# Statistical anonymity: Quantifying reidentification risks without reidentifying users

**Gecia Bravo-Hermsdorff** [1] **Robert Busa-Fekete** [1] **Lee M. Gunderson** [2] **Andrés Munõz Medina** [1] **Umar Syed** [1]

## Abstract

Data anonymization is an approach to privacy-preserving data release aimed at preventing participants reidentification, and it is an important alternative to differential privacy in applications that cannot tolerate noisy data. Existing algorithms for enforcing $k$-anonymity in the released data assume that the curator performing the anonymization has complete access to the original data. Reasons for limiting this access range from undesirability to complete infeasibility. This paper explores ideas — objectives, metrics, protocols, and extensions — for reducing the trust that must be placed in the curator, while still maintaining a statistical notion of $k$-anonymity. We suggest trust (amount of information provided to the curator) and privacy (anonymity of the participants) as the primary objectives of such a framework. We describe a class of protocols aimed at achieving these goals, proposing new metrics of privacy in the process, and proving related bounds. We conclude by discussing a natural extension of this work that completely removes the need for a central curator.

## 1. Releasing private data (Background)

As the use of big data continues to permeate modern society, so does the sharing of our personal data with centralized third-parties. For example, the U.S. Census Bureau shares aggregated population statistics with lawmakers (Abowd, 2018), and hospitals share medical information with insurance companies (Crellin & BCE, 2011). If unregulated, this type of information poses a threat to individual privacy. A trivial way to completely protect the privacy of individuals would be to simply not share any of their information, but such an absolutist approach is neither feasible nor useful. A sensible compromise is to develop methods that balance the usefulness of the data against the privacy lost by the individuals.

Two common frameworks for privacy-preserving data release are: *differential privacy*, i.e., DP (and its various extensions, e.g., Rényi differential privacy) and $k$*-anonymity* (and its various extensions, e.g., $t$-closeness).

### 1.1. A quick (incomplete) summary of DP

In the central model of differential privacy (Dwork et al., 2006), a trusted curator stores the database, and an analyst[1] issues queries about the database to a curator, who returns noisy responses. Such an approach requires the users to trust the curator with the entirety of their private data. Several models have been proposed to relax this requirement.

In the local model, each user adds noise to their own data and responds to the analyst directly (Evfimievski et al., 2003). In the shuffle model, each user encrypts their noisy data (such that only the analyst may read them), and sends them to a trusted shuffler. The shuffler then randomly permutes these encrypted messages before forwarding them to the analyst (Cheu et al., 2019).

### 1.2. A quick (incomplete) summary of $k$-anonymity

A dataset satisfies $k$-anonymity if for every individual whose data is contained in the dataset, their data are indistinguishable from that of at least $k-1$ other individuals (also presented in this dataset). Since $k$-anonymity was first introduced (Sweeney, 2002), efficient algorithms for anonymizing a database (while preserving the maximum amount of information possible) have received increasing interest. Local suppression algorithms aim to achieve this by redacting specific (feature, user) entries of the database (Meyerson & Williams, 2004), while global suppression algorithms redact the same set of features for every user (El Emam et al., 2009).

[1] Google Research, New York, US [2] Gatsby Unit, University College London, UK.
Correspondence to: *Gecia Bravo-Hermsdorff* <gecia@google.com>.

[1] Note that here the "analyst" and the "public" are the same entity since the data observed by the analyst could be seen by anyone else.

(Meyerson & Williams, 2004) showed that the problem of optimally anonymizing a database by either local and global suppression is NP-hard. In light of these results, several approximation algorithms have been proposed, particularly for local suppression (Aggarwal et al., 2005; Gkoulalas-Divanis et al., 2014). Similar to the central model of differential privacy, these algorithms/curators require access to the entire private data. Unlike differential privacy, variants of $k$-anonymity that reduce the trust that participants must place in the curator remain relatively unexplored.

## 2. Why we focus on $k$-anonymity (Motivation)

Differential Privacy (DP) (Dwork et al., 2014) is a measure of privacy loss (typically denoted by $\varepsilon$) that holds true *no matter what* (e.g., even if additional information is released in the future). As a result of this strong propriety, any DP algorithm must be stochastic (e.g., by adding noise to the data). This, however, can be undesirable in a variety of applications (see Section 2.3 for examples).

In contrast, while $k$-anonymity can be satisfied without adding noise to the data, its the privacy guarantee are contingent on the availability of auxiliary information (see (Narayanan & Shmatikov, 2008) for a famous example involving Netflix).

### 2.1. There is no panacea for private data

As differential privacy offers an upper bound on each instance of privacy loss that holds regardless of anything else, it has a simple composition rule that can be invoked without further assumptions. Perhaps for this reason, DP is currently the *de facto* academic definition of privacy.

As the issues surrounding privacy become increasingly pressing societal issues, it seems natural that the entities managing our private data would like to offer meaningful privacy guarantees. Unfortunately, despite being the "gold standard", DP is often touted with essentially meaningless parameters (Domingo-Ferrer et al., 2021). For example, the US census of 2020 claims a "mathematical algorithm to ensure that the privacy of individuals is sufficiently protected" with a "budget" of $\varepsilon = 19.61$ (US Census Press Release CB21-CN.42). Setting aside a conspicuous similarity with the natural logarithm of the US population,[2] the guarantee being made is essentially meaningless: "your participation in the census will not change the likelihood of any outcome by more than a factor of 331 million."

Given this clear rift in communication between theory and practice, it is fruitful to also consider privacy notions that might have fewer "translation" issues, despite their technically "weaker" guarantees.

### 2.2. Natural extensions of $k$-anonymity

For simplicity, consider the following setting: A database is to be released containing i.i.d. samples from the population, and the values can be split into two disjoint sets "Quasi-Identifiers" (QI) and "Sensitive Attributes" (SA). QI are not known to an adversary *a priori*, but could be learned (for some cost) via exogenous means. SA are features that are not known to the adversary, cannot be learned exogenously, and would be detrimental(valuable) to the participant(adversary) if learned by the adversary.

Many "$scalar$-word" anonymity measures can be classified by the assumptions they make on the sensitive attributes (SA). The use of $k$-anonymity assumes that all SA are completely incomparable, while $l$-diversity (Machanavajjhala et al., 2007) allows for the possibility of identical SA (but is still blind to the magnitude of differences). Other metrics, such as $t$-closeness, $\delta$-disclosure, and $\beta$-likeness, allow for a more general similarity metric between different SA (Khan et al., 2021).

The main goal of this paper is to understand the trade-off between anonymization guarantees to the participants and the trust they must place in the entity performing the anonymization. We believe that $k$-anonymity is a suitable notion to use as a proof-of-concept to introduce such statistical relaxation. Extending this framework to more nuanced measures of anonymity would be of considerable practical interest.

### 2.3. Application examples

Essentially, we consider a setting in which the private variables (the Sensitive Attributes) are incomparable (i.e., there is no metric of similarity) and unique (no two private variables are identical). In such a setting, $k$-anonymity is equivalent to $l$-diversity, and extensions such as $t$-closeness and $\beta$-likeness do not make sense (as the SA have no notion of similarity).

For example, consider a database containing X-rays images (SA), along with some (Quasi-Identifying) demographics of the patients. The latter could likely be obtained by an adversary with minimal effort, whereas the former is essentially impossible to directly measure (without explicit cooperation from the individual). Given the exposing nature of these SA, it is not a stretch to think about an adversary using them for their personal gain at the expense of the owner of the images. Moreover, the details of everyone's insides are rather unique.

Another application is that of preventing *browser fingerprinting* (Laperdrix et al., 2020). Malicious websites engaged in browser fingerprinting query detailed information about a

[2] The US population in 2020 is estimated at 331 million, and $\ln\big((331 \pm 1) \cdot 10^6\big) \approx 19.617 \pm 0.002$.

user's device (e.g., which fonts they have installed). If these details are sufficiently unique, they can be used to covertly track a user across different the web. While certain system details can be made less amenable to fingerprinting by adding noise to them (e.g., window size/resolution), the option of returning noisy responses is often not practical (e.g., uncertain browser type). Several browsers have proposed to prevent fingerprinting by ensuring that the information queried by a website is always $k$-anonymous, and blocking the query otherwise. However, the only way to completely guarantee $k$-anonymity is to grant a central curator access to the full data of every user.

## 3. The big picture (What we did)

In seeking a version of $k$-anonymity that does not require a fully-trusted curator, it appears necessary to allow for some fraction of the database that does not satisfy $k$-anonymity. It is therefore tempting to make the following analogy with differential privacy:

$$\begin{tikzcd}
\textbf{Central } \varepsilon\textbf{-DP} \arrow[r] \arrow[dr, phantom, "\Downarrow"] & \textbf{Local } (\varepsilon, \delta)\textbf{-DP} \quad \text{(with shuffler)} \\
\textbf{Central } k\textbf{-Anon} \arrow[r] & \textbf{Local } (k, Q)\textbf{-Anon} \quad \text{(with shuffler)} \\
& (\textit{this paper})
\end{tikzcd}$$

where the "amount of privacy" is quantified by $\varepsilon$ and $k$, and the "error rates" by $\delta$ and $Q$. Here, we propose several ideas related to the bottom right.

