Attacks on Deidentification's Defenses

Aloni Cohen*
University of Chicago

Abstract

Quasi-identifier-based deidentification techniques (QI-deidentification) are widely used in practice, including k-anonymity, ℓ-diversity, and t-closeness. We present three new attacks on QI-deidentification: two theoretical attacks and one practical attack on a real dataset. In contrast to prior work, our theoretical attacks work even if every attribute is a quasi-identifier. Hence, they apply to k-anonymity, ℓ-diversity, t-closeness, and most other QI-deidentification techniques.

First, we introduce a new class of privacy attacks called *downcoding attacks*, and prove that every QI-deidentification scheme is vulnerable to downcoding attacks if it is minimal and hierarchical. Second, we convert the downcoding attacks into powerful *predicate singling-out (PSO)* attacks, which were recently proposed as a way to demonstrate that a privacy mechanism fails to legally anonymize under Europe's General Data Protection Regulation. Third, we use LinkedIn.com to reidentify 3 students in a k-anonymized dataset published by EdX (and show thousands are potentially vulnerable), undermining EdX's claimed compliance with the Family Educational Rights and Privacy Act.

The significance of this work is both scientific and political. Our theoretical attacks demonstrate that QI-deidentification may offer no protection even if every attribute is treated as a quasi-identifier. Our practical attack demonstrates that even deidentification experts acting in accordance with strict privacy regulations fail to prevent real-world reidentification. Together, they rebut a foundational tenet of QI-deidentification and challenge the actual arguments made to justify the continued use of k-anonymity and other QI-deidentification techniques.

*We thank Kobbi Nissim for many helpful discussions; Ryan Sullivan for early discussions on EdX; and Gabe Kaptchuk, anonymous reviewers, and especially Mayank Varia for generous feedback. This work was primarily done at Boston University's Hariri Institute of Computing and School of Law. This work was supported by the DARPA SIEVE program under Agreement No. HR00112020021 and the National Science Foundation under Grant Nos. CNS-1915763 and SaTC-1414119. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not reflect the views of our funders.

1 Introduction

Quasi-identifier-based deidentification (QI-deidentification) is widely used in practice. The most well known QI-deidentification techniques are is k-anonymity [26]. Throughout this work we usually speak about k-anonymity specifically, but everything applies without modification to ℓ-diversity [18], t-closeness [17], and many other QI-deidentification refinements.

A relatively small number of data points suffice to distinguish individuals from the general population. For example, in the 2010 census 44% of the population was unique based only on census block, age, and sex [1]. Turning this insight into a privacy notion, k-anonymity aims to capture a sort of anonymity of a crowd.

A data release is k-anonymous if any individual row in the release cannot be distinguished from $k-1$ other individuals in the release using certain attributes called *quasi-identifiers*. Quasi-identifiers are sets of attributes that are potentially available to an attacker from other sources, combinations of which may uniquely distinguish an individual within the dataset. k-anonymity requires that the equivalence class of every record—the set of records with identical quasi-identifiers—is of size at least $k \geq 2$. A common choice for k is 5.[1] ℓ-diversity, t-closeness, and many QI-deidentification techniques refine k-anonymity in the sense that they collapse to k-anonymity when every attribute is treated as a quasi-identifier (Sec. 2.1).

Real world reidentification attacks, including on the Netflix and AOL datasets [4, 20], led to a policy debate about the QI-deidentification. Critics argued that the distinction between quasi-identifying attributes and other attributes—foundational to the whole approach—was untenable [21, 22]. Defenders argued that deidentification experts are good at determining what information is externally available [5, 6].

[1] For example, U.S. Department of Education's FAQ on disclosure avoidance states that "statisticians consider a cell size of 3 to be the absolute minimum although larger minimums (e.g., 5 or 10) may be used to further mitigate disclosure risk" (`https://studentprivacy.ed.gov/resources/frequently-asked-questions-disclosure-avoidance`). Based on this language, EdX chose $k = 5$.

The debate left unspoken and unexamined the core tenet of QI-deidentification: that if *every* attribute is treated as a quasi-identifier, then k-anonymity provides meaningful protection. Our work is the first to directly challenge that tenet.

Motivation Why bother attacking QI-deidentification? After all, the security and privacy research communities don't put much stock in these techniques. For example, it is well known that contrived mechanisms can formally satisfy k-anonymity but provide no protection. Even so, many policymakers and practitioners are convinced that QI-deidentification is effective in the real world.

Our goal in this paper is to rebut the actual arguments that QI-deidentification practioners use to justify its continued use. We rebut three arguments that—until this work—have gone unchallenged. First, that no attacks have been shown against datasets deidentified by experts and in accordance with strict privacy regulations, let alone simple attacks. Second, that k-anonymity provides meaningful protection when every attribute is a quasi-identifier. Third, that although QI-deidentification doesn't meet cryptographic standards of security, it suffices to meet the obligations in data protection regulation. We briefly elaborate these three arguments next.

Rhetorically, trust in QI-deidentification hinges on the wholesale dismissal of existing attacks as unconvincing. Practitioners dismiss many attacked datasets as "improperly deidentified" [6]. "Proper de-identification" must be done by a "statistical expert" and in accordance with procedures outlined in regulation [12], the increasing availability of QI-deidentification software notwithstanding. This argument has proven very effective in policy spheres. Moreover, practitioners dismiss attacks carried out by privacy researchers *because* they are privacy researchers. That these attacks are published in "research based articles within the highly specialized field of computer science" is used to argue that re-identification requires a "highly skilled 'expert' " and therefore is of little concern [5].

Technically, trust in QI-deidentification hinges on an unspoken, unexamined tenet:

> *QI-deidentification's tenet: If* every *attribute is treated as quasi-identifying, then k-anonymity provides meaningful protection.*

Treating every attribute as quasi-identifying defines away one major critique of QI-deidentification—namely, that the ex ante categorization of attributes as quasi-identifying or not is untenable and reckless. Moreover, when all attributes are quasi-identifying, the distinctions among k-anonymity, ℓ-diversity, and t-closeness collapse (Section 2.1). Prior attacks against k-anonymity fail in this setting.

Legally, the use of QI-deidentification hinges on the gap between the protection required by regulation and the protection desired by the academic research community. Practitioners claim only that QI-deidentification meets regulatory standards, not security researchers' stringent standards. For example, cryptographic security definitions typically make no assumptions about the techniques or auxiliary knowledge available to an adversary. However, the European Union's General Data Protection Regulation (GDPR) restricts the adversary's techniques by protecting only against "means reasonably likely to be used" by an attacker.[2] Likewise, the United State's Family Educational Rights and Privacy Act (FERPA) restricts the adversary's knowledge by protecting only against an attacker lacking "personal knowledge of the relevant circumstances."[3]

Contributions We present three attacks on QI-deidentification schemes: two theoretical attacks and one real world reidentification attack. Together, these attacks undermine the above justifications for the continued use of QI-deidentification.

First, we introduce a new class of privacy attack called *downcoding*, which recovers large fractions of the data hidden by QI-deidentification without any auxiliary knowledge. In short, downcoding undoes hierarchical generalization. A downcoding attack takes as input a dataset generalized and recovers some fraction of the generalized data. We call this downcoding as it corresponds to recoding records down a generalization hierarchy.

We prove that every QI-deidentification scheme is vulnerable to downcoding attacks if it is *minimal* and *hierarchical*. QI-deidentification is hierarchical if it works by generalizing attributes according to a fixed hierarchy (e.g., city→country→continent). QI-deidentification is minimal if no record is generalized more than necessary to achieve the privacy requirement, in a weak, local sense. Our downcoding attacks are powered by a simple observation: *minimality leaks information*. Figures 1 and 2 give simple examples of downcoding and of leakage from minimality, respectively.

Second, we convert our downcoding attacks into powerful *predicate singling-out (PSO)* attacks. PSO attacks were recently proposed as a way to demonstrate that a privacy mechanism fails to legally anonymize under the GDPR [2, 7]. We introduce a stronger type of PSO attack called *compound* PSO attacks and prove that minimal hierarchical QI-deidentification enables compound PSO attacks, greatly improving over the prior work.

Our downcoding and PSO attacks are the first attacks on QI-deidentification that work even when every attribute is a quasi-identifier. As such, they apply to QI-deidentification beyond k-anonymity, and refute the foundational tenet of QI-deidentification.

Third, we used LinkedIn.com to reidentify 3 students in a k-anonymized dataset published by Harvard and MIT from their online learning platform EdX. Despite being "properly" k-anonymized by "statistical experts" in accordance

[2] GDPR, Article 4

[3] 34 CFR §99.3

with FERPA, we show that thousands more students are potentially vulnerable to reidentification and disclosure.

Not only do these attacks rebut the arguments described above, they also show that QI-deidentification fails to satisfy three properties of a worthwhile measure of privacy of a computation, even without resorting to contrived mechanisms. Namely, we show that QI-deidentification mechanisms used in practice aren't robust to post-processing, do not compose, and rely on distributional assumptions on the data for their security.

Organization Section 2 discusses related work. Section 3 introduces notation and defines k-anonymity, along with hierarchical and minimal k-anonymity. Section 4 defines downcoding attacks and proves that minimal hierarchical k-anonymous mechanisms enable them. Section 5 defines compound predicate singling-out attacks and proves that minimal hierarchical k-anonymous mechanisms enable them. Section 6 describes the EdX dataset and shows that it is vulnerable to reidentification. Section 7 concludes that our attacks rebut the three core arguments that support the continued use of QI-deidentification in practice. The appendix includes additional details and proofs.

2 Related Work

Samarati and Sweeney proposed k-anonymity for statistical disclosure limitation in 1998 [25–27]. As new attacks were discovered, k-anonymity gave rise to more refined QI-deidentification techniques including ℓ-diversity, t-closeness, and many others (below).

Samarati was the first to study minimality for k-anonymity [25]. Our downcoding attacks build on prior work on minimality attacks [8, 28]. These works demonstrate that minimality can be used to infer sensitive attributes and violate ℓ-diversity, but not k-anonymity. They introduce two defenses against their attacks. One is yet another refinement of k-anonymity called m-confidentiality [28]. The second claims that certain anonymization algorithms offer protection for free (i.e., "methods which only inspect the QI attributes to determine the [equivalence classes]") [8]. In contrast, we use minimality to downcode, a new attack that violates k-anonymity itself and that defeats both defenses from prior work.

Predicate singling-out (PSO) attacks were recently introduced in the context of data anonymization under Europe's General Data Protection Regulation (GDPR) [7]. They were proposed as a mathematical test to show that a privacy mechanism fails to legally anonymize data under Europe's General Data Protection Regulation (GDPR) [2, 7]. The prior work gives a simple but weak PSO attack against a large class of k-anonymous mechanisms. We give much stronger PSO attacks against a restricted class of k-anonymous mechanisms.

Prior work shows that k-anonymity does not compose: multiple k-anonymous datasets can completely violate privacy when combined [13]. We show for the first time that composition failures can occur in real world uses of k-anonymity.

Differential privacy (DP) [11] presents one alternative to QI-deidentification, especially DP synthetic data [16] or local DP [10]. Switching to DP requires accepting that the resulting data will not provide the one-to-one correspondence with underlying records that makes QI-deidentification so attractive to users and laypeople.

2.1 Syntactic de-identification beyond k-anonymity

We reviewed the deidentification definitions included in the most comprehensive survey we could find [14]. Our downcoding attacks apply to any *refinement* of k-anonymity: namely, any definition that collapses to k-anonymity when every attribute is quasi-identifying. These include:

- k-anonymity and variants: k^m-, (α, k)-, p-sensitive-, (k, p, q, r)-, and (ε, m)-anonymity
- ℓ-diversity and variants: entropy-, recursive-, disclosure-recursive, multi-attribute-, ℓ^+-, and (c, ℓ)-diversity
- t-closeness and variant (n, t)-closeness
- m-invariance, m-confidentiality

Our downcoding attacks don't apply to Anatomy (which doesn't generalize quasi-identifiers at all) or differential privacy (which eschews the quasi-identifier framework all together). We have not determined whether the following definitions – which bound some posterior probability given the deidentified dataset – refine k-anonymity in the relevant sense: δ-presence, ε-privacy, skyline privacy, (ρ_1, ρ_2)-privacy, (c, k)-safety, and ρ-uncertainty.

We leave testing our downcoding attacks on actual deidentification software packages for future work. Free to use software packages include ARX Anonymization, μ-Argus, sdcMicro, University of Texas Toolkit, Amnesia, Anonimatron, Python Mondrian. All but Python Mondrian implement hierarchical algorithms. ARX Anonymization, sdcMicro, and Amenesia offer some version of local recoding (footnote 5). To the best of our knowledge, none guarantee minimality.

3 Preliminaries

3.1 Notation

Generally, fixed parameters are denoted by capital letters (e.g., number of dimensions D) and indices use the corresponding lowercase letter (e.g., $d = 1, \ldots, D$). For $a, b \in \mathbb{N}$, let $[a, b] = \{a, a+1, \ldots, b\}$ and $[b] = [1, b]$.

$\mathcal{U}^D$ is a D-dimensional data universe, where $\mathcal{U}$ is the attribute domain. For simplicity we take all attribute domains

$\mathbf{X} =$

ZIP	Income	COVID
91010	\$125k	Yes
91011	\$105k	No
91012	\$80k	No
20037	\$50k	No
20037	\$20k	No
20037	\$25k	Yes

$\mathbf{Y} =$

ZIP	Income	COVID
9101⋆	\$75–150k	⋆
9101⋆	\$75–150k	⋆
9101⋆	\$75–150k	⋆
20037	\$0–75k	⋆
20037	\$0–75k	⋆
20037	\$0–75k	⋆

$\mathbf{Z} =$

ZIP	Income	COVID
91010	\$125–150k	⋆
9101⋆	\$100–125k	⋆
9101⋆	\$75–150k	⋆
20037	\$0–75k	No
20037	\$0–75k	⋆
20037	\$25k	Yes

Figure 1: An example of downcoding. $\mathbf{Y}$ is a minimal hierarchical 3-anonymized version of $\mathbf{X}$ (treating every attribute as part of the quasi-identifier and leaving the generalization hierarchy implicit). $\mathbf{Z}$ is a downcoding of $\mathbf{Y}$: it generalizes $\mathbf{X}$ and strictly refines $\mathbf{Y}$.

