From Isolation to Identification

Giuseppe D'Acquisto[1], Aloni Cohen[2(✉)], Maurizio Naldi[3], and Kobbi Nissim[4]

[1] LUISS University, Rome, Italy
gdacquisto@luiss.it
[2] University of Chicago, Chicago, USA
aloni@g.uchicago.edu
[3] LUMSA University, Rome, Italy
m.naldi@lumsa.it
[4] Georgetown University, Washington, DC, USA
kobbi.nissim@georgetown.edu

Abstract. We present a mathematical framework for understanding when successfully distinguishing a person from all other persons in a data set—a phenomenon which we call *isolation*—may enable *identification*, a notion which is central to deciding whether a release based on the data set is subject to data protection regulation. We show that a baseline degree of isolation is unavoidable in the sense that isolation can typically happen with high probability even before a release was made about the data set and hence identification is not enabled. We then describe settings where isolation resulting from a data release may enable identification.

AQ1

AQ2

Keywords: privacy · identification · isolation · data protection

1 Introduction

The notion of *identification* is central to privacy and data protection regulation. For example, the GDPR regulates the processing of personal data: "any information relating to an identified or identifiable natural person."[1] But the GDPR leaves "identifiable" undefined. What constitutes identification?

It is an old idea that identification may be possible when attributes together uniquely distinguish a individual within a population. In 1986, Dalenius wrote that "it is well known" that "the data for some variables may, for some individuals, be unique and publicly known" and expose those individuals to record linkage attacks [6]. The U.S. National Institute of Standards and Technology even defines identification as "the process of using claimed or observed attributes of an entity to single out the entity among other entities in a set of identities."[2] Sweeney carried out this process in her re-identification of MA Governor Weld in a dataset of state employee health records. Moreover, she showed that very few attributes are needed to uniquely distinguish most US residents: 5-digit ZIP, gender, and date of birth suffice for 87% of respondents in the 1990 Census [15].

[1] Regulation (EU) 2016/679 (General Data Protection Regulation), Article 4.
[2] For variations and sources, see https://csrc.nist.gov/glossary/term/identification.

J. Domingo-Ferrer and M. Önen (Eds.): PSD 2024, LNCS 14915, pp. 1–15, 2024.
https://doi.org/10.1007/978-3-031-69651-0_1

But there is an important a difference between distinguishing a person within a dataset or a sample—which this paper calls *isolation*—and distinguishing a person within a population [3,8]. Consider the study by De Montjoye et al. showing that four random time-location pairs were enough to isolate 95% of individuals' records in a dataset on 1.5M people. Sánchez et al. reply: "With a nonexhaustive sample, an individual's sample uniqueness/unicity does not imply population uniqueness and, hence, does not allow unequivocal reidentification" [14, citing [2]]. With this specific claim, we agree – absent other information, it's not clear to what extent these isolations amount to identification. Of course, if one can also check whether a given person was in the sample, then isolation in the sample plus the fact that person was in the sample results in isolation in the population.

1.1 This Work's Contributions

Identification and Isolation. While our goal is to provide a better understanding of identification to help design approaches for releasing information while protecting individual privacy, we do not attempt to define what identification means in a mathematically formal way. We see identification as an inherently-fuzzy legal concept for which a satisfactory mathematical treatment may not even exist. What we do is take a closer look at the related phenomenon of isolation.

In a little more detail, we consider a setting where an information holder produces a data release $\mathbf{R}$ based on a table $\mathbf{X}$ of PII. An adversarial information receiver tries to use the release $\mathbf{R}$ to *identify* one or more of the data items in $\mathbf{X}$. We say that the information receiver *isolates* in $\mathbf{X}$ if they succeed in producing a description[3] that matches exactly one person in $\mathbf{X}$. That isolation and identification are related follows from observing that isolation has been a major stepping stone towards identification in linkage attacks, where, typically, an entry of a presumably deidentified dataset is first isolated and then re-identified via linkage with an dataset containing identifying information (see, e.g., [15]).

We provide novel additions to the discussion of the relationship between identification and isolation:

1. We argue that some baseline degree of isolation is unavoidable and does not enable identification. In Sects. 3.1 and 3.2 we show that an information receiver having knowledge of the probability distribution underlying the data in $\mathbf{X}$ can isolate individuals in it prior to seeing any data release. (In Appendix A we extend the results to the case of an information receiver having only partial knowledge of the distribution.)
2. In Sect. 4 we ask when a data release leads to identification. We identify *isolation gain*—a measure of how an information receiver's confidence in an isolation attempt grows once they receive a release $\mathbf{R}$ based on the dataset. In Sect. 4.2 we revisit examples from the re-identification literature, analyzing them through this lens.

[3] For example: $(25 \leq \text{Age} \leq 28) \wedge (10,000 \leq \text{Salary} \leq 50,000)$.

Postulates About Identification. As a surrogate for defining identification mathematically, our analysis proceeds from two postulates about identification which we believe are simple, intuitive, and uncontroversial.

Postulate 1. *If a release* $\mathbf{R}$ *contains no information derived from the dataset* $\mathbf{X}$*, then* $\mathbf{R}$ *itself cannot be used to identify any individual in* $\mathbf{X}$*.*

In particular, the baseline degree of isolation (item 1 above) is unavoidable does not constitute "singling out" as used in the GDPR.[4]

Postulate 2. *A description of an individual record in* $\mathbf{X}$ *may enable identification if it is specific enough to uniquely distinguish the corresponding individual in the underlying population. A release* $\mathbf{R}$ *from which such a description is derived may enable identification.*

We stress that these postulates *do not* characterize identification. The postulates describe extreme cases leaving a bulk of real-world data analyses somewhere in gray area in between. Even so, these postulates are useful as they allow us to describe the outer bounds of identification from data release, or a lack thereof.