### 3.1. Main contribution

In this paper, we analyze ideas for anonymizing a database, while only granting the curator access to a partial view of the database. In such a setting, to publish any data, one must balance:

1. *Trust*: amount of information provided to the curator.
2. *Privacy*: anonymity of the participants.

As $k$-anonymity can no longer be strictly guaranteed for all users, we quantify privacy using the *exposure*: a new statistical version of $k$-anonymity, defined as the expected fraction of users who are not $k$-anonymous in the published database. We then upper bound the exposure with high probability.

## 4. Problem setting and notation

A private database is represented by a matrix $M$ consisting of $n$ rows and $d$ columns. Each *row* in the matrix corresponds to a *user*, and each *column* corresponds to a *feature* (e.g., age or gender).

$M_{ij}$ denotes the value of feature $j$ for user $i$. For any subset of features $J \subseteq [d]$, $M_J$ denotes the submatrix containing only columns $J$. $V_j = \{M_{ij} : i \in [n]\}$ denotes the set of possible values for column $j$, $V_J = \bigotimes_{j \in J} V_j$ the set of possible joint values for columns $J$, and $V = \bigotimes_{j \in [d]} V_j$ the set of all possible joint values.

A user in a database $M$ is *$k$-anonymous* if there are at least $k$ rows in $M$ that are identical to that user's row (e.g., $k = 1$ implies that their row is unique). If every user is $k$-anonymous, we say that $M$ is $k$-anonymous. *Local* suppression algorithms achieve this by redacting specific *entries* of $M$, while *global* suppression algorithms redact entire *columns*. In this paper, we consider global suppression, though we posit analogues of local suppression in the discussion (Section 8.1).

## 5. Anonymization protocol

We proposed a two-step protocol similar in spirit to the shuffle model of differential privacy (Cheu et al., 2019).

In the first round, the users send messages to the curator. The curator learns only the marginal distribution of individual features (i.e., nothing about their correlations). Using this information, the curator selects an appropriate set of features to be released to the analyst.

In the second round, the users send messages to the analyst. The analyst learns the full joint distribution between those features selected by the curator.

### 5.1. The first step

First, the curator and shuffler both create public/private key pairs, sending the public keys to the users (see Figure 1). Each user uses these public keys to encode one message for each feature. As identical values would result in identical encoded messages, the users first concatenate their value with a random string.[3] The users then send these encoded messages to the shuffler.

The shuffler randomly permutes these encrypted messages, decodes their portion of the encryption, and sends them to the curator.

The curator receives the messages, decodes them, and remove the salt. The curator then selects a subset of features that are safe to give to the analyst. Such a decision can be made by using, for example, composition rules (Theorems 2 and 3) or statistical modeling (Section 7).

### 5.2. The second step

Using the same shuffling mechanism, the users now communicate with the analyst. Each user creates a *single* message, encoding the entire subset of features that have been deemed

[3] Thereby adding some cryptographic "salt" (Park et al., 2001) to the receipt, if you will.

"safe" to release by the curator.

Within this setting, we analyze two situations:

1. *Fixed database* (Section 6): the *same users* participate in the first and second step of the protocol.
2. *Statistical database* (Section 7): *different users* participate in the first and second step of the protocol.

# 6. Fixed database setting

We start by analyzing the simpler setting, where there is a fixed database, with the same set of users sending their individual entries to the curator and their redacted rows to the analyst. Therefore, the curator decides which feature the users should redact when sending their data to the analyst based on complete knowledge of the marginal distribution of the features. The X-ray example we mentioned in Section 2.3 is an application for this setting.

## 6.1. Exposure (new privacy measure)

**Definition 1 (Exposure).** For a given probability threshold $t \in [0,1]$, the *exposure* $Q(t)$ is defined as the fraction of users that are less than $(tn)$-anonymous in matrix/database $M$:

$$Q(t) = \sum_{v \in V} \mathbf{p}(v) \cdot \mathbf{1}\left\{\mathbf{p}(v) < t\right\}, \qquad (1)$$

where

$$\mathbf{p}(v) = \frac{|\{i \in [n] : M_{ij} = v_j \text{ for all } j \in [d]\}|}{n} \qquad (2)$$

is the empirical (observed in $M$) joint distribution of possible outputs $V$.

For subset of columns $J$ or a single column $j$, we define

$$Q_J(t), \mathbf{p}_J(v), Q_j(t), \mathbf{p}_j(v),$$

as the restriction of the above definitions to those column(s).

When the database to be released to the analyst is different than the one observed by the curator (Section 7), we put a hat on quantities belonging to the latter, e.g., $\widehat{Q}_J$ and $\widehat{\mathbf{p}}_J$.

A central question of our paper is:

> *Given a subset of columns $J \subseteq [d]$,*
> *empirical distribution $\mathbf{p}_j$ for each $j \in J$,*
> *and probability threshold $t \in [0,1]$,*
> *estimate the value of $Q_J(t)$.*

In the next sections we answer this question in two parts:

- In Section 6.2, we upper and lower bound $Q_J$, the exposure of a set of features, in terms of the $Q_j$'s, the exposure of the individual features $j \in J$.
- In Section 7.1, we upper and lower bound $Q_j$, the exposure of an individual feature, in terms of $\widehat{Q}_j$, its observed exposure in another sample.

## 6.2. Composition theorems (from marginals to joint)

As a simple adversarial example to illustrate the hardness of relating the exposure of individual columns to the exposure of their joint, consider the following $(n+1)$-by-2 binary database $M$:

$$M = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 0 & 1 \\ 0 & 1 \\ \vdots & \vdots \end{bmatrix} \begin{matrix} \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} \frac{n}{2} \\ \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} \frac{n}{2} \end{matrix}$$

It is easy to see that this matrix is $\frac{n}{2}$-anonymous for column 1 and column 2 individually. However, notice that the "middle" user is unique, so the entire matrix $M$ is only 1-anonymous.

This example shows that even when the anonymity of each column is high, there is no general strict guarantee that can be provided for the anonymity of all users. It also suggests that, when the anonymity of the individual columns is high, the fraction of users for which the anonymity is violated is small. The next theorems precisely quantify how large this fraction can be.

First, we upper bound the exposure of the empirical joint distribution of features $\mathbf{p}_J$ in terms of the exposure and support size of each individual feature distribution $\mathbf{p}_j$.

**Theorem 2 (Composition with known support sizes).** *For any subset of columns $J \subseteq [d]$ , probability thresholds $\{t_j\}_{j \in J}$, and any $j^\star \in J$:*

$$Q_J\left(\prod_{j \in J} t_j\right) \le \sum_{j \in J} Q_j(t_j) + \sum_{j \in J \setminus \{j^\star\}} t_j |V_j|$$

This bound is particularly useful when the support size of each column is small. When their support size are large or unknown, its usefulness deteriorates. The next theorem replaces the support size with a free parameter $c$ that can be optimized by the curator to decide which columns should be redacted when doing the global suppression.

**Theorem 3 (General composition rule).** *For any subset of columns $J \subseteq [d]$, probability thresholds $\{t_j\}_{j \in J}$, and free parameter $c \in (0,1)$:*