Old	Rich
1	1
0	0
1	0
0	0

Old	Rich
$\bigstar_1$	$\bigstar_5$
$\bigstar_2$	$\bigstar_6$
$\bigstar_3$	0
$\bigstar_4$	0

Old	Rich
1	$\bigstar_7$
0	0
1	$\bigstar_8$
0	0

Figure 2: An example of minimality and inferences from minimality. Attributes are binary and $\bigstar = \{0,1\}$. The middle and right datasets are both minimal hierarchical 2-anonymous versions of the left dataset with respect to $Q = \{\text{Old, Rich}\}$. The right dataset is also globally optimal: it generalizes as few attributes as possible. Minimality implies that every pair of redacted entries in the same column in matching rows must contain both a 0 and 1. Hence, $\{\bigstar_1, \bigstar_2\} = \{\bigstar_3, \bigstar_4\} = \{\bigstar_5, \bigstar_6\} = \{\bigstar_7, \bigstar_8\} = \{0,1\}$, allowing downcoding. Only one bit of information does not follow directly from minimality of the middle table: whether or not $\bigstar_1 = \bigstar_5$.

to be identical, though in reality they are usually distinct (e.g., the EdX dataset).

A *record* $\mathbf{x} = (x_1, \ldots, x_D)$ is an element of the data universe. A *generalized record* $\mathbf{y}$, denoted $(y_1, \ldots, y_D)$, is a subset of the data universe specified by the Cartesian product $y_1 \times \cdots \times y_D$, where $y_d \subseteq \mathcal{U}$ for every $d \in [D]$. Note that a record $\mathbf{x}$ naturally corresponds to the generalized record $(\{x_1\}, \ldots, \{x_D\})$, a singleton. We say $\mathbf{y}$ *generalizes* $\mathbf{x}$ if $\mathbf{x} \in \mathbf{y}$ (i.e., $\forall d,\ x_d \in y_d$). For example, $\mathbf{y} = (\mathit{Female}, 1970\text{–}1975)$ generalizes $\mathbf{x} = (\mathit{Female}, 1972)$. For generalized records $\mathbf{z} \subseteq \mathbf{y}$, we say that $\mathbf{y}$ *generalizes* $\mathbf{z}$ and $\mathbf{z}$ *refines* $\mathbf{y}$. If $\mathbf{z} \subsetneq \mathbf{y}$, the generalization/refinement is *strict*.

A *dataset* $\mathbf{X}$ is an N-tuple of records $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$. $\mathbf{X}$ can be viewed as a matrix with $\mathbf{X}_{n,d}$ the dth coordinate of $\mathbf{x}_n$. A *generalized dataset* $\mathbf{Y}$ is an N-tuple of generalized records $(\mathbf{y}_1, \ldots, \mathbf{y}_N)$. For (generalized) datasets $\mathbf{Y}, \mathbf{Z}$, we write $\mathbf{Z} \preceq \mathbf{Y}$ if $\mathbf{z}_n \subseteq \mathbf{y}_n$ for all n. We extend the meaning of generalization and refinement accordingly. We write $\mathbf{Z} \prec \mathbf{Y}$ when at least one containment is strict. We call $\mathbf{y}_n$ the record in $\mathbf{Y}$ *corresponding to* $\mathbf{z}_n$, and vice-versa. Note that $\preceq$ is a partial order on datasets of N records from a given data universe.[4]

[4] More generally, we could consider datasets whose rows are permuted relative to one another. Define $\mathbf{Z} \preceq \mathbf{Y}$ if there exists a permutation $\pi : [N] \to [N]$ such that $\mathbf{z}_n \subseteq \mathbf{y}_{\pi(n)}$ for all n, choosing some canonical π arbitrarily if more than one exists. Then $\preceq$ is a partial order over equivalence classes of datasets induced by $\mathbf{Y} \sim \mathbf{Y}' \iff \exists \pi\, \forall n\ \mathbf{y}_n = \mathbf{y}'_{\pi(n)}$. We omit this additional complexity for clarity. We believe all our results would hold, mutatis mutandis.

3.2 k-anonymity

Formally, $\mathbf{Y}$ is k-anonymous if any individual row in the release cannot be distinguished from $k-1$ other individuals [26]. This requirement is typically parameterized by a subset Q of the attribute domains $Q \subseteq \{\mathcal{U}_d\}_{d \in [D]}$ called a *quasi-identifier*. We denote by $\mathbf{y}(Q)$ the restriction of $\mathbf{y}$ to Q. For $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$, we denote by $I(\mathbf{Y}, \mathbf{y}, Q) \triangleq \{n : \mathbf{y}_n(Q) = \mathbf{y}(Q)\}$ the indices of records in $\mathbf{Y}$ that match $\mathbf{y}$ on Q (including $\mathbf{y}$ itself). Let $\mathsf{EA}(\mathbf{Y}, \mathbf{y}, Q) = |I(\mathbf{Y}, \mathbf{y}, Q)|$. This is called the *effective anonymity* of $\mathbf{y}$ in $\mathbf{Y}$ with respect to Q.

Definition 3.1 (k-anonymity)**.** For $k \geq 2$, $\mathbf{Y}$ is *k-anonymous* with respect to Q if for all $\mathbf{y} \in \mathbf{Y}$, $\mathsf{EA}(\mathbf{Y}, \mathbf{y}, Q) \geq k$. An algorithm $M : \mathbf{X} \mapsto \mathbf{Y}$ is *k-anonymizer* if for every $\mathbf{X}$, $\mathbf{Y} \leftarrow M(\mathbf{X})$ is k-anonymous *(anonymity)* and generalizes $\mathbf{X}$ *(correctness)*. We omit Q when $Q = \mathcal{U}^D$ is the whole data universe.

A few remarks are in order. First, beyond correctness and anonymity, k-anonymity places no restriction on the output $\mathbf{Y}$. Second, the term quasi-identifier is inconsistently defined in the literature. Our definition of a quasi-identifier as the collection of multiple attributes is from Sweeney [27]. Quasi-identifier is commonly used to refer to one of the constituent attributes—including by the authors of the EdX dataset [19]. So each [27]-quasi-identifier consists of multiple [19]-quasi-identifers. We adopt the quasi-identifier-as-a-set definition because it simplifies the discussion of the EdX dataset in Section 6. The distinction disappears in Sections 4 and 5: our downcoding and PSO attacks work even when every attribute is part of the quasi-identifier (i.e., $Q = \mathcal{U}^D$).

3.2.1 Hierarchical k-anonymity

It is easy to contrive k-anonymizers that reveal $\mathbf{X}$ completely. Directing our attention to more natural and widespread mechanisms, we focus on *hierarchical* k-anonymizers.

A common way of k-anonymizing data is to generalize an attribute domain $\mathcal{U}$ according to a data-independent *generalization hierarchy* H which specifies how a given attribute may be *recoded*.[5] Many natural ways of generalizing data fits this mold: using nesting geographies (e.g., city→state→country); dropping digits of postal codes (e.g., $91011 \to 9101\star \to 910\star\star$); grouping ages into ranges of 5, 10, 25, or 50 years; suppressing attributes or whole records altogether; and the techniques used to create the EdX dataset.

Formally, a generalization hierarchy H defines a structured collection of permissible subsets $\mathbf{y}$ of an attribute domain $\mathcal{U}$ (Figure 5). H is a rooted tree labelled by subsets of $\mathcal{U}$, where the subsets on any level of H form a partition of $\mathcal{U}$ and the partition on every level is a strict refinement of the partition above. The label of the root is $\mathcal{U}$ itself, and the leaves are all labelled with singletons $\{x\}$. Identifying H with the set of all its labels, we write $y \in H$ if there is some node in H labelled by y. We extend the hierarchy H to the data universe $\mathcal{U}^D$ coordinate-wise, writing $\mathbf{y} \in H^D$ if $y_d \in H$ for all $d \in [D]$.

Definition 3.2 (Hierarchical k-anonymity)**.** $\mathbf{Y}$ *respects* H if $\mathbf{y} \in H^D$ for all $\mathbf{y} \in \mathbf{Y}$. An algorithm $M : (\mathbf{X}, H) \mapsto \mathbf{Y}$ is a *hierarchical* k-anonymizer if for all $\mathbf{X}$ and all hierarchies H, $M_H : \mathbf{X} \mapsto M(\mathbf{X}, H)$ is a k-anonymizer and its output $\mathbf{Y} = M(\mathbf{X}, H)$ respects H.

Observe that one can always implement hierarchical k-anonymity by simply outputting N copies of $\mathcal{U}^D$. But a privacy technique that completely destroys the data is not useful, which leads us to consider data quality.

We consider *minimal* mechanisms [25]. A mechanism is minimal if no record is generalized more than necessary to achieve the privacy requirement (in a local way). For example, suppose a k-anonymous $\mathbf{Y}$ contains a location attribute. If there is a subset of records whose location "USA" can be changed to "California" without violating k-anonymity, then the mechanism that produced $\mathbf{Y}$ would not be minimal. Anoter example is given in Figure 2. We call this property *minimality* because it is equivalent to requiring minimality with respect to the partial ordering $\preceq$. Unlike global optimality, minimality is computationally tractable.

Definition 3.3 (Hierarchical minimality)**.** $M : (\mathbf{X}, H) \mapsto \mathbf{Y}$ is *minimal* if $\mathbf{Y}$ is always minimal in the set of all H-respecting, k-anonymous $\mathbf{Y}$ that generalize $\mathbf{X}$, partially ordered by $\preceq$. That is, for all strict refinements $\mathbf{Z} \prec \mathbf{Y}$, either: (a) $\mathbf{Z}$ is not k-anonymous, (b) $\mathbf{Z}$ does not respect H, or (c) $\mathbf{Z}$ does not generalize $\mathbf{X}$.

[5] Hierarchical algorithms differ on whether they use *local recoding* or *global recoding*. Using local recoding, attributes in different records can be generalized to different levels of the hierarchy. Using global recoding, all records must use the same level in the hierarchy for any given attribute. We consider local recoding which produces higher quality datasets in general.

4 Downcoding attacks on syntactic privacy techniques

We study a new class of attacks on hierarchical k-anonymity called *downcoding attacks* and prove that all minimal hierarchical k-anonymizers are vulnerable to downcoding attacks. Our downcoding attacks are powerful yet computationally straightforward. The attacks apply as is to ℓ-diversity, t-closeness, and the many QI-deidentification techniques in Section 2.1. They demonstrate that even when every attribute is treated as a quasi-identifier, any privacy offered by QI-deidentification depends on unstated distributional assumptions about the dataset.

4.1 Overview

In short, downcoding undoes hierarchical generalization. A downcoding attack takes as input a dataset generalized and recovers some fraction of the generalized data. We call this downcoding as it corresponds to recoding records down a generalization hierarchy. Our downcoding attacks are powered by a simple observation: *minimality leaks information*. Figures 1 and 2 give simple examples of downcoding and of leakage from minimality, respectively.

We prove that there exist data distributions and hierarchies such that every minimal hierarchical k-anonymizer is vulnerable to downcoding attacks. Hence any privacy provided by QI-deidentification is subject to distributional assumptions.

The downcoding attack adversary A gets as input a QI-deidentified dataset $\mathbf{Y}$ which is the output of an unknown mechanism M on an unknown dataset $\mathbf{X}$. A also knows anything published with $\mathbf{Y}$, namely N, k, and the hierarchy H. (Without H data users would be unable to interpret $\mathbf{Y}$.) Finally we also allow the adversary to depend on the data distribution U. One interpretation is that the security that a mechanism affords against downcoding attacks depends on limiting the attacker's knowledge, which is not good security practice. Moreover, in many settings U can be efficiently learned from an independent sample $\mathbf{X}'$.

Formally, we construct a distribution U over $\omega(\log n)$ attributes and a generalization hierarchy H such that every minimal hierarchical algorithm enables downcoding attacks on datasets drawn i.i.d. from U. Our first attack uses a natural data distribution (i.e., clustered heteroskedastic data in Section 4.4) and a tree-based hierarchy, and allows an attacker to completely recover a constant fraction of the deidentified records with high probability. Our second attack uses a less natural data distribution and hierarchy, and allows an attacker to recover $3/8$ths of every record with 99% probability.

Even with the assumptions on M and the knowledge of A, our attacks are far more general that typical attacks against QI-deidentification. For example, the attacks that motivated t-closeness as a refinement of ℓ-diversity don't even apply to a single well-defined mechanism [17]. They show only that it

is possible for a mechanism to produce ℓ-diverse outputs that are vulnerable. In contrast, we show attacks on a large and well-defined class of mechanisms. Moreover, our attacks work against all QI-deidentification definitions simultaneously, not any one alone.

4.2 Definition

Let $\mathbf{Y}$ be a k-anonymous version of a dataset $\mathbf{X}$ with respect to generalization hierarchy H. A downcoding attack takes $\mathbf{Y}$ as input and outputs a strict refinement $\mathbf{Z}$ of $\mathbf{Y}$ that simultaneously respects H and generalizes $\mathbf{X}$.

Definition 4.1 (Downcoding attack)**.** Let $\mathbf{Y}$ be a hierarchical k-anonymous generalization of a (secret) dataset $\mathbf{X}$ with respect to some hierarchy H. $\mathbf{Z}$ is a *downcoding* of $\mathbf{Y}$ if $\mathbf{X} \preceq \mathbf{Z}$, $\mathbf{Z} \prec \mathbf{Y}$, and $\mathbf{Z} \in H$.

Observation 4.1. *If* $\mathbf{Y}$ *is minimal and* $\mathbf{Z}$ *is a downcoding of* $\mathbf{Y}$*, then* $\mathbf{Z}$ *violates k-anonymity.*

We consider three measures of an attack's strength: How *many* records are refined? How *much* are records refined? How *often* records refined? Recall that if $\mathbf{Z} \prec \mathbf{Y}$, then $\mathbf{z}_n \subseteq \mathbf{y}_n$ for all n and $\mathbf{z}_n \subsetneq \mathbf{y}_n$ for at least one n.

