzog, *Am. Univ. L. Rev.* **60**, 1635–1671 (2010).
67. G. Conti, E. Sobiesk, *Malicious Interface Design: Exploiting the User* (19th International Conference on World Wide Web, ACM, Raleigh, 2010), pp. 271–280.
68. A. Goldfarb, C. Tucker, *Mark. Sci.* **30**, 389–404 (2011).
69. T. B. White, D. L. Zahay, H. Thorbjørnsen, S. Shavitt, *Mark. Lett.* **19**, 39–50 (2008).
70. H. J. Smith, T. Dinev, H. Xu, *Manage. Inf. Syst. Q.* **35**, 989–1016 (2011).
71. L. Brandimarte, A. Acquisti, G. Loewenstein, *Soc. Psychol. Personal. Sci.* **4**, 340–347 (2013).
72. C. Jensen, C. Potts, C. Jensen, *Int. J. Hum. Comput. Stud.* **63**, 203–227 (2005).
73. C. Jensen, C. Potts, *Privacy Policies as Decision-Making Tools: An Evaluation of Online Privacy Notices* (SIGCHI Conference on Human factors in computing systems, ACM, Vienna, 2004), pp. 471–478.
74. A. M. McDonald, L. F. Cranor, *I/S: J. L. Policy Inf. Society.* **4**, 540–565 (2008).
75. C. Wedekind, M. Milinski, *Science* **288**, 850–852 (2000).
76. M. Rigdon, K. Ishii, M. Watabe, S. Kitayama, *J. Econ. Psychol.* **30**, 358–367 (2009).
77. S. Kiesler, J. Siegel, T. W. McGuire, *Am. Psychol.* **39**, 1123–1134 (1984).
78. A. N. Joinson, *Eur. J. Soc. Psychol.* **31**, 177–192 (2001).
79. S. Weisband, S. Kiesler, *Self Disclosure On Computer Forms: Meta-Analysis And Implications* (SIGCHI Conference Conference on Human Factors in Computing Systems, ACM, Vancouver, 1996), pp. 3–10.
80. R. Tourangeau, T. Yan, *Psychol. Bull.* **133**, 859–883 (2007).
81. T. Postmes, R. Spears, *Psychol. Bull.* **123**, 238–259 (1998).
82. C. B. Zhong, V. K. Bohns, F. Gino, *Psychol. Sci.* **21**, 311–314 (2010).
83. J. Suler, *Cyberpsychol. Behav.* **7**, 321–326 (2004).
84. A. F. Shariff, A. Norenzayan, *Psychol. Sci.* **18**, 803–809 (2007).
85. B. Moore, *Privacy: Studies in Social and Cultural History* (Armonk, New York, 1984).
86. *Records, Computers and the Rights of Citizens* (Secretary's Advisory Committee, U.S. Dept. of Health, Education and Welfare, Washington, DC, 1973).
87. E. Hargittai, in *Social Stratification*, D. Grusky Ed. (Westview, Boulder, 2008), pp. 936-113.
88. P. Ariès, G. Duby (Eds.), *A History of Private Life: From Pagan Rome to Byzantium* (Harvard Univ. Press, Cambridge, 1992).
89. R. F. Murphy, *Am. Anthropol.* **66**, 1257–1274 (1964).
90. A. Westin, in *Philosophical Dimensions of Privacy: An Anthology*, F.D. Schoeman Ed. (Cambridge Univ. Press, Cambridge, UK, 1984), pp. 56–74.
91. I. Altman, *J. Soc. Issues* **33**, 66–84 (1977).
92. M. A. Hayat, *Inf. Comm. Tech. L.* **16**, 137–148 (2007).
93. A. Enkin, "Privacy," www.torahmusings.com/2012/07/privacy (2014).
94. J. Rykwert, *Soc. Res. (New York)* **68**, 29–40 (2001).
95. C. B. Whitman, in *Individualism and Holism: Studies in Confucian and Taoist Values*, D. J. Munro, Ed. (Center for Chinese Studies, Univ. Michigan, Ann Arbor, 1985), pp. 85–100.

THE BELMONT REPORT

Office of the Secretary

Ethical Principles and Guidelines for the Protection of Human Subjects of Research

The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research

April 18, 1979

AGENCY: Department of Health, Education, and Welfare.

ACTION: Notice of Report for Public Comment.

SUMMARY: On July 12, 1974, the National Research Act (Pub. L. 93-348) was signed into law, there-by creating the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. One of the charges to the Commission was to identify the basic ethical principles that should underlie the conduct of biomedical and behavioral research involving human subjects and to develop guidelines which should be followed to assure that such research is conducted in accordance with those principles. In carrying out the above, the Commission was directed to consider: **(i)** the boundaries between biomedical and behavioral research and the accepted and routine practice of medicine, **(ii)** the role of assessment of risk-benefit criteria in the determination of the appropriateness of research involving human subjects, **(iii)** appropriate guidelines for the selection of human subjects for participation in such research and **(iv)** the nature and definition of informed consent in various research settings.

The Belmont Report attempts to summarize the basic ethical principles identified by the Commission in the course of its deliberations. It is the outgrowth of an intensive four-day period of discussions that were held in February 1976 at the Smithsonian Institution's Belmont Conference Center supplemented by the monthly deliberations of the Commission that were held over a period of nearly four years. It is a statement of basic ethical principles and guidelines that should assist in resolving the ethical problems that surround the conduct of research with human subjects. By publishing the Report in the Federal Register, and providing reprints upon request, the Secretary intends that it may be made readily available to scientists, members of Institutional Review Boards, and Federal employees. The two-volume Appendix, containing the lengthy reports of experts and specialists who assisted the Commission in fulfilling this part of its charge, is available as DHEW Publication No. (OS) 78-0013 and No. (OS) 78-0014, for sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402.

Unlike most other reports of the Commission, the Belmont Report does not make specific recommendations for administrative action by the Secretary of Health, Education, and Welfare. Rather, the Commission recommended that the Belmont Report be adopted in its entirety, as a statement of the Department's policy. The Department requests public comment on this recommendation.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research

Members of the Commission

Kenneth John Ryan, M.D., Chairman, Chief of Staff, Boston Hospital for Women.

Joseph V. Brady, Ph.D., Professor of Behavioral Biology, Johns Hopkins University.

Robert E. Cooke, M.D., President, Medical College of Pennsylvania.

Dorothy I. Height, President, National Council of Negro Women, Inc.

Albert R. Jonsen, Ph.D., Associate Professor of Bioethics, University of California at San Francisco.

Patricia King, J.D., Associate Professor of Law, Georgetown University Law Center.

Karen Lebacqz, Ph.D., Associate Professor of Christian Ethics, Pacific School of Religion.

*** David W. Louisell, J.D., Professor of Law, University of California at Berkeley.

Donald W. Seldin, M.D., Professor and Chairman, Department of Internal Medicine, University of Texas at Dallas.

*** Eliot Stellar, Ph.D., Provost of the University and Professor of Physiological Psychology, University of Pennsylvania.

*** Robert H. Turtle, LL.B., Attorney, VomBaur, Coburn, Simmons & Turtle, Washington, D.C.

*** Deceased.

Table of Contents

Ethical Principles & Guidelines for Research Involving Human Subjects

Scientific research has produced substantial social benefits. It has also posed some troubling ethical questions. Public attention was drawn to these questions by reported abuses of human subjects in biomedical experiments, especially during the Second World War. During the Nuremberg War Crime Trials, the Nuremberg code was drafted as a set of standards for judging physicians and scientists who had conducted biomedical experiments on concentration camp prisoners. This code became the prototype of many later codes [1] intended to assure that research involving human subjects would be carried out in an ethical manner.

The codes consist of rules, some general, others specific, that guide the investigators or the reviewers of research in their work. Such rules often are inadequate to cover complex situations; at times they come into conflict, and they are frequently difficult to interpret or apply. Broader ethical principles will provide a basis on which specific rules may be formulated, criticized and interpreted.

Three principles, or general prescriptive judgments, that are relevant to research involving human subjects are identified in this statement. Other principles may also be relevant. These three are comprehensive, however, and are stated at a level of generalization that should assist scientists, subjects, reviewers and interested citizens to understand the ethical issues inherent in research involving human subjects. These principles cannot always be applied so as to resolve beyond dispute particular ethical problems. The objective is to provide an analytical framework that will guide the resolution of ethical problems arising from research involving human subjects.

