Mário S. Alvim, Natasha Fernandes, Annabelle McIver, Carroll Morgan, and Gabriel H. Nunes

Flexible and scalable privacy assessment for very large datasets, with an application to official governmental microdata

Abstract: We present a systematic refactoring of the conventional treatment of privacy analyses, basing it on mathematical concepts from the framework of *Quantitative Information Flow* (*QIF*). The approach we suggest brings three principal advantages: it is *flexible*, allowing for precise quantification and comparison of privacy risks for attacks both known and novel; it can be *computationally tractable* for very large, longitudinal datasets; and its results are *explainable* both to politicians and to the general public. We apply our approach to a very large case study: the Educational Censuses of Brazil, curated by the governmental agency INEP, which comprise over 90 attributes of approximately 50 million individuals released longitudinally every year since 2007. These datasets have only very recently (2018–2021) attracted legislation to regulate their privacy — while at the same time continuing to maintain the openness that had been sought in Brazilian society. INEP's reaction to that legislation was the genesis of our project with them. In our conclusions here we share the scientific, technical, and communication lessons we learned in the process.

Keywords: privacy, formal methods, quantitative information flow, very large datasets, longitudinal datasets

DOI Editor to enter DOI
Received ..; revised ..; accepted ...

Mário S. Alvim: UFMG, Brazil, e-mail: msalvim@dcc.ufmg.br
Natasha Fernandes: Macquarie University, Australia, e-mail: natasha.fernandes@mq.edu.au
Annabelle McIver: Macquarie University, Australia, e-mail: annabelle.mciver@mq.edu.au
Carroll Morgan: UNSW and Trustworthy Systems, Australia, e-mail: carroll.morgan@unsw.edu.au
Gabriel H. Nunes: UFMG, Brazil, e-mail: ghn@nunesgh.com

1 Introduction

Privacy preservation in the release of governmental data about individuals has led recently to legislation in many contexts. Notable examples include the European *General Data Protection Regulation* (GDPR) [57], the United States' *Confidential Information Protection and Statistical Efficiency Act* (CIPSEA) [29], and the Australian review of its *Privacy Act* [25]. There are, however, three principal problems concerning this kind of legislation.

One problem is that legislation usually addresses *known* privacy issues (since they are what brought the issues to the public eye), but when new ways of violating privacy are found (which can happen overnight), the original legislation must still apply (since changing legislation is difficult and time-consuming). A second problem is that, since such legislation is formulated at the level of governments or higher, the data affected can be huge and longitudinal. And thirdly, the legislation must be couched in terms that politicians and the public understand, even though achieving compliance to it is (eventually) a question of mathematics and computer code. It is crucial, therefore, to have a link between those two worlds, one that identifies meaningful threats while minimizing possible waste of resources on non-threats.

In this paper we consider all three issues, grounding our approach on decision- and information-theoretic principles of *Quantitative Information Flow* (*QIF*) [4, 9, 36, 54]. *QIF* has been successfully applied to a variety of privacy and security analyses, including searchable encryption [31], intersection and linkage attacks against k-anonymity [19], and differential privacy [7]. In the context of the present work, we name the three challenges introduced above *flexibility*, *scalability*, and *explainability*, and now consider each one in turn. We then put our approach to a real-world test: a thorough formal analysis of privacy issues in the official Educational Censuses of Brazil, the very large INEP [1] datasets.

1 The Anísio Teixeira National Institute of Educational Studies and Research: https://www.gov.br/INEP

1.1 The challenge of flexibility

The first challenge is to ensure that all meaningful threats are recognised, whilst minimizing resources wasted on non-threats. And the problem here is that current attack practices are either *ad hoc* or constrained to particular scenarios (as discussed ahead in Sec. 2.2).

The impact of focussing on known scenarios is illustrated by the very comprehensive *ARX* tool [44–47]: it supports the analysis of re-identification risk under El Emam's "prosecutor", "journalist", and "marketer" attack models [17, 18, 44]. *ARX* has been remarkably successful in many applications – including e.g. in the MIRACUM network in Germany with data of about 3 million patients with 70 million facts [48], and in a Norwegian re-identification analysis of medical data with over 5 million records [59]. However, *ARX* could not manage INEP's censuses: the tool is limited to datasets of at most $2^{31}-1$ cells,[2] that is ∼23 million records of 92 attributes each. That is smaller than INEP's dataset even for a single year. Furthermore, *ARX* provided only a fixed selection of privacy degradation measures, all of them related to re-identification and not e.g. supporting direct assessment of attribute-inference risks. But, more importantly, *ARX* was not designed to support the full expressiveness of *QIF* analyses, including reasoning about longitudinal attacks in which the adversary has uncertainty about the linkage of a particular individual's records across datasets, and so it could not be naturally extended to encompass attack models other than those hard-coded in the tool already. Other popular tools face similar issues (as discussed in Sec. 8).

1.2 The challenge of scalability

Our concrete example –and the motivation for this work– was INEP's longitudinal collection of official educational-statistics datasets for the whole of Brazil. Updated yearly since 2007, those contain *microdata* (i.e. for individuals) for (nearly) every student in the country, and at all levels (from elementary to graduate schools). Once processed, the data are released to the Internet where they are freely available. Even just one year's data contain about 90 attributes for approximately 50 million students — around 25% of the entire Brazilian population. This collection of official longitudinal microdata is conspicuously huge (even on the world stage). It is used for governmental planning, especially in the allocation of the budget of the Ministry of Education's National Fund of Educational Resources,[3] and by civil society both in Brazil and abroad in many ways, including in demographic research [6, 13], and policy-making and -monitoring [10, 35, 51, 52].

However, a new privacy law [28] inspired by the European GDPR came into effect in Brazil in 2021, and INEP was suddenly forced to perform a thorough exploration of possible vulnerabilities in their datasets. Although previous analyses had provided anecdotal examples of re-identification risks [49], still there had been no systematic analysis of how widespread these risks actually were in the *full* dataset collection. This demands the consideration of the adversary's confidence in the accuracy of her linkage of records across the datasets in the longitudinal collection, which directly affects also the accuracy of her inferences and, consequently, leakage. This task is relatively easy if individuals' identifiers are persistent across datasets, but becomes significantly more challenging otherwise. Originally, INEP intended to consider the latter case. Thus the challenge of scale was to analyse *all* the data, including its longitudinal aspects. That is why we were contacted by INEP.

1.3 The challenge of explainability

The third challenge is that the university scientists who discover a vulnerability [4] in *mathematical* terms must be able to explain the threat it actually poses to those affected, and to do that in *everyday* terms they understand. *"There is a potential decrease of conditional Shannon Entropy"*, for example, may not convince government ministers that "something must be done" — but *"This inference attack might cost the data curator $N"* could concentrate their minds wonderfully.

Our *QIF* approach has two conspicuous features addressing that kind of explainability. First (Sec. 3), *QIF* analyses relate directly to specific adversarial attacks: what is observed, what it might cost (the adversary) to attempt those observations, and what she might gain if she succeeds. If, e.g., government data scientists are concerned about re-identification, in *QIF* terms that can

2 Noting this limitation, our team contacted *ARX*'s curators and discussed an update to overcome it [24]. The fix has been submitted and is now under evaluation.

3 *Fundeb*: https://www.fnde.gov.br/financiamento/fundeb

4 Note that the term "vulnerability" used throughout refers to the "risk" (to the secret); this is the terminology which has been adopted in the *QIF* literature [4].

be expressed as an adversary whose intent is to identify anyone at all. That is in contrast to threat measures that are based on traditional information theory (e.g. Shannon) and whose definitions were designed for quite different purposes (i.e. efficiency of encodings).

Second (Sec. 6.1), once preliminary results are delivered, the government might then be able to clarify their concerns, to make them more precise based on what they have just learned: having seen the general risk posed by re-identification, they might *then* be able to see that the concern is not just an adversary who could identify "anyone" but, rather, that it is a specific minority group that is now possibly at risk. The modular way in which *QIF* describes threats allows computer programs built on *QIF* principles to be re-run immediately with different parameters, instead of having to be re-coded, re-tested and only then re-deployed: for *QIF*-structured tools a single change in the "intent parameter" might be enough, and the response to the more specific question could be very quickly given. Quick responses to new questions suggested by earlier answers is a key factor in the explainability of anything.

1.4 Overview of the INEP case-study

Here we look briefly at the genesis of our project [5, 30]. More technical detail is given in Sec. 5.1.

In Brazil, the issues of transparency and of privacy in the governmental release of data about individuals are regulated through two complementary laws, further detailed in Appendix A, but whose essence is as follows. On the one hand, a *transparency law* from 2011, known as LAI [27], adopts a philosophy of "transparency by default" and requires that information be publicly available on the Internet: any exceptions must be properly justified. On the other hand, a new *privacy law*, LGPD [28], restricts the release of data on individuals, prescribing sanctions in the case of non-compliance.

In this context, we were contacted by INEP to search for privacy vulnerabilities in their *already published* datasets: a longitudinal collection of ∼50 million records per year, each with ∼90 attributes. Their current measures had focussed only on *de-identification*, a known problem, but the legislation itself did not limit the kind of leaks that might be exploited in the future. And here is where the issues of *flexibility* and *scalability* arise.

For example, it was known from the literature that when non-unique attributes are released unaltered (e.g. date of birth, city of residency, gender), then those attributes can act as *quasi-identifiers* (QIDs), that is, in combination they can effect a de-anonymization [12, 39, 53, 55]. As mentioned, anecdotal evidence of such risks had already been identified in INEP's datasets [49], but they were unquantified and narrow in scope.

More significantly, though, was the possibility of other attacks not considered by INEP even anecdotally, e.g. attribute attacks where knowing an individual's city of residence could be used to infer ethnicity. The legislation was broad enough to target those as well — and of course possibly other attacks that *no-one* had invented yet. INEP was thus forced to be prepared to look for breaches they had not yet considered, and across the longitudinal collection. That is, whatever we provided to INEP had to be flexible enough, and longitudinally scalable, to handle and quantify *future* risks too.

The issue of explainability was also formidable in two ways. First, we had to be able to convince INEP that they were at risk even in cases they thought they were not. That meant putting into everyday terms –and quickly– a quantified risk that *anyone* could understand (and care about): *"Do you know that with 80% probability we can from the existing data identify who your children are, and where they go to school?"* But this had to come from a rigorous mathematical analysis.

The second part of this challenge was that whatever changes INEP was convinced (eventually) to make would likely face strong resistance from the public and lead suddenly to different, new questions — and so, again, properly justifying and communicating any change would have to be done carefully and quickly. As a high-profile example, the US Census Bureau has faced serious resistance from stakeholders when discussing changes on the current balance between transparency and privacy in their data-publishing methods [22, 38].

1.5 Our principal contributions

The main contributions of this paper are the items below, addressing the challenges we have identified:

1. We re-factorize attacks along three orthogonal axes: (i) *the information sought by the adversary* (membership-inference, re-identification, or attribute-inference); (ii) *the adversary's target* (fixed-individuals, or collective targets); and (iii) *the adversary's access to datasets* (single datasets, or longitudinal collections). As well as comprehensively covering the relevant operational scenarios from the literature, this re-factorization identifies some new ones (Sec. 2.2).

2. We use the re-factorization above within a coherent formal framework grounded on *QIF*. We devise a non-traditional instantiation of the role of the adversary's prior knowledge and the channel in the *QIF* model that allows, at the same time, for: (i) a realistic capture of INEP's scenario – in which datasets were already of public knowledge even before any attack was performed; and (ii) tractable computations of analyses (Sec. 3).
3. We illustrate the flexibility and scalability of our approach with extensive experimental evaluations of both re-identification and attribute-inference attacks in INEP's extremely large longitudinal collection of Educational Censuses datasets (Sec. 4 and 5). To the best of our knowledge, these analyses are the largest and most thorough in scope ever performed on publicly available governmental microdata, and they reveal several insights about the privacy issues of such large releases.

Additionally, we provide a free, optimized tool of our attacks and privacy analyses (Sec. 4.3).

