A formal quantitative study of privacy in the publication of official educational censuses in Brazil

Gabriel H. Nunes[1], **Mário S. Alvim**[1], **Annabelle McIver**[2]

[1]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais
Belo Horizonte, MG, Brazil

[2]Department of Computing – Macquarie University
Sydney, NSW, Australia

Abstract. *We present a summary of the work done in the dissertation "A formal quantitative study of privacy in the publication of official educational censuses in Brazil", including its contributions and impacts so far. The dissertation presents a systematic refactoring of the conventional treatment of privacy analyses, based on mathematical concepts from the framework of* Quantitative Information Flow *(QIF). This brings three principal advantages: flexibility, allowing for precise quantification and comparison of privacy risks for attacks both known and novel; computational tractability for very large, longitudinal datasets; and explainable results both to politicians and to the general public. We apply our approach to a very large case study: the educational censuses in Brazil, which comprise over 90 attributes of approximately 50 million individuals released longitudinally every year since 2007.*

1. Introduction

The dissertation *A formal quantitative study of privacy in the publication of official educational censuses in Brazil* [Nunes 2021], defended and approved in April 28, 2021, provides a thorough quantitative study of privacy risks in the release of the official Brazilian Educational Censuses published annually by INEP. [1] Supported by Brazil's LAI transparency law, INEP has published detailed information on every student in the country since 2007, creating a public longitudinal collection of $\sim$50 million records per year, each with $\sim$90 attributes. But since the enactment of Brazil's LGPD privacy law, the release of data on individuals has been restricted and sanctions may be prescribed in the case of non-compliance. In this context, INEP and the Federal University of Minas Gerais (UFMG) have signed the Decentralized Execution Term 8750, [2] whose goal was to improve INEP's data publication in order to maintain the highest possible utility for data analysts while guaranteeing privacy to individuals represented in the data. The results presented to INEP have already sparked changes in their data release [INEP 2022a, INEP 2022b] and fostered discussions in society on the balance between privacy and transparency.

In this document, we present a summary of the work done in [Nunes 2021], which provides a formal, extensive quantitative evaluation of privacy risks relative to INEP's data publication and evaluates the application of differential privacy, the current state-of-the-art disclosure control method, both based on mathematical concepts from the framework of *Quantitative Information Flow* (*QIF*) [Alvim et al. 2020a]. Additional information on its impacts are given in an accompanying document on sub-products.

[1]Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira.
[2]Termo de Execução Descentralizada 8750 (TED 8750).

1.1. Case study: INEP databases

INEP is responsible for the development and maintenance of educational statistics systems and assessment projects, as well as the dissemination of this information, according to Law 9 448 of 1997 and Decree 6 317 of 2007. As per the Decree, both Basic and Higher Education establishments, whether public or private, are required to provide the information requested by INEP, while also ensuring the confidentiality of personal data collected and prohibiting its use for other purposes. The databases released by INEP constitute our case study and could previously be found on the agency's website. [3] In this dissertation we focus on two of those studies, both released as microdata, i.e. data at the record level.

- **School Census** (*Censo Escolar*): with annual frequency, it is the main educational statistical survey in the country, which covers both Basic and Professional Education. It includes information on students of all ages, usually from 0 to 18 years old, on adults who have previously abandoned formal education and have joined a Youth and Adult Education Program, and on teachers and schools.
- **Higher Education Census** (*Censo da Educação Superior*): with annual frequency, it is the most complete statistical survey on Higher Education Institutions in the country. It includes information on students of undergraduate and graduate levels, on lecturers and professors, and on the institutions.

A preliminary analysis of the databases on students and instructors resulted in the identification of two disclosure control (DC) methods implemented by INEP:

- *de-identification*, by which direct identifiers are removed from the records, e.g. name, government-assigned unique-numbers, and detailed addresses;
- *pseudonymization*, by which INEP assigns to each record a unique, artificially-created identification code, e.g. a non-changing number across releases of databases that allows following an individual through different years.

Even so, several other attributes are available, including date and city of birth, gender, ethnicity, and the unique school or Higher Education Institution code. As discussed in Section 2.2 [Nunes 2021], it is known from the literature that combining such attributes can constitute enough information to uniquely re-identify individuals, rendering those DC methods adopted by INEP insufficient [Dalenius 1986, Samarati and Sweeney 1998, Sweeney 2000, Narayanan and Shmatikov 2008].