This statement consists of a distinction between research and practice, a discussion of the three basic ethical principles, and remarks about the application of these principles.

Part A: Boundaries Between Practice & Research

A. Boundaries Between Practice and Research

It is important to distinguish between biomedical and behavioral research, on the one hand, and the practice of accepted therapy on the other, in order to know what activities ought to undergo review for the protection of human subjects of research. The distinction between research and practice is blurred partly because both often occur together (as in research designed to evaluate a therapy) and partly because notable departures from standard practice are often called "experimental" when the terms "experimental" and "research" are not carefully defined.

For the most part, the term "practice" refers to interventions that are designed solely to enhance the well-being of an individual patient or client and that have a reasonable expectation of success. The purpose of medical or behavioral practice is to provide diagnosis, preventive treatment or therapy to particular individuals [2]. By contrast, the term "research' designates an activity designed to test an hypothesis, permit conclusions to be drawn, and thereby to develop or contribute to generalizable knowledge (expressed, for example, in theories, principles, and statements of relationships). Research is usually described in a formal protocol that sets forth an objective and a set of procedures designed to reach that objective.

When a clinician departs in a significant way from standard or accepted practice, the innovation does not, in and of itself, constitute research. The fact that a procedure is "experimental," in the sense of new, untested or different, does not automatically place it in the category of research. Radically new procedures of this

description should, however, be made the object of formal research at an early stage in order to determine whether they are safe and effective. Thus, it is the responsibility of medical practice committees, for example, to insist that a major innovation be incorporated into a formal research project [3].

Research and practice may be carried on together when research is designed to evaluate the safety and efficacy of a therapy. This need not cause any confusion regarding whether or not the activity requires review; the general rule is that if there is any element of research in an activity, that activity should undergo review for the protection of human subjects.

Part B: Basic Ethical Principles

B. Basic Ethical Principles

The expression "basic ethical principles" refers to those general judgments that serve as a basic justification for the many particular ethical prescriptions and evaluations of human actions. Three basic principles, among those generally accepted in our cultural tradition, are particularly relevant to the ethics of research involving human subjects: the principles of respect of persons, beneficence and justice.

1. Respect for Persons. — Respect for persons incorporates at least two ethical convictions: first, that individuals should be treated as autonomous agents, and second, that persons with diminished autonomy are entitled to protection. The principle of respect for persons thus divides into two separate moral requirements: the requirement to acknowledge autonomy and the requirement to protect those with diminished autonomy.

An autonomous person is an individual capable of deliberation about personal goals and of acting under the direction of such deliberation. To respect autonomy is to give weight to autonomous persons' considered opinions and choices while refraining from obstructing their actions unless they are clearly detrimental to others. To show lack of respect for an autonomous agent is to repudiate that person's considered judgments, to deny an individual the freedom to act on those considered judgments, or to withhold information necessary to make a considered judgment, when there are no compelling reasons to do so.

However, not every human being is capable of self-determination. The capacity for self-determination matures during an individual's life, and some individuals lose this capacity wholly or in part because of illness, mental disability, or circumstances that severely restrict liberty. Respect for the immature and the incapacitated may require protecting them as they mature or while they are incapacitated.

Some persons are in need of extensive protection, even to the point of excluding them from activities which may harm them; other persons require little protection beyond making sure they undertake activities freely and with awareness of possible adverse consequence. The extent of protection afforded should depend upon the risk of harm and the likelihood of benefit. The judgment that any individual lacks autonomy should be periodically reevaluated and will vary in different situations.

In most cases of research involving human subjects, respect for persons demands that subjects enter into the research voluntarily and with adequate information. In some situations, however, application of the principle is not obvious. The involvement of prisoners as subjects of research provides an instructive example. On the one hand, it would seem that the principle of respect for persons requires that prisoners not be deprived of the opportunity to volunteer for research. On the other hand, under prison conditions they may be subtly coerced or unduly influenced to engage in research activities for which they would not otherwise volunteer. Respect for persons would then dictate that prisoners be protected. Whether to allow

prisoners to "volunteer" or to "protect" them presents a dilemma. Respecting persons, in most hard cases, is often a matter of balancing competing claims urged by the principle of respect itself.

2. Beneficence. — Persons are treated in an ethical manner not only by respecting their decisions and protecting them from harm, but also by making efforts to secure their well-being. Such treatment falls under the principle of beneficence. The term "beneficence" is often understood to cover acts of kindness or charity that go beyond strict obligation. In this document, beneficence is understood in a stronger sense, as an obligation. Two general rules have been formulated as complementary expressions of beneficent actions in this sense: **(1)** do not harm and **(2)** maximize possible benefits and minimize possible harms.

The Hippocratic maxim "do no harm" has long been a fundamental principle of medical ethics. Claude Bernard extended it to the realm of research, saying that one should not injure one person regardless of the benefits that might come to others. However, even avoiding harm requires learning what is harmful; and, in the process of obtaining this information, persons may be exposed to risk of harm. Further, the Hippocratic Oath requires physicians to benefit their patients "according to their best judgment." Learning what will in fact benefit may require exposing persons to risk. The problem posed by these imperatives is to decide when it is justifiable to seek certain benefits despite the risks involved, and when the benefits should be foregone because of the risks.

The obligations of beneficence affect both individual investigators and society at large, because they extend both to particular research projects and to the entire enterprise of research. In the case of particular projects, investigators and members of their institutions are obliged to give forethought to the maximization of benefits and the reduction of risk that might occur from the research investigation. In the case of scientific research in general, members of the larger society are obliged to recognize the longer term benefits and risks that may result from the improvement of knowledge and from the development of novel medical, psychotherapeutic, and social procedures.

The principle of beneficence often occupies a well-defined justifying role in many areas of research involving human subjects. An example is found in research involving children. Effective ways of treating childhood diseases and fostering healthy development are benefits that serve to justify research involving children -- even when individual research subjects are not direct beneficiaries. Research also makes it possible to avoid the harm that may result from the application of previously accepted routine practices that on closer investigation turn out to be dangerous. But the role of the principle of beneficence is not always so unambiguous. A difficult ethical problem remains, for example, about research that presents more than minimal risk without immediate prospect of direct benefit to the children involved. Some have argued that such research is inadmissible, while others have pointed out that this limit would rule out much research promising great benefit to children in the future. Here again, as with all hard cases, the different claims covered by the principle of beneficence may come into conflict and force difficult choices.

3. Justice. — Who ought to receive the benefits of research and bear its burdens? This is a question of justice, in the sense of "fairness in distribution" or "what is deserved." An injustice occurs when some benefit to which a person is entitled is denied without good reason or when some burden is imposed unduly. Another way of conceiving the principle of justice is that equals ought to be treated equally. However, this statement requires explication. Who is equal and who is unequal? What considerations justify departure from equal distribution? Almost all commentators allow that distinctions based on experience, age, deprivation, competence, merit and position do sometimes constitute criteria justifying differential treatment for certain purposes. It is necessary, then, to explain in what respects people should be treated equally. There are several widely accepted formulations of just ways to distribute burdens and benefits. Each formulation mentions some relevant property on the basis of which burdens and benefits should be distributed. These formulations are **(1)** to each person an equal share, **(2)** to each person according to individual need, **(3)** to each person according to individual effort, **(4)** to each person according to societal contribution, and **(5)** to each person according to merit.

Questions of justice have long been associated with social practices such as punishment, taxation and political representation. Until recently these questions have not generally been associated with scientific research. However, they are foreshadowed even in the earliest reflections on the ethics of research involving human subjects. For example, during the 19th and early 20th centuries the burdens of serving as research subjects fell largely upon poor ward patients, while the benefits of improved medical care flowed primarily to private patients. Subsequently, the exploitation of unwilling prisoners as research subjects in Nazi concentration camps was condemned as a particularly flagrant injustice. In this country, in the 1940's, the Tuskegee syphilis study used disadvantaged, rural black men to study the untreated course of a disease that is by no means confined to that population. These subjects were deprived of demonstrably effective treatment in order not to interrupt the project, long after such treatment became generally available.