$$Q_J\left(c \prod_{j \in J} t_j\right) \le \sum_{j \in J} Q_j(t_j) + c$$

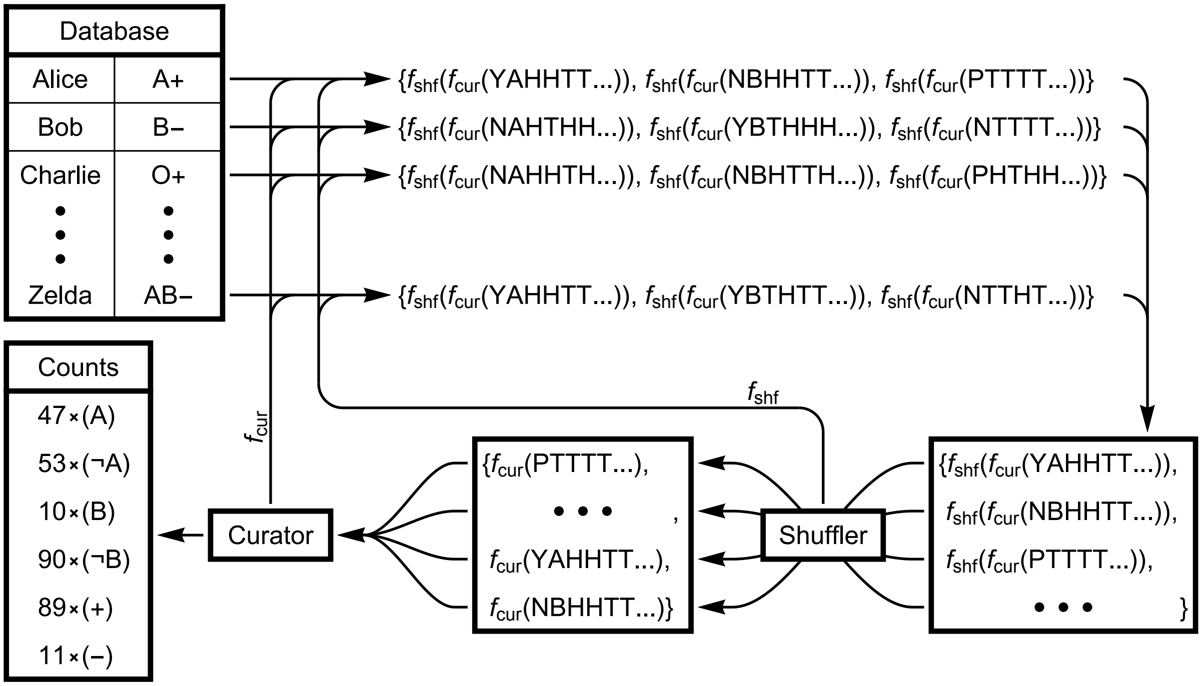
Database
Alice
A+
Bob
B−
Charlie
O+
Zelda
AB−
{fshf(fcur(YAHHTT...)), fshf(fcur(NBHHTT...)), fshf(fcur(PTTTT...))}
{fshf(fcur(NAHTHH...)), fshf(fcur(YBTHHH...)), fshf(fcur(NTTTT...))}
{fshf(fcur(NAHHTH...)), fshf(fcur(NBHTTH...)), fshf(fcur(PHTHH...))}
{fshf(fcur(YAHHTT...)), fshf(fcur(YBTHTT...)), fshf(fcur(NTTHT...))}
Counts
47×(A)
53×(¬A)
10×(B)
90×(¬B)
89×(+)
11×(−)
fcur
fshf
Curator
{fcur(PTTTT...),
• • • ,
fcur(YAHHTT...),
fcur(NBHHTT...)}
Shuffler
{fshf(fcur(YAHHTT...)),
fshf(fcur(NBHHTT...)),
fshf(fcur(PTTTT...)),
• • • }

Figure 1: **Anonymization protocol to obtain only the counts of individual features in a database.**
Schema of how the curator can learn the counts of the individual features (i.e., the marginal distributions) without learning the full joint distribution (i.e., their correlations). This protocol requires only the use of public-key cryptography as a primitive, and is example of a mix network (Sampigethaya & Poovendran, 2006).
For illustration, we use a fictitious database of blood types. A person's blood type (the joint variable) is determined by the value of 3 binary features (the marginal variables): the presence or absence of antigens $A$, $B$, and $Rh(D)$ in their red blood cells. As two examples, for Charlie's type of $O+$, the "$O$" corresponds to the absence of both the $A$ and $B$ antigens, and the "$+$" to the presence of antigen $Rh(D)$, whereas Zelda's type of $AB-$ corresponds to the presence of $A$ and $B$ and the absence of $Rh(D)$.
**1.** The protocol begins with the curator and the shuffler each generating their own public/private key pair, and sending the public key ($f_{\mathrm{cur}}$ and $f_{\mathrm{shf}}$, respectively) to the participants for encrypting their data.
**2.** Participants prepare each of their features by: putting the value for the feature in a standardized form (e.g., YA, for presence of $A$, N for absence of $Rh(D)$, etc); adding a random suffix (e.g., HHTHT...); then encrypting using first $f_{\mathrm{shf}}$, and then $f_{\mathrm{cur}}$. The participant then send these messages (one for each feature) to the shuffler.
**3.** The shuffler secretly shuffles all the messages, then decrypts each of them using their own key, and passes these (now singly encrypted) messages to the curator.
**4.** The curator decrypts the message, obtaining the counts for each feature.

These lower bounds on the probability threshold $t$ for the joint distribution given by Theorems 2 and 3 depend on the product of the probability thresholds for the marginal distributions. Clearly, this product decays exponentially with the number of marginal distributions $|J|$, which means the guarantees from these theorems become weaker when $|J|$ is large. Also notice that these guarantees require an additional *slack*, either in terms of the support size or the tunable parameter $c$. The following theorem show that this slack term is necessary in general.

**Theorem 4** (**The "slack" term is necessary**).
*Let $c \in (0, 1]$ such that $\frac{1}{c}$ is an integer. There exists a matrix $M$, subset of columns $J \subseteq [d]$, and probability thresholds probability thresholds $\{t_j\}_{j \in J}$, such that:*

$$Q_J\left(c \prod_{j \in J} t_j\right) \geq \sum_{j \in J} Q_j(t_j)$$

### 6.3. Applications to real-world data

As a concrete example, we applied the above bounds (Theorems 2 and 3) to a dataset from the UCI repository (Murphy, 1992), containing data extracted from the 1994 US Census

(Kohavi, 1996). It contains 32561 users and 14 features (in our analysis we used 4 of those features, with support sizes ranging from 2 to 9). Figure 2 displays both the true exposure and our bound for the exposure obtained by using both Theorems 2 and 3, and taking the minimum.

## 7. Statistical database setting

We now turn attention to the setting where the values of the individual features and the data to be released do not come from the same set of users. To model such a setting, we assume that the data from both sets of users are sampled from the same underlying distribution. The browser fingerprinting example we mentioned in Section 2.3 is an application for this setting.

### 7.1. Estimation of the exposure

In this case, the curator can estimate the distribution of values across users. This distribution can then be used to estimate the exposure for the set of users who will release the data to the analyst. It is natural to ask what guarantees can be given on the exposure of this data. Our first result is a bound on the plug-in estimator of the exposure $\widehat{Q}$.

**Theorem 5** (**Plug-in estimator for the exposure).**
*Let $\gamma > 0$ and $\delta \in [0, 1]$.*
*If the number of samples is $n \geq \frac{\log \frac{1}{\delta} + \log |V_j|}{2\gamma^2}$,*
*then with probability at least $1 - \delta$:*

$$\widehat{Q}_j(t - \gamma) - \gamma |V_j| \leq Q_j(t) \leq \widehat{Q}_j(t + \gamma) + \gamma |V_j|.$$

This theorem provides us with a way to quantify the difference between the empirical estimator of exposure $\widehat{Q}_j$ and the true exposure $Q_j$. Notice however that due to the discontinuity of $\widehat{Q}_j$ the bound can become vacuous even for small values of $\gamma$. The following lower bound shows that this issue is true of any estimator of the exposure, not only the plug-in estimator.

**Theorem 6** (**Hardness of estimating the exposure).**
*Let $\Delta_s = \{\sum_{i=1}^{s} p_i = 1 : p_i \geq 0 \ \forall i\}$ be the set of distributions over $\{1, \ldots, s\}$ and $\mathcal{F}_n$ be the the set of measurable functions mapping $\{1, \ldots, s\}^n \mapsto [0, 1]$. Then*

$$\lim_{n \to \infty} \mathcal{R}_n \geq \frac{1}{16s}$$

*where the minimax risk $\mathcal{R}_n$ of exposure is defined as*

$$\inf_{f \in \mathcal{F}_n} \sup_{\mathbf{p} \in \Delta_s} \mathbb{E}_{x_1, \ldots, x_n \sim \mathbf{p}} \left[ \sup_{t \in [0,1]} \left| f(x_1, \ldots, x_n) - Q_{\mathbf{p}}(t) \right| \right].$$

This formalizes the fact that the exposure function cannot be estimated with arbitrary small error. More concretely, think of the case when the threshold is equal to one of the probabilities $p_i$. In this case, one would need to estimate the probability of that observation with zero error to get an estimate of the true exposure with also zero error, which clearly requires infinitely many samples. This issue is inherent to several measures that are based on thresholded statistic of a cumulative distribution function (CDF) such as quantiles (Chen & Zhang, 2020). The next section offers a possible remedy.