- Δ_N: How *many* records are refined? For $\Delta_N \in \mathbb{N}$, we write $\mathbf{Z} \prec_{\Delta_N} \mathbf{Y}$ if there exist at least Δ_N distinct n for which $\mathbf{z}_n \subsetneq \mathbf{y}_n$. That is, $\mathbf{Z}$ strictly refines at least Δ_N records in $\mathbf{Y}$. An attacker prefers larger Δ_N.
- Δ_D: How *much* are the records refined? For $\Delta_D \in \mathbb{N}$, we write $\mathbf{z} \subsetneq_{\Delta_D} \mathbf{y}$ if there exist at least Δ_D distinct d for which $z_d \subsetneq y_d$. We write $\mathbf{Z} \prec_{\Delta_D} \mathbf{Y}$ if $\mathbf{z}_n \subsetneq \mathbf{y}_n \implies \mathbf{z}_n \subsetneq_{\Delta_D} \mathbf{y}_n$. That is, either $\mathbf{z}_n = \mathbf{y}_n$ or it $\mathbf{z}_n$ strictly refines $\mathbf{y}_n$ along at least Δ_D dimensions. An attacker prefers larger Δ_D.
- $\beth$:[6] How *often* are records refined? Consider the probability experiment $\mathbf{X} \sim U^N$, $\mathbf{Y} \leftarrow M(\mathbf{X}, H)$, and $\mathbf{Z} \leftarrow \mathsf{A}(\mathbf{Y})$ where U is a distribution over data records, M is a k-anonymizer, and A is a downcoding adversary. $\beth(\Delta_N, \Delta_D) \in [0,1]$ is the probability that $\mathbf{Z}$ downcodes with parameters at least Δ_N and Δ_D. For any fixed Δ_N and Δ_D, an attacker prefers larger $\beth$.

4.3 Minimal k-anonymizers enable downcoding attacks

Downcoding may seem impossible: How can one strictly refine $\mathbf{Y}$ using only the information contained in $\mathbf{Y}$ itself? Our attacks leverage *minimality*. The mere fact that $\mathbf{Y}$ is a minimal hierarchical generalization of $\mathbf{X}$ reveals more information about $\mathbf{X}$ that we use for strong downcoding attacks. See Figure 2 for a simple example.

A general-purpose hierarchical k-anonymizer M works for every generalization hierarchy H. Our theorems state that there exist data distributions U and corresponding hierarchies H such that every minimal hierarchical k-anonymizer M is vulnerable to downcoding. By Observation 4.1, these attacks defeat the k-anonymity of M.

Theorem 4.2. *For all* $k \geq 2$*,* $D = \omega(\log N)$*, there exists a distribution* U *over* $\mathbb{R}^D$*, and a generalization hierarchy* H *such that all minimal hierarchical k-anonymizers* M *enable downcoding attacks with* $\Delta_N = \Omega(N)$*,* $\Delta_D = 3D/8$*, and* $\beth(\Omega(N), 3D/8) > 1 - \mathrm{negl}(N)$.

Theorem 4.3. *For all constants* $k \geq 2$*,* $\alpha > 0$*,* $D = \omega(\log N)$*, and* $T = \lceil N^2/\alpha \rceil$*, there exists a distribution* U *over* $\mathcal{U}^D = [0,T]^D$*, and a generalization hierarchy* H *such that all minimal hierarchical k-anonymizers* M *enable downcoding attacks with* $\Delta_N = N$*,* $\Delta_D = D$*, and* $\beth(N, D) > 1 - \alpha$*. The attack also works for* $k = N$ *and* $D = \omega(N \log N)$.

Each of the theorems has some advantages over the other. The attacker in Theorem 4.3 manages to recover every attribute of every record $\mathbf{x} \in \mathbf{X}$ except with probability α. However the parameters of the construction depend polynomially on $1/\alpha$. Theorem 4.2 removes this dependency, at the expense of attacking only a constant fraction of records and attributes—still a serious failure of k-anonymity. The more significant advantage of Theorem 4.2 is that the data distribution and generalization hierarchy are both very natural (Example B.2). In contrast, the distribution and hierarchy in the proof of Theorem 4.3 are more contrived.

Full proofs of both Theorems 4.2 and 4.3 are in Appendix B. Both proofs follow the same structure at a very high level. We prove a structural result on minimal, hierarchical k-anonymous mechanisms for a specially constructed hierarchy H (Claims B.1 and B.3). This structural result states that if $\mathbf{X}$ satisfies certain conditions then $\mathbf{Y}$ must take a restricted form which allows the downcoding adversary to construct $\mathbf{Z}$. To prove the theorem, we construct a data distribution U such that random $\mathbf{X} \sim U^N$ will satisfy the conditions of the structural result with probability close to 1.

4.4 Example: Clustered Gaussians

The proof of Theorem 4.2 shows that distributions satisfying certain properties are vulnerable to downcoding attacks. Example B.2 describes a family of clustered Gaussian distributions that satisfy those properties. Here we give an instantiation of this family of distributions for $k = 10$ and describe the corresponding hierarchy and downcoding adversary.

We sample $N = 100$ records $\mathbf{x}$ i.i.d. as follows. Pick $\mathsf{size} = \mathsf{big}$ with probability $1/10$, and $\mathsf{size} = \mathsf{sml}$ otherwise. Pick a cluster $t \in \{1, \ldots, 10\}$ uniformly at random. Sample each attribute of $\mathbf{x}$ i.i.d. from the cluster centered at $c_t = 130t$ depending on size: If $\mathsf{size} = \mathsf{sml}$ sample from $\mathrm{N}(c_t, 1)$ distribution. If $\mathsf{size} = \mathsf{big}$ sample from the $\mathrm{N}(c_t, 100)$.

The hierarchy H consists of the interval $[A_1, A_{11})$ subdivided into intervals $[A_t, A_{t+1})$. As depicted in Figure 3, each

[6] $\beth$ is pronounced "dah-let" and is the fourth letter of the Hebrew alphabet.

$[A_t, A_{t+1})$ is further subdivided into $[B_t, D_t)$ and its complement $[A_t, B_t) \cup [D_t, A_{t+1})$. The key property is that half of the mass of $\mathrm{N}(0, 100)$ lies in the corresponding interval $[B_t, D_t)$. For the above parameters: $A_t = c_t - 65$, $B_t = c_t - 6.6$, and $D_t = c_t + 6.6$.

The adversary A is described in Algorithm 1. It takes as input $\mathbf{Y}$, k, and a description of H. It looks at each group of generalized records $\widehat{\mathbf{Y}}_t$ of the output. If the number of records in $\widehat{\mathbf{Y}}_t$ is not k, then the whole group of records is copied to the output $\mathbf{Z}$ unchanged (i.e., no downcoding on these records). If $\widehat{\mathbf{Y}}_t$ has exactly k records, then by k-anonymity these records are all identical copies of some $\mathbf{y}^t$. Some of $\mathbf{y}^t$'s entries may be aggregated to $[A_t, A_{t+1})$. If it's many more or many less than half the entries, then the whole group of records is copied to the output $\mathbf{Z}$ unchanged (i.e., no downcoding on these records). Otherwise, the k records in $\widehat{\mathbf{Y}}_t$ all get downcoded as described in the algorithm.

It follows from Example B.2 that for $k = 10$, the distribution described above, and $\mathbf{Y}$ produced by any minimal hierarchical k-anonymizer, A will downcode a constant fraction of the records in $\mathbf{Y}$ (with constant probability).

Algorithm 1: Adversary A for the example in Section 4.4 (see also Fig. 3).

Data: $\mathbf{Y}$, k
Result: $\mathbf{Z}$
for *cluster* $t = 1, \ldots, T$ **do**
  Let $\widehat{\mathbf{Y}}_t$ be the records with an entry in $[A_t, A_{t+1})$;
  if $|\widehat{\mathbf{Y}}_t| \neq k$ **then**
    Copy every $\mathbf{y} \in \widehat{\mathbf{Y}}_t$ into $\mathbf{Z}$;
    continue;
  /* $\widehat{\mathbf{Y}}_t$ is k exact copies of some $\mathbf{y}^t$ */
  $\mathsf{big}_t \leftarrow \{d : y^t_d = [A_t, A_{t+1})\}$;
  $b_t \leftarrow |\mathsf{big}_t|$;
  if $|b_t - D/2| > D/8$ **then**
    Write k copies of $\mathbf{y}^t$ to $\mathbf{Z}$;
  else
    Write $k-1$ copies of $[B_t, D_t]$ to $\mathbf{Z}$;
    Write $\mathbf{z}^t$ to $\mathbf{Z}$, where

$$z^t_d = \begin{cases} [B_t, D_t) & d \notin \mathsf{big}_t \\ [A_t, B_t) \cup [D_i, A_{i+1}) & d \in \mathsf{big}_t \end{cases}$$

5 Predicate singling-out attacks on syntactic privacy techniques

Our downcoding attacks yield powerful predicate singling-out (PSO) attacks against minimal hierarchical k-anonymous mechanisms. PSO attacks were recently proposed as a way to demonstrate that a privacy mechanism fails to legally anonymize under Europe's General Data Protection Regulation [2, 7]. Our new attacks undermine the use k-anonymity and other QI-deidentification techniques for GDPR compliance, challenging prevailing European guidance on anonymization [23].

In this section, we recall the prior work on PSO attacks and define a generalization called *compound PSO attacks*. We prove that minimal hierarchical k-anonymizers enable compound PSO attacks.

5.1 Background on PSO attacks

Predicate singling-out attacks were recently introduced by Cohen and Nissim in the context of data anonymization under Europe's General Data Protection Regulation (GDPR) [7]. They were proposed as a mathematical test to show that a privacy mechanism fails to legally anonymize data under GDPR [2, 7]. A mechanism M legally anonymizes under GDPR if it suffices to transform regulated *personal data* into unregulated *anonymous data*. That is, if $M(\mathbf{X})$ is free from GDPR regulation regardless of what $\mathbf{X}$ is. If a mechanism enables PSO attacks, then it does not legally anonymize under GDPR [2].

Informally, M enables PSO attacks if given $M(\mathbf{X})$, an adversary is able to learn an extremely specific description ψ of a single record in $\mathbf{X}$. Because ψ is so specific, it not only distinguishes the victim in the dataset $\mathbf{X}$, but likely also in the greater population. Hence PSO attacks can be a stepping stone to more blatant attacks.

Formally, we consider a dataset $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ sampled i.i.d. from distribution U over universe $\mathcal{U}^D$. The PSO adversary A is a non-uniform probabilistic Turing machine which takes as input $M(\mathbf{X})$ and produces as output a *predicate* $\psi : \mathcal{U}^D \to \{0,1\}$. ψ *isolates* a record in a dataset $\mathbf{X}$ if there exists a unique $\mathbf{x} \in \mathbf{X}$ such that $\psi(\mathbf{x}) = 1$. Equivalently, if $\psi(\mathbf{X}) = \sum_{\mathbf{x} \in \mathbf{X}} \psi(\mathbf{x})/n = 1/n$. The strength of a PSO attack is related to the *weight* of the predicate ψ output by A: $\psi(U) \triangleq \mathbb{E}(\psi(\mathbf{x}))$ for $\mathbf{x} \sim U$. We simplify the definitions from [7] to their strongest setting: where $\psi(U) < \mathrm{negl}(n)$.

To perform a PSO attack, A outputs a single negligible-weight predicate ψ that isolates a record $\mathbf{x} \in \mathbf{X}$ with non-negligible probability.

Definition 5.1 (Predicate singling-out attacks (simplified) [7])**.** *M enables predicate singling-out (PSO) attacks* if there exists U, A, and $\beta(n)$ non-negligible such that

$$\Pr_{\substack{\mathbf{X} \leftarrow U^n \\ \psi \leftarrow \mathsf{A}(M(\mathbf{X}))}} [\psi(\mathbf{X}) = 1/n \wedge \psi(U) < \mathrm{negl}(n)] \geq \beta(n).$$

Cohen and Nissim give a simple PSO attack against a large class of k-anonymizers which they call *bounded*. A k-anonymizer is bounded if there is some maximum $k_{\max}$ such that for all $\mathbf{X}$, the effective anonymity of every row of $\mathbf{Y}$ is at

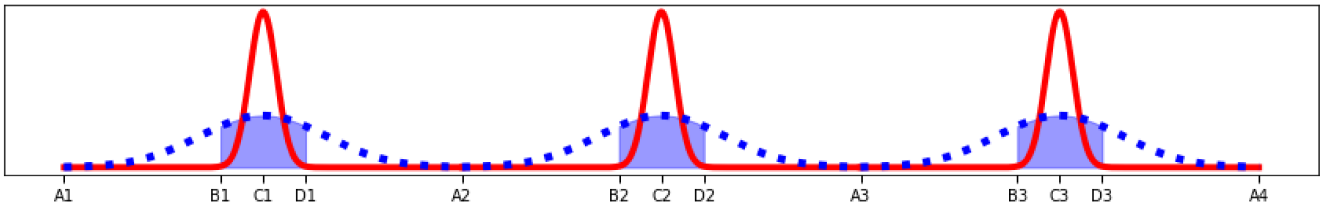

Figure 3: The marginal distribution of each attribute for the example described in Section 4.4 (depicting 3 of 10 clusters, not to scale). A key property is that half of the mass of the t^{th} blue dotted distribution lies in the interval $[B_t, D_t)$.

most $k_{\max}$. The attacker outputs $L = O(N)$ disjoint negligible-weight predicates ψ. If M is bounded, each ψ isolates a row in $\mathbf{X}$ with probability about $\eta/e \gg 0$ independently, where $\eta \in [0,1]$ is a parameter that depends on M and U.

5.2 Compound predicate singling-out attacks

PSO attacks can be unsatisfying. For example, the attack from [7] outputs L predicates and at best about L/e manage to actually isolate a record in the dataset $\mathbf{X}$. Moreover, which predicates isolate and which don't is impossible for the attacker to know without additional information. So even though there exists many isolated records with high probability, the attacker doesn't know which ones or how many. In contrast, consider an attacker that outputs $L = N$ predicates, each of which isolates a distinct record in $\mathbf{X}$. It is obvious the new attacker is stronger, but in a way that isn't captured by the definition of predicate singling-out.

We define a generalization of PSO attacks called *compound* PSO attacks. Whereas PSO attacks only require that a record is isolated with non-negligible probability, compound PSO attacks require many records to be isolated often.

To perform a compound PSO attack, A outputs multiple negligible-weight predicates $\Psi = \{\psi_1, \ldots, \psi_L\}$ each of which isolates a distinct record $\mathbf{x} \in \mathbf{X}$ with probability at least $1 - \alpha$. The strength of the attack is measured by L and α, with $L \to n$ and $\alpha \to 0$ reflecting stronger attacks. Vanilla PSO attacks correspond to the setting $L = 1$ and $\alpha = 1 - \beta$.