Against this historical background, it can be seen how conceptions of justice are relevant to research involving human subjects. For example, the selection of research subjects needs to be scrutinized in order to determine whether some classes (e.g., welfare patients, particular racial and ethnic minorities, or persons confined to institutions) are being systematically selected simply because of their easy availability, their compromised position, or their manipulability, rather than for reasons directly related to the problem being studied. Finally, whenever research supported by public funds leads to the development of therapeutic devices and procedures, justice demands both that these not provide advantages only to those who can afford them and that such research should not unduly involve persons from groups unlikely to be among the beneficiaries of subsequent applications of the research.

Part C: Applications

C. Applications

Applications of the general principles to the conduct of research leads to consideration of the following requirements: informed consent, risk/benefit assessment, and the selection of subjects of research.

1. Informed Consent. — Respect for persons requires that subjects, to the degree that they are capable, be given the opportunity to choose what shall or shall not happen to them. This opportunity is provided when adequate standards for informed consent are satisfied.

While the importance of informed consent is unquestioned, controversy prevails over the nature and possibility of an informed consent. Nonetheless, there is widespread agreement that the consent process can be analyzed as containing three elements: information, comprehension and voluntariness.

Information. Most codes of research establish specific items for disclosure intended to assure that subjects are given sufficient information. These items generally include: the research procedure, their purposes, risks and anticipated benefits, alternative procedures (where therapy is involved), and a statement offering the subject the opportunity to ask questions and to withdraw at any time from the research. Additional items have been proposed, including how subjects are selected, the person responsible for the research, etc.

However, a simple listing of items does not answer the question of what the standard should be for judging how much and what sort of information should be provided. One standard frequently invoked in medical practice, namely the information commonly provided by practitioners in the field or in the locale, is inadequate since research takes place precisely when a common understanding does not exist. Another standard, currently popular in malpractice law, requires the practitioner to reveal the information that reasonable persons would wish to know in order to make a decision regarding their care. This, too, seems insufficient since the research subject, being in essence a volunteer, may wish to know considerably more about risks gratuitously undertaken than do patients who deliver themselves into the hand of a clinician for

needed care. It may be that a standard of "the reasonable volunteer" should be proposed: the extent and nature of information should be such that persons, knowing that the procedure is neither necessary for their care nor perhaps fully understood, can decide whether they wish to participate in the furthering of knowledge. Even when some direct benefit to them is anticipated, the subjects should understand clearly the range of risk and the voluntary nature of participation.

A special problem of consent arises where informing subjects of some pertinent aspect of the research is likely to impair the validity of the research. In many cases, it is sufficient to indicate to subjects that they are being invited to participate in research of which some features will not be revealed until the research is concluded. In all cases of research involving incomplete disclosure, such research is justified only if it is clear that **(1)** incomplete disclosure is truly necessary to accomplish the goals of the research, **(2)** there are no undisclosed risks to subjects that are more than minimal, and **(3)** there is an adequate plan for debriefing subjects, when appropriate, and for dissemination of research results to them. Information about risks should never be withheld for the purpose of eliciting the cooperation of subjects, and truthful answers should always be given to direct questions about the research. Care should be taken to distinguish cases in which disclosure would destroy or invalidate the research from cases in which disclosure would simply inconvenience the investigator.

Comprehension. The manner and context in which information is conveyed is as important as the information itself. For example, presenting information in a disorganized and rapid fashion, allowing too little time for consideration or curtailing opportunities for questioning, all may adversely affect a subject's ability to make an informed choice.

Because the subject's ability to understand is a function of intelligence, rationality, maturity and language, it is necessary to adapt the presentation of the information to the subject's capacities. Investigators are responsible for ascertaining that the subject has comprehended the information. While there is always an obligation to ascertain that the information about risk to subjects is complete and adequately comprehended, when the risks are more serious, that obligation increases. On occasion, it may be suitable to give some oral or written tests of comprehension.

Special provision may need to be made when comprehension is severely limited -- for example, by conditions of immaturity or mental disability. Each class of subjects that one might consider as incompetent (e.g., infants and young children, mentally disable patients, the terminally ill and the comatose) should be considered on its own terms. Even for these persons, however, respect requires giving them the opportunity to choose to the extent they are able, whether or not to participate in research. The objections of these subjects to involvement should be honored, unless the research entails providing them a therapy unavailable elsewhere. Respect for persons also requires seeking the permission of other parties in order to protect the subjects from harm. Such persons are thus respected both by acknowledging their own wishes and by the use of third parties to protect them from harm.

The third parties chosen should be those who are most likely to understand the incompetent subject's situation and to act in that person's best interest. The person authorized to act on behalf of the subject should be given an opportunity to observe the research as it proceeds in order to be able to withdraw the subject from the research, if such action appears in the subject's best interest.

Voluntariness. An agreement to participate in research constitutes a valid consent only if voluntarily given. This element of informed consent requires conditions free of coercion and undue influence. Coercion occurs when an overt threat of harm is intentionally presented by one person to another in order to obtain compliance. Undue influence, by contrast, occurs through an offer of an excessive, unwarranted, inappropriate or improper reward or other overture in order to obtain compliance. Also, inducements that would ordinarily be acceptable may become undue influences if the subject is especially vulnerable.

Unjustifiable pressures usually occur when persons in positions of authority or commanding influence -- especially where possible sanctions are involved -- urge a course of action for a subject. A continuum of

such influencing factors exists, however, and it is impossible to state precisely where justifiable persuasion ends and undue influence begins. But undue influence would include actions such as manipulating a person's choice through the controlling influence of a close relative and threatening to withdraw health services to which an individual would otherwise be entitled.

2. Assessment of Risks and Benefits. — The assessment of risks and benefits requires a careful arrayal of relevant data, including, in some cases, alternative ways of obtaining the benefits sought in the research. Thus, the assessment presents both an opportunity and a responsibility to gather systematic and comprehensive information about proposed research. For the investigator, it is a means to examine whether the proposed research is properly designed. For a review committee, it is a method for determining whether the risks that will be presented to subjects are justified. For prospective subjects, the assessment will assist the determination whether or not to participate.

The Nature and Scope of Risks and Benefits. The requirement that research be justified on the basis of a favorable risk/benefit assessment bears a close relation to the principle of beneficence, just as the moral requirement that informed consent be obtained is derived primarily from the principle of respect for persons. The term "risk" refers to a possibility that harm may occur. However, when expressions such as "small risk" or "high risk" are used, they usually refer (often ambiguously) both to the chance (probability) of experiencing a harm and the severity (magnitude) of the envisioned harm.

The term "benefit" is used in the research context to refer to something of positive value related to health or welfare. Unlike, "risk," "benefit" is not a term that expresses probabilities. Risk is properly contrasted to probability of benefits, and benefits are properly contrasted with harms rather than risks of harm. Accordingly, so-called risk/benefit assessments are concerned with the probabilities and magnitudes of possible harm and anticipated benefits. Many kinds of possible harms and benefits need to be taken into account. There are, for example, risks of psychological harm, physical harm, legal harm, social harm and economic harm and the corresponding benefits. While the most likely types of harms to research subjects are those of psychological or physical pain or injury, other possible kinds should not be overlooked.

Risks and benefits of research may affect the individual subjects, the families of the individual subjects, and society at large (or special groups of subjects in society). Previous codes and Federal regulations have required that risks to subjects be outweighed by the sum of both the anticipated benefit to the subject, if any, and the anticipated benefit to society in the form of knowledge to be gained from the research. In balancing these different elements, the risks and benefits affecting the immediate research subject will normally carry special weight. On the other hand, interests other than those of the subject may on some occasions be sufficient by themselves to justify the risks involved in the research, so long as the subjects' rights have been protected. Beneficence thus requires that we protect against risk of harm to subjects and also that we be concerned about the loss of the substantial benefits that might be gained from research.

The Systematic Assessment of Risks and Benefits. It is commonly said that benefits and risks must be "balanced" and shown to be "in a favorable ratio." The metaphorical character of these terms draws attention to the difficulty of making precise judgments. Only on rare occasions will quantitative techniques be available for the scrutiny of research protocols. However, the idea of systematic, nonarbitrary analysis of risks and benefits should be emulated insofar as possible. This ideal requires those making decisions about the justifiability of research to be thorough in the accumulation and assessment of information about all aspects of the research, and to consider alternatives systematically. This procedure renders the assessment of research more rigorous and precise, while making communication between review board members and investigators less subject to misinterpretation, misinformation and conflicting judgments. Thus, there should first be a determination of the validity of the presuppositions of the research; then the nature, probability and magnitude of risk should be distinguished with as much clarity as possible. The method of ascertaining risks should be explicit, especially where there is no alternative to the use of such vague categories as small or slight risk. It should also be determined whether an investigator's estimates of the probability of harm or benefits are reasonable, as judged by known facts or other available studies.