### 7.2. Statistical exposure (new privacy metric)

As we have just shown, estimating the exposure of a random database from samples is hard. We leverage the assumption that the observed and released databases are sampled independently from the same distribution $\mathbf{p}$, and instead estimate the *statistical exposure* (see also Appendix D):

**Definition 7** (**Statistical exposure).** The *statistical exposure*, $\mathcal{Q}_{\mathbf{p}}(n, k)$, is the probability that a random user in a database of size of $n$ sampled i.i.d. from the discrete probability distribution $\mathbf{p}$ is less than $k$-anonymous:[4]

$$\mathcal{Q}_{\mathbf{p}}(n, k) := \sum_{i=1}^{|V|} p_i I_{1-p_i}\big(n - (k-1), k-1\big), \qquad (3)$$

where the function $I$ is the regularized incomplete beta function:

$$I_p(a, b) \equiv \frac{B(a, b; p)}{B(a, b; 1)},$$
$$B(a, b; p) \equiv \int_0^p z^{a-1} (1 - z)^{b-1} dz.$$

The main advantage of the statistical exposure is that, unlike exposure, it can be accurately estimated with access to an estimate of $\mathbf{p}$ as shown by the following theorem.

**Theorem 8** (**Estimating the statistical exposure).**
*Let $\mathbf{p}$ be the true distribution, and $\widehat{\mathbf{p}}$ its empirical frequency. Then, for all $k$ and $n$:*

$$|\mathcal{Q}_{\mathbf{p}}(n, k) - \mathcal{Q}_{\widehat{\mathbf{p}}}(n, k)| \leq C\sqrt{n} \|\mathbf{p} - \widehat{\mathbf{p}}\|_\infty, \qquad (4)$$

*where $C$ is a constant that depends linearly on the support size of $\mathbf{p}$ (see Appendix G).*

Figure 3 illustrates the difficulty of estimating the exposure from samples, and that the statistical exposure is a more reliable estimator with less variance.

[4] The CDF of the binomial distribution, $\mathbf{p}(x \leq k)$, is given by $I_{1-p}\big(n - k, k + 1\big)$. For a given user in a database of size $n$ to be less than $k$-anonymous, there must be at most $k - 2$ with the same features out of $n - 1$ other i.i.d. samples, hence the term $I_{1-p}\big((n-1)-(k-2), (k-2)+1\big) = I_{1-p}\big(n-(k-1), k-1\big)$.

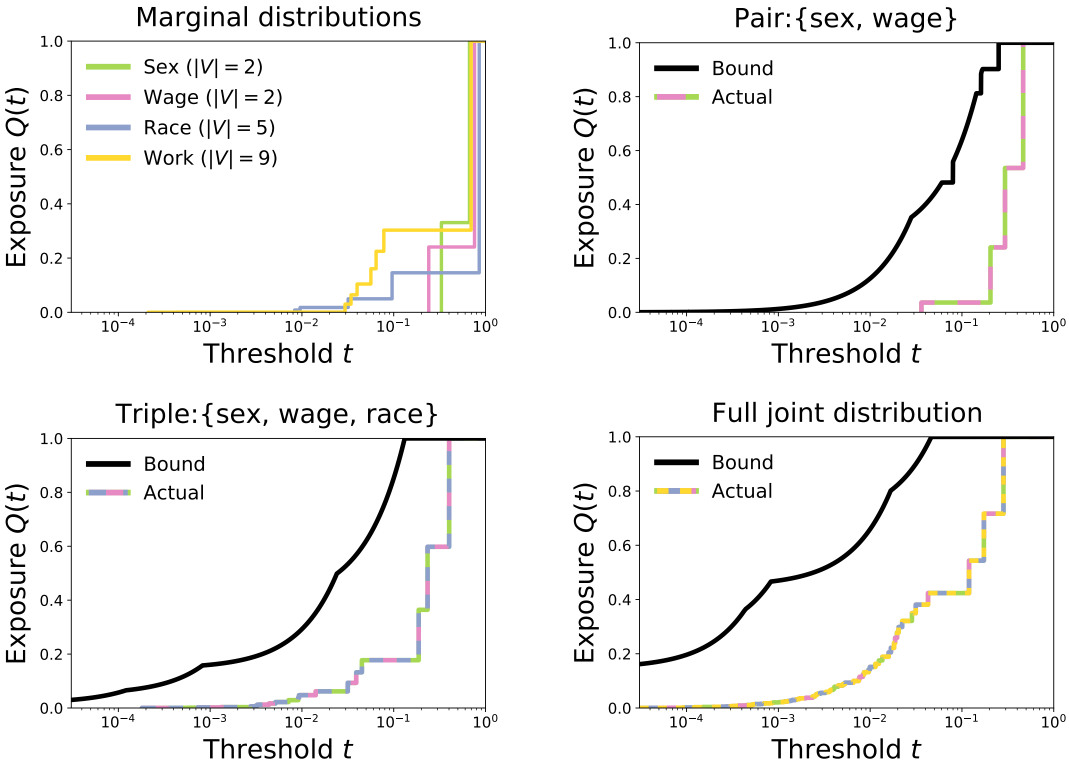

Figure 2: **Examples of exposure curves for real-world data.**
Here, we consider 4 features from the 1994 US Census dataset (Kohavi, 1996). In order of appearance, the number of different output values for that feature are: 2 (sex), 2 (wage, thresholded at $50k$ per year), 5 (race), and 9 (work, e.g., Federal, self-employed, never worked). Upper left shows the exposure curves for the frequencies of these individual features. The remaining plots show how the exposure curve (lighter/colored lines) changes as we consider the joint distribution of an increasing number of features. In black is the minimum of all possible upper bounds given by Theorems 2 and 3.

### 7.3. Relationship with Shannon entropy (a metric used for browser fingerprinting)

The browser fingerprinting community has frequently used the Shannon entropy as a metric of how identifiable users are given the information provided by different Application Programming Interfaces (APIs) (e.g., screen size, browser type, font installed). Given a distribution over values in a database $\mathbf{p} = (p_1, \dots, p_m)$, its Shannon entropy $H(\mathbf{p})$ is defined as:

$$H(\mathbf{p}) = -\sum_{i=1}^{m} p_i \log(p_i) = \mathbb{E}_{I\sim\mathbf{p}}[\log(p_i)].$$

One appealing property of the entropy is that marginal entropies can be used to bound the entropy of the full distribution via inclusion-exclusion principles. In addition, it can be accurately estimated with even when the number of samples is less than the support size of the distribution (Jiao et al., 2015).

However, a common (erroneous) interpretation of entropy is that if a database consists of $n$ users and the entropy is $B$ bits, then each value in the database is shared by $\sim n/2^B$ users. While encouraging, this interpretation is true only when $\mathbf{p}$ is uniform (or close to it). The following proposition shows a more accurate interpretation of the entropy by using the exposure.

**Proposition 9** (**Exposing the entropy**).
*Let $\mathbf{p}$ denote a distribution and $Q(t)$ denote its exposure at threshold $t$. The following relations between entropy and $Q(t)$ hold:*

$$H(\mathbf{p}) = \int_0^1 \frac{Q(t)}{t} dt \qquad \textit{and} \qquad Q(t) \leq -\frac{H(\mathbf{p})}{\log(t)}.$$

Concretely, consider the scenario of a database with $n = 2^{16} \sim 65000$ users and a distribution $\mathbf{p}$ with $H(\mathbf{p}) = 8$. The common interpretation would suggest that the majority of users cannot be identified up to $2^8 = 256$ users. On the other hand if $t = 2^{-16}$, then $Q(t)$ corresponds to the fraction of users who can be uniquely identified, and the above bound implies that this could be up to $50\%$ of the users. In the Appendix Section H, we show this bound is tight.

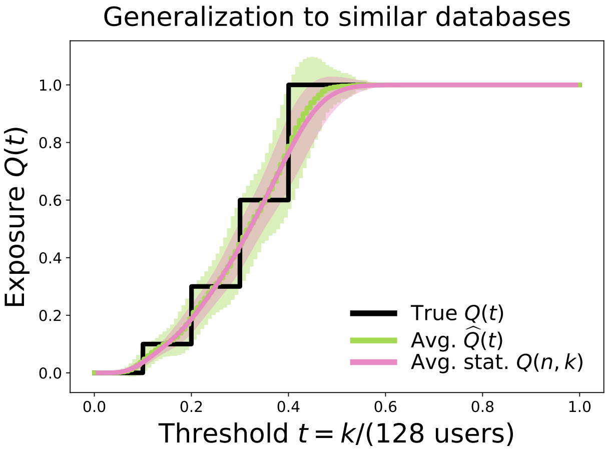

Figure 3: **The statistical exposure is a more robust estimator of the true (normalized) anonymity.**
We ran 1000 simulations of a database with 128 users sampled i.i.d. from a multinomial distribution $\mathbf{p}$ over 4 outputs, and computed the plug-in estimators for the exposure (green) and the statistical exposure (pink). When using one database to estimate the exposure of another one drawn from the same distribution, one expects fluctuations. By taking these into account, the statistical exposure has a smaller standard deviation (shading) than the plug-in estimator for the exposure.