Definition 5.2 (**(α, L)-compound-PSO attacks**)**.** *M enables (α, L)-compound predicate singling-out attacks* if there exists U, A such that

$$\Pr\left[\forall \psi, \psi' \in \Psi : \begin{array}{c} \psi(\mathbf{X}) = 1/n \wedge \psi(U) < \text{negl}(N) \\ \wedge (\psi \wedge \psi')(U) = 0 \wedge |\Psi| \geq L \end{array}\right] \geq 1 - \alpha(N)$$

in the probability experiment $\mathbf{X} \sim U^N$, $\Psi \leftarrow \mathsf{A}(M(\mathbf{X}))$.

In the language of compound attacks, the prior work gives an $(1 - O(e^{-L}), L)$-compound-PSO attack against bounded k-anonymizers for $L < cN$ and some $c > 0$.

Our compound PSO attacks are much stronger. Theorem 5.1 gives a $(\text{negl}(N), \Omega(N))$-compound-PSO attack, and Theorem 5.2 gives a $(\text{poly}(1/N), N)$-compound-PSO-attack. In both attacks, the adversary fails only if the dataset $\mathbf{X}$ is atypical in some way. If the dataset is typical, the compound PSO attack always succeeds regardless of what the mechanism M does. The tradeoff is that our new attacks only work on minimal hierarchical k-anonymizers (instead of all bounded k-anonymizers) and with more structured data distributions U (instead of any U with moderate min-entropy).

Theorem 5.1. *For all $k \geq 2$, $D = \omega(\log N)$, there exist a distribution U over $\mathbb{R}^D$, a generalization hierarchy H, such that all minimal hierarchical k-anonymizers M enable $(\text{negl}(N), \Omega(N))$-compound-PSO attacks.*

Theorem 5.2. *For all constants $k \geq 2$, $\alpha > 0$, $D = \omega(\log N)$, and $T = \lceil N^2/\alpha \rceil$, there exists a distribution U over $\mathcal{U}^D = [0,T]^D$, a generalization hierarchy H, such that all minimal hierarchical k-anonymizers M enable (α, N)-compound-PSO attacks. The attack also works for $k = N$ and $D = \omega(N \log N)$.*

These theorems mirror Theorems 4.2 and 4.3, inheriting their advantages and disadvantages. Proofs for both attacks follow the same general structure, using the corresponding downcoding attacks in non-black-box ways (Appendix B). The key observation is that some of the downcoded records in the downcoding attacks immediately give the predicates needed to predicate single-out.

Algorithm 2 illustrates the compound-PSO adversary for the example of clustered Gaussians described in Section 4.4. Compare to the downcoding adversary in Algorithm 1. Instead of outputting a complete dataset $\mathbf{Z}$ (as in the downcoding attack), we simply output descriptions of certain records within $\mathbf{Z}$. Namely, $\mathsf{matches}(\mathbf{z}) : \mathbf{x} \mapsto \{0,1\}$ is the predicate that outputs 1 if and only if $\mathbf{x}$ is consistent with $\mathbf{z}$ (i.e., $\mathbf{x} \subseteq \mathbf{z}$).

6 Reidentifying EdX students using LinkedIn

597,692 individuals registered for 17 online courses offered by Harvard and MIT through the EdX platform [15]. We show that thousands of these students are potentially vulnerable to reidentification. The EdX dataset represents an egregious failure of k-anonymity in practice and in a case where the dataset was "properly deidentified" by "statistical experts" in accordance with regulations, undermining one of the main arguments used to justify the continued use of QI-deidentification [12].

EdX collected data about students' demographics, engagement with course content, and final course grade. EdX sought

Algorithm 2: Compound-PSO adversary for the example in Section 4.4 (compare with Alg. 1).

Data: $\mathbf{Y}$, k
Result: Ψ
for *cluster* $t = 1, \ldots, T$ **do**
  Let $\widehat{\mathbf{Y}}_t$ be the records with an entry in $[A_t, A_{t+1})$;
  if $|\widehat{\mathbf{Y}}_t| \neq k$ **then**
    continue;
  $\mathsf{big}_t \leftarrow \{d : y^t_d = [A_t, A_{t+1}]\}$;
  $b_t \leftarrow |\mathsf{big}_t|$;
  if $|b_t - D/2| > D/8$ **then**
    continue;
  else
    $\Psi \leftarrow \Psi \cup \{\mathsf{matches}(\mathbf{z}^t)\}$, where

$$z^t_d = \begin{cases} [B_t, D_t) & d \notin \mathsf{big}_t \\ [A_t, B_t) \cup [D_i, A_{i+1}) & d \in \mathsf{big}_t \end{cases}$$

to make the data public to enable outside research but considered it protected by the Family Educational Rights and Privacy Act (FERPA), a data privacy law restricting the disclosure of certain educational records [19]. "To meet these privacy specifications, the HarvardX and MITx research team (guided by the general counsel, for the two institutions) opted for a k-anonymization framework" [3]. A value of $k = 5$ "was chosen to allow legal sharing of the data" in accordance with FERPA. Ultimately, EdX published the 5-anonymized dataset with 476,532 students' records.

We show that thousands of these students are potentially vulnerable to reidentification. As a proof of concept, we reidentified 3 students out of 135 students for whom we searched for matching users on LinkedIn. Each of the reidentified users failed to complete at least one course in which they were enrolled, a private fact disclosed by the reidentification attack.

The limiting factor of this attack was not the privacy protection offered by k-anonymity itself, but the fact that many records in the raw dataset were missing demographic variables altogether. In order to boost the confidence of our attack, we restricted our attention to *unambiguously* unique records. To demonstrate the possibility of attribute disclosure, we further restricted our attention to students that had enrolled in, but failed to complete, a course on EdX.

6.1 The Harvard-MIT EdX Dataset

$\mathbf{X}_{\mathsf{ed}}$ has 476,532 rows, one per student.[7] Each row contains the student's basic demographic information, and information about the student's activities and outcomes in each of 16 of the 17 EdX courses.

The demographics included self-reported level of education, gender, and year of birth, along with a country inferred from the student's IP address. Many students chose not to report level of education, gender, and year of birth at all, so these columns are missing many entries. For each course, $\mathbf{X}_{\mathsf{ed}}$ indicates whether the student enrolled in the course, their final grade, and whether they earned a certificate of completion. $\mathbf{X}_{\mathsf{ed}}$ also includes information about students' activities in courses including how many forum posts they made.

$\mathbf{X}_{\mathsf{ed}}$ was 5-anonymized with respect to 17 overlapping quasi-identifiers separately: $Q_1, \ldots, Q_{16}$, and Q_* defined next. Recall that each quasi-identifier is a subset of attributes, not a single attribute (Def. 3.1).

- $Q_i = \{$gender, year of birth, country, enrolled in course i, number of forum posts in course $i\}$
- $Q_* = \{$enrolled in course 1, …, enrolled in course 16$\}$.

Anonymization was done hierarchically. First, locations were globally coarsened to countries or continents. Then other attributes or whole records were suppressed as needed.

6.2 Uniques in the EdX dataset

Table 1 summarizes the results of all analyses described in this section. Let $Q_{\mathsf{all}} = Q_* \cup Q_1 \cup \cdots \cup Q_{16}$. $\mathbf{X}_{\mathsf{ed}}$ is very far from 5-anonymous with respect to Q_{all}. We find that 7.1% of students (33,925 students) in $\mathbf{X}_{\mathsf{ed}}$ are unique with respect to Q_{all} and 15.3% have effective anonymity less than 5.

Despite EdX's goals, $\mathbf{X}_{\mathsf{ed}}$ was not even 5-anonymous with respect to Q_*: 245 students were unique and 753 had effective anonymity less than 5! We suspect this blunder is due to k-anonymity's fragility with respect to post-processing. The raw data was first 5-anonymized with respect to Q_* and afterwards with respect to $Q_1, \ldots, Q_{16}$. Some rows in the dataset were deleted in the latter stage, ruining 5-anonymity for Q_*.

We emphasize that the creators of the EdX dataset never intended or claimed to provide 5-anonymity with respect to Q_{all}. But they admit that each of the attributes in Q_{all} is potentially public. In our view, the union of quasi-identifiers should also be considered a quasi-identifier and any exception should be justified. No justification is given.

6.2.1 Unambiguous uniques in the EdX dataset

A naive interpretation of the 7.1% unique students is that an attacker who knows Q_{all} would be able to definitively learn the grades of 7.1% of the students. But there is a major source of ambiguity: missing information. Gender, year of birth, and level of education were voluntarily self-reported by students. Many students chose not to provide this information: 14.9% of students records are missing at least one of these attributes. It is missing in the raw data, not just the published data. Thus,

[7] The dataset as published was such that each row represented a student-course pair, with a separate row for each course in which a student enrolled. Records corresponding to the same student shared a common UID. $\mathbf{X}_{\mathsf{ed}}$ as described above is the result of aggregating the information by UID. See the appendix for additional background on the EdX dataset.

	EA		$\mathsf{EA}_{\mathsf{amb}}$	
Aux info	$=1$	<5	$=1$	<5
Q_*	245	753	245	753
Q_{all}	33,925	73,136	9,125	22,491
Q_{posts}	120	216	120	216
	(1.7%)	(3.0%)	(1.7%)	(3.0%)
Q_{acq}	31,797	69,543	7,108	19,203
$Q_{\mathsf{acq}+}$	41,666	98,201	7,512	20,402
Q_{resume}	5,542	10,939	732	2,310
	(34.2%)	(67.4%)	(4.5%)	(14.2%)

Table 1: Number of students by effective anonymity (EA) or ambiguous effective anonymity ($\mathsf{EA}_{\mathsf{amb}}$) with respect to various choices of attacker auxiliary information (Q_*, Q_{all}, etc.), as described in this section. Numbers in parentheses are the value as a percentage of the relevant subset of the full dataset: for Q_{posts}, the 7,251 students with at least one forum post; for Q_{resume}, the 16,224 students with at least one certificate. $\mathsf{EA} = \mathsf{EA}_{\mathsf{amb}}$ for Q_*, Q_{posts}.

a female Italian born in 1986 might appear in the dataset with any or all three attributes missing.

This makes the 7.1% result difficult to interpret. From an inferential standpoint, the relevant question is not how many students have unique quasi-identifiers, but how many are *unambiguously* unique. We compute the ambiguous effective anonymity $\mathsf{EA}_{\mathsf{amb}}$ (defined in App. A.1) of each record by treating any missing attribute values as the set of all possible values for that attribute. This number may be much lower than 7.1%. We stress that this ambiguity comes from missing data, not from k-anonymity.

We find that 1.9% of students (9,125 students) are unambiguously unique with respect to Q_{all} and 4.7% have ambiguous effective anonymity less than 5. Over 9,000 students are unambiguously identifiable in the dataset to anybody who knows all the quasi-identifiers, without knowing whether the students chose to self-report their gender, year of birth, or level of education. This allows an attacker to draw meaningful inferences about them.

6.2.2 Limiting the attacker's knowledge

Students in the EdX dataset are vulnerable to reidentification by adversaries who have much less auxiliary information than Q_{all}. We consider the (ambiguous) effective anonymity for three attackers who could plausibly reidentify students in the EdX dataset: a prospective employer, a casual acquaintance, and an EdX classmate. The results are summarized in Table 1.

In Section 6.3, we carry out the prospective employer attack using LinkedIn. This demonstrates that some students in the EdX dataset can be reidentified by anybody.

Prospective employer Consider a prospective employer who is interested in discovering whether a job applicant failed an EdX course. An applicant is likely to list EdX certificates on their resume. The employer very likely knows $Q_{\mathsf{resume}} =$ {gender, year of birth, location, level of education, certificates earned in courses 1–16}. Q_{resume} only includes those certificates actually earned, but omits courses in which a student enrolled but did not earn a certificate.

5,546 students in $\mathbf{X}_{\mathsf{ed}}$ have effective anonymity 1 with respect to Q_{resume}, and 10,942 have effective anonymity less than 5. These numbers may seem small, but they constitute 34.2% and 67.4% of the 16,224 students in the dataset that earned any certificates whatsoever. Moreover, 732 students are unambiguously unique—333 of whom failed at least one course, and 38 of whom failed three or more courses. Thus, *2.1% of students (333 students) who earned certificates of completion failed at least one course and have unambiguous effective anonymity 1* with respect to Q_{resume}.

Casual acquaintance Casual acquaintances might, in the course of normal conversation, discuss their experiences on EdX. They would likely discuss which courses they took, and would naturally know each other's ages, genders, and locations. So acquaintances know $Q_{\mathsf{acq}} =$ {gender, year of birth, location, enrollment in courses 1–16} $\subseteq Q_{\mathsf{all}}$. 6.7% of students in $\mathbf{X}_{\mathsf{ed}}$ have effective anonymity 1 with respect to Q_{acq}, and 14.6% have effective anonymity less than 5.

Moreover, acquaintances typically know each other's level of education too, even though this is not included in Q_{all}. If we augment the acquaintance's knowledge with level of education $Q_{\mathsf{acq}+} = Q_{\mathsf{acq}} \cup \{\text{education}\}$, then things become even worse. 8.7% students in $\mathbf{X}_{\mathsf{ed}}$ have effective anonymity 1 with respect to $Q_{\mathsf{acq}+}$, and 20.6% have effective anonymity less than 5.

EdX classmate Each EdX course had an online forum for student discussions. Because these posts were public to all students enrolled in a given course, the number of forum posts made by any user was deemed publicly available information. But ignoring composition, EdX did not consider the combination of forum post counts made by a user across courses.

Consider an attacker who knows $Q_{\mathsf{posts}} =$ {number of forum posts in courses 1–16} $\subseteq Q_{\mathsf{all}}$. 120 students in $\mathbf{X}_{\mathsf{ed}}$ are unambiguously unique with respect to Q_{posts}, and 216 have ambiguous effective anonymity less than 5. These numbers may seem minute, but they constitute 1.7% and 3.0% of the 7251 students in the dataset that made any forum posts whatsoever. Effective anonymity and ambiguous effective anonymity are always the same for this attacker because Q_{posts} excludes the demographic columns that are missing many entries.