Finally, assessment of the justifiability of research should reflect at least the following considerations: **(i)**Brutal or inhumane treatment of human subjects is never morally justified. **(ii)** Risks should be reduced to those necessary to achieve the research objective. It should be determined whether it is in fact necessary to use human subjects at all. Risk can perhaps never be entirely eliminated, but it can often be reduced by careful attention to alternative procedures. **(iii)** When research involves significant risk of serious impairment, review committees should be extraordinarily insistent on the justification of the risk (looking usually to the likelihood of benefit to the subject -- or, in some rare cases, to the manifest voluntariness of the participation). **(iv)** When vulnerable populations are involved in research, the appropriateness of involving them should itself be demonstrated. A number of variables go into such judgments, including the nature and degree of risk, the condition of the particular population involved, and the nature and level of the anticipated benefits. **(v)** Relevant risks and benefits must be thoroughly arrayed in documents and procedures used in the informed consent process.

3. Selection of Subjects. — Just as the principle of respect for persons finds expression in the requirements for consent, and the principle of beneficence in risk/benefit assessment, the principle of justice gives rise to moral requirements that there be fair procedures and outcomes in the selection of research subjects.

Justice is relevant to the selection of subjects of research at two levels: the social and the individual. Individual justice in the selection of subjects would require that researchers exhibit fairness: thus, they should not offer potentially beneficial research only to some patients who are in their favor or select only "undesirable" persons for risky research. Social justice requires that distinction be drawn between classes of subjects that ought, and ought not, to participate in any particular kind of research, based on the ability of members of that class to bear burdens and on the appropriateness of placing further burdens on already burdened persons. Thus, it can be considered a matter of social justice that there is an order of preference in the selection of classes of subjects (e.g., adults before children) and that some classes of potential subjects (e.g., the institutionalized mentally infirm or prisoners) may be involved as research subjects, if at all, only on certain conditions.

Injustice may appear in the selection of subjects, even if individual subjects are selected fairly by investigators and treated fairly in the course of research. Thus injustice arises from social, racial, sexual and cultural biases institutionalized in society. Thus, even if individual researchers are treating their research subjects fairly, and even if IRBs are taking care to assure that subjects are selected fairly within a particular institution, unjust social patterns may nevertheless appear in the overall distribution of the burdens and benefits of research. Although individual institutions or investigators may not be able to resolve a problem that is pervasive in their social setting, they can consider distributive justice in selecting research subjects.

Some populations, especially institutionalized ones, are already burdened in many ways by their infirmities and environments. When research is proposed that involves risks and does not include a therapeutic component, other less burdened classes of persons should be called upon first to accept these risks of research, except where the research is directly related to the specific conditions of the class involved. Also, even though public funds for research may often flow in the same directions as public funds for health care, it seems unfair that populations dependent on public health care constitute a pool of preferred research subjects if more advantaged populations are likely to be the recipients of the benefits.

One special instance of injustice results from the involvement of vulnerable subjects. Certain groups, such as racial minorities, the economically disadvantaged, the very sick, and the institutionalized may continually be sought as research subjects, owing to their ready availability in settings where research is conducted. Given their dependent status and their frequently compromised capacity for free consent, they should be protected against the danger of being involved in research solely for administrative convenience, or because they are easy to manipulate as a result of their illness or socioeconomic condition.

[1] Since 1945, various codes for the proper and responsible conduct of human experimentation in medical research have been adopted by different organizations. The best known of these codes are the Nuremberg Code of 1947, the Helsinki Declaration of 1964 (revised in 1975), and the 1971 Guidelines (codified into Federal Regulations in 1974) issued by the U.S. Department of Health, Education, and Welfare Codes for the conduct of social and behavioral research have also been adopted, the best known being that of the American Psychological Association, published in 1973.

[2] Although practice usually involves interventions designed solely to enhance the well-being of a particular individual, interventions are sometimes applied to one individual for the enhancement of the well-being of another (e.g., blood donation, skin grafts, organ transplants) or an intervention may have the dual purpose of enhancing the well-being of a particular individual, and, at the same time, providing some benefit to others (e.g., vaccination, which protects both the person who is vaccinated and society generally). The fact that some forms of practice have elements other than immediate benefit to the individual receiving an intervention, however, should not confuse the general distinction between research and practice. Even when a procedure applied in practice may benefit some other person, it remains an intervention designed to enhance the well-being of a particular individual or groups of individuals; thus, it is practice and need not be reviewed as research.

[3] Because the problems related to social experimentation may differ substantially from those of biomedical and behavioral research, the Commission specifically declines to make any policy determination regarding such research at this time. Rather, the Commission believes that the problem ought to be addressed by one of its successor bodies.

The Menlo Report

Ethical Principles Guiding Information and
Communication Technology Research

August 2012

Homeland
Security

Science and Technology

Executive Summary

This report proposes a framework for ethical guidelines for computer and information security research, based on the principles set forth in the 1979 Belmont Report, a seminal guide for ethical research in the biomedical and behavioral sciences. Despite its age, the Belmont Report's insightful abstraction renders it a valuable cornerstone for other domains. We describe how the three principles in the Belmont report can be usefully applied in fields related to research about or involving *information and communication technology*. ICT research raises new challenges resulting from interactions between humans and communications technologies. In particular, today's ICT research contexts contend with ubiquitously connected network environments, overlaid with varied, often discordant legal regimes and social norms. We illustrate the application of these principles to information systems security research — a critical infrastructure priority with broad impact and demonstrated potential for widespread harm — although we expect the proposed framework to be relevant to other disciplines, including those targeted by the Belmont report but now operating in more complex and interconnected contexts.

We first outline the scope and motivation for this document, including a historical summary of the conceptual framework for traditional human subjects research, and the landscape of ICT research stakeholders. We review four core ethical principles, the three from the Belmont Report (Respect for Persons, Beneficence, and Justice) and an additional principle *Respect for Law and Public Interest*. We propose standard methods to operationalize these principles in the domain of research involving information and communication technology: identification of stakeholders and informed consent; balancing risks and benefits; fairness and equity; and compliance, transparency and accountability, respectively. We also describe how these principles and applications can be supported through assistive external oversight by ethical review boards, and internal self-evaluation tools such as an Ethical Impact Assessment.

The intent of this report is to help clarify how the characteristics of ICT raise new potential for harm and to show how a reinterpretation of ethical principles and their application can lay the groundwork for ethically defensible research.

Authors and Working Group Participants

This report is the product of a series of workshops and meetings held over a period of sixteen months. The participants at these meetings are listed alphabetically below. In addition, the authors thank the dozen or so ICTR community members whose feedback was invaluable to assuring that this document reflects the ground truth sentiments of the professionals at the front lines of ICT research ethics.

- Michael Bailey, University of Michigan
- Aaron Burstein, University of California Berkeley
- KC Claffy, CAIDA, University of California San Diego
- Shari Clayman, DHS Science & Technology
- David Dittrich, Co-Lead Author, University of Washington
- John Heidemann, University of Southern California, ISI
- Erin Kenneally, CAIDA, University of California San Diego, Co-Lead Author
- Douglas Maughan, DHS Science & Technology
- Jenny McNeill, SRI International
- Peter Neumann, SRI International
- Charlotte Scheper, RTI International
- Lee Tien, Electronic Frontier Foundation
- Christos Papadopoulos, Colorado State University
- Wendy Visscher, RTI International
- Jody Westby, Global Cyber Risk, LLC

Contents

Contact details:
For general inquiries about this publication, please use the following details:
Email: Menlo_Report@hq.dhs.gov
Internet: http://www.dhs.gov/csd-resources

A   Introduction – Focus and Motivations

This report attempts to summarize a set of basic principles to guide the identification and resolution of ethical problems arising in research of or involving *information and communication technology* (ICT).1 ICT is a general umbrella term that encompasses networks, hardware and software technologies that involve information communications pertaining to or impacting individuals and organizations. ICT has increasingly become integrated into our individual and collective daily lives, mediating our behaviors and communications and presenting new tensions that challenge the applications of these guiding principles.