This vulnerability in the microdata released by INEP was previously observed by two Brazilian researchers [Queiroz and Motta 2015]. In their work, Queiroz and Motta were able to re-identify one of the authors among 383 683 records of lecturers from the Higher Education Census of 2013 by using only the date of birth, gender, and the Higher Education Institution name. However, they did not provide an analysis of how common this re-identification risk was in that database, i.e. how vulnerable other data holders were.

In the same work, Queiroz and Motta suggested the use of two other DC methods to increase the privacy of individuals, the syntactic methods known as k-anonymity [Samarati and Sweeney 1998] and distinct l-diversity [Machanavajjhala et al. 2007], both discussed in Section 2.2.2 [Nunes 2021]. Those techniques were applied by them with assistance of the ARX Data Anonymization Tool [Prasser et al. 2014] and the results were

[3] https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados

significant at first glance. However, the absolute number of records subject to a *high* risk of re-identification was still considerable even for the most strict parameters values. Moreover, any syntactic method is subject to both linkage [Samarati and Sweeney 1998] and composition [Ganta et al. 2008] attacks, i.e. the use of auxiliary information to degrade privacy, particularly problematic for INEP given the annual frequency of their studies.

Given the recent enactment of Brazil's LGPD privacy legislation, mitigating such vulnerabilities is necessary and urgent for INEP. Previously available work, such as the one by Queiroz and Motta, provides only very limited evidence of the actual privacy risks and how widespread they are. Furthermore, no analysis exists for the longitudinal aspect of INEP's databases and alternatives to syntactic DC methods. Therefore, in this dissertation we also analyze the viability of a semantic method, differential privacy.

Finally, Brazil's LAI transparency legislation, enacted ten years ago, has developed an expectation of access to microdata in the country, both from researchers and policy makers, which increases the difficulty in convincing stakeholders of any changes in the utility that should be provided by the published databases.

1.2. Objectives

Our main goal in this dissertation was to provide a formal, extensive quantitative evaluation of privacy risks relative to INEP's current DC methods, grounded on the solid, formal theoretical framework of *Quantitative Information Flow* (*QIF*) [Alvim et al. 2020a]. Moreover, we evaluate the application of differential privacy, the current state-of-the-art DC method, and analyze the balance between privacy and utility in two of its variants. Particularly, we have the following research questions:

RQ1 What are the data holders' risks in a single database given INEP's DC methods?

- (a) What are the re-identification risks, i.e. complete identity disclosure?
- (b) What are the attribute-inference risks, i.e. inference of an attribute's value?
- (c) Would the removal of attributes from the databases be effective as an additional DC method?

RQ2 What are the data holders' risks when considering the longitudinal aspect of the databases, i.e. the yearly release of new databases, given INEP's DC methods?

- (a) What are the re-identification risks?
- (b) What are the attribute-inference risks?
- (c) Would the removal of attributes from the databases be effective as an additional DC method?

RQ3 What is the effectiveness of differential privacy to mitigate data holders' risks and to maintain data utility?

- (a) What is the effectiveness for the "oblivious" differential privacy model, in which a trustworthy party with access to the original, raw data from respondents is responsible for controlling the queries performed?
- (b) What is the effectiveness for the "local" differential privacy model, in which the data is changed at the record level independently of the existence of a trustworthy party?

1.3. Contributions

We propose a new classification of attacks against releases of databases in Section 3.1 [Nunes 2021] that better covers the space of possible attacks in comparison to the literature. We also develop a database model for single databases in Section 3.2 that we further extend to account for longitudinal databases in Section 5.1, in the educational censuses' context. Those database models and the following attack models were developed on the same framework, which has allowed us to directly compare their respective results.

We formally model collective-target attacks on single databases in Chapter 4 and on longitudinal databases in Chapter 5, which support our conclusions for **RQ1** and **RQ2**, respectively, for both re-identification and attribute-inference privacy risks and considering both deterministic and probabilistic metrics for the adversary success. Moreover, we perform experiments on databases released by INEP and provide results for all of those scenarios, demonstrating the existent vulnerabilities in those databases.