Who knows Q_{posts}? 20 students *in the dataset itself* enrolled in all 16 courses and could have compiled forum post counts across all courses for all other EdX students. To any one of these 20 students the 120 students with distinguishing

forum posts are uniquely identifiable. Such an attacker can then learn these 120 students ages, genders, level of educations, locations, and their grades in the class.

In fact, each of the 120 vulnerable students can be unambiguously uniquely distinguished by 23–70 classmates; 60 students by 40–49 classmates each. This enables more classmates to act as attackers than just the 20 who took all courses. This is because distinguishing a student using forum posts doesn't require being enrolled in all 16 courses. For each of the 120 vulnerable students, we find which subsets of their forum posts suffices to distinguish them. This analysis amounts to checking whether these students remain unambiguously unique if some subset of their forum post counts are redacted.

6.3 Reidentifying EdX students on LinkedIn

On LinkedIn.com, people show off the courses they completed. They may be unwittingly revealing which courses they gave up on. 2.1% of students who earned certificates of completion (333 students) failed at least one course and have unambiguous effective anonymity 1 with respect to Q_{resume}.

We reidentified three of these 333 students, with a rough confidence estimate of 90–95%.

6.3.1 Method

People routinely post Q_{resume} on LinkedIn where it is easily searchable and accessible for a small fee. We paid $119.95 for a 1 month Recruiter Lite subscription to LinkedIn. Recruiter Lite provides access to limited search tools along with the ability to view profiles in one's "extended network": 3rd degree connections to the account holder on the LinkedIn social network. It is also possible to view public profiles outside one's extended network with a direct link, for example from a Google search. A real attacker could build a larger extended network or pay for a more powerful Recruiter account.

We performed the attack as follows. We restricted our attention to 135 students in $\mathbf{X}_{\mathsf{ed}}$ who were unambiguously unique using only certificates earned plus at most one of gender, year of birth, and location, and who also had no missing demographic attributes. We manually searched for LinkedIn users that listed matching course certificates on their profile by searching for course numbers (e.g., "HarvardX/CS50x/2012"). We attempted to access the profiles for the resulting users, whether they were in our extended network or by searching on Google. If successful, we checked whether the LinkedIn user lists exactly the same certificates as the EdX student, and whether the demographic information on LinkedIn was consistent with the EdX student. If everything matched, we consider this a reidentification.

6.3.2 Results

We reidentified 3 of the attempted 135 EdX students, each of whom registered for but failed to complete an EdX course. Two were unambiguously unique using only certificates of completion. In each case, the EdX student's gender matched the LinkedIn user's presenting gender based on profile picture and name. In each case, the LinkedIn user's highest completed degree in 2013 matched the EdX student's listed level of education.

1. Student 1's EdX record lists location ℓ_1 and year of birth as y_1. The matching LinkedIn user began a bachelors degree in year $y_1 + 20$ and was employed in country ℓ_1 in 2013.
2. Student 2's EdX record lists location ℓ_2 and year of birth as y_2. The matching LinkedIn user began a bachelors degree in year $y_1 + 18$ and was in country ℓ_2 for at part of 2013.
3. Student 3's EdX record lists location ℓ_3 and year of birth as y_3. The matching LinkedIn user graduated high school in year $y_3 + 19$, attended high school and currently works in country ℓ_3. In 2013 the LinkedIn user was employed by an international firm with offices in ℓ_3 and other countries.

6.3.3 Confidence

We cannot know for sure whether our purported reidentifications on LinkedIn are correct because were instructed by our IRB not to contact the reidentified EdX students.

In this section, we estimate that our reidentifications are correct with 90–95% confidence. Moreover, an error is most likely a result of our imperfect ability to corroborate location and year of birth on LinkedIn, not a result of the protection afforded by k-anonymity. Our analysis is necessarily very rough. A precise error analysis is impossible. We omit details to avoid imparting any other impression.

We consider two main sources of uncertainty. First is the limited information available on LinkedIn profiles, especially age and location. We inferred a range of possible ages by extrapolating from educational milestones. We inferred a set of possible locations based on listed activities around 2013. Both methods are imperfect. The locations in EdX were inferred from IP address and are likely imperfect. LinkedIn users or EdX students can report their attributes inconsistently. Note that they cannot lie about earning EdX certificates: this data comes from EdX itself and the LinkedIn certificates are digitally signed and cryptographically verifiable.[8] We estimate the probability of error on at least one attribute inferred from LinkedIn is on the order of 5–10%.

The second source of error is suppressed student records. Of the 597,692 students enrolled in EdX courses over the relevant period, only 476,532 appear in the published dataset. 121,160 students (20.3%) are completely suppressed. We matched students $\mathbf{x}_{\mathsf{edx}}$ in EdX with users on LinkedIn $\mathbf{x}_{\mathsf{li}}$ using $Q_{\mathsf{resume}} = \{$gender, year of birth, location, level of education, certificates earned in courses 1–16$\}$ as well as we could. An

[8] An example certificate is available here: https://verify.edx.org/cert/26121b8dec124bc094d324f51b70e506. Instructions for verifying the signature are here: https://verify.edx.org/cert/26121b8dec124bc094d324f51b70e506/verify.html

error will occur if a suppressed student $\mathbf{x}'_{\mathsf{edx}}$ is the true match for the LinkedIn user. For this to happen, $\mathbf{x}'_{\mathsf{edx}}$ and $\mathbf{x}_{\mathsf{edx}}$ must agree on Q_{all}. If $\mathbf{x}_{\mathsf{edx}}$ is unique in the complete dataset, no error occurs.

We do a back of the envelope calculation of the chance of error from record suppression under two simplifying assumptions. First, that random student records are suppressed.[9] Second, that the number of certificates of completion that a user earns is statistically independent of their other attributes (assuming they registered for enough courses). We compute 99.5%-confidence upper bounds for two parameters: the probability p that a random EdX student matches our reidentified EdX student; the probability q that a random EdX student earns the same number of certificates as our reidentified EdX student. $121{,}160 \cdot pq$ is a very coarse estimate of the probability that a supressed student record causes an error. For the three students we reidentified, this comes out to 0.1–1%.

A much less likely source of error is suppression of individual courses from a student's record. Such an error will occur if some courses for the purported match $\mathbf{x}_{\mathsf{edx}}$ were suppressed, and there is some other EdX student $\mathbf{x}'_{\mathsf{edx}}$ that is the true match for the LinkedIn user $\mathbf{x}_{\mathsf{li}}$. This requires course suppression in $\mathbf{x}_{\mathsf{edx}}$ and also $\mathbf{x}'_{\mathsf{edx}}$ (because $\mathbf{x}_{\mathsf{edx}}$ was unambiguously unique on Q_{resume} in the published EdX data). All in all, we consider course suppression to be a much less likely source of error than student suppression or imperfect attribute inference on LinkedIn.

6.4 EdX was "properly" deidentified

El Emam, et al., criticize prior reidentification studies as using data that were "improperly deidentified" because they did not "follow[] existing standards" [12]. They thus conclude that there is no convincing evidence of real-world failure of QI-deidentification techniques in a regulated context.

In contrast, the EdX dataset incontrovertibly followed existing standards. FERPA is the relevant regulation. It requires the published information to not enable identification of any student with reasonable certainty. The EdX dataset was specifically created to comply with FERPA, following Department of Education guidance and overseen by general council for Harvard and MIT [3].

Moreover, the EdX dataset arguably followed the HIPAA Expert Determination–the standard used by El Emam, et al. The Expert Determination standard requires three things:[10] (1) Deidentification be performed by "a person with appropriate knowledge . . . and experience". (2) The person determines that the risk of reidentification is "very small". (3) The person "documents the methods and results of the analysis that justify such determination." The creation of the EdX data was overseen by Harvard professors in computer science and statistics with specific expertise in privacy and inference. They find a "low probability that the dataset will be re-identified" and their methods and analysis are well-documented [19]. The main deviation from the Expert Determination standard is the difference between "very small" and "low" reidentification risk.

[9] There are more sophisticated techniques for estimating the probability of error under this assumption [24]. But in EdX omitted records are "outliers and highly active users because these users are more likely to be unique and therefore easy to re-identify" [19]. As such, using the more sophisticated techniques would not give more meaning to our very coarse estimates.

[10] https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

7 Conclusions

In short, we show that k-anonymity – and QI-deidentification generally – fails *on its own terms*. Our attacks rebut three primary arguments that QI-deidentification's practioners make to justify its continued use. First, we reidentify individuals in EdX dataset; it was "properly de-identified" by a "statistical experts" and in accordance with procedures outlined in regulation, meeting the high bar set by El Emam, et al. [12]. Second, our downcoding attacks demonstrate that even if every attribute is treated as quasi-identifying, k-anonymity and its refinements may provide no protection. Ours are the first attacks in either of these two settings. Third, our attacks also undermine the claim that QI-deidentification meets regulatory standards for deidentification. The compound PSO and reidentification attacks challenge k-anonymity's status under GDPR and FERPA respectively.

Moreover, our attacks show that QI-deidentification violates three properties of a worthwhile privacy notion, even in practice. Namely, avoiding distributional assumptions, robustness against post-processing, and smooth degradation under composition. We expand on these next.

Downcoding attacks prove that whatever privacy is provided by QI-deidentification crucially depends on unstated assumptions on the data distribution. One possible pushback is that our downcoding attacks use specially constructed distributions and hierarchies, not naturally occurring ones. But even a contrived counterexample proves that there is some unnoticed distributional assumption that is critical for security. Moreover, the distributions and hierarchies in Theorem 4.2 are not so unnatural when considering that data are made, not found (to quote danah boyd). Say an analyst wants to k-anonymize a high dimensional dataset. One natural approach is to find a low-dimensional projection with clusters of about k rows each, and then construct the generalization hierarchy over this representation. The result could easily satisfy conditions that enable our downcoding attack or a direct extension.

Robustness against post-processing requires that further processing of the output, without access to the data, should not diminish privacy. Downcoding proves that QI-deidentification is not robust to post-processing. Our attacks recover specific secret information about a large fraction of a dataset's entries with probability close to 1. Also, the EdX dataset also proves

that k-anonymity is not robust to post-processing for purely syntactic reasons. The result of removing rows from a k-anonymous dataset may not satisfy k-anonymity as defined. We see this in the EdX data: it is not in fact 5-anonymous with respect to quasi-identifier Q_* (courses), despite claims otherwise. This fragility to post-processing is not so much a privacy failure as a syntactic weakness of the definition itself.

Smooth degradation under composition requires that a combination of two or more private applications mechanisms should also be private, albeit with worse parameters. The EdX dataset proves that QI-deidentification is not robust to composition, even when done by experts in accordance with strict privacy regulations. Ganta et al. present theoretical composition attacks, showing that if the same dataset is k-anonymized with different quasi-identifiers the original data can be recovered [13]. With the EdX dataset the possibility became reality. To the best of our knowledge, this is the first example of such a failure in practice.

The most important open question raised by this work is to characterize the power of downcoding attacks. What properties of a data distribution and generalization hierarchy enable downcoding? Is vulnerability to downcoding testable? In what settings is downcoding provably impossible? Can one demonstrated downcoding in the wild? We leave these questions for future work.

Responsible disclosure and data availability After reidentifying one EdX student, we reported the vulnerability to Harvard and MIT who promptly replaced the dataset with a heavily redacted one. Our IRB determined that this research was not human subjects research and did not need IRB approval. However, we were instructed by the IRB not to contact the reidentified LinkedIn users. The code used in our analysis of the EdX dataset is at `https://github.com/a785236/EdX-LinkedIn-Reidentification`, but we do not distribute the dataset itself to protect the students' privacy.

References

[1] John M. Abowd. Supplemental Declaration, 2021. State of Alabama v. US Department of Commerce.

[2] Micah Altman, Aloni Cohen, Kobbi Nissim, and Alexandra Wood. What a hybrid legal-technical analysis teaches us about privacy regulation: The case of singling out. *BUJ Sci. & Tech. L.*, 27:1, 2021.

[3] Olivia Angiuli, Joe Blitzstein, and Jim Waldo. How to de-identify your data. *Communications of the ACM*, 58(12):48–55, 2015.

[4] Michael Barbaro and Tom Zeller. A face is exposed for AOL searcher no. 4417749. *New York Times*, Aug 2006.

[5] Ann Cavoukian and Daniel Castro. *Big data and innovation, setting the record straight: de-identification does work*. Information and Privacy Commissioner, Ontario, 2014.

[6] Ann Cavoukian and Khaled El Emam. *De-identification protocols: essential for protecting privacy*. Information and Privacy Commissioner, Ontario, 2014.

[7] Aloni Cohen and Kobbi Nissim. Towards formalizing the gdpr's notion of singling out. *Proceedings of the National Academy of Sciences*, 117(15):8344–8352, 2020.

[8] Graham Cormode, Divesh Srivastava, Ninghui Li, and Tiancheng Li. Minimizing minimality and maximizing utility: analyzing method-based attacks on anonymized data. *Proceedings of the VLDB Endowment*, 3(1-2):1045–1056, 2010.

[9] Jon P Daries, Justin Reich, Jim Waldo, Elise M Young, Jonathan Whittinghill, Andrew Dean Ho, Daniel Thomas Seaton, and Isaac Chuang. Privacy, anonymity, and big data in the social sciences. 2014.

[10] John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 429–438. IEEE, 2013.

[11] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[12] Khaled El Emam, Elizabeth Jonker, Luk Arbuckle, and Bradley Malin. A systematic review of re-identification attacks on health data. *PloS one*, 6(12):e28071, 2011.

[13] Srivatsava Ranjit Ganta, Shiva Prasad Kasiviswanathan, and Adam Smith. Composition attacks and auxiliary information in data privacy. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 265–273. ACM, 2008.

[14] Olga Gkountouna. A survey on privacy preservation methods, 2011. `http://www.dblab.ece.ntua.gr/~olga/papers/olga_tr11.pdf`.

[15] Andrew Ho, Justin Reich, Sergiy Nesterko, Daniel Seaton, Tommy Mullaney, Jim Waldo, and Isaac Chuang. HarvardX and MITx: The first year of open online courses, fall 2012-summer 2013. 2014.

[16] James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In *International conference on learning representations*, 2018.

[17] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 106–115. IEEE, 2007.

[18] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. *L*-diversity: Privacy beyond *k*-anonymity. *TKDD*, 1(1):3, 2007.