ICT research (ICTR) involves the collection, use and disclosure of information and/or interaction with this ubiquitously connected network context which is overlaid with varied, often discordant legal regimes and social norms. The challenge of evaluating the ethical issues in ICTR stems in large part from the attributes of ICT: scale, speed, tight coupling, decentralization and wide distribution, and opacity. This environment complicates achieving ethically defensible research for several reasons. It results in interactions with humans that are often indirect, stemming from an increase in either logical or physical "distance" between researcher and humans to be protected over research involving direct intervention. The relative ease in engaging multitudes of distributed human subjects (or data about them) through intermediating systems speeds the potential for harms to arise, and extends the range of stakeholders who may be impacted. Also, legal restrictions and requirements have expanded considerably since the 1980s, and ICTR is unquestionably subject to a variety of laws and regulations that address data collection and use. While it is true that these individual complications are shared by traditional biomedical and behavioral research, this report seeks to manage the tension resulting from the simultaneous confluence of these complicating factors that occur with regularity in ICTR.

There is a need to interpret and extend the traditional ethical framework to enable ICT researchers and oversight entities to appropriately and consistently assess and render ethically defensible research.2 Such a framework should also support current and potential institutional mechanisms that are well served to implement it, such as a research ethics board (REB). We build on the foundation set by the *Belmont Report*, which articulates three fundamental ethical principles and guiding applications of these principles for protecting human subjects of biomedical and behavioral research: respecting persons; balancing potential benefits and harms; and equitably apportioning benefits and burdens across research subjects and society.3 The guidelines in this report are applicable to research that has the potential to harm humans, regardless of whether those humans are the direct research subjects or are indirectly at risk of harm from interactions with ICT. This report explains how the traditional framework fits within the context of the computer science sub-discipline of information security research. Specifically, this domain addresses ICT vulnerabilities, digital crime, and information assurance for critical infrastructure systems. These are areas where harms are not well understood yet are potentially significant in scope and impact. The framework proposed herein is germane to other disciplines that involve the use of ICT, including those targeted by the Belmont Report that now operate in ICT contexts.

A.1   Target Audience for this Report

This report offers guidance primarily for ICT researchers (including academic, corporate, and independent researchers), professional societies, publication review committees, and funding agencies. Secondarily, this report aims to assist those who administer and apply these principles, such as oversight authorities (e.g., REBs), policy makers, attorneys, and others who shape and implement human subject protection policies and procedures.

This report does not recommend particular enforcement mechanisms. To the extent that enforcement of ethical practices is inconsistent across and within academic and non-academic ICTR, we intend this report to improve consistency in ethical analyses and self-regulation for both individuals and organizations striving toward ethically defensible research.

A.2   Historical Context

Despite a long history of well-publicized abuses, it took over a decade for the ethical standards prescribed in the Belmont Report to first be defined in the Code of Federal Regulations (CFR). Language from 45 CFR 46, which covers biomedical and behavioral research

funded by the Department of Health and Human Services (HHS), was later adopted by all executive branch departments in what is known as the *Common Rule*.[4] It ushered in a government-wide requirement for REB oversight of research protocols to protect human research subjects. Prior to this point, there was no regulated oversight mechanism and biomedical and behavioral researchers relied on subjective, ad hoc, and inconsistent ethical compasses to guide their decision making.

In parallel during the 1970s, a U.S. Defense Advanced Research Projects Agency (DARPA) project was designing and implementing a communications architecture to support cooperative time-sharing of computational resources across large government-funded laboratories. Although this network architecture would eventually evolve into the global Internet, the community at the time was small, trusted, non-commercial, and research-oriented. This burgeoning Internet was not under constant attack from around the world. It did not provide access to numerous databases containing millions of personally-identifying records. It was not an integral part of providing and maintaining critical services or communications. A tiny number of people accessed the Internet during those early years compared to the billions of users who engage in this environment on a regular and almost unconscious basis today.

Early ICT research evolved without significant concern for human subjects, leading to instances where ethical considerations were either absent or misapplied because researchers failed to understand their relevance, or lacked any standards for assessment, accountability, or oversight. Cases include interactive studies of malicious software and platforms, engagement in active counterattack measures, exploitation and disclosure of systems vulnerabilities, and collection and sharing of sensitive information. The demonstrated potential for harm in ICTR illustrates the need to re-conceptualize the traditional *human subject protection* paradigm that underpins ethical oversight in other fields.

ICTR challenges us to re-conceptualize the traditional *human subject protection* paradigm that underpins ethical oversight. The foremost misunderstandings and disagreements about the applicability and scope of this protection in ICTR stem largely from how the Common Rule was written and has historically been interpreted. *Specifically,* human subject means, "a living individual

about whom an investigator (whether professional or student) conducting research obtains (1) Data through intervention or interaction with the individual, or (2) Identifiable private information" (45 CFR 46.102(f)). Key terms here are "intervention" and "private information." Intervention does not just mean physical procedures, but also "manipulations of the [subject's] environment that are performed for research purposes," which could include manipulation of their computing devices, or automated appliances in the home. Private information is not just medical records, but "information about behavior that occurs in a context in which an individual can reasonably expect that no observation or recording is taking place, and information which has been provided for specific purposes by an individual and which the individual can reasonably expect will not be made public." This could include electronic communications, or data captured by malicious actors recording online financial transactions in order to commit fraud. Taken as a whole, the intent of the Common Rule is to protect persons who might be harmed from involvement in research, not simply with whether humans are participating in research. Confusion starts because of the wording above and linkage of the terms human and *research subject,* and continues with the determination of risk and how to *protect humans* within a research study.

An evolved paradigm for applying ethical principles to protect humans who may be impacted by research considers activities having human-harming potential rather than simply looking at whether the research does or does not involve human subjects. Examples of potentially human-harming ICT artifacts that researchers may interact with include avatars in online virtual worlds, malware controlling compromised machines, embedded medical devices controlling biological functions, or process controllers for critical infrastructure. The significant changes brought about by ICT since the commencement of formal regulated research necessitates a reconceptualization of the application of ethical principles for research involving ICT.

B   Restatement of Belmont Principles in the ICTR Context

In framing the principles and applications for evaluating and applying ethics in ICTR the Menlo Report explicitly adopts the Belmont principles and acknowledges the Common Rule regime which implemented that model. As such, this Report deliberately does not explore

alternate ethical paradigms, and while not discounting that there may be novel implementations of the Belmont Report principles and applications that should be considered it makes no definitive recommendations in that regard. However, this Report does highlight areas within the Common Rule that are more consequential or problematic for ICTR.

The first three rows of Table 1 summarize the three core principles and their application as outlined in the Belmont Report.5 We offer an additional principle to guide ethical considerations in ICTR research, listed in the fourth line of Table 1 We call this principle *Respect for Law and Public Interest* because it addresses the expansive and evolving, yet often varied and discordant, legal controls relevant to communication privacy and information assurance (i.e., the confidentiality, availability, and integrity of information and information systems). While respect for the law and public interest is implicit in Belmont's application of Beneficence, several challenging factors suggest these issues merit explicit consideration in the ICTR context: the myriad laws that may be germane to any given ICTR; conflicts and ambiguities among laws in different geo-political jurisdictions; the difficulty in identifying stakeholders, a necessary prerequisite to enforcing legal obligations; and possible incongruence between law and public interest.

C Application of the Principles

The challenges of ICTR risk assessment derive from three factors: the researcher-subject relationships, which tend to be disconnected, dispersed, and intermediated by technology; the proliferation of data sources and analytics, which can heighten risk incalculably; and the inherent overlap between research and operations. In order to properly apply any of the principles listed above in the complex setting of ICT research, it is first necessary to perform a systematic and comprehensive stakeholder analysis.

Principle	Application
Respect for Persons	Participation as a research subject is voluntary, and follows from informed consent; Treat individuals as autonomous agents and respect their right to determine their own best interests; Respect individuals who are not targets of research yet are impacted; Individuals with diminished autonomy, who are incapable of deciding for themselves, are entitled to protection.
Beneficence	Do not harm; Maximize probable benefits and minimize probable harms; Systematically assess both risk of harm and benefit.
Justice	Each person deserves equal consideration in how to be treated, and the benefits of research should be fairly distributed according to individual need, effort, societal contribution, and merit; Selection of subjects should be fair, and burdens should be allocated equitably across impacted subjects.
Respect for Law and Public Interest	*Engage in legal due diligence; Be transparent in methods and results; Be accountable for actions.*

Table 1: Proposed guidelines for ethical assessment of ICT Research.