Ethics considerations. All results in this paper were obtained in a formal cooperation with INEP, at their request, and fully communicated to them. The agreement permits publication of all vulnerabilities found, including all those identified in this paper. Following Brazil's transparency law, the datasets and all results found are freely available to any citizen.

Plan of the paper. Sec. 2 explains our re-factorization of privacy attack models; Sec. 3 introduces the *QIF* framework and shows how it enables the re-factorization; Sec. 4 describes our case study with INEP's datasets; Sec. 5 explains the vulnerabilities discovered; Sec. 6 presents lessons learned; Sec. 7 considers prospects; and Sec. 8 discusses further related work.

2 Rationalizing the landscape of privacy attack models

There are currently many different approaches to the classification of privacy attacks: according to the adversary's goals (e.g. membership, identity, or attribute disclosures) [20, 23, 46]; according to her target and prior knowledge (e.g. the prosecutor, journalist, and marketer models) [17, 18, 44] etc. They have been adopted in practice by the popular ARX data-anonymization tool [46], among others. However, their motivation comes from a small number of *concrete* scenarios, rather than being organized systematically along independent dimensions and, as a result, the identified attacks might fail to cover the threat landscape. (For example, attribute-inference attacks on longitudinal collections).

And so this section rationalizes existing attack models into a unified classification which not only covers various attack models already known, but identifies some new ones. We begin by visiting the existing models.

2.1 An empirical classification of models

Some works focus on re-identification of individuals in a microdata release [17, 18, 44]. Re-identification attacks are (considered to be) of three types depending on the adversary's prior knowledge and target: [5]

- The *Prosecutor* attack model: the adversary tries to re-identify a specific individual (target) whose data is known to be in the dataset of interest.
- The *Journalist* attack model: the adversary tries to re-identify a specific individual whose data is *not necessarily* known to be in the dataset of interest.
- The *Marketer* attack model: the adversary tries to re-identify as many individuals as possible in the dataset of interest.

Yet there are other works that classify according to the type of information sought [20, 23, 46]:

- The *Membership-inference* model: the goal is only to infer whether individuals' data appear in a dataset.
- The *Re-identification* model: the goal is to link data records to the individuals to whom they refer.
- The *Attribute-inference* model: the goal is to infer the value of a sensitive attribute for individuals, regardless of whether they were re-identified.

Because the scenarios above pertain to 3 main adversarial features – her *prior knowledge*, her *targets*, and the *information* she wishes to obtain – we are now able to suggest a unified classification of attack models.

2.2 An orthogonal classification of models

We re-factorize attacks along three orthogonal axes:

5 These are described in [18] as "risks", but here we use the term "models".

	Single-dataset (S)		Longitudinal (L)	
	Ind. (I)	Col. (C)	Ind. (I)	Col. (C)
Memb. (M)	[IMS]	CMS	IML	CML
Re-id. (R)	[IRS]	[CRS]	IRL	CRL
Attr. (A)	IAS	CAS	**IAL**	**CAL**

Table 1. Re-factorization of attack models and their acronyms.

- **Axis I: The information sought by the adversary.** We consider (M) *membership-inference*, (R) *re-identification*, and (A) *attribute-inference*.
- **Axis II: The adversary's target.** We consider (I) *individual-targets*, where her goal is to obtain sensitive information on a specific individual; and (C) *collective-targets*, where her goal is to obtain sensitive information on as many individuals as possible, no matter who they might be.
- **Axis III: The adversary's access to datasets.** We consider (S) *single-dataset access*, of a single dataset corresponding to a specific point in time; and (L) *longitudinal-dataset access*, where several versions are accessible, each for a different time.

The above axes yield $3\times2\times2 = 12$ possible combinations of attack models, given acronyms in Tbl. 1. Thus the prosecutor model corresponds to IRS, the journalist model to IMS, and the marketer model to CRS (all bracketed in the table). But our re-factorization covers many other relevant scenarios as well, such as e.g. attribute-inference attacks on longitudinal collections (CAL and IAL, underlined in the table).

In the next section we show how the above adversarial features are naturally represented in the threat model provided by the *QIF* framework.

3 Quantitative information flow: what it is, and how it induces rationalization

Sec. 2 just above surveyed traditional threats to privacy, and in particular their extensive nomenclature (prosecutor, journalist, marketer, etc.). *Quantitative Information Flow* (*QIF*) provides a mathematical model in which those varying points of view can almost all be seen as aspects of the same thing, thus streamlining the conceptual approach required: we can therefore focus on the small number of *technical* elements that cause the threats, and treat them in a unified way. *QIF* can streamline the computations as well, as we see below, and that helps with scalability.

The philosophy of *QIF*. The *QIF* framework's focus is to capture the adversary's knowledge, goals, and capabilities, and from that quantify the leakage of information caused by a corresponding optimal inference attack. The framework is grounded on sound information– and decision-theoretic principles enabling the rigorous assessment of how much information leakage a system allows *in principle*, and independently from the adversary's computational power [4]. Hence, *QIF* guarantees hold no matter the particular tactic or algorithm the adversary employs to execute the attack, as what is measured is exactly how much sensitive information is leaked by the best possible such tactic or algorithm.

Overview of privacy models in *QIF*. *QIF* (we will see) separates (1) the adversary's *knowledge* from (2) the description of the "leak" she is trying to exploit, and that leak description is again separated from (3) her *intentions and capabilities*. The first (1) is modeled as a probabilistic "prior"; the second (2) is modeled by Bayesian reasoning to produce a "hyper-distribution"; and the third (3) is modeled as a "gain function" that gives what could almost be regarded as monetary values. We introduce those in turn.

A *prior* (1) is a probability distribution over unknown (but sought after) data, and it models the adversary's knowledge about that data even before any leak occurs: *how likely is it that this person is unmarried? How likely is it that this row of the dataset describes Warren Buffett?* A *gain function* (3) for an adversary gives a numerical (expected) valuation of the benefit to her of learning that information: the gain function of an adversary seeking a partner would be high in the first case, but low in the second; but if she wants to raid a bank account, it would be the opposite. Varying the gain function is how we formalize the attacks from Tbl. 1.

A *hyper-distribution* (2) summarizes mathematically how an adversary uses Bayesian reasoning to exploit an information leak: we write "hyper" for short. The specifics of the information leak are described by a channel: for each possible secret that the adversary wants to learn there is some probability that a particular output (coming from the leak) is observed by the adversary. Combining the channel probabilities with the prior enables posterior reasoning using Bayes' rule so that the adversary is able to revise her knowledge about the potential value of the secret, and better align her intent with what she has just learned. The hyper organizes this reasoning as a marginal probability over observa-

id	*lang.*	*gend.*	*age*
1	English	M	>30
2	Port.	M	≤30
3	German	F	≤30
4	German	M	≤30

(a) Original dataset.

language	*prior*
English	1/4
Port.	1/4
German	1/2

(b) Prior on *language*, the adversary's knowledge *before* the leak.

outers ▶	1/2	1/4	1/4	0
gender, age ▶	30 ∨| M	30 ∧ M	30 ∨| F	30 ∧ F
language ▼				
English	0	1	0	0
Portuguese	1/2	0	0	0
German	1/2	0	1	0

(c) Hyper-distribution over *language* given *gender* and *age*, modeling the adversary's knowledge *after* the leak – the outers constitute the probabilities for each observation and the posterior probability distributions in each column summarize what the adversary has learned about the language.

Table 2. Summary of the *QIF* analysis for leaking information about native language from a table of microdata.

tions and, for each observation, a posterior probability distribution over the secret values.

To be concrete for a moment, we mention that a popular gain function is the "Bayes Vulnerability" which rewards a correct guess of a secret's value with 1 if the guess is correct and 0 otherwise. It (and other functions like it) was just what was needed by INEP to provide the Brazilian government with hard scientific evidence to estimate the vulnerability of re-identification such as *"There is an 80% chance that a randomly selected individual can be re-identified in the currently published microdata."* A further benefit of this approach is that these *QIF*-categorized concepts, which can be distilled and explained in terms that INEP care about, can be computed *at scale* if carefully worked out and optimized.

The components of a *QIF* model, with an example. We now return to the more technical aspects of the *QIF* model and how it relates to datasets, how a *secret* is a value of some type $\mathcal{X}$, and a secret (data) release, which in *QIF* is called a *channel*, is a (probabilistic) function from $\mathcal{X}$ to some set of observations $\mathcal{Y}$, and how an *adversary* is abstracted to a (gain) function that can be applied to the hyper, induced by a channel, to determine the advantage accruing to the adversary from using that channel.

In Tbl. 2a we have a 4-row dataset giving for each individual the native language spoken (English, Portuguese, German), the gender (M, F) and the age ($\leq, >30$). The adversary is trying to guess the native language of the person she is about to meet (but has not yet seen), and she assumes the person selected is equally (i.e. uniformly) likely to be any one of the four in the dataset. We describe her with a gain function yielding \$4 if she guesses right, and \$0 if she guesses wrong. The adversary's prior on language (i.e. her knowledge about the sought secret even before meeting the person) is shown in Tbl. 2b, and clearly she will guess German (the most likely language): an expected gain of \$2.

The full procedure for converting the dataset in Tbl. 2a into a hyper as shown in Tbl. 2c is given in Appendix B and used in detail in Sec. 4.2 with a more realistic example. We continue with the small example here to illustrate the systematization that *QIF* allows.

If now our adversary *sees* the person before guessing, the gender and age are leaked. We illustrate the *QIF* approach by showing that her expected gain increases to \$3. From Fig . 2c she sees a "young" man with probability 1/2 and the posterior probabilities for language become 1/2 for both Portuguese and German: so she will guess one of those. If however she sees an old man, with probability 1/4, she will guess (definitely) English; and if she sees a young woman, she will guess German. (There are no old women in the dataset.) Her expected gain is now $\$4 \times (1/2 \times 1/2 + 1/4 \times 1 + 1/4 \times 1) = \$3$. Therefore, the leak has the effect of increasing our particular adversary's expected gain from \$2 to \$3.

The example illustrates further orthogonal decomposition (beyond Sec. 2.2) that *QIF* enables:

1. The dataset(s) and their structure are separated from the attacks that might be mounted: they are simply "there". The datasets used in a particular longitudinal attack are aggregated by some method: if there is a persistent unique identifier for all individuals across all datasets, the aggregation can be done with a simple left outer join keyed on that attribute. (In Sec. 7 we discuss general alternatives for when such an attribute is not available.)
2. The "selection prior" (on records, often uniform, as above), is separated from the actual prior (induced by the attack, here that the language spoken is twice as likely to be German as either of the other two).
3. The selections of "what attribute(s) are sought" and "what attributes are leaked" are separated from the adversary's other characteristics: they determine only what become the rows and columns of the synthesized channel matrix.
4. The posterior inferences the leaks might enable (revised-belief distributions over the secret) are separated from their worth to the adversary (i.e. are captured independently in the gain function).
5. Indeed the worth to the adversary of the information a leak delivers (gain function) is *completely* in-

dependent of all other factors, in particular of the prior, and of how many datasets were involved.

The flexiblity of the *QIF* framework. *QIF* models are flexible by design: once a data release is modeled as a channel, it is easy to switch between various attack scenarios by changing the probability distribution modeling the adversary's prior knowledge, and the gain function modeling her goals and capabilities. Moreover, even in scenarios where the adversary's prior knowledge, goals, and/or capabilities are not fully known, the framework provides quantified worst-case estimates of damage based on the theory of "channel capacities" [1].

Computing leakage with a *QIF* model. *QIF* is a *way of modeling attacks*, not an *implemented tool* in itself. Existing general-purpose implementations of the *QIF* framework make leakage computation tractable in a range of small to medium-sized scenarios, without the need to write new code for new attack models: it suffices to simply change some parameters.[6] Alternatively, *QIF* concepts could be implemented in existing anonymization tools such as ARX. However, not all *QIF* features are native to these tools, and capturing all attack models allowed by *QIF* in them may become a challenge (e.g. dealing with non-uniform priors on records, or adopting information measures not hard-coded into the tools).[7]

As is typical of information– and decision-theoretic frameworks, scalability to very large scenarios is a challenge in *QIF*. In such cases it may be necessary to write and optimize specialized code, as we had to do for INEP's scenario (see Sec. 4.3 ahead). This is in itself a contribution of this work: to show that *QIF* can, indeed, scale.

