Database de-identification for Electronic Health Records

Leonardo Mazzone

MInf Project (Part 1) Report
Master of Informatics
School of Informatics
University of Edinburgh

2019

Abstract

Researchers and governments today have access to an abundant quantity of health information that is being used to advance the medical profession and provide better healthcare. However, the availability of this data comes into direct conflict with individual privacy, thus requiring de-identification before data release. The best-known non-trivial techniques to address this problem in a generic domain are k-anonymity and differential privacy. The first is straightforward to apply and provides intuitive but formally weak guarantees. The second has strong cryptographic properties, but traditionally, differentially-private methods are based on the perturbation of data and their calibration presents challenges in practice. This project analyzes the applicability of such techniques to the healthcare domain by looking at MIMIC-III, a large medical dataset, implementing an efficient k-anonymization algorithm, and building a collection of alternative de-identification tools inspired from it. In particular, it introduces a very promising practical approach to combining the convenience and interpretability of k-anonymity, and the stronger promises of differential privacy, but without the lack of authenticity traditionally associated to the latter. This paper concludes that the techniques investigated lead to very encouraging results, but that they cannot be used in isolation, and further work is needed to find a good privacy/utility trade-off for very rich collections of healthcare data.

Acknowledgements

I would like to thank Dr. Markulf Kohlweiss for his support throughout the project, his ability to understand what I meant when I did not, and the inspired advice he has given me on many occasions.

Thanks to Dr. Ian Stark and Santiago Guillen for their help and collaboration. Thanks to Dr. Korin Reid and Craneware for their interest in this work, for sharing their insight into the healthcare industry, and for their kindly-sponsored lunches.

Many thanks to Rachel and Greg, whose impressive command of the English language, genuine interest in this work, and true friendship, has led to so many insightful suggestions.

A final thanks to my parents, whose hard work is giving me the opportunity to realize myself today.

Contents

Chapter 1

Introduction

1.1 Aims of this work

Data-mining is a field of great interest in today's world, as it promises to leverage the increasing availability of data to gain substantially deeper insight into phenomena. This could prove inestimable for virtually every industry and field of scientific research, allowing, for instance, to spot patterns, build prediction systems, and optimize processes and resources. This is particularly relevant to the field of health informatics, that aims to apply traditional data-mining and machine learning techniques to the improvement of healthcare. Such endeavors are facilitated both by the advancements of deep learning methods and by the increasing utilization of Electronic Health Record (EHR) systems, initially applied to archive information on patients and perform administrative healthcare tasks, but today a very valuable resource for various clinical informatics applications [44].

However, the same data that powers these promising applications could be abused to learn confidential information, thus constituting a privacy breach, with potential adverse consequences on the subjects involved.

There is a vast literature concerning the definition of privacy with respect to databases and database queries. Some solutions to protect the privacy of individuals participating in a database that have been proposed rely on adding noise (and thus partially falsifying data), or on suppressing data. In general, the trade-off between information loss and privacy risk is an intrinsic feature of this problem. That is to say, the privacy of data inescapably depends on its degradation. The challenge is then to identify the compromise that is the least detrimental to the performance of the task at hand. This means that there is no "one size fits all" solution. For some applications, a poorly selected de-identification transformation will render the data useless, or alternatively still make re-identification straightforward.

This project is aimed at discussing the impact of this trade-off on large healthcare datasets, and attempts to draw some more general conclusions. In doing so it implements a collection of privacy-preserving data transformation schemes, and analyses the

feasibility, convenience and risks for different subjects involved in the exchange of data (data holders, patients, and analysts).

1.2 Summary of contributions

- Literature review of existing approaches to data de-identification and discussion of their suitability to healthcare data-mining applications.
- Identification of privacy vulnerabilities in MIMIC-III, a publicly available dataset of Electronic Health Records.
- Implementation and evaluation of OLA, an efficient k-anonymization algorithm that is optimal with respect to a monotonic information loss metric.
- Conception and development of Inverse OLA, an algorithm that reverses the standard process of k-anonymization, by first looking for potential data releases with a good utility, and then optimizing its de-identification.
- Conception and development of ϵ-safe LA, an algorithm based on OLA that provides strong guarantees, similar to differential privacy but without problematic data perturbation.
- Experimental assessment on the UCI Adult and MIMIC-III datasets of these privacy-preserving techniques for data release.
- Formulation of metric to evaluate the risk of re-identification given stronger assumptions than those entailed by k-anonymity, and experimental evaluation.
- Comparison of information-theoretic data loss measures with respect to the performance of a classification task.

1.3 Report structure

Chapter 2 discusses the legal and technical specificities of the medical data publishing context, reviews the current landscape of health informatics research, introduces Craneware and clarifies their role within this project. It further introduces k-anonymity and differential privacy.

Chapter 3 describes MIMIC-III, a popular and large clinical dataset, and discusses the risk of re-identification of its records, motivating subsequent efforts.

Chapter 4 formalizes the problem of data anonymization, formulates useful assumptions to aid the selection of suitable methods, and explores the implementation and the properties of three anonymization algorithms: *Optimal Lattice Anonymization*, *Inverse OLA* and *ϵ-safe LA*.

Chapter 5 introduces a framework that will be used in the subsequent empirical evaluation to assess the performance of privacy-preserving data-release techniques, and defines a measure of privacy that generalizes k-anonymity.

Chapter 6 motivates and describes a set of experiments on the UCI Adult and MIMIC-III datasets and presents their outcome.

Chapter 7 discusses the results of previous experiments attempting to draw conclusions, provides suggestions for future work and summarizes the accomplishments of the project.

Chapter 2

Background

2.1 Legal framework

Health data is often collected and digitalized without the opportunity for subjects to opt-out. This information might be required in order to provide treatment, and might be transferred across health providers as individuals move between them. However, it is possible to perform other higher-level tasks not benefitting the single individual, but society at large, such as medical research and public-health monitoring. Finally, health information could be transferred for commercial reasons, including billing. Health information has the potential to be extremely sensitive, i.e.: its release without consent can have devastating consequences for its owner. For this reason, several countries have written legislation which aims to define the appropriate ways to collect, transfer and release medical data so as to protect the health information of individuals. In this section a brief summary of such legislation in the USA and the European Union is presented. These legal frameworks are bound to have a profound effect on the adoption of different anonymization protocols.