C.1 Stakeholder Perspectives and Considerations

Stakeholder identification includes consideration of several factors: the degree to which information involved in the research identifies individuals (including their digital identities), groups and organizations and what behaviors, communications, or relationships are associated with such identification. Harms related to exposing the identity of research subjects engaging in sensitive behaviors, communications, or relationships, which they they assume to be private, can extend beyond the direct research subject to family, friends or other community relations. While this is also true of some research where the subject is the primary party at risk, in ICTR these harms may often be broader because ICT can amplify both the disclosure as well as the number of stakeholders impacted.

Further, ICTR often involves stakeholders that are non-research entities who rely on information and systems that are involved in the research and who may be harmed by its unavailability or corruption. Groups or organizations (e.g., companies or networks) may warrant different consideration from that of individuals, especially when applying the principles of Beneficence and Justice. Research involving ICT can be complex when the risks and benefits associated with multiple stakeholders require identification and balancing. ICT Researchers In commercial, academic, and government sectors,

ICT researchers have a vested interest in pursuing, sharing, and applying empirically grounded scientific knowledge. Research in economics, network science, security, and social behavior may inform operations, policies, and business models.

Human Subjects, Non-Subjects, and ICT Users

Traditional biomedical and behavioral research requires protection of natural persons and certain data that identifies them. In ICTR, the target of research may be an information system or associated data, which complicates the assessment of potential harm to users of that system or data. Primary considerations include the ability to interact with ICT without suffering harms such as disruption of access, loss of privacy, or unreasonable constraints on protected speech or activities. Victims of computer crimes are potential human non-subjects of research.

Malicious Actors A subset of ICTR involves criminal activity or potential exploitation of vulnerabilities in the design or implementation of ICT. The disclosures of some types of research results have a greater potential for misuse and thus greater value to malicious actors. This can provide a blueprint for widespread and wide-ranging harm by disclosing system vulnerability details of legitimate or malicious applications (the former by providing exploitation knowledge and the latter by illuminating countermeasures). Malicious actors avail themselves of published research results for nefarious purposes, which can result in harm that outweighs the intended research benefits. Consideration of this stakeholder's interest, therefore, involves understanding and avoiding or minimizing these potentially harmful impacts.

Network/Platform Owners and Providers Network owners or providers are typically commercial entities who are vested in safeguarding their physical and intellectual property, pursuing innovation and wealth, and building business and customer relationships. They are concerned about obligations associated with such representation. As intermediaries between a research and end users, they may be in a position of authority to serve as proxies for consent on behalf of their customers when it is otherwise impracticable for the researcher to individually obtain informed consent from end users.

Government: Law Enforcement Public law enforcement is mandated to advance criminal justice by protecting individuals and fostering public safety. Law enforcement also has an interest in research that improves its strategic, tactical, or operational efficacy in preventing, investigating, and responding to illegal activities. Examples include countering new and complex criminal ecosystems and instruments of crime such as botnets.

Government: Non-Law Enforcement Local, state, and federal government agencies are responsible for providing public services, protecting the rights of their citizens, and establishing law and policy governing social conduct. Research is an important vehicle through which the government can promote social good and innovation. For example, cybercrime research may enhance understanding of infrastructure risks, online social networks, or

economic markets of criminal enterprises; influence the deployment of commercial countermeasure technologies; and inform the interpretation or reform of relevant laws and policies. Acknowledging the different scope of their mission, the military and Intelligence Community (IC) is another subset of this stakeholder group.

Society ICTR implicates the collective rights and interests of owners and users of networks and data to know, influence, and choose how and when to engage with information communications networks and systems. Society benefits from knowledge that improves policies, laws and the administration of justice, and the well-being of the lives of its citizens. Society may likewise be harmed through actions that negatively impact information systems infrastructures, or through the collection, use, or disclosure of information that may assist criminals as much if not more than ICT system developers and operators.

C.2   Respect for Persons

In the Belmont Report, the principle of Respect for Persons reflects two tenets: individuals should be treated as autonomous agents, and persons with diminished autonomy are entitled to protection. This principle has been applied by involving as research subjects only those with sufficient understanding or awareness to provide informed consent, or by obtaining *informed consent* from legally authorized representatives (e.g., parents of minors, relatives of unconscious patients, or guardians of those incapable of deciding for themselves). In the ICTR context, the principle of Respect for Persons includes consideration of the computer systems and data that directly interface, integrate with, or otherwise impact persons who are typically not research subjects themselves.

C.2.1 Informed Consent

Informed consent is a process during which the researcher accurately describes the project and its risks to subjects and they accept the risks and agree to participate or decline. Subjects must be free to withdraw from research participation without negative consequences. Researchers obtain informed consent when research activity has the potential to harm individuals with whom a researcher interacts or about whom the researcher obtains identifiable private information. Research involving ICT also raises the potential for harms to secondary stakeholders who, while not the direct subjects of research, may have the right to autonomy.

Researchers should inform subjects that they may not benefit from the research, although society may benefit in the future. Researchers should be mindful that leveraging intended benefits to coerce or entice consent from subjects fails the voluntary participation element of informed consent. Examples include suggesting that research participants will receive improved or enhanced services, or that services will be degraded or withheld if a subject declines participation in or withdraws from a study. Informed consent for one research purpose or use should not be considered valid for other research purposes. When an individual is identified with a group or organization, individual consent does not imply consent from other members of the group. Finally, informed consent for one research purpose or use should not be considered valid for different research purposes.

The process of informed consent is intended to respect the autonomy of research subjects. The process involves three components: notice, comprehension, and voluntariness. Notice is typically achieved through a clearly written consent document that details the intended benefits of research activities and the risks to research subjects. The language level is kept to 8th grade or lower to improve the ability of subjects to comprehend the benefits and risks. The consent document stresses that participation is voluntary and that subjects are free to withdraw from research participation without negative consequences.

Research involving ICT also raises the potential for harms to secondary stakeholders who, while not the direct subjects of research, may also have the right to autonomy. When considering informed consent, we suggest researchers and REBs carefully explore the complex interconnected relationships between users and the myriad of organizations which provide ICT services. Decisions about mechanisms for obtaining informed consent, or requesting waivers of informed consent, may be impacted by whether entities have obtained valid authorization from their users — via explicit agreements or contractual terms of service —

for participation in research activities. Such authorization, whether supportive or restrictive of research, should be appropriately balanced when considering informed consent.

When a researcher believes that obtaining informed consent makes the pursuit of research objectives impossible, the application process allows for researchers to seek waivers from an ethical review board. REBs make the determination of whether or not the Common Rule criteria of 45 CFR 46.116 and 45 CFR 46.117 allowing for alteration or elimination of informed consent have been met. These requirements ensure that: (1) The research involves no more than minimal risk to the subjects; (2) The waiver or alteration will not adversely affect the rights and welfare of the subjects; (3) The research could not practicably be carried out without the waiver or alteration; and (4) Whenever appropriate, the subjects will be provided with additional pertinent information after participation.

There are justifiable reasons why it may be impracticable for research to be carried out without a waiver or alteration of the informed consent process. Because of the difficulty in identifying all individuals from whom consent should be sought or in practicably obtaining consent, researchers or REBs may frequently conclude that seeking a waiver of informed consent or waiver of documentation of informed consent are the only options. For example, it may be infeasible to identify, or obtain consent from millions of users whose everyday communication generates traffic across a heavily aggregated backbone link in a traffic modeling study. Or it can be impossible to attempt to inform the owners of hundreds of thousands of compromised home computers that are being used as a single instrument of criminal activity (i.e., a botnet) under study. The Common Rule criteria for a waiver of documentation of informed consent in minimal or no-risk situations allows for less formal consent than a signed consent form, including verbal consent from a legally authorized representative rather than the research subjects themselves. REBs may also require some form of notification to research subjects, even if the REB does not require signed consent forms.