## 8. Extensions

### 8.1. Locally redacting entries

In this paper, we considered global suppression; the curator redacts the same set of features for all users. Local suppression methods allow for a more targeted preservation of privacy, allowing for greater utility of the released database. To implement this within the presented framework, the curator would issue a set of conditional statements (e.g., “If you have value $v^\star$ for feature 1, redact it”).

One must be careful however, as the act of redacting a value itself contains some information. Indeed, in the example above, if only one user has value $v^\star$ for feature 1, nothing has changed. This could be overcome by asking the other users to randomly redact their value for that feature as well.

### 8.2. A hierarchical protocol

Throughout this paper, we have considered a two-step process: the curator learns about the marginal distributions, then informs the users what to redact when giving their information to the analyst. The set of marginal distributions is a rather coarse-grained view of the full joint distribution. To obtain a more precise picture, the curator requires some knowledge of the correlations between different feature values. This could be accomplished by allowing the curator to ask a sequence of queries with increasing complexity. At each step, the curator uses their current understanding of the distribution to decide which queries are likely “safe” to ask.

Concretely, consider quantifying correlations of increasing order. In the first round, the curator asks for individual feature values, thereby learning the expected frequency of each (the mean). In the second round, the curator asks the users to provide specific *pairs* of feature values, thereby learning something about the *covariance* between them. From the first round, the curator knows that some feature values are shared by only a few users. Asking about pairs involving those values is likely to cause significant privacy loss, so the curator specifically does not ask them of the users. Using covariance information from the second round, the curator asks for specific *triples* of feature values. This process could then continue until there are no “safe” queries. Granted with this more fine-grained picture of the distribution, the curator would make a decision as to how the users should release their data to the analyst.

### 8.3. Distributed private data aggregation

By taking a statistical approach to reducing trust requirements, the doors open for many exciting applications. For instance, imagine a group of users, all with their own private data, would like to know something about their collective statistics without compromising the privacy of any individual. The protocol in the previous sections could be used nearly verbatim, only now the participants, the curator, and the analyst, are all the same entity. With zero trust invested in anything but the protocol, such a method of distributed private data aggregation could prove to be a very useful tool.

## 9. Coda

To close, we remark that, as with any new idea, frameworks that claim to guarantee some level of privacy should be treated with caution. Does the metric capture the notion you are trying to quantify? It is difficult to judge “how private” something is if it is measured incorrectly. Are there proven bounds for this metric? Guarantees aren’t worth much if they are frequently false. To combat misrepresentation, either malicious or accidental, it is imperative that the problems being solved are appropriately practical, and that metrics used to evaluate performance are appropriately statistical.

## References

Abowd, J. M. The US Census Bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2867–2867, 2018.

Adell, J. and Jodrá, P. Exact Kolmogorov and total variation distances between some familiar discrete distributions. *Journal of Inequalities and Applications*, 2006(1):64307, 2006.

Aggarwal, G., Feder, T., Kenthapadi, K., Motwani, R., Panigrahy, R., Thomas, D., and Zhu, A. Approximation algorithms for $k$-anonymity. *Journal of Privacy Technology (JOPT)*, 2005.

Chen, Z. and Zhang, A. A survey of approximate quantile computation on large-scale data. *IEEE Access*, 8:34585–34597, 2020.

Cheu, A., Smith, A., Ullman, J., Zeber, D., and Zhilyaev, M. Distributed differential privacy via shuffling. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 375–403. Springer, 2019.

Crellin, W. and BCE, C. R. A survey of reimbursement practices of private health insurance companies for pharmaceuticals not covered. *Australian Health Review*, 35: 210, 2011.

Domingo-Ferrer, J., Sánchez, D., and Blanco-Justicia, A. The limits of differential privacy (and its misuse in data release and machine learning). *Communications of the ACM*, 64(7):33–35, 2021.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

Dwork, C., Roth, A., et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014.

El Emam, K., Dankar, F. K., Issa, R., Jonker, E., Amyot, D., Cogo, E., Corriveau, J.-P., Walker, M., Chowdhury, S., Vaillancourt, R., et al. A globally optimal $k$-anonymity method for the de-identification of health data. *Journal of the American Medical Informatics Association*, 16(5): 670–682, 2009.

Evfimievski, A., Gehrke, J., and Srikant, R. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 211–222, 2003.

Gkoulalas-Divanis, A., Loukides, G., and Sun, J. Publishing data from electronic health records while preserving privacy: A survey of algorithms. *Journal of Biomedical Informatics*, 50:4–19, 2014.

Jiao, J., Venkat, K., Han, Y., and Weissman, T. Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 61(5):2835–2885, 2015.

Khan, M., Foley, S., and O'Sullivan, B. From k-anonymity to differential privacy: A brief introduction to formal privacy models. 2021.

Kohavi, R. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pp. 202–207, 1996.

Laperdrix, P., Bielova, N., Baudry, B., and Avoine, G. Browser fingerprinting: A survey. *ACM Transactions on the Web (TWEB)*, 14(2):1–33, 2020.

Le Cam, L. An approximation theorem for the poisson binomial distribution. *Pacific Journal of Mathematics*, 10 (4):1181–1197, 1960.

Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3–es, 2007.

US Census Press Release CB21-CN.42. https://web.archive.org/web/20211021091202/https://www.census.gov/newsroom/press-releases/2021/2020-census-key-parameters.html.

Meyerson, A. and Williams, R. On the complexity of optimal $k$-anonymity. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 223–228, 2004.

Murphy, P. M. UCI repository of machine learning databases [machine-readable data repository]. *Technical report, Department of Information and Computer Science, University of California*, 1992.

Narayanan, A. and Shmatikov, V. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pp. 111–125. IEEE, 2008.

Park, D., Kim, J., Boyd, C., and Dawson, E. Cryptographic salt: A countermeasure against denial-of-service attacks. In *Australasian Conference on Information Security and Privacy*, pp. 334–343. Springer, 2001.

Sampigethaya, K. and Poovendran, R. A survey on mix networks and their secure applications. *Proceedings of the IEEE*, 94(12):2142–2181, 2006.

Sweeney, L. $k$-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.*, 10(5): 557–570, 2002.

## A. Composition rule with known support size (Theorem 2)

For this proof, we use the following two lemmas:

**Lemma 10 (Logsum inequality).**
*Let $f : \mathbb{R}_+ \to \mathbb{R}$ be a function such that $g(p) = pf(p)$ is a concave function.*
*If $((a_i, b_i))_{i=1}^n > 0$, then:*

$$\sum_{i=1}^{n} a_i f\left(\frac{a_i}{b_i}\right) \leq \left(\sum_{i=1}^{n} a_i\right) f\left(\frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}\right)$$

*Proof.* Let $B = \sum_{i=1}^{n} b_i$, then we have:

$$\sum_{i=1}^{n} a_i f\left(\frac{a_i}{b_i}\right) = B\sum_{i=1}^{n} \frac{b_i}{B}\frac{a_i}{b_i} f\left(\frac{a_i}{b_i}\right) = B\sum_{i=1}^{n} \frac{b_i}{B} g\left(\frac{a_i}{b_i}\right)$$

Since $g$ is a concave function, we can use Jensen's inequality to bound the previous expression by:

$$\leq Bg\left(\sum_{i=1}^{n} \frac{b_i}{B}\frac{a_i}{b_i}\right) = Bg\left(\frac{1}{B}\sum_{i=1}^{n} a_i\right) = \left(\sum_{i=1}^{n} a_i\right) f\left(\frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}\right)$$

□

**Lemma 11.**
*Let $G_q$ be a function $G_q \colon [0, 1] \to [0, 1]$ defined as:*

$$G_q(p) = \min\left(1, \frac{q(1-p)}{p(1-q)}\right)$$

*Then, for every $q \in [0, 1]$:*

- *The function $p \mapsto pG_q(p)$ is concave*
- *$pG_q(p) \geq p$ (for $p < q$), and*
  *$pG_q(p) < q$ (always).*

*Proof.* It is easy to see that $pG_q(p) = \min\left(p, \frac{q(1-p)}{1-q}\right)$ is concave, as it is the minimum of two linear functions. These two linear functions are equal at $q$, so the maximum of $pG_q(p)$ is $q$, with argument $p = q$. □

Now, we first prove the theorem for two marginal distributions: creatively named "1" and "2":

$$Q_J(q_1q_2) \leq Q_1(q_1) + Q_2(q_2) + q_2|V_2|, \tag{5}$$

where $|V_2|$ is the size of the support of marginal 2.