[19] MITx and HarvardX. HarvardX-MITx Person-Course Academic Year 2013 De-Identified dataset, version 2.0, 2014.

[20] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *IEEE Symposium on Security and Privacy*, 2008.

[21] Arvind Narayanan and Vitaly Shmatikov. Myths and fallacies of "personally identifiable information". *Commun. ACM*, 53(6):24–26, June 2010.

[22] Paul Ohm. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, 57:1701–1777, 2010.

[23] Article 29 Data Protection Working Party. *Opinion 05/2014 on Anonymisation Techniques*.

[24] Luc Rocher, Julien M Hendrickx, and Yves-Alexandre De Montjoye. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature communications*, 10(1):1–9, 2019.

[25] Pierangela Samarati. Protecting respondents identities in microdata release. *IEEE transactions on Knowledge and Data Engineering*, 13(6):1010–1027, 2001.

[26] Pierangela Samarati and Latanya Sweeney. Generalizing data to provide anonymity when disclosing information (abstract). In Alberto O. Mendelzon and Jan Paredaens, editors, *Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*. ACM Press, 1998.

[27] Latanya Sweeney. K-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[28] Raymond Chi-Wing Wong, Ada Wai-Chee Fu, Ke Wang, and Jian Pei. Anonymization-based attacks in privacy-preserving data publishing. *ACM Transactions on Database Systems (TODS)*, 34(2):1–46, 2009.

A Additional background on the EdX dataset

We summarize the EdX dataset—the chosen quasi-identifiers, the implementation of k-anonymization, and the resulting published dataset $\mathbf{X}_{\mathsf{ed,raw}}$ based on documentation included with the dataset [19] and in two articles describing the creation of the dataset itself [3, 9].

The raw dataset consisted of 841,687 rows for 597,692 students. Each row corresponded to the registration of a single student in a single course and included the information described above. IP addresses were used to infer a student's location even when a student chose not to self-report their location. EdX considered username and IP address to be identifying. Each username was replaced by a unique 7-digit identification number (UID). A username appearing in multiple rows was replaced by the same UID in each. IP addresses were redacted.

The dataset, with a row corresponding to a student-course pair, was k-anonymized according to two different quasi-identifiers Q and Q_* (each a subset of the attributes). $Q =$ {gender, year of birth, country, course, number of forum posts}. "The last one was chosen as a quasi-identifier because the EdX forums are somewhat publicly accessible and someone wishing to re-identify the dataset could, with some effort, compile the count of posts to the forum by username" [19]. Separately, the set of courses that each student enrolled in were considered to form a quasi-identifier: $Q' =$ {enrolled in course 1, ..., enrolled in course 16}. The data was k-anonymized first according to Q_* and then according to Q. Additionally, ℓ-diversity was enforced for the final course grade, with $\ell = 2$.

After aggregating the rows by UID, $\mathbf{X}_{\mathsf{ed}}$ can be seen as k-anonymized with respect to 17 overlapping quasi-identifiers: Q_* as before and $Q_1, \ldots, Q_{16}$, where $Q_i =$ {gender, year of birth, country, enrolled in course i, number of forum posts in course i}.

The final published result $\mathbf{X}_{\mathsf{ed,raw}}$ includes 641,138 course registrations by 476,532 students across 16 courses.

As published, the EdX dataset had 641,138 rows, each representing to a single course registration for one of 476,532 distinct students. But the object of our privacy concerns is a student, not a student-course pair. We aggregated the rows corresponding to the same UID. We call the result $\mathbf{X}_{\mathsf{ed}}$.

The creators of the EdX dataset failed to identify which attributes are publicly available—the very thing that experts are supposed to be good at. Specifically, level of education and certificates of course completion are not included in any of the quasi-identifiers despite both being readily available on LinkedIn. The exclusion of certificates is particularly indefensible: at the same time as the EdX dataset was being created, EdX and LinkedIn collaborated to allows LinkedIn users to include cryptographically unforgeable certificates of completion on their profiles.

[10] Different rows of the same student often listed different countries. Al-

A.1 Ambiguous effective anonymity

We consider a relaxation of the notion of effective anonymity which we call *ambiguous effective anonymity*. Let $I_{\mathsf{amb}}(\mathbf{Y},\mathbf{y},Q) \triangleq \{n : \mathbf{y}_n(Q) \cap \mathbf{y}(Q) \neq \emptyset\}$. The ambiguous effective anonymity of $\mathbf{y}$ in $\mathbf{Y}$ with respect to Q is $\mathsf{EA}_{\mathsf{amb}}(\mathbf{Y},\mathbf{y},Q) = |I_{\mathsf{amb}}(\mathbf{Y},\mathbf{y},Q)|$. Ambiguous effective anonymity helps us reason about what an attacker can infer from the dataset.

Definition A.1 (Unambiguous uniqueness)**.** We say $\mathbf{y} \in \mathbf{Y}$ is *unique* with respect to Q if $\mathsf{EA}(\mathbf{Y},\mathbf{y},Q) = 1$, and *unambiguously unique* if $\mathsf{EA}_{\mathsf{amb}}(\mathbf{Y},\mathbf{y},Q) = 1$.

$\mathsf{EA}_{\mathsf{amb}}$ is never less than EA, and is very often greater in EdX. The presence of unambiguously unique records in a supposedly-anonymized dataset indicates a clear failure of syntactic anonymity. Considering ambiguous effective anonymity makes critiquing k-anonymity much harder. We are giving k-anonymity the benefit of all the additional ambiguity that comes from missing data rather than from the anonymizer itself.

B Deferred Proofs

B.1 Proof of Theorem 4.2

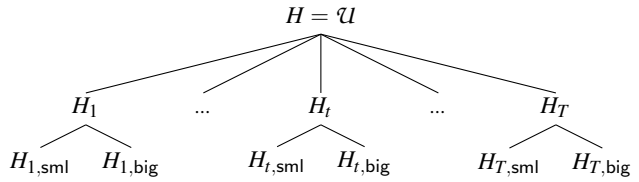

Figure 4: The generalization hierarchy used in the proof of Theorem 4.2. The attribute domain is an interval in $\mathbb{R}$, as are each H_t and $H_{t,\mathsf{sml}}$. Each set $H_{t,\mathsf{big}} = H_t \setminus H_{t,\mathsf{sml}}$ is the union of two intervals

Claim B.1. *Let H be a hierarchy with T nodes at the second level: $H_1,\ldots,H_t$ (as in Figure 4). Let $\mathbf{X} \in \mathcal{U}^D$ be a dataset, M be a minimal hierarchical k-anonymizer, and $\mathbf{Y} \leftarrow M(\mathbf{X},H)$. For $t \in [1,T]$, let $\mathbf{X}_t = \mathbf{X} \cap H_t^D$ and let $\mathbf{Y}_t$ be the records in $\mathbf{Y}$ corresponding to the records in $\mathbf{X}_t$. If $\mathbf{X} = \cup_t \mathbf{X}_t$, then for all but at most one $t \in \{t : |\mathbf{X}_t| = k\}$, $\mathbf{y} \subseteq H_t^D$ for all $\mathbf{y} \in \mathbf{Y}_t$.*

Note that as defined, the generalized records in $\mathbf{Y}_t$ are not necessarily contained in H_t^D. The claim says that if $\mathbf{X}$ consists of data in the T clusters $\mathbf{X}_1,\ldots,\mathbf{X}_T$, then the records in $\mathbf{Y}_t$ will be contained in H_t^D for almost all clusters of size exactly k.

most always there were only two different values, one of which was "Unknown/Other." In this case, we used the other value for the student's unified record. In all other cases, we used "Unknown/Other."

Proof of Claim B.1. First, we show that for any $\mathbf{y} = (y_1,\ldots,y_D)$, a single coordinate of $\mathbf{y}$ is generalized to $H = \mathcal{U}$ if and only if *every* coordinate in $\mathbf{y}$ is generalized to H. Namely, if $y_d = H$ for some d, then $\mathbf{y} = H^D$.

Suppose for contradiction that there exists $\mathbf{y} = (y_1,\ldots,y_D)$ corresponding to $\mathbf{x} \in \mathbf{X}$ such that $y_1 = H$ but $y_2 \subseteq H_t$ for some t. By the assumption on $\mathbf{X}$, there exists t' such that $\mathbf{y} \in \mathbf{Y}_{t'}$. Because $y_2 \subseteq H_t$, $t' = t$ and hence $\mathbf{y} \in \mathbf{Y}_t$. By k-anonymity, there are at least $k-1$ additional records $\mathbf{y}' \in \mathbf{Y}$ such that $\mathbf{y}' = \mathbf{y}$. Repeating the previous argument, $\mathbf{y}' \in \mathbf{Y}_t$.

Let $\mathbf{y}^* = (H_t, y_2, \ldots, y_D) \subsetneq \mathbf{y}$. Construct $\mathbf{Y}^*$ by replacing all copies of $\mathbf{y}$ in $\mathbf{Y}$ with $\mathbf{y}^*$. It is immediate that $\mathbf{Y}^*$ is k-anonymous and respects the hierarchy. By the assumption that $y_1 = H$, $\mathbf{Y}^*$ strictly refines $\mathbf{Y}$. Additionally, $\mathbf{Y}^*$ generalizes $\mathbf{X}$, because all altered rows were in $\mathbf{Y}_t$. This contradicts the minimality of the k-anonymizer M. Therefore we have proved that if $y_d = H$ for some d, then $y_d = H$ for all d.

Next, we show that for all but at most one $t \in \{t : |\mathbf{X}_t| = k\}$, there exists $\mathbf{y} \in \mathbf{Y}_t$ such that $\mathbf{y} \subseteq H_t^D$. By the preceding argument, it suffices to show that $\mathbf{y} \neq H^D$. Suppose for contradiction there exists $t \neq t'$ such that for all $\mathbf{y} \in \mathbf{Y}_t \cup \mathbf{Y}_{t'}$, $\mathbf{y} = H^D$. Construct $\mathbf{Y}'$ by replacing each $\mathbf{y} \in \mathbf{Y}_{t'}$ with $H_{t'}^D \subsetneq \mathbf{y}$. It is easy to see that $\mathbf{Y}'$ respects the hierarchy, satisfies k-anonymity, generalizes $\mathbf{X}$, and strictly refines $\mathbf{Y}$. This contradicts the minimality of the k-anonymizer M.

To complete the proof, let $t \in \{t : |\mathbf{X}_t| = k\}$ and suppose there exists $\mathbf{y} \in \mathbf{Y}_t$ such that $\mathbf{y} \subseteq H_t^D$. By k-anonymity, there must be at least $k-1$ distinct $\mathbf{y}' = \mathbf{y} \subseteq H_t^D$. By assumption on $\mathbf{X}$, each such $\mathbf{y}'$ must be an element of $\mathbf{Y}_t$. Because $|\mathbf{Y}_t| = |\mathbf{X}_t| = k$, every element of $\mathbf{Y}_t$ is equal to $\mathbf{y} \subseteq H_t^D$. □

Proof of Theorem 4.2. **Data distribution** Records $\mathbf{x} \sim U$ are noisy versions of one of $T = N/k$ cluster centers $\mathbf{c}_t \in \mathbb{R}^D$. Each coordinate x_d of $\mathbf{x}$ is $c_{t,d}$ masked with i.i.d. noise with variance σ^2. The variance is usually *small*, but is *large* with probability $1/k$ (variances $\sigma^2_{\mathsf{sml}} \ll \sigma^2_{\mathsf{big}}$). The generalization hierarchy H is shown in Figure 4. H divides the attribute domain $\mathcal{U}$ into T components H_t, each of which is further divided into small values $H_{t,\mathsf{sml}}$ and large values $H_{t,\mathsf{big}}$.

We set the parameters so that w.h.p. all of the following hold. First, the data is clustered: $\Pr_{\mathbf{x}}[\exists t,\ \mathbf{x} \in H_t^D] > 1 - \mathrm{negl}(N)$. Second, every coordinate x_d of a small-noise (variance σ^2_{sml}) record is small: $x_d \in H_{t,\mathsf{sml}}$. Third, the coordinates of large-noise (variance σ^2_{big}) records are large or small ($x_d \in H_{t,\mathsf{big}}$ or $x_d \in H_{t,\mathsf{sml}}$, respectively) with probability $1/2$ independent of all other coordinates . In particular, if $\mathbf{x}$ is generated using large noise then $\mathbf{x} \notin H^D_{t,\mathsf{sml}}$ with high probability. An example of a distribution U and hierarchy H satisfying the above is given in Example B.2. In that example, the cluster centers $\mathbf{c}_t$ are masked with i.i.d. Gaussian noise.

The adversary The adversary A takes as input $\mathbf{Y}$ and produces the output $\mathbf{Z}$ as follows. For $t \in [T]$, let $\widehat{\mathbf{Y}}_t = \mathbf{Y} \cap H_t^D$. If $|\widehat{\mathbf{Y}}_t| \neq k$, copy every $\mathbf{y} \in \widehat{\mathbf{Y}}_t$ into the output $\mathbf{Z}$. Otherwise

$|\widehat{\mathbf{Y}}_t| = k$. By k-anonymity $\widehat{\mathbf{Y}}_t$ consists of k copies of a single generalized record $\mathbf{y}^t$. Let $\mathsf{big}_t = \{d : y^t_d = H_t\}$ be the large coordinates of $\mathbf{y}^t$, and let $B_t = |\mathsf{big}_t|$ be the number of large coordinates. If $|B_t - D/2| > D/8$, then A writes k copies of $\mathbf{y}^t$ to the output $\mathbf{Z}$. Otherwise A writes $k-1$ copies of $H^D_{t,\mathsf{sml}}$ and one copy of $\mathbf{z}^t = (z^t_1, \ldots, z^t_D)$ to the output, where

$$z^t_d = \begin{cases} H_{t,\mathsf{big}} & d \in \mathsf{big}_t \\ H_{t,\mathsf{sml}} & d \notin \mathsf{big}_t \end{cases} \tag{1}$$

Analysis It is immediate from the construction that $\mathbf{Z} \preceq \mathbf{Y}$. Moreover, it is easy to arrange the records in $\mathbf{Z}$ so that $\mathbf{z}_n \subseteq \mathbf{y}_n$ for all $n \in [N]$. By construction, $\mathbf{z}_n \subsetneq \mathbf{y}_n$ implies that $\mathbf{z}_n$ differs from $\mathbf{z}_n$ differs from $\mathbf{y}_n$ on at at least $B_t \geq 3D/8$ coordinates.