Some research involving retrospectively collected identifiable data may not be possible if consent must be obtained from all individuals identifiable in the data. In such situations, respect for persons is maintained by REBs instead focusing on data protections and/or removal of identifying information that is not germane to research as alternative means of minimizing potential harm and granting a waiver of informed consent for the research. Data that has already been de-identified and can be approved for exemption from REB review falls into a a special regulatory category of "pre-existing public data." REBs have some flexibility in how they define and interpret this class of data and some institutions maintain a list of pre-approved sources of such data that researchers may freely use. Data that is not on such pre-approved lists that contains fields that can identify individuals — even though it may be accessible to the general public — may not be considered "pre-existing public data." Researchers should therefor consult with their REB to discuss whether the data they wish to use falls under their institution's "pre-existing public data" exemption criteria, or whether they can qualify for a waiver of informed consent to re-use existing data in conformance with REB requirements. Prospective research is the more problematic case, where informed consent may be required by an REB unless it can be shown there is no risk what so ever.

As a contingency of granting a waiver of informed consent, REBs often require that the researcher notify subjects post hoc of their involvement in research, and demonstrate respect for autonomy by allowing subjects to direct the destruction of the data collected about them. Research involving deception may be performed by providing misleading data in the consent form, or with consent having been waived and no subject knowledge of the research activity at all. In either case, an REB may require debriefing in order to mitigate harm resulting from loss of trust in researchers by those subjects who were deceived. Research of criminal activity often involves deception or clandestine research activity, so requests for waivers of both informed consent and post hoc notification and debriefing may be relatively common as compared with research studies of non-criminal activity.

C.3   Beneficence

In the Belmont Report, the Beneficence principle reflects the concept of appropriately balancing probable harm and likelihood of enhanced welfare resulting from the research. Translating this principle to ICTR demands a framework for systematic identification of risks and benefits for a range of stakeholders, diligent analysis of how harms are minimized and benefits are maximized, preemptive planning to mitigate any realized harms, and implementing these evaluations into the research methodology.

C.3.1   Identification of Potential Benefits and Harms

Similar to traditional human-centered research, ICT researchers should identify benefits and potential harms from the research for all relevant stakeholders, including society as a whole, based on objective, generally accepted facts or studies. Since communication technologies intermediate so much of our lives, designing, conducting and evaluating ICTR may demand attention to potential societal benefits and harms related to: systems assurance (confidentiality, availability, integrity); individual and organizational privacy; reputation, emotional well-being, or financial sensitivities; and infringement of legal rights (derived from constitution, contract, regulation, or common law). Challenges identifying harms in ICTR environments stem from the scale and rapidity at which risk can manifest, the difficulty of attributing research risks to specific individuals and/or organizations, and our limited understanding of the causal dynamics between the physical and virtual worlds. As with all exploratory research, it can be challenging to articulate benefits such that subjects can make informed decisions. In ICTR our ability to qualitatively and quantitatively foresee the probable benefits is particularly immature.

One helpful approach to identifying harms is to review the laws and regulations that apply to an ICTR activity, and analyze the underlying individual and public interests that the research might negatively impact. While researchers are not expected to render legal conclusions or have legal subject matter expertise, they are obligated to respect what is written in the law and understand the underlying societal norms those laws represent. However, as the development of the law and technology occur at a different trajectory and pace, relying exclusively on the law may overlook important harms not expressly addressed by law. Similarly, it is not

the role of researchers to judge guilt or innocence, but they should consider how malicious actors might avail themselves of published research results for nefarious purposes, and assess whether that potential harm might outweigh the intended research benefits.

C.3.2   Balancing Risks and Benefits

A simplistic interpretation of Beneficence is the maximization of benefits and minimization of harms. Beneficence does not require that all harm be completely eliminated and every possible benefit be identified and fully realized. Rather, researchers should systematically assess risks and benefits across all stakeholders. In so doing, researchers should be mindful that risks to individual subjects are weighed against the benefits to society, not to the benefit of individual researchers or research subjects themselves. Ideally, researcher actions are measured using the objective standard of a *reasonable researcher*, who exercises the knowledge, skills, attention, and judgment that the community requires of its members to protect their interests and the interests of others. As researchers gain a greater understanding of how to reason about and apply ethical principles, community norms and expectations about what is *reasonable* will evolve. From the subjective perspective of the researcher, especially in light of evolving community standards, the elements of "integrity" are instructive: (1) discerning what is right and what is wrong, (2) acting on what you have discerned, even at personal cost; and (3) saying openly that you are acting on your understanding of right and wrong." 6

When ICT is involved, burdens and risks can extend beyond "the human subject," making the quantification of potential harm more difficult than with direct intervention. It can be difficult to balance risks and benefits with novel research whose value may be speculative or delayed, or whose realized harm may be perceived differently across stakeholders. If there are plausible risks, researchers bear the burden of illuminating those risks and their consideration of how those risks will be managed, and not simply rely on outside reviewers or REBs to identify and oversee those risks.

In a direct intervention research scenario, balancing is partially addressed through the informed consent process. When a study involves minimal risk and a researcher can give valid scientific reasons for altering or eliminating the consent requirement, post-research

debriefing may be required to respect individual autonomy. Balancing benefit and harm gets complicated when both deception and waiver of informed consent are involved, as may occur when studying social engineering using email (i.e., phishing). A researcher may seek to justify a waiver of the debriefing requirement under a *relative degree of harm* rationale, whereby deceived research subjects could suffer more harm from knowing researchers had deceived them than they would suffer from malicious actions. This in turn would be balanced by an REB against the knowledge developed through research intended to ameliorate the malicious harm. The process of comprehensive stakeholder analysis can assist both researchers and REBs to consider how best to balance benefit and harm in conformance with Common Rule waiver justification requirements (see 45 CFR 46.116 and .117).

While it is incumbent upon a researcher to identify and minimize potential harms, even with reasonable measures to detect and reduce them, harms may still occur. REBs must evaluate such risks in the context of what at-risk individuals actually experience in normal ICT usage, and in light of researchers' pursuit of generalizable knowledge that is vital to understanding the problem studied. For example, a researcher studying live malicious software may need to run the software on his own platform and observe its interactions with the criminals controlling it. Even with multiple layers of protection, the malicious software under study could still accidentally infect other computers. The risks posed by these accidental infections must be considered in light of everyday events that users encounter – programs crashing, malicious software accessing and infecting networked computers, and electronic communications being exposed – and must be balanced with potential benefits of understanding the behavior of the malicious software. Ethically defensible Beneficence lies on a spectrum between unequivocal adherence to averting all risk, which can have a chilling effect on beneficial research, and acting without regard to risk, which can be harmful to individuals and society.

C.3.3 Mitigation of Realized Harms

Some research involves greater than minimal risk, yet still has the potential to yield benefit to society and is allowed to be carried out. Despite appropriate precautions and attempts to balance risks and benefits in ICTR, such research may cause unintended side effects that harm stakeholders. Data breaches are one such form of

harm, but others may exist from disruption of information systems. Research of greater than minimal risk that has been approved by an REB must undergo continuation review regularly in accordance with the period set for the study by the specific REB, but no less than annually. While reporting of adverse events is part of regular status reports, "serious adverse events" may need to be reported immediately to an REB for possible actions. This can include the REB requiring a halt to research activities. For the same reasons that benefit is hard to calculate in ICTR, determining what could constitute a "serious adverse event" in the ICTR context is unclear.

In anticipation, researchers should consider preempting the escalation of realized harms by notifying affected parties or otherwise engaging mitigation actions. To that end, researchers should develop mitigation procedures and checklists, such as a contact list of parties to notify, if such unintended consequences ensue. Other potential harms that are reasonably foreseeable may have a low probability of occurring, but have a high impact. Researchers should anticipate such worst-case scenarios and make appropriate preparations to respond in a manner and scope that shows due diligence on the part of the researcher. It may be necessary and prudent to involve the researchers' own institutional risk management and oversight authorities and media relations in addition to the REB.

ICTR may involve records containing sensitive data about individuals, evidence of criminal activity, or that could potentially cause disruption to millions of computers around the world. ICT researchers must be aware of these harms as not only primary risks, but also secondary, collateral risks (e.g., to customers of primary data subjects or computer owners) and be prepared to responsibly inform affected stakeholders. In many cases, it is impracticable to notify all affected individuals, but it may be feasible to notify service providers or other entities who have the authority and capability – derived from their relationship with the affected stakeholders – to mitigate harm. A mitigation strategy should admit the variance in capacity and/or willingness of the notified entity to understand and act on the notification. Research records that identify individuals pose a risk of disclosure as long as those records exist, and may fall under REB oversight because of the risk posed. Researchers should be prepared to continually protect these records for as long as those records exist and are under researchers' control. Upon completion or

termination of approved research activities (allowing for a reasonable retention period approved by REBs in order to satisfy obligations of scientific reproducibility), the risky data should be destroyed. If records are maintained, the data should continue to be protected at the same level as was implemented during research under the same REB-approved mechanisms.