1.2 Related Work

As discussed above and in Sect. 4.2 below, a long line of work studies re-identification from anonymized or de-identified data releases [2–4,6,7,11,12,14,15].[5] A recent work of particular relevance is that of Rocher, Hendricks, and De Montjoye [12], who address the gap between sample- and population-uniqueness by showing that it is often possible to empirically estimate the probability that an isolating set of attributes uniquely distinguishes an individual within the underlying population, (an estimation task with a long history [3,8]).

Postulate 1 is implicit in Ruggles and Van Riper's critique of the US Census Bureau's reconstruction of the 2010 Decennial Census: empirically comparing those results to a baseline that "would be expected by chance" [13]. Jarmin et al. analyze disclosure risk assessment frameworks in part using as a santity check that they should "deem releasing uniformative statistics not a disclosure risk" [10].

We use Postulate 1 to derive a baseline level of isolation that does not amount to disclosure of any sort. A line of work does a version of this by instead excluding an individual from the data analysis [9, and citations therein]. Most recently, Francis and Wagner put forward a "non-member framework"[6], empirically apply it to prior attacks, and discuss its relevance to the concepts of identifiability and anonymization under the GDPR to past re-identification studies [9].

[4] General Data Protection Regulation, Recital 26. See also: Article 29 Working Party, *Opinion 05/2014 on Anonymisation Techniques.*

[5] See citations in [12] for more.

[6] "If an individual is not present in a dataset, and is independent of all other individuals in the dataset, then the release of that dataset does not violate that individual's privacy.".

Our approach builds on prior work by Cohen and Nissim [5] and Altman et al. [1]. These works introduce an abstract framework for isolation and analyze its relation with the GDPR notion of singling out, based on what we call Postulate 1. The current paper connects isolation with identification by introducing Postulate 2, extends the prior work by considering multiple isolation attempts, and presents its findings in more concrete settings. We also consider heuristic isolation strategies for the setting where the information receiver does not have sufficient knowledge of the probability measure underlying the data.

2 The Isolation Problem

Consider a scenario where some personal data is stored by an information holder, and an information receiver wishes to uncover that information (or part of it). The information holder owns a table (the "ground truth" table) with a row for each individual,

$$\mathbf{X} = (\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n).$$

As an example, each row $\underline{x}_i$ may take value in some space $D \subset R^d$. We assume that we can build a probability space on D by defining the space of elementary events and a probability measure P, so that we can assign a probability value to all events (i.e., subsets of D) forming a Borel field. The table $\mathbf{X}$ is assumed to contain the result of n i.i.d. random samples draw from P.

We define the following basic statistic, computed over $\mathbf{X}$:

Definition 1 (counting function). *For a generic subdomain $B \subseteq D$, the number of instances in $\mathbf{X}$ falling in it is given by the counting function:*

$$H_{\mathbf{X}}(B) = \sum_{i=1}^{n} I(\underline{x}_i \in B).$$

The information holder releases aggregate data about $\mathbf{X}$. For concreteness, we assume that a data release $\mathbf{R}$ consists of a collection of pairs $\mathbf{R} = \{(A_i, m_i)\}$ where each of the pairs in $\mathbf{R}$ declares a subdomain $A_i \subseteq D$ and a number m_i where $m_i = H_{\mathbf{X}}(A_i) + \epsilon_i$. In the case of absence of noise, $\epsilon_i = 0$ for all i and therefore m_i is guaranteed to be the exact number of individuals in $\mathbf{X}$ that fall into the subdomain A_i. In the case of a noisy release, the noise variable ϵ_i is assumed to be drawn from a known distribution.

Initially, the release is empty (i.e., initially $\mathbf{R}_0 = \emptyset$) and the only information known to the information receiver is the data subspace D, the underlying probability measure P, and the number of entries n in $\mathbf{X}$.[7] The information receiver's initial knowledge $(D, P, n, \mathbf{R}_0)$ about $\mathbf{X}$ is updated with the aggregate counts $\mathbf{R} = \{(A_i, m_i)\}$ once they are released. We assume the receiver also knows whether the release is noiseless or noisy, along with the noise distribution in the latter case.

[7] Our results are robust to a substantial relaxation of these assumptions, in particular, knowledge of the distribution P and the number of records n may be approximate. See Remark 3 and Appendix A.

2.1 Isolation

We define *isolation* to happen when the information receiver outputs a description matching exactly one row in $\mathbf{X}$, formally:

Definition 2 (isolation). *We say that $B \subseteq D$ isolates in $\mathbf{X}$ if $H_{\mathbf{X}}(B) = 1$.*

Guessing an isolating B. The information receiver may use their knowledge about $\mathbf{X}$—namely, the data domain D, probability measure P, number of entries n, and a release $\mathbf{R}$—to guess one or more sets that isolate in $\mathbf{X}$. That is, the receiver comes up with subsets $B_1, B_2, \ldots, B_k \subseteq D$ that the information receiver hopes isolate many of the entries in $\mathbf{X}$.

Verifying that a Guess B Isolates. If the information receiver has query access to a noiseless release mechanism then they can ask for the release $(B, H_{\mathbf{X}}(B))$, and hence check whether the guess B indeed consists an isolation.[8,9]

Remark 1. The information receiver may try to isolate in $\mathbf{X}$ even before receiving any release (i.e., only given D, P, n, and $\mathbf{R}_0 = \emptyset$). With a release $\mathbf{R} \neq \emptyset$ made about $\mathbf{X}$, the receiver can improve its confidence in guessing an isolating B. The information receiver's rate of successful isolation given $\mathbf{R}_0$ can be thought of as a baseline to which their isolation ability given a release $\mathbf{R} \neq \emptyset$ can be compared.