2.1.1 European Union

In the European Union, since May 2018 health information is protected under the General Data Protection Regulation (GDPR) [1]. Recital 26 of the GDPR defines anonymized data as "data rendered anonymous in such a way that the data subject is not or no longer identifiable." Where data is correctly anonymized, it falls outside of the scope of the GDPR. However, the definition could be too strict to be achievable in practice in many cases. This led the European legislators to additionally define pseudonymization, in article 4(5), as "the processing of personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information." Pseudonymization both guards companies legally against the cases in which a malicious actor manages to re-identify the data using background information and allows companies to "tokenize" the data, i.e.: replace personal identifiers with unique tokens that allow to link individuals across different databases without knowing

their identity explicitly. It reduces the duties for data holders as compared to fully identifiable data. For example, it could grant an exception to the "purpose limitation principle", according to which data should be processed and stored only for the purposes clearly-defined when the data is collected (which could be severely limiting). In general, the GDPR mandates "data protection by design", or the principle of including privacy protections for individuals at the design stage of a product that processes personal information. One way to achieve data protection by design is by means of pseudonymization.

Another important aspect of the GDPR is that it requires data holders to notify without delay all the persons affected in the case of a data breach. Who exactly has been affected by a breach and to what extent is very often a complex question and answering it is a real challenge for data holders.

2.1.2 United States of America

In the USA, the management of identifiable health data is governed by the Health Insurance Portability and Accountability Act of 1996 (HIPAA) [2], where it is called "Protected Health Information" (PHI). PHI cannot be shared by a data holder without explicit written individual consent, with the exception of the following cases:

" (1) To the Individual (2) Treatment, Payment, and Health Care Operations (3) Opportunity to Agree or Object (4) Incident to an otherwise permitted use and disclosure (5) Public Interest and Benefit Activities (6) Limited Data Set for the purposes of research, public health or care operations. "

However, de-identified health information do not fall under the jurisdiction for PHI. According to HIPAA, there are two ways of de-identifying health data:

- **Expert determination** that the re-identification risk is "small", to be performed by a qualified statistician.
- **The "Safe Harbor" method** is based on the removal of 18 types of data for "the individual or relatives, employers, or household members of the individual." The 18 types include the name, social security number, IP addresses, full-face photographs, precise geographical information and dates such as birthdates, and other uniquely identifying attributes.

It is desirable to meet either of the two requirements because seeking explicit consent from patients can lead to reduced recruitment and selection bias, which are harmful to health research.

2.1.3 "Reasonable" standards

In EU and US health privacy statutes, and in general in many other jurisdictions, no precise metric to quantify the acceptable risk of a privacy breach is provided. Instead, they mandate to adhere to some "reasonableness" standard. This is, for instance, fairly

evident in the case of the HIPAA "expert determination". On one hand this is likely to favor the application of straightforward and repeatable rules where available, such as "Safe Harbor". On the other hand, this means that precedents have a huge importance when determining acceptable levels of risk, and thus it is likely that institutions would prefer well-established techniques to more effective but novel ones, in order to protect themselves from legal disputes and to safeguard their reputation. In this way a data custodian can point to previous court cases, regulatory orders and such to justify their decisions [11].

2.2 Health informatics landscape

With the goal of understanding the applications and needs of the domain, I have performed a relatively thorough literary review of the health informatics research carried out in the last few years, principally that which bases the evaluation of their methods on MIMIC-III, the dataset which has also been chosen for the evaluation of privacy-preserving methods in this project (described in Chapter 3). It is important to note that for the nature of the data contained by MIMIC-III, pertaining to Intensive Care Unit patients, a lot of the scrutinized research was focused on severe illness and emergency procedures. I believe however that the insight that can be gained from the kind of questions asked and the data analysis/machine learning methods used can be generalized to the wider context of health informatics. This section can by no means concisely provide a thorough report on a field that is getting richer and more diverse at a surprising pace. However, notable examples are mentioned that will help highlight some of the most popular and promising research directions.

In [33] the authors advocate for the importance of data-driven health research focused on "precision medicine", an up-and-coming approach to disease prevention and treatment that "exploits the multiple distinct characteristics of each individual (in gene, environment, and lifestyle) to maximize effectiveness". This orientation is closely intertwined with the field of genomics, which is based on different, extremely high-dimensional types of data that go beyond the scope of this project. Additionally, "most tests based on genomics are still too slow to be useful in the ICU, and the data they generate do not readily inform clinical decision-making" [33]. In line with this, several studies have emerged that attempt to use Electronic Health Records data to solve tasks as different as the clustering of patients, the early detection of conditions, and the prediction of effective personalized treatment.

For instance, the authors in [40] build a decision support tool for the management of mechanical ventilation, based on reinforcement learning. They predict the optimal time-to-extubation and a personalized regime of sedation dosage and ventilator support. Many similar efforts exist: for example, to determine the optimal sequence of drugs to be administered in HIV therapy or cancer treatment, or for the regulation of insulin for diabetes patients. In general, such reinforcement learning approaches assign a value to the clinical outcome of different treatment decisions given the state of the patient.

Sepsis, a condition that arises when a response to infection injuries the body's own

tissues and organs, is the main cause of mortality in hospitals worldwide [21]. In [12] the authors attempt to build a prediction system to detect early the development of sepsis in patients. Their classifier is based on hand-crafted features constructed from the patients' vitals (e.g.: pulse pressure, white blood cells count) at multiple points in time. The work in [27] uses the reinforcement learning framework to train an agent for establishing a sequence of procedures to optimize the outcome for patients affected by sepsis. They encode patients' data as multidimensional discrete time series, comprising of vital signs, laboratory values, and procedures received. They additionally include demographic data and information about the patients' survival, crucial to determine the merit of different policies.

Longitudinal datasets record multiple data-points pertaining to the same patient at different points in time, across different records. Systematic and large datasets of longitudinal data have a key importance: events such as administration of medicines and consequent responses are used to determine the right treatment, and patient events in combination with demographic information can allow to discover patterns in different cohorts of individuals [9]. In all the research highlighted so far, time-series (and hence, longitudinal datasets) play a fundamental role. In [9] the authors introduce a novel LSTM network unit[1] designed to handle irregular time intervals in longitudinal patients records for the task of patients sub-typing, i.e. their grouping into disease characterizing subtypes. In [46], an attention network [50] is built that matches or improves the state-of-the-art performance on several benchmark tasks [22] for the analysis of clinical time-series.