We also formally model global and local differential privacy mechanisms in Chapter 6 to analyze how each of those DC methods affect both privacy and utility, particularly in the context of possibly correlated-databases. Moreover, we implement those models and provide experimental results for those analyses. Those experiments were performed on a sample from one of INEP's databases and demonstrate the impossibility of having optimal noise-adding mechanisms with the best possible privacy or utility in all scenarios. This contribution supports our conclusions for **RQ3**.

Moreover, we provide the largest, most thorough study of actual privacy threats in official government data releases in Brazil. In doing so, our results demonstrate the privacy risks to which more than fifty million current students in the country are subject.

Our work has also contributed to INEP's current efforts to tackle privacy issues given the recent enactment of Brazil's LGPD privacy legislation. Particularly, our work was fundamental to Reports 1 and 2, and contributed to Report 5, of the Decentralized Execution Term 8750 (TED 8750) signed between INEP and UFMG. [4]

Finally, our work presented in Chapter 6 was published in the *31st International Conference on Concurrency Theory* (CONCUR) [Alvim et al. 2020b], where we demonstrated our theoretical results with experiments performed on data from the *Correctional Offender Management Profiling for Alternative Sanctions* (COMPAS) tool.

1.4. Related work

Disclosure control (DC) has been a topic of interest among statisticians since at least the decade of 1970. Back then, there was already the worry about accidental disclosures caused by an ever increasing volume and detail of statistics produced, which has sparked the discussion on disclosure "in the context of a reasoned balance between the right to privacy and the need to know" [Dalenius 1977]. In his 1977 paper, Dalenius argued that the *elimination* of disclosure was unfeasible, since it would impose restrictions on what data could be published to the point of eliminating statistics altogether in the process. Hence the proposition of *controlling* instead of *eliminating* disclosures.

[4]This partnership between INEP and UFMG was initially sought by the former given the new privacy legislation in Brazil and the previous awareness of existent privacy vulnerabilities in the released databases [Queiroz and Motta 2015]. All the TED 8750 Reports were made available by INEP after the publication of the School Census of 2021 [INEP 2022a, INEP 2022b, INEP 2022c].

Data de-identification, i.e. the removal of any direct identifiers of individuals, is a natural first step towards disclosure control, particularly in the case of re-identification. However, as discussed in Section 2.2 [Nunes 2021], Dalenius considered data de-identification to be a necessary but insufficient measure [Dalenius 1986]. In fact, it was demonstrated to be unsuccessful in [Samarati and Sweeney 1998] and exemplified in [Sweeney 2000], where it was shown that 87% of the United States population in the Census of 1990 could be uniquely identified by using only a combination of the ZIP code, gender, and date of birth. As a solution, [Samarati and Sweeney 1998] proposed the k-anonymity DC method, which in turn is known nowadays to be vulnerable to composition attacks, as any syntactic method [Ganta et al. 2008].

Given the known possible vulnerabilities in data releases and the increasingly complex legal landscape on privacy across different jurisdictions, as discussed in Section 1.2 [Nunes 2021], non-experts in need of anonymization and data vulnerabilities analyses tools also increased. In this context, some DC software were proposed, from which we highlight two that are open source and continuously supported: the `sdcMicro` package [Meindl et al. 2021] for the R programming language and the ARX Data Anonymization Tool [Prasser and Kohlmayer 2021].