3 Optimal Isolation Without Any Release

We now consider strategies an information receiver may use to isolate one or more individuals in the dataset $\mathbf{X}$. We analyze the information receiver's isolation ability prior to any release $\mathbf{R}$, i.e., with $\mathbf{R}_0 = \emptyset$. The analysis is done under the assumption that the information receiver has knowledge of the data domain D, the underlying probability measure P, and the number of elements n in $\mathbf{X}$.[10]

Remark 2. The same analysis carried out in this section holds for the case of a non-empty release $\mathbf{R}$ by replacing the probability measure P with the probability measure resulting from conditioning P on the release $\mathbf{R}$.

Subsections 3.1 and 3.2 discuss the guessing strategies the information receiver may use assuming they have perfect knowledge of the underlying probability measure P. In Appendix A we discuss how these strategies can be used by an information receiver that does not have complete knowledge of P.

[8] Query access to the release mechanism can also be used in other ways, see Remark 4 below.

[9] Access to alternative sources of information may also be used for boosting the information receiver's confidence that a guess B isolates.

[10] The number of records n is included as part of the the receiver's prior knowledge since n can often be inferred (exactly or approximately) from public information. Examples include (i) where a survey design specifying n was made public prior to the collection of information, and (ii) where n was made public in previous surveys (e.g., a census). For a reader who considers n as part of the release, this section should be understood as demonstrating that the release of n alone suffices for isolation.

3.1 A Single Isolation Guess

We begin with the analysis of how the information receiver may make one guess B so as to maximize their probability of a successful isolation.

For $B \subseteq A$, $P(B)$ is the probability that an individual sampled according to P falls into the subdomain B. The probability that B contains exactly one individual from the n individuals in D (i.e., that some individual from the n individuals in D has been isolated) is given by the expression

$$p^{\mathsf{iso}}(B) := \Pr_{\mathbf{X}\sim P^n}[H_{\mathbf{X}}(B) = 1] = n \cdot P(B) \cdot (1 - P(B))^{n-1}. \tag{1}$$

Observe that $p^{\mathsf{iso}}(B)$ depends only on $P(B)$. The information receiver can choose the subset B so that $P(B)$ maximizes $p^{\mathsf{iso}}(B)$. All that remains is to determine the value $P(B)$ that achieves the maximum.

Theorem 1. *The probability of isolation $p^{\mathsf{iso}}(B)$ achieves its maximum value of $\left(1 - \frac{1}{n}\right)^{n-1} \approx \frac{1}{e} \approx 0.37$ when $p(B) = \frac{1}{n}$.*

Proof. To simplify notation, let $p^{\mathsf{iso}} = p^{\mathsf{iso}}(B)$ and $p = P(B)$. We compute the first and second derivatives of Eq. (1):

$$\frac{\partial p^{\mathsf{iso}}}{\partial p} = n(1-p)^{n-2}(1-np), \quad \text{and} \quad \frac{\partial^2 p^{\mathsf{iso}}}{\partial p^2} = -n(n-1)(1-p)^{n-3}(2-np).$$

The first derivative is positive on $p \in [0, \frac{1}{n})$, zero at $p = \frac{1}{n}$, and negative on $p \in (\frac{1}{n}, 1]$. The second derivative is negative for $p < \frac{2}{n}$. Hence, $p = \frac{1}{n}$ maximizes p^{iso}, with maximum value $n \cdot \frac{1}{n} \cdot \left(1 - \frac{1}{n}\right)^{n-1} = \left(1 - \frac{1}{n}\right)^{n-1}$.

Theorem 1 may be interpreted as follows: if the information receiver can make a single guess B, the receiver maximizes the probability that B isolates in $\mathbf{X}$ by choosing B such that $p(B) = \frac{1}{n}$, in which case the receiver's guess is successful with probability about 0.37.

Remark 3. If instead of picking B such that $p(B) = \frac{1}{n}$ the information receiver picks B such that $p(B) = \frac{c}{n}$ then the isolation probability drops to

$$n \cdot \frac{c}{n} \cdot \left(1 - \frac{c}{n}\right)^{n-1} = \frac{c}{1 - \frac{c}{n}} \left(1 - \frac{c}{n}\right)^{n} \approx \frac{c}{e^c}.$$

- This implies that the result of Theorem 1 is only mildly sensitive to errors in the information receiver's knowledge of P and n. For example, if the receiver's knowledge is off by a factor of at most 2 (i.e., $\frac{1}{2n} \leq P(B) \leq \frac{2}{n}$ or, equivalently, $\frac{1}{2} \leq c \leq 2$), then $p^{\mathsf{iso}}(B) \gtrapprox 0.27$.
- If, however, the information receiver chooses to make a guess with $p(B) = \frac{c}{n}$ where c is very small, then we get $e^c \approx 1$ and and the isolation probability is $p^{\mathsf{iso}}(B) \approx c$, i.e., also very small.

3.2 Multiple Isolation Guesses

We now consider an information receiver that makes $k > 1$ guesses $B_1, \ldots, B_k \subseteq D$ trying to isolate multiple individuals in $\mathbf{X}$. We will require that the guesses B_i are mutually disjoint, so it is impossible for two of them to isolate the same element in $\mathbf{X}$. The receiver's goal is to maximize the total number isolated elements in $\mathbf{X}$ (i.e., number of guesses in $B_1, \ldots, B_k$ which successfully isolate in $\mathbf{X}$), i.e., $\sum_{i=1}^{k} I(H_{\mathbf{X}}(B_i) = 1)$. As discussed in Sect. 2.1, if after choosing $B_1, \ldots, B_k$ the information receiver can issue them as queries then the receiver would learn with certainty which of them isolates.