A usual pre-processing step on the data includes resampling events with means in set time intervals and using interpolation techniques to impute missing values. This is done to cope with the irregularity and sparseness of measurements such as vitals and lab results [40]. In [44] a processing pipeline is developed that transforms EHR into structured predictor variables and thus curb the pre-processing burden necessary to harmonize EHR in previous research. Both of these suggestions could provide interesting stimuli to anonymization protocols; in the first case exploiting resampling to eliminate useless identifying differences in time-series, in the second case by building a useful representation that does not require full EHRs to be shared.

The selection of patients based on the study's criterion of utility has been found to be a common necessity. For instance, EHRs might need to be filtered by a query such as "patients with sepsis", or by even more specific ones, e.g. "patients kept under ventilator support for more than 24 hours", or "patients that had previously been administered antibiotics". Failure to filter all the irrelevant clinical cases for a study according to possibly very specific requirements, is likely to destroy the utility of a dataset even before any data analysis has to be performed. This is in turn likely to complicate the specification of an anonymization scheme which aims to preserve utility.

The digitalization of clinical information into Electronic Health Records has become so widespread, oddly, for billing purposes, and only recently have they served as a basis for medical research [44]. Because of their original purpose, structured fields in

[1] LSTM stands for Long Short-Term Memory, and it is a type of neural network used to model time series [23].

EHRs are not as good as a source of clinical information as hand-written clinical notes associated to them. For this reason, several efforts exist that attempt to use free-text notes for learning. In [10] an LSTM neural network was applied to the extraction of clinical concepts. That is, the task the authors attempt to solve is the identification and classification of concepts into categories (such as tests and treatments) from natural language transformed into word embeddings. A more comprehensive review of word embeddings for medical natural language processing applications is offered by [52]. The work in [39] introduces a neural network with "condensed memory" for diagnostic inferencing from the free-text notes in MIMIC-III, and uses raw text from Wikipedia as the knowledge source.

2.3 The role of Craneware

Craneware is a prominent US company developing software to enable healthcare providers improve margins and enhance patient outcomes, e.g. software to analyze insurance claims and predict cost. They have provided support in developing some of the key starting points of this project, and plenty of useful insight and feedback. They have stressed how increasingly important EHRs are, beyond medical research, to improve the efficiency of healthcare providers. Useful applications include:

- Forecasting expenses.
- Predicting readmission rates, measured in terms of patients being admitted to the hospital within 30 days of discharge. This is of interest because insurance providers in the US will not cover expenses incurred after readmission, that will have to be covered by the hospital.
- Calculating the length of stay for admission, often utilized as a metric for physician efficiency comparisons.

These enquiries will have requirements that differ slightly from those discussed in Section 2.2 and provide an interesting basis for further analysis of the utility of anonymized healthcare datasets.

2.4 k-anonymity

Classical de-identification techniques are based on the deletion of *identifiers*, i.e.: attributes that can univocally identify one record. Examples include a name or social security number. Sweeney showed that this is an insufficient means of protecting data [49]. She managed to re-identify medical data about the then governor of Massachusetts by linking an anonymized dataset with publicly available voters registration records. A similar famous incident happened with some data released in 2006 by Netflix [37]. By cross-referencing information in multiple datasets it is possible to break such a simple scheme because a collection of fields has the potential to uniquely identify individuals as well. Call this ensemble of fields the *quasi-identifiers*. In response

to such short-comings, Sweeney defined k-anonymity [49] as the guarantee that each combination of quasi-identifiers will be represented by at least k records. Define an *equivalence class* as a group of records whose quasi-identifiers have all identical values. Then k-anonymity can also be stated as the requirement for all equivalence classes to have size at least k. The hope is to protect records by "hiding them in a crowd" of other $k-1$ individuals with the same quasi-identifiers combination. The assumption is that it is possible for the custodian to correctly identify the quasi-identifiers. This is only possible if the linkage attack is performed with external information that is also available to the custodian. It might be problematic because different attackers might have access to different resources. Marking more attributes than necessary as quasi-identifiers becomes quickly unfeasible because the probability of their combination being unique becomes very large very quickly and the de-identification mechanism must, in order to keep the data k-anonymous, increase the information loss excessively. This problem is known as the *curse of dimensionality* for k-anonymity [5].

There are at least two widely known attacks on k-anonymity:

- **Homogeneity Attack**: Suppose all the values are identical for a sensitive field (as opposed to the quasi-identifiers) of an equivalence class in a k-anonymous release. Then it is still possible to infer precisely the value of the sensitive field, once an individual has been confidently mapped to that equivalence class.
- **Background knowledge attack**: There could exist a relation, known to the attacker, between different quasi-identifiers and sensitive fields that allows to restrict the value of a sensitive field to a small set of options. An example of this is using the known statistical fact that Japanese have a very low incidence of cardiac diseases [32] to discard some records in an equivalence class and break k-anonymity through probabilistic inference. The severity of this problem is exacerbated by the growing power of machine learning and its increasing accessibility to a wide audience.

An attempt to provide stronger protection against these types of attacks produced l-diversity [32]. l-diversity requires that in each equivalence class there are at least l "well represented" values for a sensitive field, where "well represented" can be specified by various conditions, the simplest one of them being the presence of at least l distinct values per equivalence class. A more complex definition is based on a measure of the entropy of the sensitive values in an equivalence class.

t-closeness [30] is another variant that requires that the distribution of values for a sensitive field in each equivalence class is similar to that of the entire database (more precisely, their distance[2] is thresholded by t).

t-closeness is regarded to have superseded l-diversity because the latter presents at least two major problems. First of all, having l distinct sensitive values in an equivalence class does not exclude that they could be semantically very similar, thus effectively rekindling the problem of homogeneity attacks. Furthermore l-diversity could be a hard condition to achieve that is stronger than necessary in cases where not all values for a sensitive field are equally revealing (e.g. a rare positive indicator for a disease

[2]Several definitions of distance can be used, for example, Euclidean distance.