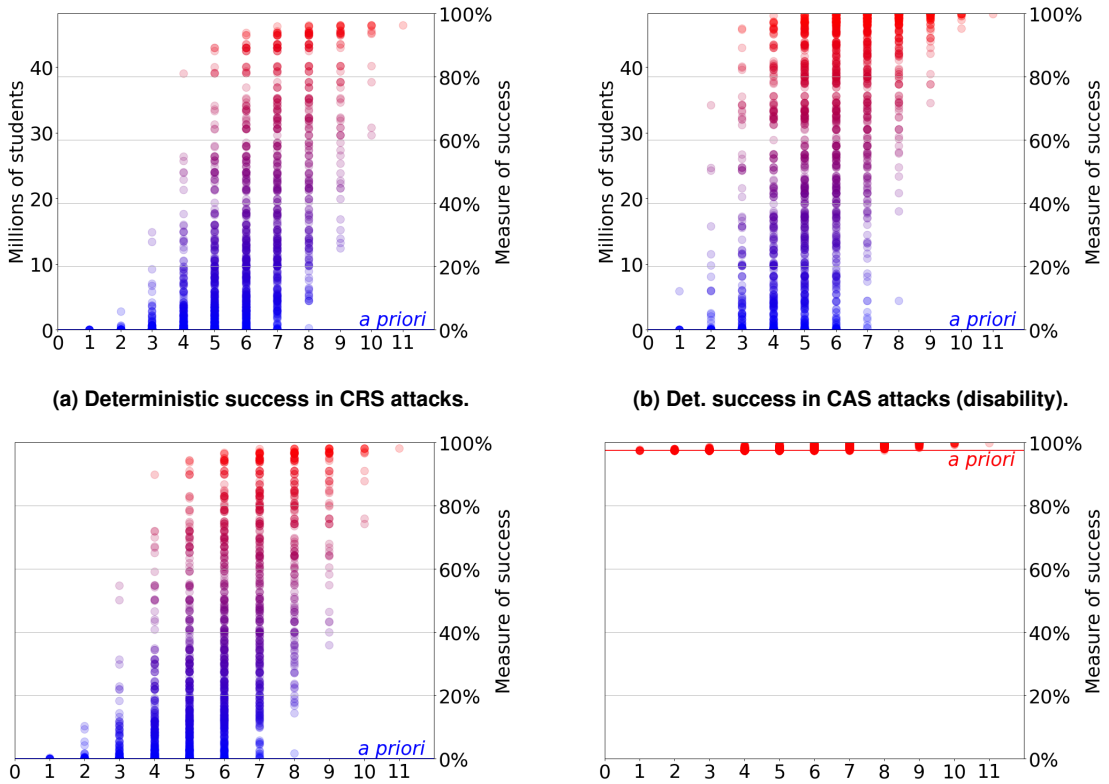
Moreover, a known but less discussed DC problem relates to the continued release of data. [Fung et al. 2010] discuss different approaches to this problem, including *multiple views publishing*, which consists of releasing databases with different sets of attributes from the same collection of data; *sequential releases with new attributes*, which consists of considering possible increments of attributes to the database release; and *incrementally updated data records*, which can be either a *continuous data publishing*, i.e. every subsequent database release contains all the previous releases in addition to the new records, or a *dynamic data republishing*, i.e. any records can be inserted, deleted, or updated as new databases are released. In any case, all the proposed solutions presented are syntactic in nature and hence vulnerable to composition attacks [Ganta et al. 2008].

Finally, regarding our contributions in Chapter 6 [Nunes 2021] where we formalize privacy and utility analyses for two different implementations of differential privacy, we have the *no-free-lunch* theorem from [Kifer and Machanavajjhala 2011]. Through this theorem, the authors showed that it is not possible to provide privacy and utility without considering how the data is generated, i.e. how it is collected and treated prior to release. This result debunks the idea that differential privacy does not consider assumptions about the data in order to guarantee privacy and was the first analysis of differential privacy limitations in databases where there is correlation between secrets.

2. Results

We were able to provide extensive answers to our research questions, as summarized here.

RQ1 The experiments performed on single databases in Section 4.2 [Nunes 2021] have demonstrated that INEP's use of de-identification and pseudonymization as the only DC methods for protecting data holders' privacy is clearly insufficient. For instance, an adversary with knowledge of only three quasi-identifying attributes, i.e. day and month of birth and school code, could re-identify with absolute certainty up to 30.92% of the records on the School Census of 2018, which is equivalent to approximately 14 896 149 students. Furthermore, if the adversary could increase their knowledge with only the attribute for



(a) Deterministic success in CRS attacks.

(b) Det. success in CAS attacks (disability).

(c) Probabilistic success in CRS attacks.

(d) Prob. success in CAS attacks (disability).

Figure 1. 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. The eleven quasi-identifiers used were: `day of birth`, `month of birth`, `year of birth`, `gender`, `ethnicity`, `nationality`, `country code`, `city of birth code`, `city of residency code`, `school code`, and `administrative dependency of the school` (i.e. whether public or private).

year of birth, they would be able to re-identify with absolute certainty up to 81.13% of the records, or approximately 39 085 531 students. Figure 1 summarizes the results for the School Census of 2018, including all the 2 047 possible combinations of the eleven quasi-identifying attributes we have analyzed.