Remark 4. Our analysis considers an *oblivious* information receiver which makes all the guesses $B_1, \ldots, B_k$ at once. Indeed, this is the information receiver's only choice if they cannot issue queries and obtain more releases. However, if the information receiver may issue queries, then an *adaptive* strategy would be more effective for multiple isolations. In such a strategy the receiver would choose guess B_{i+1} after seeing $H_{\mathbf{X}}(B_i)$. As an example, an adaptive adversary may choose to use a divide-and-conquer strategy to isolate elements in $\mathbf{X}$.

Remark 5. We ignore other, less direct, modes of isolation. For example, if B_i is a strict subset of B_j and $H_{\mathbf{X}}(B_i) = H_{\mathbf{X}}(B_j) - 1$ then their difference $B_j \setminus B_i$ isolates.

Observation 1. *If the number of guesses $k \leq n$ then we can use Theorem 1: the information receiver may choose k disjoint subsets where $P(B_i) = \frac{1}{n}$ for all $i = 1, \ldots, k$. In expectation about $\frac{k}{e}$ of the guesses B_i would isolate.*

The information receiver's optimal strategy in the case $k > n$ is characterized by the following theorem:

Theorem 2. *When $k \geq n$, the expected number of isolations achieves its maximum when $P(B_i) = \frac{1}{k}$ for all $i \in 1, \ldots, k$.*

Proof. To simplify notation, let $p_i = P(B_i)$. For $0 \leq p \leq 1$, define $f(p) = n \cdot p \cdot (1-p)^{n-1}$. For $\mathbf{p} = (p_1, \ldots, p_k)$, define $F(\mathbf{p}) = \sum_{i=1}^{k} f(p_i)$. By Eq. (1), we can rewrite the expected number of isolations as $F(\mathbf{p})$:

$$\mathbf{E}\left(\sum_{i=1}^{k} I(H_{\mathbf{X}}(B_i) = 1)\right) = \sum_{i=1}^{k} \mathbf{E}(p^{\mathsf{iso}}(B_i)) = \sum_{i=1}^{k} f(p_i) = F(\mathbf{p}).$$

Fix $\mathbf{p} = (p_1, \ldots, p_k)$ arbitrarily and let $\mathbf{p}^* = (\frac{1}{k}, \ldots, \frac{1}{k})$. We must show that $F(\mathbf{p}) \leq F(\mathbf{p}^*)$. To do so, we will give two intermediate variables $\mathbf{p}'$ and $\mathbf{p}''$ and show that $F(\mathbf{p}) \leq F(\mathbf{p}') \leq F(\mathbf{p}'') \leq F(\mathbf{p}^*)$. We use two facts about f already shown in the proof of Theorem 1. First, f is (strictly) concave in the interval $0 \leq p \leq \frac{1}{n}$. Second, f is (strictly) increasing as $p \to \frac{1}{n}$ from either side.

Construct $\mathbf{p}'$ by clamping each p_i to the interval $[0, 1/n]$. Namely, $p'_i = \min(p_i, 1/n)$ for each $i = 1, \ldots, k$. If $p'_i = p_i$, then $f(p_i) = f(p'_i)$. Otherwise, $1/n \leq p'_i < p_i$ and hence $f(p_i) < f(p'_i)$. Therefore $F(\mathbf{p}) \leq F(\mathbf{p}')$.

Observe that $0 \leq \sum p'_i \leq 1$. Construct $\mathbf{p}''$ arbitrarily such $\sum p''_i = 1$ and $p'_i \leq p''_i \leq 1/n$ for all i. This is possible because $\sum_{i=1}^{k} \frac{1}{n} = \frac{k}{n} \geq 1$. For all i, $f(p''_i) \geq f(p'_i)$. Therefore $F(\mathbf{p}') \leq F(\mathbf{p}'')$.

By construction, $0 \leq p''_i \leq 1/n$ for all i. Because f is concave on $[0, 1/n]$,

$$F(\mathbf{p}) \leq \sum_{i=1}^{k} f(p''_i) \leq k \cdot f\left(\frac{p''_1 + \cdots + p''_k}{k}\right) = F(\mathbf{p}^*).$$

Putting it all Together. By the combination of Observation 1 and Theorem 2, setting $p(B_i) = \frac{1}{\max(k,n)}$ for all B_i maximizes the expected number of isolations. For $k \leq n$, the expected number of isolations is

$$k\left(1 - \frac{1}{n}\right)^{n-1} \approx \frac{k}{e}.$$

For $k > n$, the expected number of isolations is

$$k \cdot n \cdot p \cdot (1-p)^{n-1} = n \cdot \left(1 - \frac{1}{k}\right)^{n-1} = n \cdot \left(1 - \frac{1}{k}\right)^{k \cdot \frac{n-1}{k}} \approx n \cdot e^{-\frac{n}{k}},$$

an expression that tends to n as k grows to infinity. More concretely, in the regime $k \geq n$, to achieve αn isolations in expectation, it suffices for the information to make $k \approx \frac{n}{\ln(1/\alpha)}$ guesses. In particular, approximately $20n$ guesses suffice for $\alpha = 0.95$ and approximately $100n$ guesses suffice for $\alpha = 0.99$.

Figure 1 provides an example of how the expected number of isolations grows as a function of guesses k if $n = 5$ and $P(B_i)$ is taken based on Observation 1 and Theorem 2.

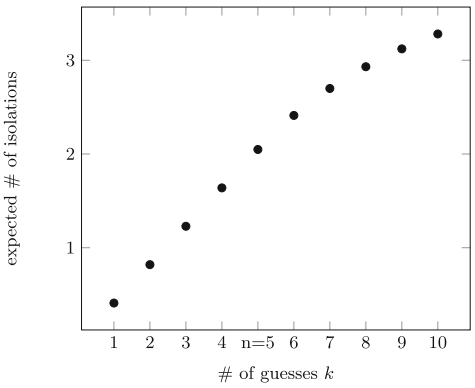

Fig. 1. Expected number of isolations as a function of k. ($n = 5$).