4 Application of *QIF* to a large-scale privacy problem: INEP's datasets

In this section we apply the rationalization of privacy analyses in the *QIF* framework, discussed in the previous sections, to the large-scale scenario of INEP's Educational Censuses. We start with the fundamentals of our attack models in *QIF*, and then provide concrete instantiations on a running example. The results obtained by the application of these models to the full extent of INEP's scenarios are reserved to Sec. 5.

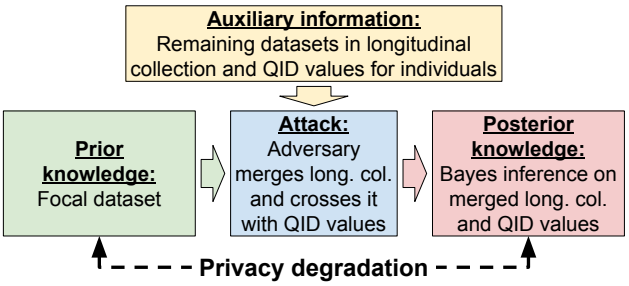


Fig. 1. General schema of an attribute-inference attack on a longitudinal collection, which generalizes all attacks in Tbl. 1.

4.1 Instantiating *QIF* to INEP's scenario

The *QIF* framework can be used to model the attacks from Sec. 3. First, we can unify single-dataset and longitudinal attacks into a single model by aggregating all available datasets along a common axis. We can also unify both re-identification and membership attacks with attribute-inference attacks by considering the sensitive attribute to infer to be, respectively, each individuals' unique identifier or a special attribute indicating the individual's presence/absence in the dataset. Hence all such attacks can be seen as instances of attribute-inference attacks on a longitudinal collection.

Using *QIF* we model an adversary using a prior $\pi{:}\mathbb{D}\mathcal{X}$ over secret values X, representing her prior knowledge. (We use $\mathbb{D}\mathcal{X}$ for the set of distributions over the set $\mathcal{X}$.) We assume there is a channel $C{:}\mathcal{X}{\to}\mathbb{D}\mathcal{Y}$ which leaks information about secrets X via observations Y. We can then represent an adversary's prior and posterior information about the secret (i.e. before and after an observation from C) using vulnerabilityfunctions, which consider the adversary's prior π and gain function g modeling her capabilities and preferences. The overall privacy degradation is then computed by comparing the vulnerability of the secret before and after the attack.

Fig. 1 schematizes our attack models in *QIF*. To accurately capture the scenario of INEP's Educational Censuses from Sec. 1 –thereby constructing an appropriate prior, channel, and vulnerability measure for the *QIF* model–, we formalize assumptions A1–A4 below.

A1: Published census data. A1-A: There is a *longitudinal collection* $\mathcal{L}_{\mathcal{D}}{=}\{D_1, D_2, \ldots, D_I\}$ of I datasets of interest. Each dataset D_i, with $1{\leq}i{\leq}I$, is

6 An example is *LibQIF*: https://github.com/chatziko/libqif/

7 It has been proven that *every* model of inference attack (including worst-case attacks [7]) is captured in the *QIF* framework [2, 3]. The primary limitation of *QIF* is computational tractability rather than generality.

defined over a (finite) attribute set $\mathcal{A}_i$. **A1-B:** There is an *attribute of unique identification* a_{id} common to all datasets in $\mathcal{L}_\mathcal{D}$, and each individual of interest holds a persistent value for this attribute across all datasets.

A2: Adversary's prior knowledge. A2-A: In order to apply Bayesian reasoning we need to attribute a prior over secrets to the adversary. In this situation, the adversary has access to the *focal dataset* $D_1 \in \mathcal{L}_\mathcal{D}$ from which she wishes to re-identify individuals. We can use the distribution of secrets in this dataset as her prior knowledge –her guess– about which secret might belong to any particular individual. We denote by $X \subset \mathcal{A}_1$ the set of secret attributes to infer, and the prior distribution by $\pi : \mathbb{D}\mathcal{X}$. **A2-B:** The adversary assumes that each individual of interest holds exactly one record in D_1, and at most one record in each other dataset in $\mathcal{L}_\mathcal{D}$. [8]

A3: Channel representing the adversary's acquisition of auxiliary information in attack execution. A3-A: The adversary combines the remaining datasets $D_2, D_3, \ldots, D_I$, called *auxiliary datasets*, with the focal dataset D_1 to produce an *aggregated dataset* $\mathcal{D}$ in which the records of individuals across all datasets are linked (see for example Tbl. 3). [9] **A3-B:** In order to find unique mappings between named individuals and other quasi-identifiers (QIDs) in the dataset, we assume that the adversary mines auxiliary information derived from e.g. other public datasets. [10] The set of QIDs is defined $Y \subseteq (\cup_i \mathcal{A}_i) \setminus X$ (so we denote the domain of possible QID values by $\mathcal{Y}$). **A3-C:** The aggregated dataset $\mathcal{D}$ can be rewritten as a channel $C : \mathcal{X} \to \mathbb{D}\mathcal{Y}$ where each entry $C_{x,y}$ is the ratio between the count of individuals with QID values $y \in \mathcal{Y}$ and secret value $x \in \mathcal{X}$, and the total count of individuals with secret value $x \in \mathcal{X}$.

A4: The attack and its privacy degradation. A4-A: The attack consists in the adversary combining her prior knowledge $\pi : \mathbb{D}\mathcal{X}$ with the channel $C : \mathcal{X} \to \mathbb{D}\mathcal{Y}$, and then applying Bayesian inference to produce posterior (conditional) distributions on secret values for each possible observed value of QID (i.e. a hyper giving a probability of inferring $x \in \mathcal{X}$ for each $y \in \mathcal{Y}$, together with the probability of y itself occurring). Combined with the adversary's mined knowledge from A3-B, this posterior knowledge can be used to guess the secret values corresponding to named individuals. **A4-B:** In a *deterministic* attack, the threat is quantified considering the proportion of individuals whose secret values can be inferred with probability 1 using the adversary's knowledge. **A4-C:** In a *probabilistic* attack, the threat can be quantified using the Bayes vulnerability function [4, 54], which gives an optimal adversary's probability of correctly inferring the secret value in one try. **A4-D:** The leakage of information caused by the attack can be quantified using either the ratio or the difference between the adversary's prior and posterior information about the secret (be it probabilistic or deterministic).

4.2 Concrete example: collective-target attribute-inference attack on a longitudinal collection (CAL)

We now illustrate the instantiation of our general *QIF* model to a concrete CAL attack. Other attacks (as in Tbl. 1) can be modeled as special cases; Appendix C exemplifies a CRL attack. We consider the following scenario, under assumptions A1–A4 above.

Example 1 (Running example based on Tbl. 3)**.** *Consider a longitudinal collection of two datasets* $\mathcal{L}_\mathcal{D} = \{D_1, D_2\}$*. The focal dataset,* D_1*, is defined on the set of attributes* $\mathcal{A}_1 = \{id, age, gender, grade, disability\}$ *and is represented in Tbl. 3a. The auxiliary dataset,* D_2*, is defined on the set of attributes* $\mathcal{A}_2 = \{id, age, grade\}$ *and is represented in Tbl. 3b. The adversary merges the datasets in* $\mathcal{L}_\mathcal{D}$*, via a left outer join keyed on the persistent attribute of unique identification id, to produce the aggregated dataset* $\mathcal{D} = D_1 \rtimes D_2$ *in Tbl. 3c.*

Recall that in a *collective-target attribute-inference attack on a longitudinal collection (CAL)*, the adversary's goal is to infer the value of a sensitive attribute for as many individuals as possible in the focal dataset D_1, no matter who they might be. Assume that in our running example the adversary wants to infer the value of the sensitive attribute $X = \{disability\}$.

Attack execution. Before the attack the adversary only has access to the focal dataset D_1, and her prior knowledge about *disability* is determined by this attribute's distribution in this dataset. Since (from Tbl. 3a) *disability* is distributed uniformly (50% "no" and 50% "yes"), the adversary's prior is uniform. Now consider that during the attack the adversary gains access to the auxiliary dataset D_2 and merges it with D_1 to obtain the aggregated dataset $\mathcal{D}$ (as in Tbl. 3c). Fur-

8 The enforcement of assumption A2-B is discussed in Sec. 5.1.
9 In the INEP Censuses analyzed, there exists a persistent unique identifier for every individual across all considered datasets, which makes the aggregation straightforward. In Sec. 7 we discuss how the *QIF* framework can capture more general scenarios.
10 Uniqueness is not necessary for the *QIF* model, but is used here to simplify the presentation of results.

id	*age*	*gend.*	*grd.*	*dis.*
1	25	F	A	no
2	25	F	A	yes
3	25	F	C	yes
4	25	M	B	yes
5	25	M	B	no
6	49	F	C	yes
7	49	F	C	yes
8	49	F	E	no
9	49	M	D	no
10	60	M	D	no

(a) Focal dataset D_1.

id	*age*	*grd.*
1	26	B
2	26	A
3	26	C
4	26	B
5	26	B
6	50	D
7	50	C
8	50	E
9	50	D
11	19	A

(b) Aux. dataset D_2.

(*id*,1)	(*age*,1)	(*gend.*,1)	(*grd.*,1)	(*dis.*,1)	(*age*,2)	(*grd.*,2)
1	25	F	A	no	26	B
2	25	F	A	yes	26	A
3	25	F	C	yes	26	C
4	25	M	B	yes	26	B
5	25	M	B	no	26	B
6	49	F	C	yes	50	D
7	49	F	C	yes	50	C
8	49	F	E	no	50	E
9	49	M	D	no	50	D
10	60	M	D	no	–	–

(c) Aggregated dataset $\mathcal{D}{=}D_1{\bowtie}D_2$, with each attribute tagged with its origin.

Table 3. Example of longitudinal collection of datasets $\mathcal{L}_{\mathcal{D}}{=}\{D_1, D_2\}$ and their aggregation $\mathcal{D}$. Note that the record with *id* 10 is only present in D_1, so attributes (*age*, 2) and (*grade*, 2) have null values in the aggregated dataset $\mathcal{D}$, whereas the record with *id* 11 is only present in D_2 and hence is absent from $\mathcal{D}$.

thermore, we assume that she obtains as auxiliary information (e.g. via other public datasets) the values of the QIDs $Y{=}\{gender, grade\}$ for all individuals in $\mathcal{D}$. Using this auxiliary information, she performs Bayesian reasoning and updates her knowledge about the secret value from the prior to a set of revised conditional distributions (given the learned value of each individual's QIDs) on *disability* s.t. each of these posterior distributions has its own probability of occurring — i.e. she updates her knowledge to a hyper on the secret value.

This whole process is modeled in *QIF* as in Tbl. 4. First the adversary extracts from $\mathcal{D}$ all co-occurrences of values for the secret and for QIDs (Tbl. 4a), and from that she derives a joint probability distribution on these values (Tbl. 4b). By marginalizing the joint distribution, we get the adversary's prior π on the secret value *disability*, and by conditioning the joint distribution on the prior we get the channel representing the adversary's information-gathering process during the attack (Tbl. 4c). The adversary's posterior knowledge is then represented by the hyper in Tbl. 4d. Finally, the overall degradation of privacy can be computed as follows.