*Proof.* By definition of exposure we have:

$$\begin{aligned} 1 - Q_J(q_1q_2) &= \sum_{v_1}\sum_{v_2} p(v_1, v_2)\mathbf{1}\left\{p(v_1, v_2) \geq q_1q_2\right\} \\ &\geq \sum_{v_1}\sum_{v_2} p(v_1, v_2)\mathbf{1}\left\{p(v_1, v_2) \geq q_1q_2\right\}\mathbf{1}\left\{p(v_1) \geq q_1\right\} \end{aligned} \tag{6}$$

Notice that if $p(v_1, v_2)/p(v_1) \geq q_2$ and $p(v_1) \geq q_1$ then $p(v_1, v_2) \geq q_1 q_2$.
Thus, we can lower bound the above expression by

$$\begin{aligned}
&\sum_{v_1}\sum_{v_2} p(v_1, v_2)\mathbf{1}\left\{\frac{p(v_1, v_2)}{p(v_1)} \geq q_2\right\}\mathbf{1}\left\{p(v_1) \geq q_1\right\} \\
&= \sum_{v_1}\sum_{v_2} p(v_1, v_2)(1 - \mathbf{1}\left\{\frac{p(v_1, v_2)}{p(v_1)} < q_2\right\})(1 - \mathbf{1}\left\{p(v_1) < q_1\right\}) \\
&= 1 + \sum_{v_1}\sum_{v_2} p(v_1, v_2)\left(\mathbf{1}\left\{\frac{p(v_1, v_2)}{p(v_1)} < q_2\right\}\mathbf{1}\left\{p(v_1) < q_1\right\} - \mathbf{1}\left\{\frac{p(v_1, v_2)}{p(v_1)} < q_2\right\} - \mathbf{1}\left\{p(v_1) < q_1\right\}\right) \\
&\geq 1 - \sum_{v_1} p(v_1)\mathbf{1}\left\{p(v_1) < q_1\right\} - \sum_{v_1}\sum_{v_2} p(v_1, v_2)\mathbf{1}\left\{\frac{p(v_1, v_2)}{p(v_1)} < q_2\right\} \\
&= 1 - Q_1(q_1) - \sum_{v_1}\sum_{v_2} p(v_1, v_2)\mathbf{1}\left\{\frac{p(v_1, v_2)}{p(v_1)} < q_2\right\},
\end{aligned}$$

where we used the fact that $\sum_{v_2} p(v_1, v_2) = p(v_1)$ for the second to last equality.

Combining this bound with (6) and rearranging terms we have:

$$Q_J(q_1 q_2) \leq Q_1(q_1) + \sum_{v_1}\sum_{v_2} p(v_1, v_2)\mathbf{1}\left\{\frac{p(v_1, v_2)}{p(v_1)} < q_2\right\} \tag{7}$$

By Lemma 11, we have the following upper bound:

$$\begin{aligned}
\sum_{v_1}\sum_{v_2} p(v_1, v_2)\mathbf{1}\left\{\frac{p(v_1, v_2)}{p(v_1)} < q_2\right\} &= \sum_{v_2}\sum_{v_1} p(v_1, v_2)\mathbf{1}\left\{\frac{p(v_1, v_2)}{p(v_1)} < q_2\right\} \\
&\leq \sum_{v_2}\sum_{v_1} p(v_1, v_2) G_{q_2}\left(\frac{p(v_1, v_2)}{p(v_1)}\right)
\end{aligned}$$

We can now apply Lemma 10 and use the fact that $\sum_{v_1} p(v_1) = 1$ and $\sum_{v_1} p(v_1, v_2) = p(v_2)$ to upper bound the previous expression by:

$$\sum_{v_2} p(v_2) G_{q_2}(p(v_2)) = \sum_{v_2} p(v_2) G_{q_2}(p(v_2))\mathbf{1}\left\{p(v_2) < q_2\right\} + \sum_{v_2} p(v_2) G_{q_2}(p(v_2))\mathbf{1}\left\{p(v_2) \geq q_2\right\}$$

However, we know that $p(v_2)G_{q_2}(p(v_2)) = p(v_2)$ for $p(v_2) < q_2$ and we can upper bound $p(v_2)G_{q_2}(p(v_2))$ by $q_2$. Therefore, we have:

$$\begin{aligned}
\sum_{v_2} p(v_2) G_{q_2}(p(v_2)) &\leq \sum_{v_2} p(v_2)\mathbf{1}\left\{p(v_2) < q_2\right\} + q_2\mathbf{1}\left\{p(v_2) \geq q_2\right\} \\
&\leq \sum_{v_2} p(v_2)\mathbf{1}\left\{p(v_2) < q_2\right\} + q_2|V_2| = Q_2(q_2) + q_2|V_2|
\end{aligned}$$

Equation 5 can be obtained by replacing this bound in (7).

To prove the general case, we use induction as follows.
Let $J$ denote an arbitrary index set of size $n \geq 2$, and let $j^*, j'$ denote two arbitrary elements of $J$.
By equation (5), we know that

$$Q_J\left(\prod_{j \in J} t_j\right) \leq Q_{J/\{j'\}}\left(\prod_{j \in J/\{j'\}} t_j)\right) + Q_{j'}(t_{j'}) + t_{j'}|V_{j'}|.$$

Since $j^* \in J/\{j'\}$, we can apply induction to the first term on the right hand side of the above equation to obtain:

$$Q_J\left(\prod_{j\in J} t_j\right) \leq \sum_{j\in J/\{j'\}} Q_j(t_j) + \sum_{j\in J/\{j'\}:\, j\neq j^*} t_j|V_j| + Q_{j'}(t'_j) + t_{j'}|V_{j'}|.$$

The proof follows by rearranging terms in the above expression.

□

## B. General composition rule (Theorem 3)

*Proof.* Let $f_j(i) = M_{ij}$ be the function that returns the value of column $j$ for row $i$.
Assume $J = \{1, \ldots, k\}$, and let $f_1 \times \cdots \times f_k$ denote the function $u \mapsto (f_1(u), \ldots, f_k(u))$.
For any function $f$ with domain $U$, let $V_f = \{f(u) : u \in U\}$ be the range of $f$.
Let $g = f_1 \times \cdots \times f_k$.

To prove the theorem, it suffices to show that if

$$\left|\left\{u \in U : \mathbf{p}_i(f_i(u)) \geq \frac{1}{2^b_i}\right\}\right| \geq (1-\delta_i)n.$$

for all $i$, then

$$\left|\left\{u \in U : \mathbf{p}_J(g(u)) \geq \frac{c}{2^{\sum_i b_i}}\right\}\right| \geq (1 - \sum_i \delta_i - c)n.$$

Let $V_i^+ = \left\{v \in V_i : \mathbf{p}_i(v) \geq \frac{1}{2^{b_i}}\right\}$ and $U_i^+ = \{u \in U : f_i(u) \in V_i^+\}$.
Clearly

$$|V_i^+| \leq 2^{b_i} \text{ and } |U_i^+| \geq (1-\delta_i)n.$$

Let $V_g^+ = (V_1^+ \times \cdots \times V_k^+) \cap V_g$ and $U_g^+ = \{u \in U : g(u) \in V_g^+\}$.
We immediately have

$$|V_g^+| \leq 2^{\sum_i b_i} \text{ and } |U_g^+| \geq \left(1 - \textstyle\sum_i \delta_i\right) n$$

the latter by taking a union bound.

Let $V_g^- = \left\{v \in V_g^+ : \mathbf{p}_J(v) < \frac{c}{2^{\sum_i b_i}}\right\}$ and $U_g^- = \{u \in U_g^+ : g(u) \in V_g^-\}$.
Since $|V_g^+| \leq 2^{\sum_i b_i}$, we have

$$|U_g^-| = \left(\textstyle\sum_{v\in V_g^-} \mathbf{p}_J(v)\right) n \leq cn.$$

Thus:

$$\left|\left\{u \in U : p_J(g(u)) \geq \frac{c}{2^{\sum_i b_i}}\right\}\right| \geq |U_g^+| - |U_g^-| \geq \left(1 - \sum_i \delta_i - c\right) n$$

completing this proof. □

## C. Tightness of these composition theorems (Theorem 4)

*Proof.* Let $a = \frac{3}{c}$, which is an integer by assumption.
Let $M$ be a matrix with $a2^k$ rows and $k$ columns, where $k$ is a positive integer whose value will be specified below.
And, let $J = \{1, \ldots, k\}$ and $J_i = \{i\}$ for each $i \in J$. In other words, $J$ contains all the columns of $M$.

For simplicity, assume that the values of the entries of $M$ belong to the set $\{0, 1, \perp\}$ ($\perp$ for redacted).
Then, partition the $a2^k$ rows of $M$ into $2^k$ groups, with $a$ rows per group. All the rows in each group will contain identical

values for every column, with $k$ exceptions.
Specifically, number the groups from $0$ to $2^k - 1$, and assign the binary encoding of $i$ to the $i^{\text{th}}$ group, with one bit per column. However, for $k$ arbitrarily chosen groups, replace one of the columns in one of the rows with $\perp$, choosing a different column for each row.