4 From Isolation to Identification

Isolation Alone is Insufficient for Identification. Theorem 1 shows that even prior to any release (i.e., with $\mathbf{R}_0 = \emptyset$) an information receiver making just a single isolation attempt would be successful in more than one in three trials. Likewise, Observation 1 and Theorem 2 show that by making many guesses $B_1, \ldots, B_k$, the information receiver can drive the expected number of isolations up to n as k grows. By Postulate 1, none of these cases amount to identification. Some other criterion beyond mere isolation is needed.

A Baseline. We again look to Theorems 1 and 2. They characterize the optimum rate of isolation achievable without any data-specific knowledge, and what strategies achieve that optimum. Those results therefore describe a *baseline* against which an information receiver's attempts at isolation may be measured. If the receiver can significantly outperform the baseline, it is indicative of some non-trivial disclosure.

We are now guided by Postulate 2. A guess B that isolates in $\mathbf{X}$ describes an individual record in the dataset (e.g., SEX=male, ZIP=02138, DOB=07/31/1945). Such a description may enable identification if it is specific enough to uniquely distinguish that the corresponding individual in the underlying population.

Isolation may enable identification when B is specific enough to uniquely distinguish an individual in a population, and also isolates in $\mathbf{X}$ better than the baseline chance. Two parameters are important: $P(B)$, the probability mass of B in the prior distribution; and $\Pr[H_{\mathbf{X}}(B) = 1 \mid \mathbf{R}]$, the probability that B isolates in $\mathbf{X}$ conditioned on the the release $\mathbf{R}$. Smaller $P(B)$ and larger $\Pr[H_{\mathbf{X}}(B) = 1 \mid \mathbf{R}]$ are stronger evidence that identification may be possible.

Consider, e.g., a simple setting where a population $\mathbf{X}^*$ of N individuals is drawn i.i.d. from P, and a subset $\mathbf{X}$ of size $n = rN$ is sampled uniformly at random, where $0 < r \ll 1$. The information receiver sees a release $\mathbf{R}$ derived from $\mathbf{X}$ and outputs a single guess B. The receiver's goal is that B should both isolate in $\mathbf{X}$ and uniquely distinguish an individual in the population $\mathbf{X}^*$. If B isolates in $\mathbf{X}$, then it uniquely distinguishes in $\mathbf{X}^*$ if no element of $\mathbf{X}^* \setminus \mathbf{X}$ is in B. Taking $P(B) \leq \frac{1}{100N} = \frac{r}{100n}$, we get that B uniquely distinguishes in $\mathbf{X}^*$ with probability

$$(1 - P(B))^{N-n} = (1 - P(B))^{(1-r)N} \geq \left(1 - \frac{1}{100N}\right)^{100N \cdot \frac{1-r}{100}} \approx e^{-\frac{(1-r)}{100}} > 0.99.$$

On the other hand, by Remark 3, the baseline chance that the information receiver produces such a B that also isolates in $\mathbf{X}$ after only receiving the empty release $\mathbf{R}_0$ is:

$$p^{\mathsf{iso}}(B) = \Pr[H_{\mathbf{X}}(B) = 1 \mid \mathbf{R}_0] \lessapprox \frac{r}{100} \ll 0.01.$$

In this example, the release $\mathbf{R}$ may enable identification if—with probability much greater than 0.01—the information receiver produces a guess B such that both $P(B) \leq \frac{1}{100N}$ and B isolates in $\mathbf{X}$.

We quantify the improvement in the probability of isolation conditioned on the release $\mathbf{R}$ using a quantity we call *isolation gain*.

Definition 3 (isolation gain). *Let $B \subseteq D$. Consider two datasets $\mathbf{X}, \mathbf{X}' \sim P^n$ and let $\mathbf{R}$ be a release derived from $\mathbf{X}$. The* isolation gain *for B is defined as:*

$$G(B) = \frac{\Pr[H_{\mathbf{X}}(B) = 1 \mid \mathbf{R}]}{\Pr[H_{\mathbf{X}'}(B) = 1]}.$$

Rephrasing the preceding discussion, the information receiver wants to make a guess B with a high isolation gain. Namely, the receiver attempts to maximize the numerator—the probability that B isolates given $\mathbf{R}$—while minimizing the denominator, which for $P(B) < \frac{1}{n}$ is equivalent to minimizing $P(B)$.[11]

4.1 How an Information Release may Enable Identification

We now consider how the information receiver might use a release $\mathbf{R}$ about a dataset $\mathbf{X}$ of size n in an attempt to identify an individual in a population of $N \gg n$. Throughout this section, we view a (noiseless) data release $\mathbf{R}$ as consisting of a collection of pairs $\mathbf{R} = \{(A_i, m_i)\}$, each pair specifying a subdomain $A_i \subseteq D$ and the count $m_i = H_{\mathbf{X}}(A_i)$. In this subsection, we assume that the A_i are chosen independently of $\mathbf{X}$ (the worst case for the information receiver). We consider an information receiver who produces a single guess B seeking to minimize $P(B)$ while maximizing the probability of isolation $\Pr[H_{\mathbf{X}}(B) = 1 \mid \mathbf{R}]$.