Deterministic degradation of privacy. Recall that deterministic success is concerned with the proportion of individuals whose value for the sensitive attribute can be inferred with absolute certainty. In this example, the adversary's deterministic prior success is 0%, since before the attack no individual's *disability* status can be inferred with certainty. After the attack, however, the adversary's knowledge is updated to the hyper in Tbl. 4d. Note that in that hyper the posteriors containing only 1 and 0 values –i.e. all columns but the one labeled as (M,B,B)– have unique QIDs and therefore allow the adversary to infer with probability 1 the *disability* status of the corresponding individuals. The adversary's deterministic posterior success is the fraction of individuals whose attribute is inferred in this way, which is exactly 80%, or 8 out of 10 (note that some posteriors in the hyper represent more than one individual, which is reflected by the posterior's weight). We describe the overall deterministic degradation of privacy additively, as 80%−0% = 80%, meaning that the execution of the attack increases the proportion of individuals with inferrable *disability* status by an absolute value of 80%.

Probabilistic degradation of privacy. Recall that probabilistic success is concerned with the chance that randomly selected individuals can have their sensitive attributes inferred, even if without certainty. In this example, the prior vulnerability of the dataset is 50%, since before the attack the adversary's prior on *disability* is uniform and therefore 50% is the maximum chance with which she can guess the secret value for an individual. After executing the attack and updating her knowledge to the hyper from Tbl. 4d, the adversary's posterior success is measured as the expected value of Bayes vulnerability (which, recall, is the probability of guessing the secret correctly in one try) taken over all posteriors distributions. Indeed, since 7 of the posteriors allow the adversary to guess the secret with probability 1 –and 6 of these posteriors occur themselves with probability $1/10$, whereas 1 occurs with probability $1/5$–, and 1 of the posteriors allows a correct guess with probability $1/2$ –and this posterior occurs itself with probability $1/10$–, the overall posterior Bayes vulnerability is $6{\cdot}1/10{\cdot}1{+}1{\cdot}1/5{\cdot}1{+}1{\cdot}1/10{\cdot}1/2{=}90\%$. We describe the overall probabilistic degradation caused by the attack multiplicatively, as $90\%/50\%{=}1.8$, meaning that the adversary's chance of inferring a randomly selected individual's *disability* status in the focal dataset increases by a factor of 1.8 – so the completion of the CAL attacks almost doubles the adversary's success in inferring the sensitive information.

QIDs ▶	(F,A,B)	(F,A,A)	(F,C,C)	(M,B,B)	(F,C,D)	(F,E,E)	(M,D,D)	(M,D,-)
disability ▼								
yes	0	1	2	1	1	0	0	0
no	1	0	0	1	0	1	1	1

(a) Co-occurrence of values for secret $X=\{(\textit{disability}, 1)\}$ and for observable QIDs $Y=\{(\textit{gender}, 1), (\textit{grade}, 1), (\textit{grade}, 2)\}$, derived from the aggregated dataset $\mathcal{D}$ from Tbl. 3c. E.g. exactly one record has *disability* status "no" and at the same time is a female with grade A in the focal dataset D_1, and grade B in the auxiliary dataset D_2.

QIDs ▶	(F,A,B)	(F,A,A)	(F,C,C)	(M,B,B)	(F,C,D)	(F,E,E)	(M,D,D)	(M,D,-)
disab.▼								
yes	0	1/10	2/10	1/10	1/10	0	0	0
no	1/10	0	0	1/10	0	1/10	1/10	1/10

(b) Joint distribution of values for secret $X=\{(\textit{disability}, 1)\}$ and for observable QIDs $Y=\{(\textit{gender}, 1), (\textit{grade}, 1), (\textit{grade}, 2)\}$, derived from the co-occurrence matrix from Tbl. 4a, and assuming a uniform distribution on the records in $\mathcal{D}$. E.g. there is a probability 1/10 that an individual does not present a disability and has QID vaues (F,A,B).

π
1/2
1/2

QIDs ▶	(F,A,B)	(F,A,A)	(F,C,C)	(M,B,B)	(F,C,D)	(F,E,E)	(M,D,D)	(M,D,-)
disab.▼								
yes	0	1/5	2/5	1/5	1/5	0	0	0
no	1/5	0	0	1/5	0	1/5	1/5	1/5

(c) Prior distribution π on the values for secret $X=(\textit{disability}, 1)$, and the channel for the CAL attack, each derived from the joint distribution from Tbl. 4b by marginalization and conditioning, respectively. E.g. the prior indicates that before the attack (i.e. without learning any QID value) the adversary believes that the probability of any individual having a disability is 1/2. On the other hand, the channel indicates that during the attack the adversary can use the fact that if an individual without a disability is the owner of a record, then the probability that that record has QID values (F,A,B) is 1/5.

outers ▶	1/10	1/10	1/5	1/5	1/10	1/10	1/10	1/10
QIDs ▶	(F,A,B)	(F,A,A)	(F,C,C)	(M,B,B)	(F,C,D)	(F,E,E)	(M,D,D)	(M,D,-)
disab.▼								
yes	0	1	1	1/2	1	0	0	0
no	1	0	0	1/2	0	1	1	1

(d) Hyper-distribution (with column labels added for clarity) representing the adversary's knowledge after completing the CAL attack. The top row ("outers") gives the probability of each possible combination of QID values being revealed, and each column gives the posterior probability distribution on secret values given that the corresponding QID values were revealed. E.g. after the attack, the adversary has a probability 1/10 of learning that an individual's QID values are (F, A, B), and in this case she assigns probability 1 to the corresponding individual having no disability.

Table 4. Step-by-step derivation of prior, channel, and hyper-distribution for CAL attack on the longitudinal collection $\mathcal{L}_{\mathcal{D}}$ from Tbl. 3, considering secret $X = \{\textit{disability}\}$ and observable QIDs $Y = \{\textit{gender}, \textit{grade}\}$.

4.3 Outline of the developed software

As explained in Sec. 1.1, no existing tool met the needs for the scope of our analyses: either they did not support all attack models we consider (especially attribute-inference), did not support longitudinal analysis, or simply could not run analyses on data as large as INEP's. Hence, we implemented and optimized our own tool.

Our software is implemented on Python 3.9.10 using `numpy` 1.22.2 and `pandas` 1.4.1 to streamline some operations. To optimize the use of hardware, we employ the Python `multiprocessing` standard library to simultaneously analyze different sets of QIDs –up to the number of available CPU threads. Instead of relying on the `pandas` built-in functions to partition a dataset based on QIDs, we perform our own sorting of the records according to a given set of QIDs and compute all the values related to that attack on a single pass through the whole dataset. The re-identification and sensitive attribute inference attacks are carried out simultaneously for each selection of QIDs and of sensitive attributes.

Under these optimizations and using 20 threads from two *Intel Xeon E5-2620 v2* processors with 96 GB DDR3-1866 RDIMM, all 2,047 single-dataset attacks performed on the School Census of 2018 were conducted in 40 hours. Due to our choice of only one set of QIDs for the longitudinal attacks, all the 4 analyses were performed in less than one hour. We describe the results of such analyses in the next section.

5 Privacy analyses of the INEP datasets

We now summarize the main results of employing the attack models from Sec. 4 to extensive experimental privacy analyses on INEP's Educational Censuses. These results were critical information for INEP's decision making, and we discuss their implications in Sec. 6.

5.1 Overall synopsis

As mentioned, INEP's Educational Censuses datasets contain microdata for every student at all levels of education in Brazil, including elementary, middle, high, professional, and college education. The datasets have been published yearly since 2007, and the only privacy protection techniques employed are *de-identification* (i.e. the removal of obvious personal identifiers, such as name or governmental-issued ID numbers) and *pseudonymization* (i.e. the substitution of such obvious personal identifiers for artificially-created ones).

Our experimental analyses focused on the *School Census*. These datasets are the largest published by INEP, concerning all students in the country enrolled at all levels of education other than college. Each yearly dataset contains microdata for approximately 50 million students, with about 90 attributes per student. For this

Year	# of records		# of attrib.	Attacks performed
	Original	Treated		
2014	56,064,675	49,491,319	85	CRL /CAL
2015	54,851,222	48,536,347	93	CRL /CAL
2016	52,356,383	48,561,221	92	CRL /CAL
2017	53,900,669	48,377,987	92	CRL /CAL
2018	51,829,413	48,176,423	92	CRS /CAS

Table 5. School Census datasets used in our experiments.

study we selected the 5 most recent datasets at the time of the analyses, as presented in Tbl. 5. We now describe the fundamentals of these analyses.

Treatment of the datasets. Since these datasets may contain duplicated entries for the same student, we treated them to meet the uniqueness Assumption A2-B from Sec. 4. For that, we randomly selected only one record for each student with a same unique pseudonymization code in each dataset [41]. Notice that this treatment can only underestimate privacy risks, so our analyses provide a lower bound on the real risks.

Selection of attributes. For computational tractability, we restricted our experimental analyses to the attributes listed in Tbl. 6, which were selected according to the criteria below. [11]

- **Selection of QIDs.** For both re-identification and attribute-inference attacks on single-datasets (CRS and CAS, respectively), performed on the Census of 2018, we considered all possible 2,047 non-empty combinations of QIDs from a set of 11 attributes which we presume to be easily obtainable by an adversary as auxiliary information. For the longitudinal attacks (CRL and CAL), on the Censuses from 2014 to 2017, we employed a fixed set of 3 QIDs that are expected to vary over the years (since attributes that tend to remain constant tend not to be particularly useful in longitudinal attacks). Tbl. 6a lists the QIDs selected for each attack.
- **Selection of sensitive attributes.** For attribute-inference attacks on both single-datasets and on longitudinal collections (CAS and CAL, respectively), we considered as sensitive: (i) the flag indicating whether the student has any disability, and (ii) the flag indicating whether the student uses public school transport (which may indicate economic status). These attributes are listed in Tbl. 6b.

Attribute	CRS / CAS	CRL / CAL
Day of birth	yes	–
Month of birth	yes	–
Year of birth	yes	–
Gender	yes	–
Ethnicity	yes	–
Nationality	yes	–
Country of birth	yes	–
City of birth	yes	–
City of residency	yes	yes
School id code	yes	yes
School type (public, private, ...)	yes	–
Education level (middle, high, ...)	–	yes

(a) Attributes selected as QIDs in each attack.

Attribute	Domain
Disability status	yes, no
Uses public school transportation	yes, no, n/a

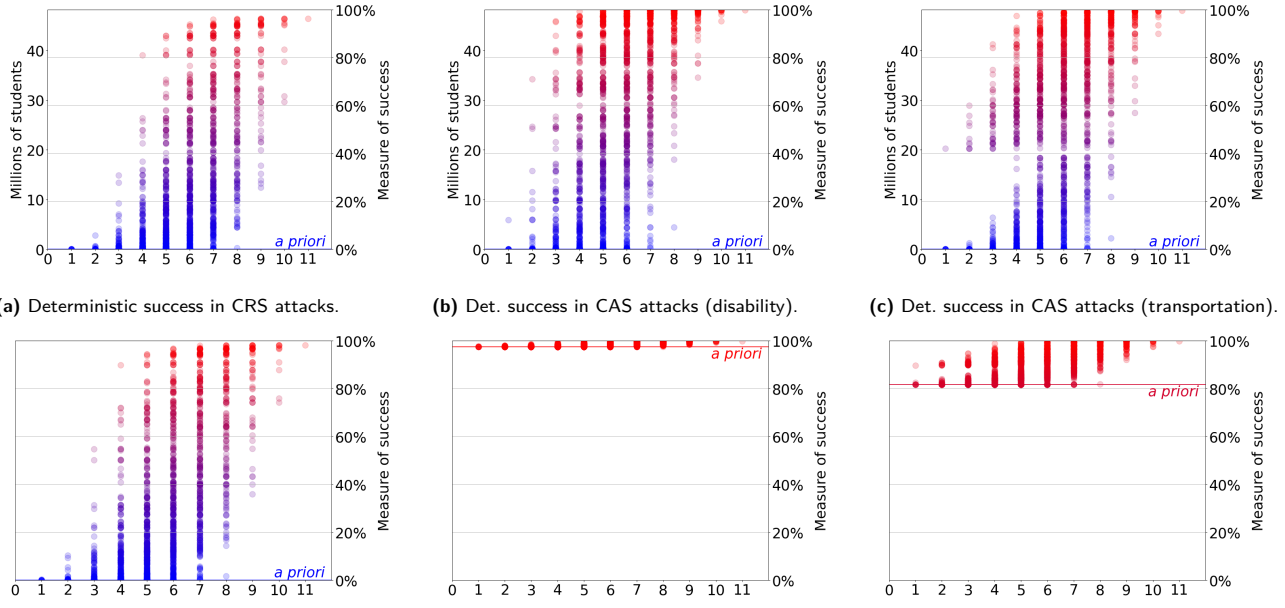
(b) Attributes selected as sensitive in attribute-inference attacks.