Those results only get worse for the data holders' privacy if we consider the adversary's probabilistic measures of success, which are designed to measure how certain an adversary would be in a given attack. According to our results, that certainty is extremely high for the average data holder in most scenarios. For instance, with knowledge of the same four quasi-identifying attributes as before, i.e. date of birth and school code, the adversary's success in correctly re-identifying a randomly selected record from the School Census of 2018 is of 89.93%. Even worse, an adversary with the goal of inferring the values for the sensitive attribute on students' disabilities on the same database and using the same quasi-identifying attributes would have a staggering 99.69% success.

Additionally, one of the main characteristics of the statistical studies released

by INEP that we have considered is their annual frequency. As discussed in Section 2.4 [Nunes 2021], most DC methods were designed to be applied to single databases open to additional vulnerabilities. Particularly, it is known from the literature that de-identification is vulnerable to both linkage [Samarati and Sweeney 1998] and composition [Ganta et al. 2008] attacks. This leads us to our second research question.

RQ2 We were able to demonstrate with the experiments performed on longitudinal databases in Section 5.2 [Nunes 2021] that even when considering only seemly innocuous information, such as city of residency, school code, and educational stage, an adversary would be able to severely degrade the data holders' privacy. For instance, with knowledge of only those three quasi-identifying attributes and access to three auxiliary databases, up to 36.31% of the students on the School Census of 2014 could be re-identified with absolute certainty, which is equivalent to approximately 17 970 297 students. Figure 2 summarizes the results for the School Censuses from 2014 to 2017.

Again, those results only get worse for the data holders' privacy if we consider the adversary's probabilistic measures of success and, particularly, an adversary with the goal of inferring the values for the sensitive attribute on students' disabilities. For instance, an adversary with knowledge of only those three quasi-identifying attributes, i.e. city of residency, school code, and educational stage, and access to the same three auxiliary databases would have a staggering 99.49% success.

Given the recent enactment of Brazil's LGPD privacy legislation, mitigating such vulnerabilities is necessary and urgent for INEP. Therefore, we have also proposed a third research question considering the knowledge from the literature that syntactic methods are inefficient to mitigate privacy risks to individuals, hence the need to investigate a semantic method such as differential privacy. Particularly, we were interested on how to mitigate data holders' risks while maintaining data utility for analyst.

RQ3 We have developed a model for analyzing the privacy and utility balance in differential privacy implementations for both global and local mechanisms in Chapter 6 [Nunes 2021]. Our results demonstrate that differential privacy is not only highly dependent on the mechanism and parameter for noise addition used but also on correlations existent in the database itself. Furthermore, our results demonstrate some challenges of properly setting a differential privacy implementation with optimal privacy and utility.

3. Conclusion

The work developed in this dissertation has directly contributed to informing INEP on how to properly analyze the vulnerabilities present in their databases released as microdata. Given the recent enactment of Brazil's LGPD privacy legislation, mitigating such vulnerabilities is necessary and urgent for INEP. Particularly, our work was fundamental to Reports 1 and 2, and contributed to Report 5, of the Decentralized Execution Term (TED) 8750 signed between INEP and UFMG. The results presented to INEP have already sparked changes in their data release [INEP 2022a, INEP 2022b] and fostered discussions in society on the balance between privacy and transparency.

Moreover, the work presented in [Nunes 2021] provides the largest, most thorough study of privacy threats in official government data releases in Brazil, comprising more than fifty million individuals, or around 25% of the country's population.

In addition, we have developed a model for analyzing the implementation of both oblivious and local differential privacy mechanisms in terms of privacy loss and utility, particularly in the context of possibly correlated-databases. Differential privacy is the golden standard in the area of DC methods and was adopted by the United States Census Bureau for the 2020 decennial census [United States Census Bureau 2019]. The analyses we have performed are particularly important given that possible correlations between attributes in a database may expose sensitive attributes to unexpected inference attacks.

Furthermore, the dissertation includes a new proposal for the categorization of attacks classification. The literature on DC provides some commonly used privacy and risk models to categorize the possible attacks against databases, as discussed in Section 2.4 [Nunes 2021]. However, the resulting categorization does not account for all possible scenarios and presents some overlapping definitions. Our proposed categorization, detailed in Section 3.1 [Nunes 2021], untangles the possible category dimensions and solves the overlapping problems. Based on this new classification, we have developed four different models of collective-target attacks against databases, which have allowed us to widely explore the vulnerabilities present in the databases released by INEP as microdata. [5]

Finally, an extension of [Nunes 2021], proposed in the project *A robust and explainable framework based on QIF for assessing big data privacy risks*, was awarded the 2021 Google Latin America Research Awards [Google 2022]. Additional information on the impacts of the dissertation, including the proposal for the extension project, are given in an accompanying document on sub-products.