We use a simple observation: if B_i isolates a single record among the m_i records in A_i, then $B = B_i \cap A_i$ isolates a single record in $\mathbf{X}$. To find such a B_i, we adapt the strategies analyzed in Sect. 3. In that section, we analyzed the optimum probability of isolation assuming only that $\mathbf{X} \sim P^n$, for any P and n. For any i, let $\mathbf{X}_i = \{\underline{x} \in \mathbf{X} : \underline{x} \in A_i\}$ be those elements of $\mathbf{X}$ contained in A_i, and P_i be the data distribution conditioned on $\underline{x} \in A_i$. The posterior distribution of $\mathbf{X}_i$ conditioned on the release $\mathbf{R}$ is m_i i.i.d. samples from P_i. (This uses the assumption that A_i is independent of $\mathbf{X}$.) Theorem 1 implies that B_i satisfying $P_i(B_i) = P(B_i|A_i) = \frac{1}{m_i}$ will maximize the probability of isolating in $\mathbf{X}_i$, as desired. The isolation gain for $B = B_i \cap A_i$ is

$$G(B) = \frac{m_i \cdot \frac{1}{m_i} \cdot (1 - \frac{1}{m_i})^{m_i - 1}}{n \cdot P(B) \cdot (1 - P(B))^{n-1}}. \tag{2}$$

Whether the above strategy is good depends on which sets A_i are in the release $\mathbf{R}$. We analyze three examples based on the value of $P(A_i)$ relative to $\frac{1}{n}$.

If $P(A_i) \ll \frac{1}{n}$, the attack is very successful: $G(B) \gg 1$. Furthermore, if $P(A_i) \ll \frac{1}{N}$, then the release may enable identification. To see why, observe that many A_i will contain exactly 1 record: $m_i = 1$. Take $B = A_i$ for any such

[11] We exclude $P(B) \geq \frac{1}{n}$, as the chance that B uniquely distinguishes an individual in a population of size $N = n/r$ in extremely small. Namely, $(1 - P(B))^{N-n} \leq e^{-(1-r)/r}$. Taking $r = 0.01$, say, the probability is about 10^{-43}.

A_i. The numerator of G is $\Pr[H_{\mathbf{X}}(B) = 1] = \Pr[m_i = 1] = 1$. By Remark 3, the denominator of G is $\Pr[H_{\mathbf{X}'}(B) = 1] \approx n \cdot P(B) \ll 1$. If $P(B) = P(A_i) \ll \frac{1}{N}$, then B uniquely distinguishes the isolated individual in the population with high probability.

If $P(A_i) = \frac{1}{n}$, the attack slightly beats the baseline: $G(B) \gtrapprox \frac{\ln(n)}{e}$. By a standard balls-in-bins analysis, there exists i such that $m_i \approx \ln(n)$ with high probability. For this i, we take $B = B_i \cap A_i$, yielding $P(B) \approx \frac{1}{n \ln(n)}$. The numerator of G is $\approx \frac{1}{e}$. The denominator is $< n \cdot P(B) \approx \frac{1}{\ln(n)}$.

If $P(A_i) \gg \frac{1}{n}$, the attack doesn't beat the baseline: $G \approx 1$ in expectation. In this case, $m_i \gg 1$ with high probability, and thus then numerator of Eq. (2) is $\approx \frac{1}{e}$. Next, observe that

$$\mathbf{E}[P(B)] = \mathbf{E}\left[\frac{P(A_i)}{m_i}\right] = \mathbf{E}\left[\frac{\mathbf{E}(H_{\mathbf{X}}(A_i))}{n \cdot H_{\mathbf{X}}(A_i)}\right] = 1.$$

For $P(B) \approx \mathbf{E}[P(B)]$,[12] the denominator of Eq. (2) is also $\approx \frac{1}{e}$.

4.2 Examples from the Re-identification Literature

We briefly consider few types of releases $\mathbf{R}$ which capture re-identification attacks from prior work.

Microdata Releases. In a typical microdata release $\mathbf{R}$, the subdomain A_i describes the attributes of record $\underline{x}_i \in \mathbf{X}$, possibly with some attributes generalized or redacted. Typically, the attributes are very rich, hence $P(A_i) \ll \frac{1}{n}$. As described in the previous section, the information receiver can achieve a high isolation gain by taking $B = A_i$. Whether B uniquely distinguishes the isolated individual in the population depends on details of the release. But, since it does not take many attributes to uniquely distinguish somebody in the population [15], it is likely that, for a rich data domain, $P(B) < \frac{1}{100N}$.

A well-known example is the "Unique in the Crowd" study by De Montjoye et al. [7] The A_i consisted of many time-location data points for each of 1.5M people. The study showed that taking $B \supseteq A_i$ to contain just four of these points sufficed to isolate for 95% of the rows i. But no evidence was given that $P(B)$ was small enough to uniquely distinguish an individual in the underlying population [14].

To get beyond mere isolation, Rocher et al. directly estimated the probability of population uniqueness for microdata releases [12]. Like us, they observed that the probability that B uniquely distinguishes the isolated individual in the population is at least $(1 - P(B))^{N-1}$. They showed that for real-world datasets, an information receiver can empirically estimate $(1 - P(B))^{N-1}$ to within a few percent by using sample of the population to learn the distribution P.[13] Using

[12] Accounting for the variance of $P(B)$ yields only an insignificant improvement.

[13] E.g., for Governor Weld's attributes used by Sweeney, they estimated $(1 - P(B))^{N-1} \approx 0.58$.

this estimate, an information receiver can choose B such that $(1 - P(B))^{N-1} \geq 0.95$, say, which implies that $P(B) \ll \frac{1}{N}$.

Other re-identification studies on microdata releases include Sweeney's re-identification of Governor Weld [15] (see below) and Narayanan and Shmatikov's re-identification using the Netflix Prize Dataset [11].

k-anonymity. A k-anonymous data release contains ℓ-many counts (A_i, m_i) subject to the constraint that $m_i \geq k$. The parameter $k > 1$ is a small constant (e.g., $k = 5, 10$). For simplicity, let us assume that $m_i = k$ for all i.