Table 6. Attributes selected for the attacks.

Experimental analyses of single-dataset attacks. Collective-target re-identification (CRS) and collective-target attribute-inference (CAS) attacks on a single-dataset were performed on the dataset of the School Census of 2018, described in Tbl. 5. Fig. 2 depicts both the deterministic and the probabilistic degradation of privacy in each of the 2,047 distinct scenarios considered for each attack, every one of them corresponding to an adversary obtaining as auxiliary knowledge a different non-empty subset of the 11 possible QID attributes listed in Tbl. 6a. Additionally, Tbl. 7 provides detailed numbers for some of the 2,047 scenarios from Fig. 2.

Experimental analyses of longitudinal attacks. Collective-target re-identification (CRL) and collective-target attribute-inference (CAL) attacks on longitudinal collections were applied to the collection of School Census datasets from 2014 to 2017, described in Tbl. 5. In all attacks the dataset of 2014 was considered the focal one, and the 2015–2017 datasets were used as auxiliary information. In order to track the evolution of risks as the longitudinal collection grows, in each scenario we assumed that the adversary begins with knowledge of just the focal dataset, and then performs a new attack as each new dataset is released through the years from 2015 to 2017. In each case, the adversary performs

11 Notice that our goal is not to define whether an attribute should be considered as a QID or sensitive. Instead, our results just illustrate privacy risks of possible real-life circumstances, and they can be reproduced for any other choice of attributes.



(a) Deterministic success in CRS attacks. **(b)** Det. success in CAS attacks (disability). **(c)** Det. success in CAS attacks (transportation).

(d) Probabilistic success in CRS attacks. **(e)** Prob. success in CAS attacks (disability). **(f)** Prob. success in CAS attacks (transport.).

Fig. 2. Adversary's success in re-identification (CRS) and attribute-inference (CAS) attacks on the School Census of 2018. In each graph, the horizontal axis indicates the number of QIDs used by the adversary, and the vertical axis indicates the adversary's success. Each dot is the posterior success of a distinct adversary having as auxiliary knowledge one of the 2,047 possible combinations of QIDs. The horizontal "*a priori*" line represents the adversary's success before the attack.

	QIDs: `DoB`, `Gender`, `CoR`	**QIDs:** `DoB`, `Gender`, `CoR`, `SC`
CRS	26.63% (∼12.8 million)	90.43% (∼43.6 million)
CAS (disability)	65.54% (∼31.6 million)	99.67% (∼48.0 million)
CAS (transportation)	54.54% (∼26.3 million)	98.77% (∼47.6 million)

Table 7. Deterministic degradation of privacy in re-identification (CRS) and attribute-inference (CAS) attacks on the School Census of 2018 for some of the 2,047 scenarios from Fig. 2. `DoB` is day, month, and year of birth, `CoR` is city of residency, and `SC` is school code. Percentages are the fraction of students whose sensitive attribute is inferrable with certainty after the attack.

an aggregation of the focal and the auxiliary datasets by linking the unique, permanent pseudonymization code provided by INEP in all datasets considered. Finally, we selected the attributes `City of residency`, `School code`, and `Educational stage` as QIDs for both experiments, as specified in Tbl. 6a. We then measured both the deterministic and the probabilistic degradation of privacy for the set composed of those three attributes, as summarized in Tbl. 8.

5.2 Vulnerabilities identified

Next we highlight key risks uncovered by our analyses. An extensive list of all results is provided in [40, 43].

A vast number of the approximately 50 million students in each of INEP's datasets are at considerable risk even against modest adversaries. As an example, in the School Census of 2018, an adversary starting with prior knowledge of only the released dataset itself would not achieve absolute certainty in any of the re-identification or attribute-inference attacks considered. However, after acquiring as auxiliary information only 3 QIDs –day and month of birth, and school code– the adversary becomes able to re-identify up to 30.92% of the records (∼14.9 million students) and infer the disability status and transportation method of 95.35% (∼45.9 million students) and 85.63% (∼41.3 million students), respectively. By adding year of birth as a fourth QID, those numbers increase to 81.13% (∼39.1 million students), 99.31% (∼47.8 million students), and 97.42% (∼46.9 million students), respectively. As for probabilistic attacks, the adversary's expected prior success in re-identifying any individual is only 0.000002%, but with the use of the same four QIDs as before, the posterior expected success increases to 89.93%. On the other hand, the expected success in inferring a random

Datasets in the long. collection	**CRL** prior success: 0.00% posterior success	**CAL** (disability) prior success: 0.00% posterior success	**CAL** (transportation) prior success: 0.00% posterior success
2014	1.44% (∼0.7 million)	57.17% (∼28.3 million)	58.07% (∼28.7 million)
2014 to 2015	12.88% (∼6.4 million)	79.21% (∼39.2 million)	68.60% (∼34.0 million)
2014 to 2016	25.26% (∼12.5 million)	87.59% (∼43.4 million)	75.32% (∼37.3 million)
2014 to 2017	36.31% (∼18.0 million)	91.28% (∼45.2 million)	79.92% (∼39.6 million)

(a) Deterministic measure of privacy degradation (i.e. proportion of students whose sensitive attribute is inferred with certainty).

Datasets in the long. collection	**CRL** prior success: 0.000002% posterior success	**CAL** (disability) prior success: 98.21% posterior success	**CAL** (transportation) prior success: 82.50% posterior success
2014	4.25%	98.71%	91.64%
2014 to 2015	20.08%	99.03%	93.03%
2014 to 2016	34.37%	99.30%	94.17%
2014 to 2017	45.60%	99.49%	95.07%

(b) Probabilistic measure of privacy degradation (i.e. probability of successful inference of the sensitive attribute in one try).

Table 8. Privacy degradation in re-identification (CRL) and attribute-inference (CAL) attacks on the longitudinal collection containing the School Census datasets from 2014 to 2017. In all attacks the focal dataset is that of 2014 (and all others are used as auxiliary datasets), and the QIDs employed are `City of residency`, `School code`, and `Educational stage`.

individual's disability or transportation method is already very high even a priori, due to the highly skewed distribution of such attributes in the population: 97.56% and 81.75%, respectively. By using again the same four QIDs, the adversary's posterior expected success increases to 99.69% and 98.82%, respectively.

There are plenty of unexpectedly powerful combinations of QIDs. The use of just city of birth and city of residency as QIDs, for instance, allows for the unique re-identification of approximately 430,000 students. The addition of ethnicity to that combination increases that number to around 800,000 uniquely re-identifiable students. Another remarkable example is the use of School Code alone (which is a unique identifier for each school in the country) as a QID in the School Census of 2018. This attribute alone allows the adversary to re-identify with absolute certainty 99 students. [12] Perhaps even more impressively, the use of the same School Code on its own allows for the inference, with absolute certainty, of the disability status of 5.9 million students in the same dataset.

Even modest longitudinal attacks can be highly damaging. As an example, an adversary starting with prior knowledge of only the released School Census of 2014 would not achieve certainty either in re-identifying or in inferring any individual's disability status or transportation method. However, by knowing only some seemingly innocuous QIDs –city of residency, school, and educational stage–, and having access to three auxiliary datasets, from years 2015–2017, the adversary can re-identify with certainty up to 36.31% of the records (∼18.0 million students), and infer the disability status and transportation method of, respectively, 91.28% (∼45.2 million students) and 79.92% (∼39.6 million students). When considering probabilistic measures, again with prior knowledge of only the School Census of 2014, the adversary's probability of re-identifying a randomly selected individual is 0.000002%, whereas the probability of inferring that random individual's disability status or transportation method is, respectively, 98.21% and 82.50%. After acquiring access to the same three QIDs and the same 3 auxil-

12 More precisely, 57 of these 99 students were already uniquely re-identifiable in the original, pre-treated dataset (i.e. before the removal of duplicate entries for each student), and the other 42 students became uniquely re-identifiable only after such treatment. On the other hand, 26 students that were unique based on School Code in the original dataset had those records removed by the treatment. In any case, it is remarkable that there are dozens of such unique students, even in a dataset as large as the one analyzed. Interestingly, our analysis shows that the majority of these unique school codes refer to institutes in rural, indigenous, or *"quilombola"* (i.e. traditional communities formed by descendants of slaves who escaped captivity) areas of the country, or to institutes of specialized education. This suggests that, although such re-identification cases may be relatively rare in the country, they disproportionately affect protected minorities.

iary datasets, the adversary's probability of correctly re-identifying a random individual increases to 45.60%, whereas those of inferring disability status or transportation method increase to 99.49% and 95.07%, respectively.

6 Lessons learned

The rigorous evaluation of privacy in INEP's Educational Censuses raised challenges in various fronts: scientific, technical, and of communication. Here we discuss the main lessons learned while overcoming these challenges.

6.1 Communication and social aspects

The main *communication* challenge in our project was to identify –and even develop– ways to effectively transmit the results of our formal analyses to INEP's agents acting at the technical, managerial, and political levels. For as mathematically sound and experimentally thorough any formal analysis might have been, it could only foster real change if it persuaded INEP's agents that the results applied specifically to their datasets so that they could report these findings to the people empowered to make decisions. Next we summarize some additional lessons not yet covered in this paper.

Results from the academic literature needed to be specifically interpreted and reproduced in the INEP setting. As academics, one of the important inputs to a research effort is to learn from the findings of other researchers working on similar problems. However, in our early interactions with INEP's staff it became clear that, despite the abundant evidence in the literature pointing to the contrary, some influential (although not all) agents remained skeptical that the literature could be relevant to INEP's own dataset. Indeed, these few skeptical agents hewed firmly to their belief that the large number of individuals in each data-release under scrutiny would automatically ensure some reasonable level of privacy — and remained unmoved even when presented with our comprehensive literature review pointing out vulnerabilities in other datasets together with some anecdotal examples of relatives of members of our research team that could be easily re-identified in the Censuses. In fact some of INEP's agents reaffirmed their belief in an intuition of "safety in a crowd," and brushed off our initial findings as a fluke.

As well as convincing INEP's agents of the need for change we also had to anticipate the possible impact of any modification in their current data-release policies on the civil society's perception of the agency's commitment to transparency.

For these reasons we found that it was not enough to expect all of our INEP counterparts to be able to recognize how well-known privacy issues described in the literature could apply to their own datasets, even if, as researchers, we were able to explain the underlying principles. We had to reproduce those attacks and demonstrate specifically the potential for future harm. This turned out to be the only way to convince both the agency (and public) of the relevance of prior research.

As described, the production of this concrete evidence posed a serious challenge; but it turned out to be critical in convincing the agency –and, hopefully, in the near future, also a public accustomed to having access to highly useful data releases– of the necessity of changes in INEP's management of finding the right balance between transparency and privacy.

Irrational adversaries and unrealistic but simple scenarios acted as a stepping stone to explaining how realistic privacy risks in the INEP datasets could potentially harm many citizens. As described in Sec. 5, great effort went into analyzing "deterministic vulnerabilities" in spite of the sometimes misleading impression of security that they imply [2]. Such measures, for example, do not distinguish between an adversary who is 99.99% accurate in her ability to identify individuals and one who is only 0.01% accurate. In a deterministic assessment neither adversary is regarded as a threat because they cannot identify individuals with absolute certainty. However early on in our discussions, the deterministic measures turned out to be the most understandable for our INEP counterparts; but we wanted to alert them to the potential harm that highly accurate, albeit imperfect adversaries could pose to a large number of citizens. As well as computing the level of risk exposure, we also explained the results using scenarios such as the following.