To prove the theorem, it remains to show that w.h.p. $\mathbf{X} \preceq \mathbf{Z}$ and $\mathbf{Z} \prec_{\Omega(N)} \mathbf{Y}$. Let $\mathbf{X}_{\mathsf{big}} = \mathbf{X} \setminus \left(\bigcup_t H^D_{t,\mathsf{sml}}\right)$ consist of all the records $\mathbf{x}$ that have at least one large coordinate (i.e., $x_d \in H_{t,\mathsf{big}}$ for some t, d).

For all $t \in [T]$, let $\mathbf{X}_t = \mathbf{X} \cap H^D_t$ and let $\widehat{\mathbf{X}}_t \subseteq \mathbf{X}_t$ be the records $\mathbf{x} \in \mathbf{X}$ that correspond to the records in $\widehat{\mathbf{Y}}_t$. (Whereas $\mathbf{X}_t$ consists of all records that are in cluster t, $\widehat{\mathbf{X}}_t$ consists of only those records that correspond to generalized records $\mathbf{y} \in \widehat{\mathbf{Y}}_t$ that can be easily inferred to be in cluster t based on $\mathbf{Y}$.) A cluster t is $\mathbf{X}$-*good* if $|\mathbf{X}_t| = k$ and $|\mathbf{X}_t \cap \mathbf{X}_{\mathsf{big}}| = 1$. A cluster t is $\widehat{\mathbf{Y}}$-*good* if $\widehat{\mathbf{Y}}_t = k$ and $|B_t - D/2| \leq D/8$.

It suffices to show that:

- $\Omega(N)$ clusters t are $\mathbf{X}$-good.
- All but at most one $\mathbf{X}$-good clusters are $\widehat{\mathbf{Y}}$-good.
- For all $\widehat{\mathbf{Y}}$-good clusters t, $\widehat{\mathbf{X}}_t \cap \mathbf{X}_{\mathsf{big}} = \{\mathbf{x}^t\}$ and $\mathbf{x}^t \subseteq \mathbf{z}^t$.

Note that if t is both $\mathbf{X}$-good and $\widehat{\mathbf{Y}}$-good, then $\widehat{\mathbf{X}}_t = \mathbf{X}_t$. But there may be t that are $\widehat{\mathbf{Y}}$-good but not $\mathbf{X}$-good.

Many clusters are X-good We lower bound $\Pr[t\ \mathbf{X}\text{-good}]$ by a constant and then apply McDiarmid's Inequality.

$$\Pr[t\ \mathbf{X}\text{-good}] = \Pr[|\mathbf{X}_t| = k] \cdot \Pr\big[|\mathbf{X}_t \cap \mathbf{X}_{\mathsf{big}}| = 1 \;\big|\; |\mathbf{X}_t| = k\big].$$

$|\mathbf{X}_t|$ is distributed according to $\mathsf{Bin}(N, k/N)$, which approaches $\mathsf{Pois}(k)$ as N grows. Using the fact that $k! \leq (k/e)^k e\sqrt{k}$ we get: $\Pr[|\mathbf{X}_t| = k] \approx (k^k e^{-k})/k! \geq 1/(e\sqrt{k}) = \Omega(1)$. $\Pr[\mathbf{x} \in \mathbf{X}_{\mathsf{big}}] = \frac{1}{k} \pm \mathrm{negl}(N)$. The events $\mathbf{x} \in \mathbf{X}_t$ and $\mathbf{x} \in \mathbf{X}_{\mathsf{big}}$ are independent. Therefore

$$\Pr\big[|\mathbf{X}_t \cap \mathbf{X}_{\mathsf{big}}| = 1 \;\big|\; |\mathbf{X}_t| = k\big] = (1 - 1/k)^{k-1} \pm \mathrm{negl}(N) > 1/e.$$

Combining the above, $\Pr[t\ \mathbf{X}\text{-good}] = \Omega(1)$. Quantitatively, for $k \leq 15$, $\Pr[t\mathbf{X}\text{-good}] \gtrsim 1/(e^2\sqrt{k}) > 1/30$.

Let $g(\mathbf{X})$ be the number of $\mathbf{X}$-good values of t. By the above, $\mathbb{E}[g(\mathbf{X})] = \Omega(N)$. Changing a single record $\mathbf{x}$ can change the value of g by at most 2. Applying McDiarmid's Inequality,

$$\Pr\left[g(\mathbf{X}) < \frac{\mathbb{E}(g(\mathbf{X}))}{2}\right] \leq \exp\left(-\frac{2\left(\frac{\mathbb{E}(g(\mathbf{X}))}{2}\right)^2}{4N}\right) < \mathrm{negl}(N).$$

Thus there are $\Omega(N)$ $\mathbf{X}$-good values of t with high probability.

Most X-good clusters are $\widehat{\mathbf{Y}}$-good Cluster t is $\widehat{\mathbf{Y}}$-good if $\widehat{\mathbf{Y}}_t = k$ and $|B_t - D/2| \leq D/8$. First we show that for all but one $\mathbf{X}$-good t, $|\widehat{\mathbf{Y}}_t| = k$. Let $\mathbf{Y}_t \supseteq \widehat{\mathbf{Y}}_t$ be the records in $\mathbf{Y}$ corresponding to the records in $\mathbf{X}_t$. (Whereas $\mathbf{Y}_t$ consists of all records that correspond to $\mathbf{X}_t$, $\widehat{\mathbf{Y}}_t$ consists of only those records whose membership in $\mathbf{Y}_t$ can be easily inferred from $\mathbf{Y}$.) Observe that $|\mathbf{X}_t| = |\mathbf{Y}_t| \geq |\widehat{\mathbf{Y}}_t|$. By construction, for all $\mathbf{x} \in \mathbf{X}$ there exists t such that $\mathbf{x} \in \mathbf{X}_t$ with high probability (i.e., $\mathbf{X} = \cup_t \mathbf{X}_t$). By Claim B.1, for all but at most one $\mathbf{X}$-good t and every $\mathbf{y} \in \mathbf{Y}_t$, $\mathbf{y} \subseteq H^D_t$. Thus $|\widehat{\mathbf{Y}}_t| = \mathbf{Y}_t = k$.

Finally we show that for all $\mathbf{X}$-good t as guaranteed by Claim B.1, $B_t \in (3D/8, 5D/8)$ with high probability. $\widehat{\mathbf{Y}}_t$ consists of k copies of the same generalized record $(y_1, \ldots, y_d)$. Since M is hierarchical, $y_d \in \{H_t, H_{t,\mathsf{sml}}, H_{t,\mathsf{big}}\}$. By the $\mathbf{X}$-goodness of t, $\mathbf{X}_t$ contains 1 large-noise record $\mathbf{x}_{\mathsf{big}}$ and $k-1$ small-noise records $\mathbf{x}'$ By correctness of the k-anonymizer M, $x_{\mathsf{big},d} \in H_{t,\mathsf{big}} \implies y_d \supseteq H_{t,\mathsf{big}}$. Minimality implies the converse: $y_d \supseteq H_{t,\mathsf{big}} \implies x_{\mathsf{big},d} \in H_{t,\mathsf{big}}$. With high probability, $x'_d \in H_{t,\mathsf{sml}} \implies y_d \supseteq H_{t,\mathsf{sml}}$. Putting it all together,

$$x_{\mathsf{big},t} \in H_{t,\mathsf{big}} \iff y_d = H_t \iff d \in \mathsf{big}_t.$$

By construction of the data distribution U, $\Pr[d \in \mathsf{big}_t] = 1/2$ independently for each $d \in [D]$. Applying Chernoff again, $\Pr[|B_t - D/2| \geq D/8] < 2e^{-\Omega(D)} < \mathrm{negl}(N)$.

Analyzing $\widehat{\mathbf{Y}}$-good clusters By construction, $\mathsf{big}_t = \{d : \exists \mathbf{x} \in \widehat{\mathbf{X}}_t \cap \mathbf{X}_{\mathsf{big}} \text{ st } x_d \in H_{t,\mathsf{big}}\}$. Because $|\mathsf{big}_t| > 0$, $|\widehat{\mathbf{X}}_t \cap \mathbf{X}_{\mathsf{big}}| > 1$. A simple Chernoff-then-union-bound argument shows that the probability that there exist distinct records $\mathbf{x}, \mathbf{x}' \in \mathbf{X}_{\mathsf{big}}$ such that $|\mathsf{big}_t| = |\{d : x_d \in H_{t,\mathsf{big}} \vee x'_d \in H_{t,\mathsf{big}}\}| < 5D/8$ is negligible. Hence $\widehat{\mathbf{X}}_t \cap \mathbf{X}_{\mathsf{big}}$ is a singleton $\{\mathbf{x}^t\}$ with high probability. $\mathbf{x}^t \subseteq \mathbf{z}^t$ follows immediately from the construction. □

Example B.2. *The following distribution U and hierarchy H suffice for the proof of Theorem 4.2. The distribution U is defined by $T = N/k$ cluster centers $c_1, \ldots, c_T \in \mathbb{R}$ and standard deviations $\sigma_{\mathsf{sml}}, \sigma_{\mathsf{big}} \in \mathbb{R}$. A record $\mathbf{x} \in \mathbb{R}^D$ is sampled as follows. Sample a cluster center $t \leftarrow [T]$ uniformly at random. Sample* $\mathsf{size} \leftarrow \{\mathsf{sml}, \mathsf{big}\}$ *with* $\Pr[\mathsf{size} = \mathsf{big}] = 1/k$. *Sample noise* $\mathbf{e} \leftarrow \mathsf{N}(0, \sigma^2_{\mathsf{size}}\mathbf{I})$. *Output* $\mathbf{x} = \mathbf{c}_t + \mathbf{e}$, *where* $\mathbf{c}_t = (c_t, \ldots, c_t) \in \mathbb{R}^D$.

The hierarchy H consists of intervals $H_t = [c_t - \Delta, c_t + \Delta]$ centered at the cluster centers c_t, for some Δ. The hierarchy further subdivides each H_t into a smaller interval $H_{t,\mathsf{sml}} =$

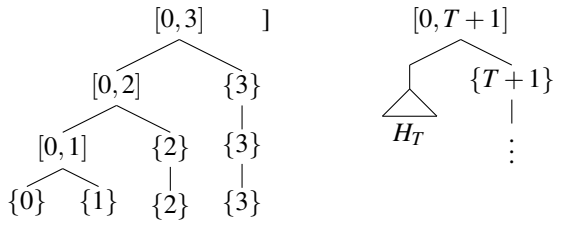

Figure 5: The generalization hierarchy used in the proof of Theorem 4.3. On the left is the hierarchy H_3 for the domain $\mathcal{U} = [03]$. On the right is a recursive construction of H_{T+1} for domain $\mathcal{U} = [0, T+1]$ from the hierarchy H_T.

$[c_t - \tau, c_t + \tau)$*, for some* $\tau < \Delta$*, and the complement* $H_{t,\mathsf{big}} = H_t \setminus H_{t,\mathsf{sml}}$.

To suffice for our proof, we require that with high probability over $\mathbf{x} \sim U$ *there exists* $t \in [T]$ *such that: (a)* $\mathbf{x} \in H_t^D$*; (b) if* $\mathsf{size} = \mathsf{sml}$*, then* $\mathbf{x} \in H_{t,\mathsf{sml}}^D$*; (c) if* $\mathsf{size} = \mathsf{big}$*, then each coordinate* x_d *is in* $H_{t,\mathsf{big}}$ *with probability* $1/2$ *independent of all other coordinates* $x_{d'}$*. Many instantiations of the parameters would work, such as:* $\sigma_{\mathsf{sml}} = 1$*,* $\tau = \log N$*,* $\sigma_{\mathsf{big}} = \zeta \log N$*,* $\Delta = \zeta \log^2 N$*, and* $c_t = 2t\Delta$*, where* $\zeta = \frac{1}{\sqrt{2} \cdot \mathrm{erf}^{-1}(1/2)} \approx 1.48$ *and* erf *is the Gaussian error function.*

B.2 Proof of Theorem 4.3

A k-anonymizer $M : \mathbf{X} \mapsto \mathbf{Y}$ groups records $\mathbf{x} \in \mathbf{X}$ into equivalence classes such that if $\mathbf{x}$ and $\mathbf{x}'$ are in the same class, then $\mathbf{Y}(\mathbf{x}) = \mathbf{Y}(\mathbf{x}')$. In general, M may have a lot of freedom to group the $\mathbf{x}$'s the equivalence classes and also to choose the $\mathbf{y}$'s that generalize each equivalence class.

Claim B.3 states that if M is minimal and generalizes using hierarchy like in Figure 5, then it has much less freedom. Namely, $\mathbf{Y}$ is fully determined by the choice of equivalence classes (with probability at least $1 - \alpha$ over the dataset $\mathbf{X}$). M can group the $\mathbf{x}$'s together, but then has no control over the resulting $\mathbf{y}$'s.

Claim B.3 and its proof are meant to be read in the context of the proof of Theorem 4.3 and freely uses its notation.

Proof of Theorem 4.3. Let $T = \lceil N^2/\alpha \rceil$ and $\mathcal{U} = [0, T]$ be the attribute domain. Records $\mathbf{x} \in \mathcal{U}^D$ are sampled according to the distribution U as follows. First sample $t(\mathbf{x}) \leftarrow [T]$ uniformly at random. Then sample each coordinate x_d of $\mathbf{x}$ i.i.d. with $\Pr[x_d = t(\mathbf{x})] = 1/2k$ and $x_d = 0$ otherwise. In other words, $\mathbf{x} \in \{0, t(\mathbf{x})\}^D$ consists of D independent samples from $t(\mathbf{x}) \cdot \mathsf{Bern}(1/2k)$.

All the $t(\mathbf{x})$ will be distinct except with probability at most $\binom{N}{2}\frac{1}{T} < \alpha/2$. If all $t(\mathbf{x})$ are distinct, we say $\mathbf{X}$ is *collision-free*. The remainder of the proof shows that the adversary succeeds with high probability conditioned on $\mathbf{X}$ collision-free.

Figure 5 defines the generalization hierarchy. It consists of intervals $[0, t]$ and singletons $\{t\}$ for $t \in [T]$.