Observe that, in each column, $a2^{k-1}$ rows are assigned one of the values in $\{0, 1\}$, and $a2^{k-1} - 1$ rows are assigned the other value, with the remaining row assigned $\perp$. Thus

$$Q_i\left(\frac{a2^{k-1}-1}{a2^k}\right) = \frac{1}{a2^k}$$

for all $i \in J$.

Also, since $a \geq 3$, we know that exactly $k$ rows are unique (specifically, the $k$ rows containing a $\perp$), and thus

$$Q_J\left(\frac{2}{a2^k}\right) = \frac{k}{a2^k}.$$

Now, choose $k$ large enough so that

$$c\left(\frac{a2^{k-1}-1}{a2^k}\right)^k \geq \frac{2}{a2^k},$$

which is possible since the limit of the ratio of the two sides of this inequality is less than $1$:

$$\lim_{k\to\infty} \frac{\frac{2}{a2^k}}{c\left(\frac{a2^{k-1}-1}{a2^k}\right)^k} = \lim_{k\to\infty} \frac{\frac{1}{a2^{k-1}}}{c\left(\frac{1}{2}-\frac{1}{a2^k}\right)^k} = \lim_{k\to\infty} \frac{1}{ac} \cdot \frac{1}{\left(\frac{1}{2}-\frac{1}{a2^k}\right)} \cdot \left(\frac{\frac{1}{2}}{\frac{1}{2}-\frac{1}{a2^k}}\right)^{k-1} = \frac{1}{3} \cdot 2 \cdot 1 = \frac{2}{3}.$$

Let $q_i = \frac{a2^{k-1}-1}{a2^k}$ for all $i \in J$.
Putting everything together, we have

$$\begin{aligned} Q_J\left(c\prod_i q_i\right) &= Q_J\left(c\left(\frac{a2^{k-1}-1}{a2^k}\right)^k\right) \geq Q_J\left(\frac{2}{a2^k}\right) = \frac{k}{a2^k} \\ &= \sum_i \frac{1}{a2^k} = \sum_i Q_i\left(\frac{a2^{k-1}-1}{a2^k}\right) = \sum_i Q_i(q_i) \end{aligned}$$

where the inequality follows because $Q_J(q)$ is monotonically non-decreasing in $q$. □

## D. Derivation of the statistical exposure

First, recall that the cumulative distribution of a Binomial distribution with parameters $n$ and $p$ is:

$$F(k; n, p) = P(X \leq k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i} = I_{1-p}\big(n-k, k+1\big).$$

The statistical exposure $\mathcal{Q}_{\mathbf{p}}(n, k)$ is defined as the probability that a random user in a database composed of $n$ users sampled i.i.d. from $\mathbf{p}$ is less than $k$-anonymous. That is:

$$
\begin{aligned}
\mathcal{Q}_{\mathbf{p}}(n,k) &= \frac{1}{n}\sum_{i=1}^{|V|}\sum_{j=0}^{k-1}\binom{n}{j} j p_i^j (1-p_i)^{n-j} \\
&= \frac{1}{n}\sum_{i=1}^{|V|}\sum_{j=1}^{k-1} j\frac{n!}{j!(n-j)!} p_i^j (1-p_i)^{n-j} \\
&= \frac{1}{n}\sum_{i=1}^{|V|}\sum_{j=1}^{k-1} n\frac{(n-1)!}{(j-1)!(n-j)!} p_i^j (1-p_i)^{n-j} \\
&= \sum_{i=1}^{|V|}\sum_{j=1}^{k-1}\binom{n-1}{j-1} p_i^j (1-p_i)^{n-j}
\end{aligned}
$$

where on the second line we start the sum over $j$ at $1$ as the term associated with $j = 0$ is zero.

Let $j' = j - 1$ and $n' = n - 1$:

$$
\begin{aligned}
\mathcal{Q}_{\mathbf{p}}(n,k) &= \sum_{i=1}^{|V|}\sum_{j'=0}^{k-2}\binom{n'}{j'} p_i^{j'+1} (1-p_i)^{(n'-1)-(j'-1)} \\
&= \sum_{i=1}^{|V|}\sum_{j'=0}^{k-2} p_i\binom{n'}{j'} p_i^{j'} (1-p_i)^{(n'-j')} \\
&= \sum_{i=1}^{|V|} p_i I_{1-p_i}\big(n' - (k-2), (k-2)+1\big) \\
&= \sum_{i=1}^{|V|} p_i I_{1-p_i}\big((n-1) - (k-2), k-1\big) \\
&= \sum_{i=1}^{|V|} p_i I_{1-p_i}\big(n - (k-1), k-1\big).
\end{aligned}
$$

## E. Error bound for the plug-in estimator of the exposure (Theorem 5)

*Proof.* Note that if $|\mathbf{p}_j(v) - \widehat{\mathbf{p}}_j(v)| \leq \gamma$ for all $v \in V_j$ then

$$
\begin{aligned}
Q_j(t) &= \sum_{v\in V_j} \mathbf{p}_j(v)\cdot \mathbf{1}\left\{\mathbf{p}_j(v) < t\right\} \\
&\leq \sum_{v\in V_j} (\widehat{\mathbf{p}}_j(v) + \gamma)\cdot \mathbf{1}\left\{\widehat{\mathbf{p}}_j(v) - \gamma < t\right\} \\
&= \sum_{v\in V_j} \widehat{\mathbf{p}}_j(v)\cdot \mathbf{1}\left\{\widehat{\mathbf{p}}_j(v) < t + \gamma\right\} + \gamma \sum_{v\in V_j} \mathbf{1}\left\{\widehat{\mathbf{p}}_j(v) < t + \gamma\right\} \\
&\leq \widehat{Q}_j(t+\gamma) + \gamma|V_j|
\end{aligned}
$$

and

$$
\begin{aligned}
Q_j(t) &= \sum_{v\in V_j} \mathbf{p}_j(v)\cdot \mathbf{1}\left\{\mathbf{p}_j(v) < t\right\} \\
&\geq \sum_{v\in V_j} (\widehat{\mathbf{p}}_j(v) - \gamma)\cdot \mathbf{1}\left\{\widehat{\mathbf{p}}_j(v) + \gamma < t\right\} \\
&= \sum_{v\in V_j} \widehat{\mathbf{p}}_j(v)\cdot \mathbf{1}\left\{\widehat{\mathbf{p}}_j(v) < t - \gamma\right\} - \gamma \sum_{v\in V_j} \mathbf{1}\left\{\widehat{\mathbf{p}}_j(v) < t - \gamma\right\} \\
&\geq \widehat{Q}_j(t-\gamma) - \gamma|V_j|.
\end{aligned}
$$

Moreover, by Hoeffding's inequality and the union bound, we have

$$\Pr\left[\max_{v\in V_j}|\mathbf{p}_j(v)-\widehat{\mathbf{p}}_j(v)|>\gamma\right]\le |V_j|\exp(-2m\gamma^2). \qquad \square$$

## F. Hardness of estimating the exposure using any estimator (Theorem 6)

We use the following form of Le Cam's theorem (Le Cam, 1960):

**Theorem 12.**
*Suppose there exist $\mathbf{p}_0$ and $\mathbf{p}_1$ from some parametric family of distributions $\mathcal{P}$ such that $KL(\mathbf{p}_0||\mathbf{p}_1)\le\frac{\log 2}{n}$. Then*

$$R_n=\inf_{\widehat{\theta}}\sup_{\mathbf{p}\in\mathcal{P}}\mathbb{E}[d(\widehat{\theta}(X_1,\dots,X_n),\theta(\mathbf{p}))]\ge\frac{d(\theta(\mathbf{p}_0),\theta(\mathbf{p}_1))}{16}$$

*where $d$ is a semi-metric.*

*Proof.* Let us fix $d$ as $d(x,y)=|x-y|$ and let $\mathcal{P}=\Delta_{s+1}$ be the set of discrete distributions represented by the $s+1$ dimensional probability simplex.
Next, we define $\mathbf{p}_0=(p_{0,1},\dots,p_{0,s+1})$ and $\mathbf{p}_1=(p_{1,1},\dots,p_{1,s+1})$ as

$$p_{0,i}=\begin{cases}\frac{1}{s^2}-\varepsilon & \text{if } i\in[s]\\ \frac{s-1+s^2\varepsilon}{s} & \text{if } i=s+1\end{cases}$$

and

$$p_{1,i}=\begin{cases}\frac{1}{s^2} & \text{if } i\in[s]\\ \frac{s-1}{s} & \text{if } i=s+1\end{cases}$$

where $\varepsilon\le\frac{1}{s^2}$.
We use

$$q=\frac{1}{s^2}-\frac{\varepsilon}{2}$$

with this at hand, we can show that the L1 difference of exposure for $\mathbf{p}_0$ and $\mathbf{p}_1$ is

$$d(Q(\mathbf{p}_0,q),Q(\mathbf{p}_1,q))=\left|\frac{1}{s}-s\varepsilon\right|$$

since $p_{0,s+1}>q$ as

$$\begin{aligned}p_{0,s+1}-q&=\frac{s-1+s^2\varepsilon}{s}-\frac{1}{s^2}+\frac{\varepsilon}{2}\\&\ge\frac{s^2-s-1}{s^2}\\&=1-\frac{s+1}{s^2}\end{aligned}$$

which is positive when $s\ge 2$ and $Q(\mathbf{p}_1,q)=0$ as $\frac{s-1}{s}\ge q=\frac{1}{s^2}-\frac{\varepsilon}{2}$.