A job agency is considering two candidates for a position. However, the way that INEP currently releases data allows the company to use additional information provided in the candidates' resumés to determine that the first candidate has a 30% chance of having a disability, whereas the second candidate has only a 5% chance. Of course now the agency has a choice whether to discriminate against the first candidate and offer the job to the second — this would be a decision made using a

probabilistic inference but causing definite harm that, if done at scale, could reach a large number of citizens.

When these risks were explained, the INEP team reported that their perception of privacy changed significantly and that there were many more threats than just being able to identify individuals precisely. This also led them to re-evaluate the simple mitigation techniques that they had been considering.

6.2 Technical and scientific aspects

The main *technical* challenge in our endeavor consisted of adapting *QIF* attack models –which are information-theoretic and, hence, provide exact, rather than approximate, results– to enable effective computational analyses at the large scale of INEP's scenario: the datasets to be analyzed covered a period of 13 years, with each year's data containing approximately 50 million records, each with up to around 90 attributes (see Tbl. 5). In particular, for each attack it was necessary to identify instantiations of the adversary's prior knowledge and of vulnerability measures that were not only meaningful and persuasive to the decision makers at INEP, but also computationally tractable. Here we describe the main lessons learned while overcoming these challenges.

Applying academic research to real-world privacy problems requires consolidation and explainability. One of our main scientific contributions outlined in Sec. 2 was to consolidate and systematize the body of knowledge on privacy threats. Unfortunately, as it turns out we were unable to apply those findings directly to the INEP datasets without a laborious consolidation step. This was because our task was to provide a comprehensive assessment of known privacy risks; but whilst the literature provided an important input, each of the documented vulnerabilities were often performed by different teams on different datasets with their own special features and unique experimental set up. Not only were their overarching lessons not accessible to the INEP agents, but it was not clear which experiments were essentially doing the same thing and which were related to a genuinely different adversarial setting.

Our systematic treatment (Sec. 2) together with its implementation in *QIF* terms (Sections 3 and 4) ensured the broad coverage required for us to be confident in the advice we provided to INEP.

Engagement with real-world problems benefits basic scientific research. The main *scientific* challenge consisted of rigorously formalizing both known and novel attacks to obtain a comprehensive evaluation of privacy vulnerabilities in a real-world setting: our coherent *QIF* framework enables rigorous quantification and comparison of privacy risks.

One of the outcomes of our project with INEP was to discover new ways to use our *QIF* framework, and a new understanding of its agility for representing and explaining complex scenarios, *and* that the hyper-distribution approach is not only useful as an abstract concept but leads to a compact representation important for scalability. As a positive side effect, we identified many ways in which our framework can be used to model further, more sophisticated threats; we discuss these prospects in the next section.

7 Conclusions and prospects

In this work we rationalized a myriad of known and novel attack models in the rigorous framework of *Quantitative Information Flow*, showing how it can express concrete attacks and quantify privacy risks at very large scale using as an example the case of INEP's Educational Censuses datasets. To the best of our knowledge, this is the largest privacy analysis ever performed on official governmental microdata –with rich records (over 90 attributes) from around 50 million individuals across many years. Our results were crucial in enabling INEP to reach well-informed decisions on the balance between privacy and transparency, which directly impacts the 25% of the Brazilian population represented in these data. Indeed, the agency is currently considering our suggestions for coping with the problem, including the publication of only aggregated data protected by some form of differential privacy or its variants, and allowing access to microdata only via safe rooms. But, beyond INEP's context, we hope that the lessons learned in our endeavor can help other practitioners to communicate more effectively with decision makers and the public.

We now consider some meaningful extensions of the analyses performed on INEP's scenario.

Scalability in general longitudinal collections. In the analyzed INEP Censuses, there exists a persistent identifier for every individual across all datasets, which renders dataset aggregation straightforward. Without such an identifier, an adversary may rely on QID values and prior knowledge about the population of interest to try to match the same individual's records across different datasets. For instance, she may link a record for a student aged "12" in one year with one aged "13" in the following year, and with the same nationality over both

years. However, the aggregated dataset so obtained will present some inherent uncertainty (e.g. because some people indeed *do* change nationality from one year to the other, which may lead to a wrong record-linkage), and any leakage analysis performed on such a dataset needs to account for that uncertainty. This effect can be naturally accounted for in the *QIF* framework with appropriate models of an adversary's prior knowledge about the population of interest, guaranteeing an accurate overall leakage assessment. Indeed, starting from 2018 INEP has discontinued the use of a unique individual identifier across datasets, and we intend to perform a formal privacy analysis on the agency's new policy.

Robust analysis of privacy risks using capacity. Our analyses considered reasonable –and modest– adversaries, and showed that even those posed significant privacy risks to data owners in INEP's datasets. More precisely, such adversaries had limited prior knowledge, and their intention was mostly to guess the secret value correctly in one try. The *QIF* framework allows for many more adversarial models, and, in a precise mathematical sense, it can cover all "reasonable" adversarial models according to a set of fundamental information-theoretic axioms [2, 3]. Furthermore, we can use a theory of channel *capacity* [1] to estimate the maximum risk INEP's data publishing can cause over *all* these reasonable adversarial models, providing a robust upper bound on the corresponding privacy risks.

Analyses of publications other than unmodified microdata. We performed our attacks on the unmodified microdata released by INEP, and protected only by de-identification and pseudonymization. As already mentioned, the institute is now considering applying mitigation techniques that will change the published data's format (e.g. to sanitized microdata, or to aggregated data protected by some form of differential privacy [16]). Because our *QIF* model is agnostic to the particular form of published data (since the adversary's posterior knowledge is represented by a hyper), it can be easily extended to analyze these scenarios, whereas other tools (such as ARX) cannot. As an example, Appendix D shows how the *QIF* framework can be used to assess privacy under the popular syntactic anonymization techniques of k-anonimity and t-closeness.

Availability. The software developed for this work, together with our privacy analyses, is available at `nunesgh/bvm-lib` [40]. The repository includes a demo with ProPublica's data from the COMPAS tool. [13]

8 Further related work

Dalenius initiated a rigorous approach to statistical disclosure control in 1977 [11]. De-identification was already known to be an insufficient measure [12], and several disclosure control methods have been proposed considering various attack models [15, 16, 32, 33, 50, 53]. Fung et al. [20] and Divanis et al. [23] provide a thorough review of available methods. Notable re-identification attacks include Sweeney's seminal analysis of the US 1990 Census [55] and Narayanan and Shmatikov's attacks on the Netflix dataset [39]. Our work, however, covers significantly larger datasets and is, to the best of our knowledge, the most comprehensive on longitudinal governmental microdata. More specifically, Sweeney's results were limited to estimates of how many people could be re-identified with certainty in the US 1990 Census, using only a few combinations of QIDs as auxiliary knowledge, whilst we considered both probabilistic and deterministic measures of success under thousands of combinations of QIDs, and also included longitudinal collections. On the other hand, Narayanan and Shmatikov's attacks on the Netflix dataset did not involve governmental data being limited to the Netflix Prize dataset and publicly available data from the Internet Movie Database (IMDb), used as auxiliary information. The US Census Bureau has identified that their published tables are vulnerable to database reconstruction attacks [14], but they do not publish their results because of legal reasons [21, 58]. Alternative tools to ARX (discussed in Sec. 1.1) include the open source and continuously supported `sdcMicro` package [37, 56] for the R programming language. This package focuses on measuring the disclosure risk in microdata and provides some well-known anonymization methods, but it suffers from similar lack of flexibility as ARX. *QIF* was pioneered by Clark, Hunt, and Malacaria [9], followed by a growing community (see e.g. [8, 34, 54]), and its principles have been organized in [4].

Acknowledgments.
Mário S. Alvim and Gabriel H. Nunes were supported by CNPq and CAPES. Carroll Morgan was supported by Trustworthy Systems. The authors are grateful to INEP for the partnership and constructive interactions, as well as to the other members of the UFMG PRICE project team: Ramon G. Gonze, Jeroen van de Graaf, Igor W. Lemes, Lucas Lopes, and José C. Oliveira Jr.

13 https://github.com/propublica/compas-analysis

References

[1] Mário S. Alvim, Konstantinos Chatzikokolakis, Annabelle McIver, Carroll Morgan, Catuscia Palamidessi, and Geoffrey Smith. Additive and Multiplicative Notions of Leakage, and Their Capacities. In *IEEE CSF*, pages 308–322. IEEE, 2014.

[2] Mário S. Alvim, Konstantinos Chatzikokolakis, Annabelle McIver, Carroll Morgan, Catuscia Palamidessi, and Geoffrey Smith. Axioms for Information Leakage. In *IEEE CSF*, pages 77–92, 2016.

[3] Mário S Alvim, Konstantinos Chatzikokolakis, Annabelle McIver, Carroll Morgan, Catuscia Palamidessi, and Geoffrey Smith. An axiomatization of information flow measures. *TCS*, 777:32–54, 2019.

[4] Mário S. Alvim, Konstantinos Chatzikokolakis, Annabelle McIver, Carroll Morgan, Catuscia Palamidessi, and Geoffrey Smith. *The Science of Quantitative Information Flow*. Springer, 2020.

[5] Mário S. Alvim, Jeroen van de Graaf, Gabriel H. Nunes, Ramon Gonze, Igor Lemes, and Lucas Lopes. TED 8750 - Privacidade nos Censos Educacionais. https://download.inep.gov.br/microdados/TED_8750-UFMG.pdf, 2021.

[6] Martin Carnoy, Luana Marotta, Paula Louzano, Tatiana Khavenson, Filipe Recch Franca Guimarães, and Fernando Carnauba. Intranational comparative education: What state differences in student achievement can teach us about improving education—the case of Brazil. *Comp. Edu. Review*, 61(4):726–759, 2017.

[7] Konstantinos Chatzikokolakis, Natasha Fernandes, and Catuscia Palamidessi. Comparing systems: Max-case refinement orders and application to differential privacy. In *IEEE CSF*, pages 442–457. IEEE, 2019.

[8] Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Prakash Panangaden. Anonymity protocols as noisy channels. *Inf. and Comp.*, 206(2-4):378–401, 2008.

[9] David Clark, Sebastian Hunt, and Pasquale Malacaria. Quantitative analysis of the leakage of confidential data. *Electron. Notes Theor. Comput. Sci.*, 59(3):238–251, 2001.

[10] Leandro Oliveira Costa and Martin Carnoy. The effectiveness of an early-grade literacy intervention on the cognitive achievement of Brazilian students. *Ed. Eval. Pol. Analysis*, 37(4):567–590, 2015.

[11] Tore Dalenius. Towards a methodology for statistical disclosure control. *statistik Tidskrift*, 15(429-444):2–1, 1977.

[12] Tore Dalenius. Finding a needle in a haystack or identifying anonymous census records. *Jrnl Off. Stats.*, 2(3):329, 1986.

[13] Danilo Leite Dalmon, Izabel Fonseca, Cláudio Pondé Avena, Martin Carnoy, and Tatiana Khavenson. Do students make greater achievement gains in some higher education institutions' programs than others? Insights from Brazil. *Higher Education*, 78(5):887–910, 2019.

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

[15] Cynthia Dwork. A firm foundation for private data analysis. *Communications of the ACM*, 54(1):86–95, 2011.

[16] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Thr of Crypto. Conf.*, pages 265–284. Springer, 2006.

[17] Khaled El Emam. *Guide to the de-identification of personal health information*. CRC Press, 2013.

[18] Khaled El Emam and Luk Arbuckle. *Anonymizing health data: case studies and methods to get you started*. O'Reilly Media, 2013.

[19] Natasha Fernandes, Mark Dras, and Annabelle McIver. Processing text for privacy: an information flow perspective. In *FM*, pages 3–21. Springer, 2018.

[20] Benjamin C. M. Fung, Ke Wang, Ada Wai-Chee Fu, and Philip S. Yu. *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques*. Chapman & Hall/CRC, 1st edition, 2010.