C.4   Justice: Fairness and Equity

In the Belmont Report, the principle of Justice is applied through fairness in the selection of research subjects, and equitable distribution of the burdens and benefits of research according to individual need, effort, societal contribution, and merit. Fairness should guide the initial selection of the subjects, as well as the apportionment of burdens to those who will most likely benefit from the research. Research design and implementation should consider all stakeholders' interests, although conflicting interests may render equal treatment impracticable. In the ICTR context, this principle implies that research should not arbitrarily target persons or groups based on attributes including (but not limited to): religion, political affiliation, sexual orientation, health, age, technical competency, national origin, race, or socioeconomic status. Neither should ICTR target specific populations for the sake of convenience or expediency.

It is important to distinguish between purposefully *excluding* groups based on prejudice or bias versus purposefully *including* entities who are willing to cooperate and consent, or who are better able to understand the technical issues raised by the researcher. The former raises Justice concerns, while the latter demonstrates efforts to apply the principles of Respect for Persons and Beneficence and still conduct meaningful research. All researchers have an obligation to not exclude/include individuals or groups from participation for reasons unrelated to the research purpose. The arbitrary targeting of subjects in ways that are not germane to pursuing legitimate research questions violates this principle.

Challenges to obtaining informed consent from users might motivate a researcher to work with a service provider who has direct contractual relationships with its network's users. These may serve as legally authorized representatives as described in the Common Rule for situations of minimal risk and requests for waivers of documentation of consent through "short form" or verbal consent. Such decisions to engage entities who are willing and able to act as legally authorized

representatives for obtaining consent and move forward with non-representative subject populations may raise fairness and equity concerns. Each provider with whom a researcher may interact will have varying levels of understanding and ability (or willingness) to act. If a researcher is required to get unanimous and uniform responses from all autonomous entities, it may be impossible to perform beneficial research. On the other hand, moving forward with risky research without the involvement, or at least awareness, of autonomous entities is undesirable as it may increase the potential for greater harm.

From an equity standpoint, open public disclosure of system vulnerabilities demands that researchers consider how the burdens and benefits of publicizing newly discovered vulnerability balance out. The burdens might be borne by the developers, yet actually might benefit malicious actors more in the short-term than developers or users of those systems. The calculation of benefits is actually a function of time, where malicious actors may act faster at exploiting vulnerability information than benevolent actors can act in mitigating the vulnerabilities.

C.5   Respect for Law and Public Interest

Respect for Law and Public Interest is implicit in the Belmont Report's application of Beneficence. In the context of ICTR, we include it as a separate principle with two applications — *Compliance and Transparency and Accountability.* The second application refers to transparency of methodologies and results, and accountability for actions. Transparency and accountability serve vital roles in many ICTR contexts where it is challenging or impossible to identify stakeholders (e.g., attribution of sources and intermediaries of information), to understand interactions between highly dynamic and globally distributed systems and technologies, and consequently to balance associated harms and benefits. A lack of transparency and accountability risks undermining the credibility of, trust and confidence in, and ultimately support for, ICT research.

There may be a conflict between simultaneously satisfying ethical review requirements and applicable legal protections. Even if a researcher obtains a waiver of informed consent due to impracticability reasons, this may not eliminate legal risk under laws that require consent or some other indication of authorization by rights holders in order to avoid liability. For example, information privacy and trespass statutes prohibit researchers

from accessing, acquiring or disclosing communications or other protected information without the consent of the communicating parties or owner of the system. Until REBs can overcome limited ICT expertise on committees and in administrative staff positions, they may not be capable of recognizing that certain ICT research data actually presents greater than minimal risk and may erroneously consider it exempt from review or subject it to expedited review procedures that bypass full committee review. As long as there is a gap in the capacity of REBs to properly evaluate research proposals just entering the review process, researchers following the guidance provided in this report can help illuminate the risks and relevant laws so as to improve the REB oversight process.

C.5.1 Compliance

Researchers should engage in due diligence to identify laws, regulations, contracts, and other private agreements that are applicable to their research, and should design and implement ICTR that respects these restrictions. While legal controls that call for compliance can be numerous and wide-ranging, those that should inform ethical assessments cluster categorically around computer crime and information security, privacy and anonymity, intellectual property, computer system assurance, and civil rights and liberties. More specifically, ICT research may implicate rights and obligations related to: identity theft; unsolicited bulk electronic mail; privacy in electronic and wire communications; notification of security breaches; copyright and other intellectual property infringement; data security and destruction; child pornography; spyware and phishing; fraudulent deception; financial privacy; economic espionage; constitutional privacy; health information security and privacy; industry standards and best practices; and contractual privacy and acceptable use policies.

Respect for public interest can often be addressed by obeying relevant laws. If applicable laws conflict with each other or contravene the public interest, researchers should have ethically defensible justification and be prepared to accept responsibility for their actions and consequences.

C.5.2 Transparency and Accountability

Transparency is a mechanism to assess and implement accountability, which itself is necessary to ensure that researchers behave responsibly. These applications interact to ultimately generate trust in ICTR by the public. Transparency-based accountability helps researchers, oversight entities, and other stakeholders avoid guesswork and incorrect inferences about whether, where, and how ethical principles are addressed. Transparency entails clearly communicating the purposes of research — why data collection and/or direct interaction with ICT is required to fulfill those purposes — and how research results will be used. It also involves clear communication of risk assessment and harm minimization related to research activities.

Accountability demands that research methodology, ethical evaluations, data collected, and results generated should be documented and made available responsibly in accordance with balancing risks and benefits. Data should be available for legitimate research, policy-making, or public knowledge, subject to appropriate collection, use, and disclosure controls informed by the Beneficence principle. The appropriate format, scope and modality of the data exposure will vary with the circumstances, as informed by Beneficence determinations.

D   Implementing the Principles and Applications

This document describes foundational ethical principles and their applications at a level intended to span a broad range of current and future research that will undoubtedly be affected by changes in ICT. For federally funded biomedical and behavioral research, the responsibility for evaluating whether a research project comports with these principles lies with REBs, which in the United States are known as Institutional Review Boards (IRBs). IRB review is a requirement for federally funded research, however researchers in the ICT field frequently either do not know of this requirement, or believe that they are not engaged in "human subjects research" and do not interact with their IRB at all. This report contends that ICTR will benefit from similar oversight, and the proposed guidelines will assist ICT researchers and oversight authorities identify, preempt and manage ethical risks. Current ICTR that does not fall under the purview of REBs would also benefit from community-derived self-regulation guided by this report. Proactively and transparently engaging in ethical assessment of ICT research will help move the research

community mindset in the direction of embedding ethics into ICTR design as productively and safely as possible, and more practically influence policy and governance at these crossroads.

Notes

[1] The term *information and communication technology* was coined by Denis Stevenson in a 1997 report to the United Kingdom government, *Information and Communication Technologies in the UK Schools: An Independent Inquiry*
http://rubble.heppell.net/stevenson/ICT.pdf

[2] This report offers pragmatic guidance in the application of these fundamental principles to ICTR, and avoids taking a position in the philosophical debate about the uniqueness of computer ethics. For an overview of the philosophical debate, see Bynum, Terrell, "Computer and Information Ethics", The Stanford Encyclopedia of Philosophy (Winter 2008 Edition), Edward N. Zalta (ed.). http://plato.stanford.edu/archives/win2008/entries/ethics-computer/

[3] *The Belmont Report*, the touchstone document guiding human subjects research in the biomedical and behavioral research fields, was named after the conference center where it was drafted in 1976 (See http://ohsr.od.nih.gov/guidelines/belmont.html) This document similarly takes its name from the city where a substantial portion of the working group meetings that resulted in this document took place in 2009-2010.

[4] Fifteen government departments and agencies performing research involving human subjects adopted 45 CFR 46 Subpart A in what is known as the *Common Rule*. Each has its own guidance on the interpretation of their section of the CFR. Refer to guidance appropriate to the funding source.

[5] *See* http://ohsr.od.nih.gov/guidelines/belmont.html

[6] Stephen L. Carter Carter, Stephen L (1996). Integrity. New York: BasicBooks/HarperCollins. pp. 7, 10. ISBN 0-06-092807-7.