Cohen and Nissim analyze an the success of the following information receiver for *arbitrary* k-anonymization algorithms [5]. Guess $B \subseteq A_i$ arbitrary subject to $P(B|A_i) = \frac{1}{k}$, for any i. They show that so long as the data distribution has a moderate amount of entropy, B isolates with probability about $(1 - \frac{1}{k})^{k-1} > \frac{1}{e}$, regardless of the k-anonymization algorithm.[14] It remains to analyze $P(B) = \frac{P(A_i)}{k}$. As for the microdata release, how small $P(A_i)$ is depends on the k-anonymization algorithm and data distribution. Most k-anonymization algorithms are designed to preserve as much richness of the input dataset $\mathbf{X}$ as possible, i.e., minimizing $P(A_i)$. For rich-enough data, it is possible to provide k-anonymity while also guaranteeing that $P(A_i) < \frac{1}{100N}$ with high probability.

Cohen gives a much more effective strategy called *downcoding*, but which requires some assumptions on the k-anonymization algorithm and data distribution [4]. The core observation is that, if the k-anonymization algorithm preserves as much of $\mathbf{X}$ as possible, the sets A_i must depend on the data. Cohen shows that for some data distributions, the A_i enable the information receiver to recover a very detailed description B_j of some fraction of the rows $\underline{x}_j \in \mathbf{X}$ ($\geq 3\%$ of the rows for $k \leq 15$). These B_j isolate in $\mathbf{X}$, and $P(B_j) < \frac{1}{100N}$ as long as $\mathbf{X}$ contains at least $3\ln(100N)$ attributes.

When Membership in $\mathbf{X}$ *is Known.* Often, $\mathbf{X}$ is not a random sample of the population. Rather, membership in $\mathbf{X}$ is correlated with some attribute of the data. This extra information can help the information receiver turn isolation into identification by excluding from B individuals not in $\mathbf{X}$. For example, Cohen's re-identification of EdX students required only a few attributes about the students. Cohen was able to exclude all individuals not in the EdX release using the certificates of completion posted by many EdX students on their LinkedIn profiles, thereby turning isolation into identification [4]. As another example, Sweeney's re-identification of Governor Weld made use of the fact that the dataset contained the hospital records for all state employees [15].

Overlapping Contingency Tables. An example that doesn't fit neatly into the above comes from Israel's Central Bureau of Statistics.[15] Very roughly, the release included a count m for subdomains A specified by any choice of up to 5 attributes. For example, there was exactly 1 male widower veteran with

[14] The proof of this fact is somewhat nuanced, as A_i can depend arbitrarily on the dataset $\mathbf{X}$.

[15] See https://www.slideserve.com/ordell/razi-mukatren-golan-salman, and https://archive.is/W20kx.

no children among the survey respondents. Alone, these subdomains had probability $P(A) \approx \frac{1}{n}$. By the analysis in the previous section, identification would seem impossible. But if only four attributes were needed to isolate a record—as for the widower above—it is easy to reconstruct the record entirely. For every additional attribute, exactly one possible value will be non-zero. In this way, the information receiver can bootstrap many isolations into possible identifications.

Acknowledgments. Work of K.N. was supported by NSF Grant No. CCF2217678 "DASS: Co-design of law and computer science for privacy in sociotechnical software systems" and a gift to Georgetown University. Work completed while K.N. visited Bocconi University, Milan.

A Isolating with a Partial Knowledge of P

The analysis in Sects. 3.1 and 3.2 assumed that the information receiver has perfect knowledge of the underlying probability measure P (but not $\mathbf{X}$ sampled from P). We now discuss what the receiver may do when they do not know P in full.

Observation 1 and Theorem 2 teach that all the information receiver needs is a partition of the data space into sets of probability weight $p^* = \frac{1}{\max(n,k)}$ and Remark 3 suggests that it suffices that the partition is close to the optimal weight for the information receiver to succeed in isolating. We now develop these ideas.

Let $\mathcal{C} = \{C_i\}_{i=1}^{\ell}$ be a partition of D where $\ell = \max(n, k)$. The information receiver may choose the partition $\mathcal{C}$ heuristically in combination with their partial knowledge about P and the data domain D. For example, $\mathcal{C}$ may partition D into high-dimensional rectangles, each described as the conjunction of one or more attribute ranges (e.g., all combinations of 5-year Age by Sex by City). Denote by $p_i = P(C_i)$ the probability of an individual falling into C_i. We show that if $\underline{p} = (p_1, \ldots, p_\ell)$ is close close enough to $p^* = (\frac{1}{\ell}, \ldots, \frac{1}{\ell})$ then (even without knowing $p_1, \ldots, p_\ell$) the information receiver succeeds in isolating.

As $\mathcal{C}$ is a partition of the data domain D we have that $\sum_{i=1}^{\ell} p_i = 1$. Hence, if we pick a partition element at random, then, in expectation, its probability weight would be exactly $p^* = \frac{1}{\ell}$:

$$\mathop{\mathbf{E}}_{i \sim U_\ell}[p_i] = \sum_{j=1}^{\ell} \Pr_{i \sim U_\ell}[i = j] \cdot p_j = \frac{1}{\ell} \sum_{j=1}^{\ell} p_j = \frac{1}{\ell},$$

where we use $i \sim U_\ell$ to denote that the expectancy is over choosing an element of the partition $i \in \{1, \ldots, \ell\}$ uniformly at random.