Claim B.3 states that the output $\mathbf{Y} \leftarrow M(\mathbf{X}, H)$ of a minimal hierarchical k-anonymizer must take a restricted form. For $\mathbf{y} \in \mathbf{Y}$, let $\mathbf{X}_\mathbf{y} = \{\mathbf{x} \in \mathbf{X} : \mathbf{Y}(\mathbf{x}) = \mathbf{y}\}$ be the records in $\mathbf{X}$ that correspond to a copy of $\mathbf{y} \in \mathbf{Y}$. The claim states that

$$\Pr\left[\max_{\mathbf{y}} |\mathbf{X}_\mathbf{y}| < 2k \;\middle|\; \mathbf{X} \text{ collision-free}\right] > 1 - \mathrm{negl}(N).$$

Moreover, if $\mathbf{X}$ is collision-free then for all $\mathbf{y} \in \mathbf{Y}$ and $d \in [D]$:

$$y_d = [0, \max_{\mathbf{x} \in \mathbf{X}_\mathbf{y}} x_d]. \tag{2}$$

Let A be deterministic adversary that on input $\mathbf{Y}$ does the following. For t, pick $\mathbf{y}^t = (y_1^t, \ldots, y_D^t) \in \mathbf{Y}$ such that $\exists d \in [D]$, $y_d^t = [0, t]$. Let $\mathbf{y}^t = \perp$ if no such d exists. By the (2), all $\mathbf{y}$ satisfying the above are identical. If $\mathbf{y}^t \neq \perp$, we define the following subsets of $[D]$:

$$\begin{aligned} D^t(\mathbf{Y}) &= \{d : y_d^t = [0, t]\} \\ D^{>t}(\mathbf{Y}) &= \{d : y_d^t = [0, t'] \text{ for } t' > t\} \\ D^{<t}(\mathbf{Y}) &= \{d : y_d^t = [0, t'] \text{ for } t' < t\}. \end{aligned}$$

If $\mathbf{y}^t \neq \perp$, A writes $\mathbf{z}^t = (z_1^t, \ldots, z_D^t)$ to the output $\mathbf{Z}$, where

$$z_d^t = \begin{cases} 0 & d \in D^{<t}(\mathbf{Y}) \\ t & d \in D^t(\mathbf{Y}) \\ [0, t] & d \in D^{>t}(\mathbf{Y}) \end{cases} \tag{3}$$

Let $T_\mathbf{X} = \{t : \mathbf{y}^t \neq \perp\}$. $|\mathbf{Z}| = |T_\mathbf{X}|$, and it is easy to see that $\Pr[|T_\mathbf{X}| = N \mid \mathbf{X} \text{ collision-free}] > 1 - \mathrm{negl}(N)$. Hence if $\mathbf{X}$ is collision-free, then $\mathbf{Z} \prec_N \mathbf{Y}$ by construction. In this case, we assume without loss of generality that the rows in $\mathbf{Z}$ are ordered in a way that $\mathbf{z}_n \subseteq \mathbf{y}_n$. It follows immediately from the construction that $\mathbf{z}_n \subsetneq_D \mathbf{y}_n$.

If $\mathbf{X}$ is collision-free, then for every $t \in T_\mathbf{X}$ there is a unique $\mathbf{x}^t = (x_1^t, \ldots, x_D^t) \in \mathbf{X}$ such that $t(\mathbf{x}^t) = t$. By (2) and the fact that $\mathbf{x}^t \in \{0, t\}^D$, $\mathbf{x}^t \subseteq \mathbf{z}^t$. Hence if $\mathbf{X}$ is collision-free, then $\mathbf{X} \preceq \mathbf{Z}$ with high probability, proving the first part of the theorem. □

The following claims is meant to be read in the context of the proof of Theorem 5.2 and freely uses notation therefrom.

Claim B.3. *For* $\mathbf{y} \in \mathbf{Y}$*, let* $\mathbf{X}_\mathbf{y} = \{\mathbf{x} \in \mathbf{X} : \mathbf{Y}(\mathbf{x}) = \mathbf{y}\}$ *be the records in* $\mathbf{X}$ *that correspond to a copy of* $\mathbf{y} \in \mathbf{Y}$*. If* $\mathbf{X}$ *is collision-free, then for all* $\mathbf{y} \in \mathbf{Y}$ *and* $d \in [D]$*:*

$$y_d = [0, \max_{\mathbf{x} \in \mathbf{X}_\mathbf{y}} x_d].$$

Moreover, $\Pr[\max_\mathbf{y} |\mathbf{X}_\mathbf{y}| < 2k \mid \mathbf{X} \text{ collision-free}] > 1 - \mathrm{negl}(N)$.

Proof. Both parts of the claim rely on the minimality of M.

Recall that $\mathbf{x} \in \{0, t(\mathbf{x})\}^D$. Let $T_d^*(\mathbf{X}_\mathbf{y}) = \{x_d : \mathbf{x} \in \mathbf{X}_\mathbf{y}\} \subseteq \cup_{\mathbf{x} \in \mathbf{X}_\mathbf{y}} \{0, t(\mathbf{x})\}$ be the set of all values in the dth column of $\mathbf{X}_\mathbf{y}$. $\mathbf{X}$ collision-free implies that either $0 \in T_d^*(\mathbf{X}_\mathbf{y})$ or $|T_d^*(\mathbf{X}_\mathbf{y})| \geq 2$ (probably both). Because M is correct and hierarchical, $T_d^*(\mathbf{X}_\mathbf{y}) \subseteq y_d \in H$. Hence, by construction of H,

$y_d = [0, t_d]$ for some $t_d \in [0, T]$. Let $t_d^* = \max_{\mathbf{x} \in \mathbf{X_y}} x_d$. Correctness requires $[0, t_d^*] \subseteq [0, t_d]$. Moreover, replacing $\mathbf{y} = [0, t_d]$ with $[0, t_d^*]$ would yield a k-anonymous, hierarchy-respecting refinement of $\mathbf{Y}$. By minimality of M, $[0, t_d] \subseteq [0, t_d^*]$. Hence, $y_d = [0, t_d^*]$.

It remains to prove the bound on $\max_{\mathbf{y}} |\mathbf{X_y}|$. Let $\mathbf{X}_0$ and $\mathbf{X}_1$ be an arbitrary partition of $\mathbf{X_y}$. For $b \in \{0,1\}$, define $\mathbf{y}'_b \subseteq \mathbf{y}$ as:

$$\mathbf{y}'_b = (y'_{b,1}, \ldots, y'_{b,D}) = ([0, \max_{\mathbf{x} \in \mathbf{X}_b} x_1], \ldots, [0, \max_{\mathbf{x} \in \mathbf{X}_b} x_D])$$

$\Pr[\exists b,\ \mathbf{y}'_b \subsetneq \mathbf{y} \mid \mathbf{X} \text{ collision-free}] > 1 - \text{negl}(N)$. To see why, observe that $\mathbf{y}'_0 = \mathbf{y} = \mathbf{y}'_1$ implies that for every coordinate d, $\max_{\mathbf{x} \in \mathbf{X}_0}(x_d) = \max_{\mathbf{x} \in \mathbf{X_y}}(x_d) = \max_{\mathbf{x} \in \mathbf{X}_1}(x_d)$. If $\mathbf{X}$ is collision-free, this implies that for all d, $\max_{\mathbf{x} \in \mathbf{X_y}}(x_d) = 0$. This occurs with probability $1 - 2^{-D} = 1 - \text{negl}(N)$ (even conditioned on collision-free).

Consider $\mathbf{Y}'$ constructed by replacing every instance of $\mathbf{y}$ in $\mathbf{Y}$ with $\mathbf{y}'_0$ or $\mathbf{y}'_1$, using $|\mathbf{X}_0|$ and $|\mathbf{X}_1|$ copies respectively. By construction, $\mathbf{Y}'$ correctly generalizes $\mathbf{X}$ and respects the hierarchy H. By the preceding argument, $\mathbf{Y}'$ strictly refines $\mathbf{Y}$ with high probability. Thus, by minimality of M, $\mathbf{Y}'$ cannot be k-anonymous. This means that for every partition $\mathbf{X}_0, \mathbf{X}_1$, one of $|\mathbf{X}_b| \leq k-1$. Therefore, $|\mathbf{X_y}| < 2k$.

□

The following claim is used to prove Theorem 5.2. It is meant to be read in the context of Theorem 4.3 and freely uses notation therefrom.

Claim B.4. *Let $k \geq 2$, $D = \omega(\log N)$, U, $\mathbf{X}$, $\mathbf{Y}$, and $D^t(\mathbf{Y})$ as defined in the proof of Theorem 4.3. Let $T' = \{t : \mathbf{z}^t \in \mathbf{Z}\}$.*

$$\Pr\left[\forall t \in T' : |D^t(\mathbf{Y})| \geq \frac{D}{4ke} \;\middle|\; \mathbf{X} \textit{ coll-free}\right] > 1 - \text{negl}(N) \tag{4}$$

Equation (4) *also holds for $k = N$, $D = \omega(N \log N)$.*

Proof of Claim B.4. The proof is an application of Chernoff and union bounds. We rewrite $D^t(\mathbf{Y})$ as $\{d : \exists n,\ \mathbf{Y}_{n,d} = [0,t]\}$.

For $\mathbf{y} \in \mathbf{Y}$, let $\mathbf{X_y}$ contain the records $\mathbf{x}$ that correspond to a copy of $\mathbf{y}$. Consider $\mathbf{x}^* \in \mathbf{X_y}$, and let $t^* = t(\mathbf{x}^*)$. We call d SUPER if $(x_d^* \neq 0)$ and $(x'_d = 0 \text{ for all } \mathbf{x}' \in \mathbf{X_y} \setminus \{\mathbf{x}^*\})$. By Claim B.3, if d is SUPER then $d \in D^t(\mathbf{Y})$. We will lower bound the number of SUPER d.

For an index set $I \subseteq [N]$, let $\mathbf{X}_I = \{\mathbf{x}_n\}_{n \in I}$. By Claim B.3,

$$\Pr[\exists I \text{ st } (|I| < 2k) \wedge (\mathbf{X_y} = \mathbf{X}_I) \mid \mathbf{X} \text{ coll-free}] > 1 - \text{negl}(N).$$

Observe that if $\mathbf{X_y} = \mathbf{X}_I$, then $\mathbf{x}^* \in \mathbf{X}_I$ and $|I| \geq k$ (by k-anonymity).

We call d GOOD with respect to $\mathbf{X}_I$ if there is a unique $\mathbf{x} \in \mathbf{X}_I$ such that $x_d \neq 0$. Let $D_I = \{d \text{ GOOD wrt } \mathbf{X}_I\}$. Observe that if $\mathbf{X_y} = \mathbf{X}_I$ and d is SUPER, then d is GOOD with respect to $\mathbf{X}_I$. Therefore

$$|D^t| \geq \min_{I : k \leq |I| < 2k, \mathbf{x}^* \in I} |D_I|$$

except with at most negligible probability (conditioned on $\mathbf{X}$ collision-free).

For fixed I, $\mathbb{E}[|D_I|] = D(1/2k)(1-1/2k)^{|I|} > D/2ke$ where the last inequality follows from $|I| < 2k$. By a Chernoff bound: $\Pr[|D_I| \leq D/4ke] \leq e^{-D/32ke}$. By the union bound:

$$\Pr\left[\min_{I : k \leq |I| < 2k, \mathbf{x}^* \in I} |D_I| \leq D/4ke\right] \leq e^{-D/32ke} \sum_{|I|=k}^{2k-2} \binom{N}{|I|}$$

For $k = O(1)$, $\sum_{|I|=k}^{2k-2} \binom{N}{|I|} < N^{2k-2}$. Putting it all together with a final union bound:

$$\Pr\left[\exists t,\ |D^t| \leq D/4ke \;\middle|\; \mathbf{X} \text{ collision-free}\right] < N^{2k-1} e^{-D/32ke}.$$

For $D = \omega(\log N)$, this probability is negligible. For $k = N$ and $D = \omega(N \log N)$, the set $\{I : k \leq |I| < 2k\}$ is a singleton, doing away with the need for a union bound. In this case the upper bound is $Ne^{-D/32Ne} < \text{negl}(N)$. □

B.3 Theorems 5.1 and 5.2

Proof outline. Proofs for both compound PSO attacks follow the same general structure, using the corresponding downcoding attacks in non-black-box ways. The compound PSO adversary A gets as input $\mathbf{Y} \leftarrow M_H(\mathbf{X})$. It emulates the appropriate downcoding adversary, which produces an output $\mathbf{Z}$ such that $\mathbf{X} \preceq \mathbf{Z} \prec \mathbf{Y}$.

$\mathbf{Z}$ contains special generalized records $\mathbf{z}^t$ indexed by some $t \in [T]$ (equations (1) and (3)), and may also contain other records. A outputs $\Psi = \{\psi^t\}$ where $\psi^t : \mathbf{x} \mapsto \mathbb{I}(\mathbf{x} \subseteq \mathbf{z}^t)$.

To complete the proof, one must show that the following hold with probability at least $1 - \alpha(N)$:

- $\psi(\mathbf{X}) = 1/N$
- $(\psi \wedge \psi')(U) = 0$
- $\psi(U) < \text{negl}(N)$
- $|\Psi| \geq L$

The first three are implied by the following:

- $\forall t$, there exists a unique $\mathbf{x}^t \in \mathbf{X}$ such that $\mathbf{x}^t \subseteq \mathbf{z}^t$.
- $\forall t \neq t'$, $\mathbf{z}^t \cap \mathbf{z}^{t'} = \emptyset$.
- $\forall t$, $\mathbf{x} \sim \mathcal{U}^D$, $\Pr_{\mathbf{x}}[\mathbf{x} \subseteq \mathbf{z}^t] < \text{negl}(N)$.

For the downcoding attack from Theorem 4.2, these properties are immediate. For the downcoding attack from Theorem 4.3, the first two are immediate and the third follows from Claim B.4.

The requirements on L and α follow from the parameters of the corresponding downcoding attacks. The downcoding attack for Theorem 4.2 yields $\Omega(N)$ records $\mathbf{z}^t$ with probability $1 - \text{negl}(N)$. The downcoding attack for Theorem 4.3 yields N records $\mathbf{z}^t$ with probability $1 - \alpha(N)$. □