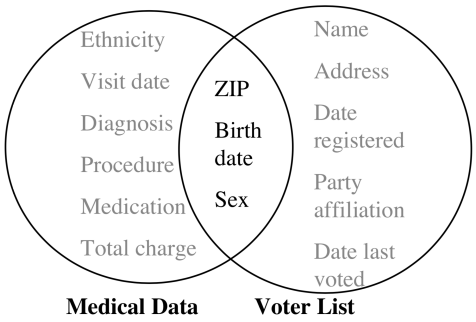


Figure 2.1: The quasi-identifiers used by Sweeney in his attack (credit to [49])

versus a common negative one). t-closeness addresses these concerns by "calibrating" the need for diverse attributes based on their distribution in the database.

This project has not directly investigated the effect of these stronger guarantees on the outcome of anonymization. This choice has been made in order to focus the field of investigation to an optimal implementation of "vanilla" k-anonymization, and then compare that to a rather different direction, namely ϵ-safe anonymity, which will be dealt with later. It is however worth mentioning that it is possible (because of the *generalization property* proved in [30]) to achieve both l-diversity and t-closeness with an extension to the k-anonymity check in the algorithm used throughout this project and described in Section 4.3.

2.5 Differential privacy

2.5.1 Definition

Unlike k-anonymity, differential privacy, proposed by Dwork [15], provides strong semantic guarantees that require very few prior assumptions. It is based on the addition of uncertainty for the attacker via randomized algorithms (or mechanisms) influencing the response to a query. In order to define it, it is useful to introduce the ℓ_1 distance between two databases $d^{(1)}$ and $d^{(2)}$ from $\mathcal{D}$, the set of all valid databases. It is a measure of how many records differ between the two and is defined as

$$\|d^{(1)} - d^{(2)}\|_1 = \sum_{i=1}^{|\mathcal{X}|} |d^{(1)}_{\mathcal{X}_i} - d^{(2)}_{\mathcal{X}_i}|$$

where $d^{(n)}_{\mathcal{X}_i}$ denotes the histogram count of records equal to $\mathcal{X}_i \in \mathcal{X}$ for database $d^{(n)}$. $\mathcal{X}$ is the set of all possible valid records, but for the purpose of this definition, it could be substituted in the summation by the union of the entries in $d^{(1)}$ and $d^{(2)}$.

If a data release is differentially private, an attacker should not be able to learn "much" more information about individuals in a database than they could learn if those individuals opted out of the database. Under some conditions, this implies that how much

previous knowledge an attacker possesses is irrelevant from the perspective of the data holder, as the privacy protection relies on the mechanism rather than on the data release. Formally, given any two databases $d^{(1)}$ and $d^{(2)}$ such that $\|d^{(1)} - d^{(2)}\|_1 \leq 1$ (meaning they only differ in one entry, at most) a randomized mechanism $\mathcal{M}$ is (ϵ, δ)-differentially private if:

$$\mathbb{P}[\mathcal{M}(d^{(1)}) \in \mathcal{S}] \leq e^{\epsilon}\mathbb{P}[\mathcal{M}(d^{(2)}) \in \mathcal{S}] + \delta$$

for any $\mathcal{S} \subseteq \text{range}(\mathcal{M})$, and where ϵ is the *privacy parameter* and δ is a relaxation constant. Higher values of ϵ entail weaker privacy guarantees by allowing the randomized outputs of the two databases to be less close. Similarly, we want δ to be negligible (usually a value less than 10^{-4} [42]), or we might be "protecting privacy" by publishing a small number of records in their entirety [17].

2.5.2 Mechanisms

The study of mechanisms that allow to obtain differential privacy is an active area of research. One of the first and most popular suggestions consists in the addition of noise drawn from a Laplace distribution [16] with mean 0 and scale $\frac{\Delta f}{\epsilon}$, where Δf is the global sensitivity of the query, i.e.: the maximum distance in a query output produced by f, over all neighboring databases in $\mathcal{D}$. Intuitively, the greater the impact of a record to the output, the more perturbation is needed to avoid leaking information. This mechanism is only suitable to numerical values. The *exponential mechanism* is an alternative that can be applied to values from arbitrary domains and will be discussed in Section 4.5.3.

There are two distinct contexts in which the release of differentially private data can operate. In the interactive context, users can adaptively query a database and obtain differentially private answers, usually after noise has been added to them through mechanisms such as the ones just introduced. Interactive querying is possible because differential privacy is composable: for $i \in [1, P]$ it is possible to apply in sequence, on the same input, P algorithms $\mathcal{M}_i$ that are (ϵ_i, δ_i)-differentially private, and the cumulative output of all algorithms will be $(\sum_{i=1}^{P} \epsilon_i, \sum_{i=1}^{P} \delta_i)$-differentially private. Other advanced composition results, that allow parameters to degrade more slowly and for more complicated forms of composition, have been proven [17]. It is then required to define a *privacy budget*, which will be eroded by each interactive query. After its exhaustion, the mechanism will not be able to answer any other new query whose result is not a simple reformulation of previous answers to old queries.

Non-interactive mechanisms are either based on the release of a perturbed version of a database, possibly in the form of a contingency table, or on the release of a new synthetic database that for some set of queries mimics the distribution of the original database. They generally have the advantage of providing non-contradictory answers to correlated queries. They also allow data recipients to independently explore the data. Alas, Dinur and Nissim [13] show that it is not possible for a table perturbed by adding noise to be able to provide accurate answers to many queries. This means that

data custodians might have to assume a useful set of queries when releasing. Another drawback is that in the non-interactive setting it is not possible for the privacy-protecting mechanism to adjust in an adaptive manner for each query.

2.5.3 Independence assumption

In order for differential privacy to protect against privacy breaches it is crucial that entries are independent, in the sense that each individual does not affect other entries in a database [26] during the data generation process. This is, for instance, not the case for datasets describing social networks, where the participation of an individual could imply edges formed between pairs of friends of such individual. If this assumption is violated, it is not possible to adequately limit inference on the participation of data subjects, and a stronger privacy definition is required, in which the background knowledge of the attacker is taken into account (thus frustrating one of the main benefits of using differential privacy in the first place).

Chapter 3

Exploration of MIMIC-III

A few alternative datasets have been considered to identify the critical privacy issues of data collections in the health domain, and to benchmark the privacy-preserving transformations proposed. Initially, it was planned to obtain a synthetic dataset generated from real sensitive data possessed by Craneware, through Synthpop [38], a package for the statistical computing environment R. Synthpop allows to produce data that exhibits similar characteristics to its input, in terms of distribution of values and relationships between variables. It should not be considered a tool to make valid inference on anonymized data, but rather an aid to test models that would later be applied to confidential data.