An important parameter of the partition is its standard deviation σ:

$$\sigma^2 := \mathop{\mathbf{Var}}_{i \sim U_\ell}[p_i] = \mathop{\mathbf{E}}_{i \sim U_\ell}\left[p_i^2\right] - \left(\mathop{\mathbf{E}}_{i \sim U_\ell}[p_i]\right)^2 = \frac{1}{\ell} \sum_{i=1}^{\ell} p_i^2 - \frac{1}{\ell^2} = \frac{1}{\ell} \cdot \|\underline{p}\|_2^2 - \frac{1}{\ell^2}.$$

If the standard deviation σ is small compared to $\frac{1}{\ell}$, say $\sigma \leq \frac{c}{\ell}$ for $c \ll 1$, then many of the elements of the partition have weight $p_i \approx \frac{1}{\ell}$. More precisely, by Chebyshev's inequality[16] we have:

$$\Pr\left[p_i \notin \left[\frac{1}{2\ell}, \frac{3}{2\ell}\right]\right] = \Pr\left[|p(C_i) - \mathbf{E}[p(C_i)]| > \frac{1}{2\ell}\right] < 4c^2.$$

Hence, if the information receiver samples guesses $B_1, \ldots, B_k$ from the partition $\mathcal{C}$ without replacement, then in expectation at least $(1-4c^2)k$ of them would satisfy $p(B_i) \in \left[\frac{1}{2\ell}, \frac{3}{2\ell}\right]$. Using Eq. 1 each of these guesses would result in isolation probability $p^{\mathsf{iso}}(B_i) \geq n \cdot \min(\frac{1}{2\ell} \cdot (1-\frac{1}{2\ell})^{n-1}, \frac{3}{2\ell} \cdot (1-\frac{3}{2\ell})^{n-1})$, and in expectation the number of isolating guesses would be at least

$$(1-4c^2) \cdot \frac{kn}{2\ell} \cdot \min\left((1-\frac{1}{2\ell})^{n-1}, 3 \cdot (1-\frac{3}{2\ell})^{n-1}\right).$$

As an example, if $c = \frac{1}{4}$ and $k = n$ (hence $\ell = k = n$) we get that in expectation the number successful isolations is at least

$$\left(1 - 4 \cdot \left(\frac{1}{4}\right)^2\right) \cdot \frac{n^2}{2n} \cdot \min\left(\left(1-\frac{1}{2n}\right)^{n-1}, 3 \cdot \left(1-\frac{3}{2n}\right)^{n-1}\right)$$
$$\approx \frac{3n}{8} \cdot \min\left(e^{-1/2}, 3e^{-3/2}\right) \approx 0.23n.$$

I.e., in expectation almost a quarter of the guesses would consist successful isolations in spite of B_i not being chosen optimally.

Remark 6. Cohen and Nissim [5] used hashing to create a structure that is equivalent to a partition $\mathcal{C}$ where p_i is very close to $\frac{1}{\ell}$ (assuming P has sufficient min-entropy). The main qualitative difference between the hashing approach and the one described in this work is that hashing destroys the structure of the data domain and makes it harder for the information receiver to make effective use of isolation (e.g., as a step towards a linkage attack) whereas in the approach described herein the information receiver may choose partitions $\mathcal{C}$ that are more suitable for their purposes.

References

1. Altman, M., Cohen, A., Nissim, K., Wood, A.: What a hybrid legal-technical analysis teaches us about privacy regulation: the case of singling out. BUJ Sci. Tech. L. **27**, 1 (2021)
2. Barth-Jones, D.: The 're-identification' of governor William weld's medical information: a critical re-examination of health data identification risks and privacy protections, then and now. Then Now (2012) (2012)
3. Bethlehem, J.G., Keller, W.J., Pannekoek, J.: Disclosure control of microdata. J. Am. Stat. Assoc. **85**(409), 38–45 (1990)

[16] Chebyshev's inequality: $\Pr[|X - \mathbf{E}[X]| \geq k] \leq \frac{\mathbf{Var}[X]}{k^2}$.

4. Cohen, A.: Attacks on deidentification's defenses. In: Butler, K.R.B., Thomas, K. (eds.) 31st USENIX Security Symposium, USENIX Security 2022, Boston, MA, USA, 10–12 August 2022, pp. 1469–1486. USENIX Association (2022). https://www.usenix.org/conference/usenixsecurity22/presentation/cohen
5. Cohen, A., Nissim, K.: Towards formalizing the GDPR's notion of singling out. Proc. Natl. Acad. Sci. USA **117**(15), 8344–8352 (2020). https://doi.org/10.1073/pnas.1914598117
6. Dalenius, T.: Finding a needle in a haystack or identifying anonymous census records. J. Official Stat. **2**(3), 329 (1986)
7. De Montjoye, Y.A., Hidalgo, C.A., Verleysen, M., Blondel, V.D.: Unique in the crowd: the privacy bounds of human mobility. Sci. Rep. **3**(1), 1–5 (2013)
8. Fienberg, S.E., Makov, U.E.: Confidentiality, uniqueness, and disclosure limitation for categorical data. J. Official stat. **14**(4), 385 (1998)
9. Francis, P., Wagner, D.: Towards more accurate and useful data anonymity vulnerability measures. arXiv preprint arXiv:2403.06595 (2024)
10. Jarmin, R.S., et al.: An in-depth examination of requirements for disclosure risk assessment. Proc. Natl. Acad. Sci. **120**(43), e2220558120 (2023)
11. Narayanan, A., Shmatikov, V.: Robust de-anonymization of large sparse datasets. In: 2008 IEEE Symposium on Security and Privacy (SP 2008), 18–21 May 2008, Oakland, California, USA, pp. 111–125. IEEE Computer Society (2008). https://doi.org/10.1109/SP.2008.33
12. Rocher, L., Hendrickx, J.M., De Montjoye, Y.A.: Estimating the success of re-identifications in incomplete datasets using generative models. Nat. Commun. **10**(1), 1–9 (2019)
13. Ruggles, S., Van Riper, D.: The role of chance in the census bureau database reconstruction experiment. Popul. Res. Policy Rev. **41**, 781–788 (2022). https://doi.org/10.1007/s11113-021-09674-3
14. Sánchez, D., Martínez, S., Domingo-Ferrer, J.: Comment on "unique in the shopping mall: on the reidentifiability of credit card metadata". Science **351**(6279), 1274–1274 (2016)
15. Sweeney, L.: Simple demographics often identify people uniquely. Health (San Francisco) **671**(2000), 1–34 (2000)