[21] Simson Garfinkel, John M Abowd, and Christian Martindale. Understanding database reconstruction attacks on public data. *Queue*, 16(5):28–53, 2018.

[22] Simson Garfinkel, John M Abowd, and Sarah Powazek. Issues Encountered Deploying Differential Privacy. In *WPES*, WPES'18, pages 133–137. ACM, 2018.

[23] Aris Gkoulalas-Divanis and Grigorios Loukides. *Medical Data Privacy Handbook*. Springer, 1st edition, 2015.

[24] Ramon G. Gonze and Fabian Prasser. https://github.com/arx-deidentifier/arx/pull/299. Private communication.

[25] Government of Australia. Privacy Act 1988. https://www.legislation.gov.au/Details/C2015C00598, 1988.

[26] Government of Brazil. Constitution of the Republic. https://www2.senado.leg.br/bdsf/handle/id/243334, 1988.

[27] Government of Brazil. Law 12 527 of November 18, 2011. http://www.planalto.gov.br/ccivil_03/_Ato2011-2014/2011/Lei/L12527.htm, 2011.

[28] Government of Brazil. Law 13 709 of August 14, 2018. http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/Lei/L13709.htm, 2018.

[29] Government of the United States of America. Confidential information protection and statistical efficiency act (cipsea). https://www.eia.gov/cipsea/cipsea.pdf, 2002.

[30] Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Nota de esclarecimento | Divulgação dos microdados. https://www.gov.br/inep/pt-br/assuntos/noticias/institucional/nota-de-esclarecimento-divulgacao-dos-microdados, 2022.

[31] Mireya Jurado, Catuscia Palamidessi, and Geoffrey Smith. A formal information-theoretic leakage analysis of order-revealing encryption. In *IEEE CSF*, pages 1–16. IEEE, 2021.

[32] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. volume 2, pages 106 – 115, 05 2007.

[33] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. L-Diversity: Privacy beyond k-Anonymity. *ACM TKDD*, 1(1):3–es, 2007.

[34] Pasquale Malacaria. Assessing security threats of looping constructs. In *POPL*, pages 225–235, 2007.

[35] Marcelo Martins, Leonardo Rosa, and Martin Carnoy. The 'quality of quantity': Achievement gains from adding a year to Brazilian primary schooling. *Mimeograph, Stanford University Graduate School of Education*, 2016.

[36] Annabelle McIver, Larissa Meinicke, and Carroll Morgan. Compositional closure for Bayes Risk in probabilistic noninterference. In *ICALP*, volume 6199, pages 223–235, 2010.

[37] Bernhard Meindl, Alexander Kowarik, and Matthias Templ. sdcMicro - Statistical Disclosure Control Methods for Anonymization of Microdata and Risk Estimation. https://sdctools.github.io/sdcMicro/index.html, 2021.

[38] Jeffrey Mervis. Can a set of equations keep U.S. census data private? *Science Insider*, 2019.

[39] Arvind Narayanan and Vitaly Shmatikov. Robust Deanonymization of Large Sparse Datasets. In *Proc. of S&P*, pages 111–125, 2008.

[40] Gabriel Henrique Nunes. BVM library. 10.5281/zenodo.6533704, April 2021.

[41] Gabriel Henrique Nunes. INEP enrollment codes. 10.5281/zenodo.6533675, April 2021.

[42] Gabriel Henrique Nunes. INEP (syntactic) anonymization. 10.5281/zenodo.6533684, April 2022.

[43] Gabriel Henrique Lopes Gomes Alves Nunes. A formal quantitative study of privacy in the publication of official educational censuses in Brazil. Master's thesis, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil, April 2021.

[44] Fabian Prasser, Johanna Eicher, Helmut Spengler, Raffael Bild, and Klaus A. Kuhn. Flexible data anonymization using ARX - Current status and challenges ahead. *Software: Practice and Experience*, 50(7):1277–1304, 2020.

[45] Fabian Prasser and Florian Kohlmayer. Putting statistical disclosure control into practice: The ARX data anonymization tool. In *Medical data privacy handbook*, pages 111–148. Springer, 2015.

[46] Fabian Prasser and Florian Kohlmayer. ARX - Data Anonymization Tool. https://arx.deidentifier.org/, 2021.

[47] Fabian Prasser, Florian Kohlmayer, Ronald Lautenschlaeger, and Klaus A Kuhn. ARX - a comprehensive tool for anonymizing biomedical data. In *AMIA Annual Symp. Proc.*, volume 2014, page 984. AMIA, 2014.

[48] Hans-Ulrich Prokosch, Till Acker, Johannes Bernarding, Harald Binder, Martin Boeker, Melanie Boerries, Philipp Daumke, Thomas Ganslandt, Jürgen Hesser, Gunther Höning, et al. MIRACUM: Medical informatics in research and care in university medicine. *Mth. Inf. Med.*, 57(S 01):e82–e91, 2018.

[49] MJ Queiroz and GHMB Motta. Privacidade e Transparência no Setor público: Um Estudo de Caso da Publicação de Microdados do INEP. In *XV SBSEG*, 2015.

[50] Sofya Raskhodnikova, Adam Smith, Homin K Lee, Kobbi Nissim, and Shiva Prasad Kasiviswanathan. What can we learn privately. In *FOCS*, pages 531–540, 2008.

[51] Leonardo Rosa, Eric Bettinger, Martin Carnoy, and Pedro Dantas. The effects of public high school subsidies on student test scores. 2020.

[52] Leonardo Rosa, Marcelo Martins, and Martin Carnoy. Achievement gains from reconfiguring early schooling: The case of Brazil's primary education reform. *Economics of Education Review*, 68:1–12, 2019.

[53] Pierangela Samarati and Latanya Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. 1998.

[54] Geoffrey Smith. On the Foundations of Quantitative Information Flow. In *FOSSACS*, volume 5504 of *LNCS*, pages 288–302. Springer, 2009.

[55] Latanya Sweeney. Simple demographics often identify people uniquely. *Health (San Francisco)*, 671(2000):1–34, 2000.

[56] Matthias Templ, Alexander Kowarik, and Bernhard Meindl. Statistical disclosure control for micro-data using the R package sdcMicro. *Jrnl. Stat. Software*, 67(4):1–36, 2015.

[57] The European Parliament and the Council of the European Union. Regulation (EU) 2016/679. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679, 2016.

[58] United States Census Bureau. Legacy Techniques and Current Research in Disclosure Avoidance at the U.S. Census Bureau. https://www.census.gov/library/working-papers/2019/adrm/legacy-da-techniques.html, 2019.

[59] Giske Ursin, Sagar Sen, Jean-Marie Mottu, and Mari Nygård. Protecting privacy in large datasets - first we assess the risk; then we fuzzy the data. *Cancer Epidemiology and Prevention Biomarkers*, 26(8):1219–1224, 2017.

A Overview of Brazilian privacy and transparency laws

Article 5 of the Brazilian Constitution from 1988 [26] establishes guiding principles on the right to both privacy (for individuals) and transparency (on matters of public concern). However, the Constitution does not provide guidance on how to balance those two principles; these are detailed in the following pieces of legislation.

The transparency law – this is Law 12 527 of 2011, known as LAI (*Lei de Acesso à Informação*, or *Access to Information Act*). It requires that public authorities guarantee broad access to information, particularly that considered to be of collective or general interest, which must be made available via the Internet (Articles 6 and 8) or otherwise (Article 7) except where such information is considered confidential (Article 22). Regarding the treatment of personal information, Article 31 establishes that it must be done in a transparent manner and with respect to individual freedoms and guarantees; however, provisions on the handling of personal information are open to subsequent regulation.

The privacy law – this is Law 13 709 of 2018, known as LGPD [28] (*Lei Geral de Proteção de Dados Pessoais*, or *General Data Protection Act*), which is based on the European General Data Protection Regulation (GDPR). The law aims to protect the fundamental rights of freedom and privacy (Article 1). It determines that the processing of sensitive personal data is permitted with consent of the data subject or their legal guardian (Articles 7, 11). In its Article 5, LGPD defines *sensitive personal data* as "personal data on racial or

ethnic origin, religious belief, political opinion, union membership or affiliation to organizations of a religious, philosophical, or political nature, data relating to health or sexual life, genetic or biometric data, when linked to a natural person"; and *anonymous data* as "data relating to an unidentifiable holder, considering the use of reasonable technical means available at the time of processing." Article 12 defines that "anonymous data" is not to be considered personal data, except when the anonymization process can be reversed with reasonable efforts. Therefore, objective factors such as the cost and time needed to reverse the anonymization process should be considered given the available technologies and disregarding the use of third party means. But again, the proper definition of what would be considered a reasonable effort, or which anonymization methods should be used, were left to subsequent regulation. Finally, LGPD is to be regulated by the National Data Protection Authority (*Autoridade Nacional de Proteção de Dados*, or ANPD), which is expected to face several challenges in harmonizing LAI with LGPD.

B Full procedure for Ex. in Sec. 3

Here we present the full *QIF* procedure to obtain the hyper-distribution (Tbl. 2c) from the original dataset (Tbl. 2a) in the example from Sec. 3. Recall that when meeting a randomly selected individual, the adversary is able to identify this person's age and gender. She then performs Bayesian reasoning on the collected information and updates her knowledge about the language from the prior to a hyper on the secret value.

This whole process occurs as in Tbl. 9. First the adversary extracts from the original dataset all co-occurrences of values for language, gender, and age (Tbl. 9a), and from that she derives a joint probability distribution on these values (Tbl. 9b). By marginalizing the joint distribution, we get the adversary's prior on language, and by conditioning the joint distribution on the prior we get the channel representing the adversary's information-gathering process during the attack (Tbl. 9c). The adversary's posterior knowledge is then represented by the hyper in Tbl. 9d, which is exactly the same as that in Tbl. 2c.

C Collective-target re-identification attack on a longitudinal collection (CRL)

Here we revisit our claim from Sec. 4.2 that CAL attacks can be seen as generalizations of all others (as in Tbl. 1).

The key idea is that in the *QIF* framework we can consider that the secret (i.e. the values the adversary is trying to make inferences about) consists in the whole collection of real, un-sanitized records for all individuals of interest, some of which may be treated and published, and some of which may not be published at all. In this model, each real record contains the accurate value for all attributes of an individual, including: (i) those which are typically removed in any private microdata release, such as personal identifiers like name or, in our case, a uniquely identifying *id* attribute; (ii) those which are published in a possibly sanitized form, such as QIDs or sensitive attributes; and (iii) a special *membership attribute* indicating whether or not the record in question is published at all in the data release.

In *QIF*, the adversary's goals and capabilities in an attack are modeled as a gain function selecting the part of the secret she is interested in. In an attribute-inference attack, such as that from Sec. 4.2, the gain function represents the adversary's goal of inferring the mapping from individuals' unique *id*s to their sensitive attribute. Note that, in that example, all individuals of interest were known to be represented in the published dataset, but that is not required in general. Indeed, *QIF* allows for the assessment of leakage of information even about individuals *not present* in the data release.

Now, notice that re-identifying individuals is the same as finding a mapping from published records to the real *id*s of their owners. Hence, a re-identification attack is just an instance of attribute-inference attacks in which the attribute to be inferred is the unique *id* of individuals associated with published records. Notice that, although the *id* values to be inferred are not present in the published dataset, they are part of the secret collection of real records, and *QIF* allows us to measure the information leaked about them. Similarly, we can model a membership attack as an attribute-inference attack in which the attribute to be inferred is the attribute in the complete, secret collection of records indicating which individuals are part of the published data.