A different alternative could have been using insurance claims for Medicare, the US insurance program targeted at Americans aged at least 65 years. This demographic restriction however could have made the exploration less meaningful as age is a determining factor in the distribution of health features and thus would have made its distribution less diverse. This choice could have also proven unnecessarily costly, because of a 250 US dollars fee necessary to access even a small portion of the full Limited Data Set[1].

Instead, the chosen dataset is MIMIC-III [25], from The Laboratory for Computational Physiology at Massachusetts Institute of Technology. Firstly, MIMIC-III includes a large quantity of diverse health data. Additionally, it has an open nature that made it accessible for this project. As an extra benefit, because of its accessibility, MIMIC-III is an extremely popular choice for research in health informatics, which facilitates a scrutiny of research trends and needs, and an investigation of how they would be affected by anonymization. Being granted access to the dataset demanded the completion of certified training on ethical procedures for research involving human participants, including knowledge of HIPAA as described in Section 2.1.2. Additionally, a data use agreement had to be signed, in which attempts to identify individual patients were forbidden. Further information on the measures in place to safeguard the data used throughout this project are explained in Appendix A.

MIMIC-III contains clinical data for 61,532 ICU (Intensive Care Unit) stays and 46,520

[1] 5%, which could nevertheless be considered abundant in terms of the absolute number of entries.

patients. It also includes information such as prescriptions, laboratory measurements, observations and notes charted by care providers, and more, comprising in total of 40 tables. Some of these tables are dictionary tables providing definitions for identifiers. A number of precautions have been taken to de-identify patients. All HIPAA protected fields with the exception of dates have been removed. Dates relative to events (for example, procedures) have been shifted randomly into the future, consistently for each patient so as to preserve the length of intervals. The date of birth of subjects over 89 years old has been shifted as well, independently of their event dates. Other precautions have been taken for free-text fields, like physician notes, such as the detection (through pattern-matching) of HIPAA protected fields in unstructured text form, and their subsequent removal.

As a preliminary exercise, let us have a look at the effect of the assurances of Safe Harbor applied to MIMIC-III. This chapter should help motivate the theoretical efforts of this project, their implementation into algorithms and all successive experiments, as it shows that standard anonymization techniques based on the removal of direct identifiers are insufficient for non-trivial sets of data. We shall not look at all the different possible attack vectors, but three examples of ways in which an adversary could accomplish re-identification are presented. We will be focusing on joins of the *patients* table with other tables in order to produce a richer view, whose increased dimensionality can be exploited to restrict the set of patients whose features can be matched.

3.1 Basic demographics

We can join the *patients* and *admission* tables and drop all the duplicates with respect to the *subject_id* attribute, in order to obtain patient records comprising the following attributes:

year_of_birth, gender, insurance, language, religion, marital_status

Unfortunately for the purpose of analysis, dates of birth in the database are unreliable. They have been shifted in the future and in the past in an undocumented way, and they are distributed with a standard deviation which is unrealistically large. The removal of age is not required by Safe Harbor. Then, to include age in my analysis, I approximate its effect by resampling years of birth from a normal distribution with a standard deviation equal to 33. The mean of this distribution is irrelevant for the sake of the size of equivalence classes, but with mean 30, the age distribution does not differ by more than 7 percent from the census on the CIA World Factbook [3] for any of the age groups defined. The quality of this analysis is inferior to one performed with real ages. For example, the result might include babies covered by Medicare (which could not happen in reality). We drop all the entries that have missing values for some of the above attributes, and obtain 23,236 patient records. Some inconsistencies are present among the attribute recordings (for example some English speakers are classified as speaking “American”) but their impact is negligible enough to ignore it. With

these premises, we can build equivalence classes of unique sets of attribute values. Across 1,000 versions of the table obtained with different age samples from the normal distribution, at least one third of patients (on average, 8,254 patients) belonged to an equivalence class of size 1. The average size of all equivalence classes was 1.91. A different assumption on the distribution of age has also been tried: with age sampled uniformly at random between 0 and 99, never has the number of size-one classes been lower than 6,900 (about 29% of patients) across 1,000 more trials. Hence, given information about these attributes, which are generally public and widely available, it seems possible to re-identify an unsettling amount of entries. If we drop ages, the situation looks less grim: the average size of equivalence classes is ~15.7 and 3.03 is the approximate percentage of re-identifiable entries. This confirms the wisdom of MIMIC-III's designers for perturbing the age field.

3.2 Time-series of procedures

MIMIC-III contains, for each patient, numerous data points representing sequential events such as diagnoses received, or procedures performed. Focusing on procedures: to look into the uniqueness of small combinations of sequences, we need to build a *pivot table* in which these sequences are stacked horizontally (one column per item in the sequence), rather than vertically (one row per item) . Conveniently, procedures in the dataset are represented using a common dictionary of numeric codes (ICD9), removing the problem of the inconsistent representation of concepts encountered for basic demographics. Procedures are associated with patient admissions, of which there can be several per patient. Looking at the granularity of single admissions, it has been seen that, given information on the first three procedures linked to an admission (and excluding all the admission with less than three events), it is possible to single out over 44 percent (14,745) of the total 33,044 admissions. The average size of an equivalence class built with the three procedures as quasi-identifiers is 1.88. Looking at only the first two procedures, it is only possible to identify about 15 percents of 42,119 admissions. Interestingly, re-identification becomes easier when looking at pairs of admission that are not adjacent. For example, information about the first and fifth procedure distributes over 28 percent of the 20,434 admissions in equivalence classes of size one. Presumably, this can be explained with some procedures being very likely to be performed together. If we erase the temporal information from sequences, and look instead at unordered sets of procedures performed, the average size of equivalence classes become 2.18 in the case of the first three data points, with 37 percent of admissions being re-identifiable. That is to say, removing time information does not seem significant for anonymization purposes, but might have a markedly negative impact on the usefulness of the data, as outlined in Section 2.2. It is possible to re-identify patients (find their numerical identifier) by matching them to any of their admissions, each containing sequences of procedures, which makes the results just presented even more concerning.