References

Alvim, M. S., Chatzikokolakis, K., McIver, A., Morgan, C., Palamidessi, C., and Smith, G. (2020a). *The Science of Quantitative Information Flow*. Springer. 1, 3

Alvim, M. S., Fernandes, N., McIver, A., and Nunes, G. H. (2020b). On Privacy and Accuracy in Data Releases (Invited Paper). In *31st International Conference on Concurrency Theory, CONCUR 2020, September 1-4, 2020, Vienna, Austria (Virtual Conference)*, volume 171 of *LIPIcs*, pages 1:1–1:18. Schloss Dagstuhl - Leibniz-Zentrum für Informatik. 4

Dalenius, T. (1977). Towards a methodology for statistical disclosure control. *statistik Tidskrift*, 15(429-444):2–1. 4

Dalenius, T. (1986). Finding a needle in a haystack or identifying anonymous census records. *Journal of official statistics*, 2(3):329. 2, 5

Fung, B. C. M., Wang, K., Fu, A. W.-C., and Yu, P. S. (2010). *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques*. Chapman & Hall/CRC, 1st edition. 5

Ganta, S. R., Kasiviswanathan, S. P., and Smith, A. (2008). Composition attacks and auxiliary information in data privacy. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 265–273. 3, 5, 7

[5] We have also developed another four different models of individual-target attacks, which are presented in Appendices C.1 and C.2 [Nunes 2021], for single databases and longitudinal databases, respectively.

Google (2022). Conheça os vencedores da 9ª edição do LARA, o programa de bolsas de pesquisa do Google. https://blog.google/intl/pt-br/novidades/iniciativas/conheca-os-vencedores-do-premio-lara-2021-o-programa-de-bolsas-de-pesquisa-do-google/. 8

INEP (2022a). Inep publica microdados do Enem 2020 e do Censo Escolar da Educação Básica 2021. https://www.gov.br/inep/pt-br/assuntos/noticias/institucional/inep-publica-microdados-do-enem-2020-e-do-censo-escolar-2021. 1, 4, 7

INEP (2022b). Nota de esclarecimento | Divulgação dos microdados. https://www.gov.br/inep/pt-br/assuntos/noticias/institucional/nota-de-esclarecimento-divulgacao-dos-microdados. 1, 4, 7

INEP (2022c). Resultados do Termo de Execução Descentralizada (TED) firmado entre o Inep e a Universidade Federal de Minas Gerais (UFMG). https://download.inep.gov.br/microdados/TED_8750-UFMG.pdf. 4

Kifer, D. and Machanavajjhala, A. (2011). No Free Lunch in Data Privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, SIGMOD '11, pages 193–204. Association for Computing Machinery. 5

Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M. (2007). L-Diversity: Privacy beyond k-Anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1(1):3–es. 2

Meindl, B., Kowarik, A., and Templ, M. (2021). sdcMicro - Statistical Disclosure Control Methods for Anonymization of Microdata and Risk Estimation. https://sdctools.github.io/sdcMicro/index.html. 5

Narayanan, A. and Shmatikov, V. (2008). Robust De-anonymization of Large Sparse Datasets. In *Proc. of S&P*, pages 111–125. 2

Nunes, G. H. L. G. A. (2021). 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. 1, 2, 4, 5, 7, 8

Prasser, F. and Kohlmayer, F. (2021). ARX - Data Anonymization Tool. https://arx.deidentifier.org/. 5