Now we provide a concrete example of how to model a *collective-target re-identification attack on a longitudinal collection* (CRL) attack as an instance of a CAL

gender, age ▶	30 ∨| M	30 ∧ M	30 ∨| F	30 ∧ F
language ▼				
English	0	1	0	0
Portuguese	1	0	0	0
German	1	0	1	0

(a) Co-occurrence of values for *language*, *gender*, and *age*, derived from the original dataset from Tbl. 2a. E.g. exactly one record (that of *id* 1) represents an English-speaking male over 30.

gender, age ▶	30 ∨| M	30 ∧ M	30 ∨| F	30 ∧ F
language ▼				
English	0	1/4	0	0
Portuguese	1/4	0	0	0
German	1/4	0	1/4	0

(b) Joint distribution for *language*, *gender*, and *age*, derived from the co-occurrence matrix from Tbl. 9a, and assuming a uniform distribution on the records in the original dataset. E.g. the probability that an individual is an English-speaking male over 30 is 1/4.

prior
1/4
1/4
1/2

gender, age ▶	30 ∨| M	30 ∧ M	30 ∨| F	30 ∧ F
language ▼				
English	0	1	0	0
Portuguese	1	0	0	0
German	1/2	0	1/2	0

(c) Adversary's prior knowledge about a randomly selected individual's *language*, and the channel that probabilistically maps *language*, *gender*, and *age*, each derived from the joint distribution from Tbl. 9b by marginalization and conditioning, respectively.

outers ▶	1/2	1/4	1/4	0
gender, age ▶	30 ∨| M	30 ∧ M	30 ∨| F	30 ∧ F
language ▼				
English	0	1	0	0
Portuguese	1/2	0	0	0
German	1/2	0	1	0

(d) Hyper-distribution (with column labels added for clarity) representing the adversary's knowledge about *language* after meeting the person and learning *gender* and *age*. This is identical to the final result of Tbl. 2c.

Table 9. Step-by-step derivation of prior, channel, and hyper-distribution for the native-language example from Tbl. 2.

attack. Recall in such an attack, the adversary's goal is to re-identify as many individuals as possible in the focal dataset D_1, no matter who they might be. Hence we consider a CAL attack in which the attribute to be inferred is the individual's identification itself, so $X = \{id\}$.

Attack execution. In the absence of further prior knowledge, before the attack the adversary considers that all individuals of interest have the same probability of being the owner of any record in the focal dataset D_1, meaning that her prior π on *id* is uniform. Consider again that during the attack the adversary obtains the values of the QIDs $Y=\{gender, grade\}$ for every individual in $\mathcal{D}$ as auxiliary information. She then performs Bayesian reasoning to update her knowledge about the secret value from the prior π to a hyper. This whole process is analogous to that presented for the CAL attack, and is modeled in *QIF* as presented in Tbl. 10. The degradation of privacy can be computed as follows.

Deterministic degradation of privacy. The deterministic prior vulnerability of the dataset is 0%, since before the attack no individual can be re-identified with certainty. After the attack, the adversary's posterior knowledge is given by the hyper of Tbl. 10d. Note that in that hyper posteriors containing only 1 and 0 values – i.e. records with *id*s 1, 2, 6, 8, 9, and 10 – have unique QIDs and can therefore be re-identified with certainty. The adversary's posterior success is the proportion of individuals identified in this way, which is exactly 6 out of 10, i.e. $1/10 \cdot 6 = 60\%$. The overall deterministic degradation of privacy is $60\% - 0\% = 60\%$, meaning that the execution of the attack increases the proportion of re-identifiable individuals by an absolute value of 60%.

Probabilistic degradation of privacy. The prior vulnerability of the dataset is given by its Bayes vulnerability (i.e. the maximum probability of guessing any secret value), which is $1/10 = 10\%$ (given the prior is uniform). After the attack, the adversary's knowledge is given by the hyper in Tbl. 10d. The posterior Bayes vulnerability is the expected maximum probability of success over all posteriors. More precisely, since in 6 of the posteriors the probability of a correct guess is 1 –and each of these posteriors occur themselves with probability 1/10–, and in 2 of the posteriors the probability of success is 1/2 –and each of them occurs with probability 1/5–, the overall posterior Bayes vulnerability is $6 \cdot 1/10 \cdot 1 + 2 \cdot 1/5 \cdot 1/2 = 80\%$. The overall probabilistic degradation of privacy caused by the attack is $80\%/10\% = 8$, meaning that the adversary's chance of re-identifying a randomly selected record in the focal dataset D_1 increases by a factor of eight after the CRL attack.

D Using the *QIF* framework in other large-scale scenarios

We now exemplify how the *QIF* framework, which grounded the INEP privacy analyses, can be generalized to other other large-scale scenarios. More precisely, we show how the *QIF* framework can be used to assess privacy under popular syntactic anonymization techniques.

QIDs ▶ / *id* ▼	(F,A,B)	(F,A,A)	(F,C,C)	(M,B,B)	(F,C,D)	(F,E,E)	(M,D,D)	(M,D,-)
1	1	0	0	0	0	0	0	0
2	0	1	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0
4	0	0	0	1	0	0	0	0
5	0	0	0	1	0	0	0	0
6	0	0	0	0	1	0	0	0
7	0	0	1	0	0	0	0	0
8	0	0	0	0	0	1	0	0
9	0	0	0	0	0	0	1	0
10	0	0	0	0	0	0	0	1

(a) Co-occurrence of values for secret $X=\{(id, 1)\}$ and for observable QIDs $Y=\{(gender, 1), (grade, 1), (grade, 2)\}$, derived from the aggregated dataset $\mathcal{D}$ from Tbl. 3c. E.g. exactly one record has *id* 1 and at the same time is a female with grade A in the focal dataset D_1, and grade B in the auxiliary dataset D_2.

QIDs ▶ / *id* ▼	(F,A,B)	(F,A,A)	(F,C,C)	(M,B,B)	(F,C,D)	(F,E,E)	(M,D,D)	(M,D,-)
1	1/10	0	0	0	0	0	0	0
2	0	1/10	0	0	0	0	0	0
3	0	0	1/10	0	0	0	0	0
4	0	0	0	1/10	0	0	0	0
5	0	0	0	1/10	0	0	0	0
6	0	0	0	0	1/10	0	0	0
7	0	0	1/10	0	0	0	0	0
8	0	0	0	0	0	1/10	0	0
9	0	0	0	0	0	0	1/10	0
10	0	0	0	0	0	0	0	1/10

(b) Joint distribution of values for secret $X=\{(disability, 1)\}$ and for observable QIDs $Y=\{(gender, 1), (grade, 1), (grade, 2)\}$, derived from the co-occurrence matrix from Tbl. 4a, and assuming a uniform distribution on the records in $\mathcal{D}$. E.g. there is a probability 1/10 that an individual has *id* 1 and has QID values (F,A,B).

π
1/10
1/10
1/10
1/10
1/10
1/10
1/10
1/10
1/10
1/10

QIDs ▶ / *id* ▼	(F,A,B)	(F,A,A)	(F,C,C)	(M,B,B)	(F,C,D)	(F,E,E)	(M,D,D)	(M,D,-)
1	1	0	0	0	0	0	0	0
2	0	1	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0
4	0	0	0	1	0	0	0	0
5	0	0	0	1	0	0	0	0
6	0	0	0	0	1	0	0	0
7	0	0	1	0	0	0	0	0
8	0	0	0	0	0	1	0	0
9	0	0	0	0	0	0	1	0
10	0	0	0	0	0	0	0	1

(c) Prior distribution π on the values for secret $X=(id, 1)$, and the channel for the CRL attack, each derived from the joint distribution from Tbl. 4b by marginalization and conditioning, respectively.

outers ▶	1/10	1/10	1/5	1/5	1/10	1/10	1/10	1/10
QIDs ▶ / *id* ▼	(F,A,B)	(F,A,A)	(F,C,C)	(M,B,B)	(F,C,D)	(F,E,E)	(M,D,D)	(M,D,-)
1	1	0	0	0	0	0	0	0
2	0	1	0	0	0	0	0	0
3	0	0	1/2	0	0	0	0	0
4	0	0	0	1/2	0	0	0	0
5	0	0	0	1/2	0	0	0	0
6	0	0	0	0	1	0	0	0
7	0	0	1/2	0	0	0	0	0
8	0	0	0	0	0	1	0	0
9	0	0	0	0	0	0	1	0
10	0	0	0	0	0	0	0	1

(d) Hyper-distribution (with column labels added for clarity) representing the adversary's knowledge after completing the CRL attack.

Table 10. Step-by-step derivation of prior, channel, and hyper-distribution for CRL attack on the longitudinal collection $\mathcal{L}_{\mathcal{D}}$ from Tbl. 3, considering secret $X = \{disability\}$ and observable QIDs $Y = \{gender, grade\}$.

The techniques considered partition the set of records into blocks of records with the same values for QIDs, and then perform *generalization*, *suppression*, or *swapping*. Here we considered as QIDs the 11 attributes in Tbl. 6a, and as sensitive the attribute *disability*. We then employed the ARX tool (extended with our update to treat datasets larger than $2^{31}-1$ cells) to anonymize the School Census of 2018 (see Tbl. 5), using the following techniques [42].[14] First, *k-anonymity* [53], which ensures that each block with the same values for QIDs has at least k records. Second, *t-closeness* [32], which ensures that the distance (according to some suitable metric, e.g. Earth Mover's Distance) between the distribution on the sensitive attribute in each block with the same QID values and the overall distribution on the sensitive attribute is bounded by a threshold t.

On each anonymized dataset we then performed the same privacy analyses of deterministic and probabilistic privacy degradation for both collective-target re-identification (CRS) and collective-target attribute-inference attacks (CAS) on *disability*. Tbl. 11 summarizes our results, and confirms the intuitions that larger values of k and smaller values of t lead to more private data releases (at a usually increasing cost on utility). [15]

14 We also initially considered *ℓ-diversity* [33], which ensures that each block with the same values for QIDs has "well-represented" values for the sensitive attribute according to some suitable metric and threshold ℓ. However, due to the skewness of the distribution, all solutions found by ARX suppressed all QID values, and we discarded them from our experiments.

15 We have used ARX in a standard configuration (which looks for an optimal solution to the k-anonymity or t-closeness problem wrt. the tool's default utility metric) to compute anonymized datasets for varying values of the parameters. Because ARX does not keep the same anonymity groupings across different values of the parameters, the resulting privacy guarantees (in terms of inferences) do not necessarily increase monotonically. For instance, note on Tbl. 11a the CAS values for k=4 and k=12, and on Tbl. 11b the CRS values for k=4 and k=12

Dataset	**CRS**	**CAS** (disability)
	prior success: 0.000000%	**prior success:** 0.000000%
	posterior success	**posterior success**
Original	96.342560% (∼46.4 mi.)	99.890918% (∼48.1 mi.)
k=4	0.000000% (0)	0.048254% (∼23,200)
k=12	0.000000% (0)	0.079773% (∼38,400)
k=20	0.000000% (0)	0.008977% (∼4,300)
t=0.1	0.000000% (0)	0.008977% (∼4,300)
t=0.3	0.000023% (∼11)	0.013258% (∼6,300)
t=0.5	0.000166% (∼79)	0.041118% (∼19,800)

(a) Deterministic measure of privacy degradation (i.e. proportion of students whose sensitive attribute is inferred with certainty).

Dataset	**CRS**	**CAS** (disability)
	prior success: 0.000002%	**prior success:** 97.556444%
	posterior success	**posterior success**
Original	98.138799%	99.946785%
k=4	0.008369%	97.556444%
k=12	0.029857%	97.556444%
k=20	0.002790%	97.556444%
t=0.1	0.002790%	97.556444%
t=0.3	0.002750%	97.556444%
t=0.5	0.006615%	97.556444%

(b) Probabilistic measure of privacy degradation (i.e. probability of successful inference of the sensitive attribute in one try).

Table 11. Comparison of privacy degradation in re-identification (CRS) and attribute-inference (CAS) attacks on the School Census of 2018 before and after sanitization by k-anonymity and by t-closeness. In all of the attacks, the QIDs employed are the 11 listed on Table 6a for CRS / CAS attacks.

and for t=0.1 and t=0.3. This is a strength of our *QIF* analysis, as it uncovers unexpected consequences of anonymization techniques not tailored towards inference attacks.