3.3 Hand-written notes

Let us now focus on hand-written notes that come with the Electronic Health Records. A qualitative analysis reveals that they contain extremely sensitive information that extends beyond the medical domain. For instance, they often reveal details such as the patient's habits and lifestyle, psychological evaluation, abuse of substances, history of violence (a large amount of records reference things like "domestic violence", or "sexual assault"), personal relationships (e.g.: "Recently got into fight with partner").

A second analysis reveals that (perhaps unsurprisingly) of the 45,089 patients for which there are nurses notes in the database, almost half of them (49.5%) can be immediately mapped to the attribute *subject_id*, based on the uniqueness of the first 20 characters of those notes. For example, this would allow a malicious actor to identify all the health information associated to a particular subject in MIMIC-III EHRs, after having obtained through some means the physical document (or a fragment) before its digitalization.

Finally, it has been found that the notes contain a wide variety of attributes that could (and probably should, in the light of previous analysis) be considered quasi-identifiers. This means that even if it were completely unrealistic for anyone to obtain the hand-written notes for a patient and use those for re-identification, or if the chances of doing so were exactly the same as those of knowing all the remaining health information about that patient, the publication of notes could still pose a threat: they allow to gather quasi-identifier values that the custodian might have decided to generalize or not include in the release in their respective (structured) fields. The publishers of MIMIC-III were aware of this, as mentioned earlier, and removed some direct identifiers such as names.

Table 3.1: Attributes inferred in unstructured text (notes on patients) from MIMIC-III, on the basis of a set of example keywords

Attribute type	**Matches**	**Keywords**
Age	564,491	"year old", "year-old", "years old", "years-old"
Sex	1,823,400	"she", "he", "male", "female", "man", "woman"
Ethnicity	4,226	"white man", "white male", "white female", "white woman" "black male", "black man", "black female", "black woman", "african american", "native american", "asian male", "asian man", "asian female", "asian woman", "caucasian", "hispanic", "indian male", "indian man", "indian female", "indian woman", "arabic man", "arabic woman", "arabic male", "arabic female"
Religion	722	"religion", "catholic", "protestant", "muslim", "jewish", "buddhist", "hindu"
Sexual orientation	445	"heterosexual", "cisgender", "homosexual", "gay", "lesbian"

The extent to which the obfuscation performed prior to the release of MIMIC-III has been successful is hard to verify, given the quantity of data and the subtleties of natural language (including the possibility of spelling mistakes), which make it hard to come up with a complete set of search queries for direct identifiers. However, for the same reasons, it would be unwise to completely rule out the possibility of sensitive data not detected and thus not having been erased.

Table 3.1 shows, among all the 2,083,180 notes, how many of them might reveal patients' attributes for some categories. The number is computed by means of a naive string search for some indicator keywords and is expected to be a conservative estimate. A more exhaustive search would be a lot more involved for the reasons outlined above, and unnecessary to prove the point.

In conclusion, it appears that the protections offered by Safe Harbor are unsatisfactory. MIMIC-III achieves a better trade-off between utility and privacy, but despite the precautions taken by the designers of the dataset, the data could still be harmful if it fell into the wrong hands. The designers are aware of this, hence the required training and legal agreement.

Chapter 4

Implementation of privacy-enhancing algorithms

4.1 Domain-specific issues and assumptions made

4.1.1 Data types

This project aims at building a toolbox of privacy-enhancing techniques that are suitable for healthcare data, and evaluate it experimentally. This kind of data, as has been seen, is heterogeneous, and the best outcome is likely to come from a combination of different methods. In order to do so, an incremental approach has been adopted. At this stage, only relatively low-dimensional structured tables are addressed. This includes patient demographics and short time-series. It momentarily excludes free-text, and rich time-series, which will need a sensible ad-hoc representation. In fact, the curse of dimensionality is highly likely not to make k-anonymity applicable to these kinds of data. Differential privacy could be applicable, but standard implementations of differential privacy have been rejected for reasons that will become clear later. This section summarizes the assumptions that have been made in the scope of this stage of the project.

EHRs can contain categorical data (e.g., diagnosis codes, procedure codes, drugs dispensed, laboratory tests ordered, and geographical information about the patient and the provider), as well as numerical quantities (e.g. body mass index) and date-time objects [44]. A good anonymization technique should be able to handle reasonably-sized structured data without applying unreasonable distortions. Where time-series are present and represented by longitudinal tables, they will be reduced so as to compact them into single rows, as more clearly explained in section 4.2. Where fields contain unstructured data, they will automatically be suppressed in a data release.

4.1.2 Table publishing vs. online querying

A challenge for the custodian is that the needs of prospective analysts could be unknown ahead of time. Different data-mining applications might have different (orthogonal) requirements. Not only, in a perfect world, would we want to address all of them; the custodian must release data in such a way that the union of all published datasets or queries answered could not be exploited to breach privacy. This prerequisite stems from the fact that different data recipients might collude. Furthermore, even the analysts themselves might not be able to know *a priori* the information they want to extract from a release. Being able to refine their approach in reaction to previous interactions with the data might prove decisive to a positive outcome.

Users of health data are accustomed to table publishing. One reason for this is that it can present unexpected distributions or errors that can only be addressed by analysts if they appropriately manipulate the release to take those complications into account, informed by insights gathered through direct access to that published data [11]. Additionally, some contexts (e.g. pattern recognition) inherently require exploration, as it is impossible to pre-determine a set of useful queries to submit. Dwork reports that "conversations with experts in this field [research statisticians] frequently involve pleas for a noisy table that will permit highly accurate answers to be derived for computations that are not specified at the outset" [14].

Unluckily, because of Dinur and Nissim's result, at least in the context of noisy tables, a choice must be made on whether to allow data recipients to have direct access to a published database (non-interactive setting), but risk restricting its usefulness a priori, or alternatively only give them the possibility of submitting queries in the interactive setting until the privacy budget runs out, at which point the database should stop responding to new queries for anyone[1].

Here I chose non-interactive methods based on publishing. In addition to the convenience of access for data analysts, they shift from the data analysts to the data custodians the responsibility of deciding over the type and precision of information that will be released. This makes the process less arbitrary in the sense that the custodians can decide which audiences they can address when releasing, and work out the appropriate compromises between the breadth of those audiences and the focused utility for some specific individuals, institutions or tasks. In the interactive context, this type of decisions can only be taken by rationing the privacy budget among different users, and if the users are not known in advance and this rationing is not performed, it will be up to each of them to be parsimonious with the total budget. Custodians can still receive feedback from users before a release, for example in the form of useful queries they can audit and approve. This will not allow the adaptive formulation of queries, but can be considered a "fairer" approach.