Next, we upper bound the KL divergence as

$$\begin{aligned}
\mathrm{KL}(\mathbf{p}_0||\mathbf{p}_1) &= s\left(\frac{1}{s^2}-\varepsilon\right)\log\frac{1/s^2-\varepsilon}{1/s^2}+\frac{s-1+s^2\varepsilon}{s}\log\left(\frac{s-1+s^2\varepsilon}{s}\cdot\frac{s}{s-1}\right)\\
&= \left(\frac{1}{s}-s\varepsilon\right)\log(1-s^2\varepsilon)+\left(1+s\varepsilon-\frac{1}{s}\right)\log\left(1+\frac{s^2\varepsilon}{s-1}\right)\\
&= \left(\frac{1}{s}-s\varepsilon\right)\left(\log(1-s^2\varepsilon)-\log\left(1+\frac{s^2\varepsilon}{s-1}\right)\right)+\log\left(1+\frac{s^2\varepsilon}{s-1}\right)\\
&= \left(\frac{1}{s}-s\varepsilon\right)\log\left(1-\frac{s^3\varepsilon}{s-1+s^2\varepsilon}\right)+\log\left(1+\frac{s^2\varepsilon}{s-1}\right)\\
&\le \left(\frac{1}{s}-s\varepsilon\right)\frac{s^3\varepsilon}{1-s-s^2\varepsilon}+\frac{s^2\varepsilon}{s-1} \qquad (8)\\
&\le \frac{s^2\varepsilon}{1-s-s^2\varepsilon}+\frac{s^2\varepsilon}{s-1}\\
&= s^2\varepsilon\left(\frac{1}{s}+\frac{1}{s-1}\right)\\
&\le \frac{2s^2\varepsilon}{s-1}
\end{aligned}$$

where (8) follows from the fact that $\log(1+x)\le x$ for $x>-1$.

Setting $\varepsilon=\frac{s-1}{s^2n}$ allows us to apply Theorem 12, and we are done. □

## G. Error bound for the plug-in estimator of the statistical exposure (Theorem 8)

*Proof.* We start by noticing that the regularized incomplete betafunction $I_{1-p_i}(n-(k-1),k-1):=F(k;n,p)_i)$ corresponds to the cumulative distribution of a binomial random variable $\mathrm{Bin}(n,p_i)$ with parameters $n$ and $p_i$.
Thus, by definition of $\mathcal{Q}$ we have:

$$\begin{aligned}
\mathcal{Q}_{\mathbf{p}}(k,n)-\mathcal{Q}_{\widehat{\mathbf{p}}}(k,n) &= \sum_{i=1}^{|V|}p_iF(k;n,p)_i)-\widehat{p}_iF(n,k,\widehat{p}_i)\\
&= \sum_{i=1}^{|V|}p_i\left(F(k;n,p)_i)-F(n,k,\widehat{p}_i)\right)+F(n,k,\widehat{p}_i)(p_i-\widehat{p}_i)
\end{aligned}$$

So, using the fact that $F(k;n,p_i))\le 1$, we have that

$$\begin{aligned}
|\mathcal{Q}_{\mathbf{p}}(k,n)-\mathcal{Q}_{\widehat{\mathbf{p}}}(k,n)| &\le \sum_{i=1}^{|V|}p_i\left|F(k;n,p)_i)-F(n,k,\widehat{p}_i)\right|+\|\mathbf{p}-\widehat{\mathbf{p}}\|_1\\
&\le \sum_{i=1}^{|V|}p_i\left|F(k;n,p)_i)-F(n,k,\widehat{p}_i)\right|+|V|\|\mathbf{p}-\widehat{\mathbf{p}}\|_\infty \qquad (9)
\end{aligned}$$

We now proceed to bound the first term in the above equation, by Section 2.2 of (Adell & Jodrá, 2006), we have:

$$\begin{aligned}
|F(k;n,p)_i)-F(n,k,\widehat{p}_i| &\le d_{\mathrm{TV}}(\mathrm{Bin}(n,p_i),\mathrm{Bin}(n,\widehat{p}_i))\\
&\le \frac{\sqrt{e}}{2}\frac{\tau(|p_i-\widehat{p}_i|)}{1-\tau(|p_i-\widehat{p}_i|)^2}\\
&\le \frac{\sqrt{e}}{2}\tau(|p_i-\widehat{p}_i|).
\end{aligned}$$

where $d_{\mathrm{TV}}$ denotes the total variation distance and $\tau(x):=x\sqrt{\frac{n+1}{p_i(1-p_i)}}$.

Replacing this bound in (9), we obtain:

$$
\begin{aligned}
|\mathcal{Q}_{\mathbf{p}}(k,n) - \mathcal{Q}_{\widehat{\mathbf{p}}}(k,n)| &\leq \frac{\sqrt{e}}{2}\sum_{i=1}^{|V|}\sqrt{\frac{p_i(n+1)}{1-p_i}}|p_i - \widehat{p}_i| + |V|\|\mathbf{p} - \mathbf{p}\|_\infty \\
&\leq \frac{\sqrt{e}}{2}\|\mathbf{p} - \widehat{\mathbf{p}}\|_\infty\left(\sum_{i=1}^{|V|}\sqrt{\frac{p_i(n+1)}{1-p_i}} + |V|\right) \\
&\leq |V|\left(\frac{\sqrt{e(n+1)}}{2\sqrt{|V|-1}} + 1\right)\|\mathbf{p} - \widehat{\mathbf{p}}\|_\infty
\end{aligned}
$$

□

## H. Relationships between exposure and entropy

### H.1. Proof of Proposition 9

*Proof.* By definition of exposure $Q(t)$ we have:

$$
\begin{aligned}
\int_0^1 \frac{Q(t)}{t}dt &= \int_0^1 \frac{1}{t}\sum_{i=1}^n p_i \mathbf{1}\left\{p_i < t\right\} dt \\
&= \sum_{i=1}^n p_i \int_0^1 \frac{\mathbf{1}\left\{p_i < t\right\}}{t}dt \\
&= \sum_{i=1}^n p_i \int_{p_i}^1 \frac{1}{t}dt = -\sum_{i=1}^n p_i \log p_i = H(\mathbf{p})
\end{aligned}
$$

This proves the first statement of the proposition. To prove the upper bound on the exposure we use the fact that

$$
\begin{aligned}
Q(t) = P_{I\sim\mathbf{p}}(p_I < t) &= P_{I\sim\mathbf{p}}(-\log(p_I) > -\log t) \\
&\leq \frac{E_{I\sim\mathbf{p}}[-\log p_I]}{-\log t} = \frac{H(\mathbf{p})}{-\log t},
\end{aligned}
$$

where we have used Marokv's inequality. □

We now show that the above bound is tight.

**Proposition 13.**
*Let $B > 0$, $1 > t > 0$ be such that $-\frac{B}{\log t} < 1 - t$ and $-\frac{B}{t\log t}$ is an integer greater or equal to 1.*
*There exists a distribution $\mathbf{p}$ such that*

$$
H(\mathbf{p}) \leq B + \frac{1}{e} \quad \textit{and} \quad Q(t) = \frac{B}{-\log t}.
$$

*Proof.* Let $n = \frac{B}{-t\log t}$, and let $\mathbf{p} \in \mathbb{R}^{n+1}$ be defined as:

$$
p_i = t \quad \text{if } i \leq n \quad \text{and} \quad p_{n+1} = 1 - nt = 1 + \frac{B}{\log t}.
$$

Note that since $1 + \frac{B}{\log t} > t$, it follows that $Q(t) = \sum_{i=1}^n p_i = nt = -\frac{B}{\log t}$.

On the other hand, the entropy of this distribution is given by:

$$
-\sum_{i=1}^n p_i \log p_i - p_{n+1}\log p_{n+1} = nt\log t - p_{n+1}\log p_{n+1} = B - p_{n+1}\log p_{n+1}
$$

The result of the proposition follows from the fact that the function $x \mapsto -x\log x$ achieves a maximum value of $\frac{1}{e}$ in $[0,1]$. □