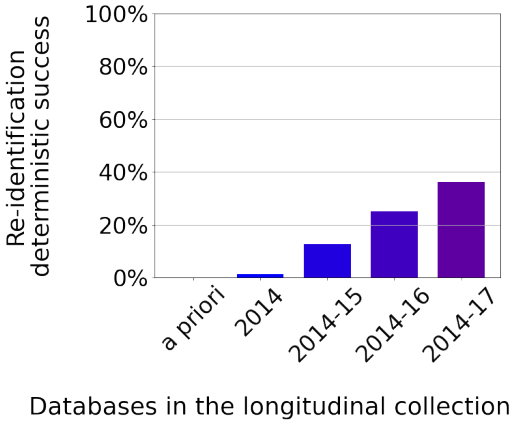
Prasser, F., Kohlmayer, F., Lautenschlaeger, R., and Kuhn, K. A. (2014). ARX - a comprehensive tool for anonymizing biomedical data. In *AMIA Annual Symposium Proceedings*, volume 2014, page 984. American Medical Informatics Association. 2

Queiroz, M. and Motta, G. (2015). Privacidade e Transparência no Setor público: Um Estudo de Caso da Publicação de Microdados do INEP. In *XV Simposio Brasileiro em Seguranca da Informacao e de Sistemas Computacionais-SBSeg*. 2, 4

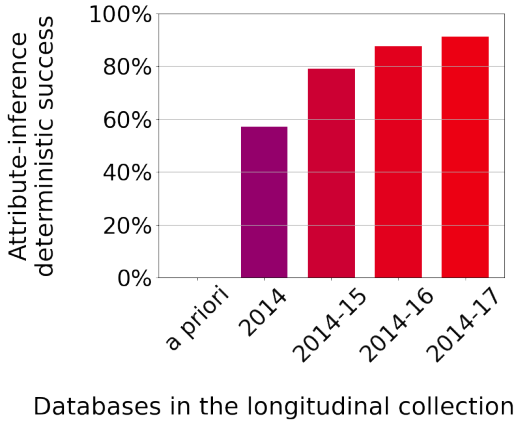
Samarati, P. and Sweeney, L. (1998). Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. 2, 3, 5, 7

Sweeney, L. (2000). Simple demographics often identify people uniquely. *Health (San Francisco)*, 671(2000):1–34. 2, 5

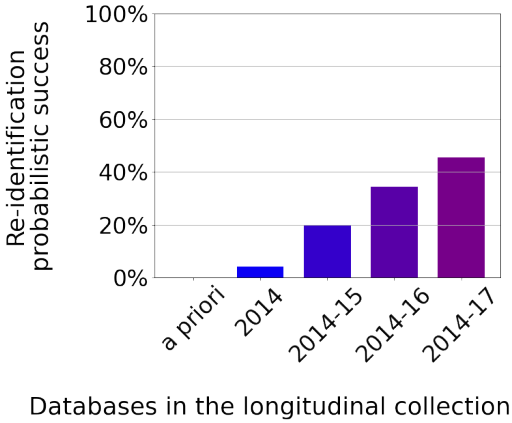
United States Census Bureau (2019). Legacy Techniques and Current Research in Disclosure Avoidance at the U.S. Census Bureau. https://www.census.gov/library/working-papers/2019/adrm/legacy-da-techniques.html. 8



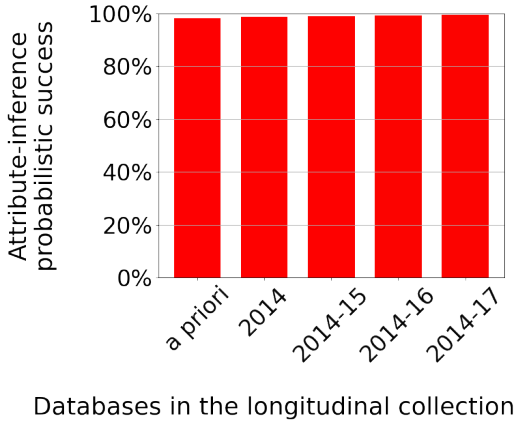
(a) Deterministic success in CRL attacks.



(b) Det. success in CAL attacks (disability).



(c) Probabilistic success in CRL attacks.



(d) Prob. success in CAL attacks (disability).

Figure 2. Adversary's success in re-identification (CRL) and attribute-inference (CAL) attacks on the School Censuses from 2014 to 2017 using the quasi-identifiers `city of residency code`, `school code`, and `educational stage code`. Here, each bar represents a different longitudinal aggregation of databases, always having the School Census of 2014 as the focal one, except for the bar with label "a priori", which indicates the adversary's *a priori* success relying only on the focal database. The height of each bar represents the adversary's deterministic or probabilistic success.