[1] Unless it is possible to reply with a result that is only dependent on previous answers to old queries.

4.1.3 Data consistency

As mentioned, anonymization mechanisms such as the ones for differential privacy sometimes rely on the addition of noise to query responses or published tables. When noise is added independently across fields or records, the response can violate some natural constraints or correlations that characterize health data (e.g. compatible drugs, or combinations of lab results and diagnoses). In practice this makes noise-based perturbation mechanisms problematic, as they could decrease the consistency of the data and erode the trust of analysts in it [11]. Similar problems are encountered, though to a lesser extent, with synthetic databases. This project wants to focus on release procedures that do not suffer from these troubles. The rejection of noisy tables or synthetic databases following from this observation, together with that of interactive mechanisms motivated in the previous paragraph seems to rule out all differentially private approaches. This limitation only holds until a relaxation of differential privacy that adds additional uncertainty through sampling will be introduced in Section 4.5.2.

4.1.4 Recording models

Several anonymization protocols (especially k-anonymization and its derivatives) will replace a field, in a subset of all records containing semantically close values for that field, with a more general value covering all of those records. This is a process known as *generalization*. An example of this is changing, for the attribute "city", multiple occurrences of cities in favor of their common country.

If generalization is performed with *global recording* [53], it is applied to all the records. On the other hand, *local recording* implies that a generalization is applied to single instances in a database. The global recording model can further be subdivided into *single-dimensional* and *multi-dimensional* recording [28]. In the single-dimensional case, a generalization is a function mapping single attribute values to their generalization. Instead, with multi-dimensional recording (where each attribute is a dimension), generalizing maps the vector-valued Cartesian product of multiple attributes to their generalization and divides the record space accordingly. Because a generalization function will map identical inputs to the same output deterministically, with single-dimensional recording, attribute values will be mapped to generalized values that define a set of non-overlapping single-dimensional regions. If the non-overlapping regions are multi-dimensional (as a result of multi-dimensional recording), the mappings for a single attribute can overlap. An illustration of this concept is presented in Figure 4.1.

Both local recording and multi-dimensional recording entail inconsistent generalization across records. Coming back to the previous example, this might mean changing the city "Beijing" to "China" for some records and to "East Asia" for others, or generalize it only in some instances. For numerical values this could lead to substituting in a non-consistent manner the age of 17 with various ranges containing it, like $[16, 22]$ and $[10, 20]$. In practice, instances of these phenomena complicate a lot the data analysis using standard techniques [18]. Therefore, global single-dimensional models have been strongly preferred and ultimately guided the choice of the algorithms dealt with in this

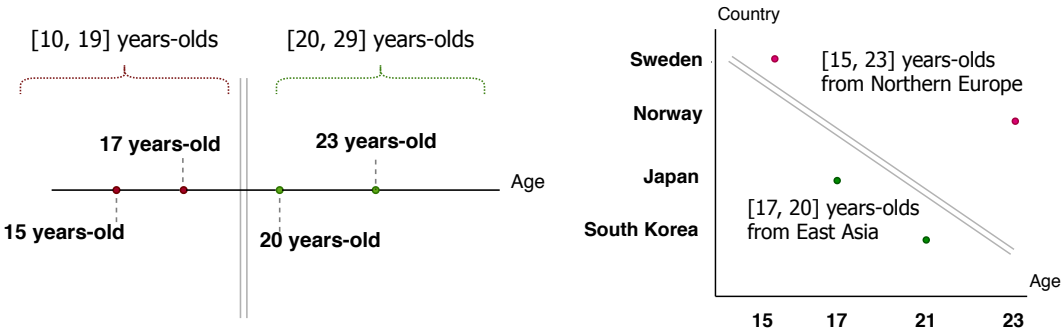


Figure 4.1: Possible generalizations of the attribute *age* for single-dimensional recording (on the left), and multi-dimensional recording (on the right).

paper.

4.2 Problem definition

We consider a data holder (or *custodian*) maintaining a database table $d \in \mathcal{D}$, consisting of $N = |d|$ ordered tuples $x^{(n)} \in \mathcal{X}$. A tuple is an ordered collection of *fields* (or columns) $x_i^{(n)}$. This is a model of data that is general enough to cover virtually all practical applications. Consider for example a dataset with multiple tables: a table $t^{(0)}$ recording individuals and tables $t^{(1)}, ..., t^{(N)}$ recording information, objects or events related to those individuals. Each tuple $y \in \bigcap_{i=1}^{N} t^{(i)}$ references a record from $t^{(0)}$ through a foreign key (an index to the appropriate row). All these tables can be compacted into one, in which all records from $t^{(0)}$ are preserved. We enrich each tuple $x \in t^{(0)}$ so that for all $y \in \bigcap_{i=1}^{N} t^{(i)}$ referencing x, we add all their attributes to x. In order to make the order consistent across tuples in $t^{(0)}$, we add attributes from other tables according to some total ordering we define on all tables (e.g. lexicographic), and we use the order defined on records within the same table. When this operation is performed, we say that tables $t^{(1)}, ..., t^{(N)}$ have been joined with $t^{(0)}$. If we have three tables, a table $t^{(0)}$ for individuals, an "intermediate" $t^{(1)}$ whose rows are associated to records in $t^{(0)}$ through foreign keys, and a table $t^{(2)}$ whose rows reference $t^{(1)}$ through foreign keys, we first want to craft $t^{(1,2)}$, product of joining $t^{(2)}$ with $t^{(1)}$, and then a final table $t^{(0,1,2)}$, product of joining $t^{(1,2)}$ with $t^{(0)}$. For time-series, after joining tables, several rows will be stacked horizontally into a single row, with one or more columns per data point. To provide a concrete example, consider a table recording data about patients, and one recording test results, whose tuples contain p values. The compacted result would consist of the patients table with $p \times k$ more fields, where k is the maximum amount of test result entries pertaining to any patient. Clearly, it is possible that some patients have performed more tests than others. Where needed, it is possible to pad some fields with null values. It will be assumed throughout that a database will only have one row per individual. The reason I preferred a simple data model to which more complex datasets can be reduced is that, to my knowledge, all algorithms described in the literature operate on a single table.

The custodian wants to produce up to R anonymized data releases $r = \texttt{Anonymize}(d)$ that can be distributed to third-parties for some data analysis tasks. The release is tabular ($r \in \mathcal{D}$), but can take different shapes. It might be a filtered and/or perturbed version of the original database, and mirror the schema of d (where for the purpose of this paper the schema is defined as the fields of a table and their semantics). Alternatively, it could consist of summary statistics based on data in d. Generically speaking, given the collection of all the releases $\{r^{(j)}\}_{j=1}^{R}$, it should be possible for legitimate data analysts to perform useful inference, whereas it should not be possible for a malicious actor to breach the privacy of individuals in the database. Depending on the anonymity guarantee of the release, such a statement might require modeling the *background knowledge* of an attacker, i.e.: all of the information, from additional sources, that they can utilize to break said guarantee.

4.3 Optimal Lattice Anonymization

As a means to produce experimental results on the effectiveness of k-anonymity and have somewhere to start for more complex models, a variety of algorithms have been considered. Most of them assume the existence of a *generalization hierarchy*, that defines different levels of granularity for the generalization of attributes that was introduced in Section 4.1.4. This means that each quasi-identifier can be generalized to multiple degrees (or *levels*), e.g. a post-code could be generalized to its city, with one step of generalization, or to its county, with a further step of generalization. The algorithm "Optimal Lattice Anonymization" (OLA) [18] was finally selected. This choice will be justified in Section 4.3.4.

4.3.1 Specifications

OLA takes an input dataset $d \in \mathcal{D}$, the maximum percentage of input records that we are comfortable suppressing in the release to achieve k-anonymity, and a list of generalization rules defining both which fields should be considered quasi-identifiers and what their generalization hierarchy is. Suppression is the process of deleting whole entries. OLA then outputs a k-anonymous release $r \in \mathcal{D}$. Besides k, it is parameterized by a function $l : \mathcal{D} \to \mathbb{R}$ representing an information loss metric to be applied to potential releases. OLA looks in the space of all possible combinations of generalization levels of the quasi-identifiers in the input. Let's call each of these combinations a *node*. OLA finds all the nodes that are k-anonymous and outputs the one associated with the lowest information loss. For this reason we can say that OLA is *globally optimal* with respect to an information loss metric (in the domain of single-dimensional recording).

4.3.2 General operation

Consider fields $\{c_i\}_{i=1}^{C}$, each one with an associated $c_i^* \in \mathbb{N}$ representing its largest generalization level. There will be $c_1^* c_2^* \cdots c_C^*$ nodes in total, i.e. the number of nodes

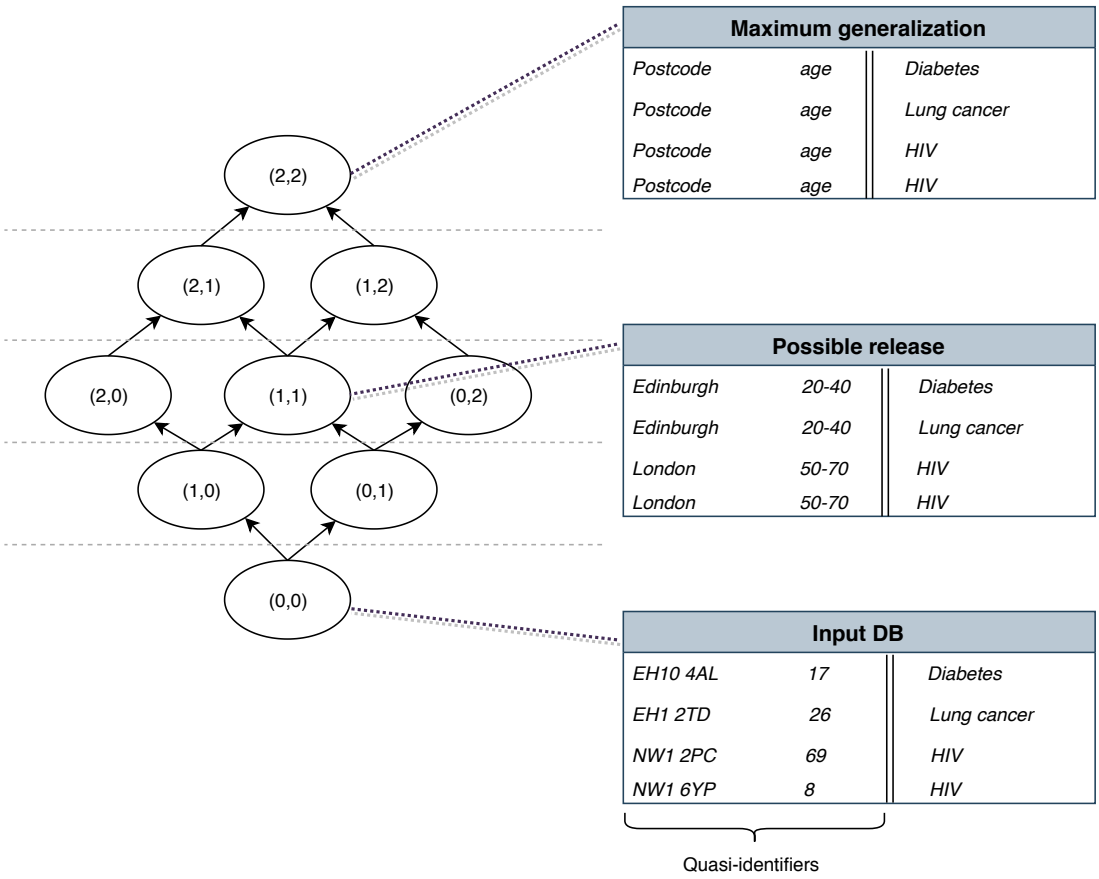


Figure 4.2: On the left, lattice of database with 3 fields (post code, age and condition) of which the first two are considered quasi-identifiers. Each quasi-identifier has a generalization hierarchy that is two levels deep. Each node specifies the generalization level on the first and second quasi-identifiers. On the right, the release associated to some nodes from the lattice. The output of the bottom node equals the original input received. In the table produced by the top node all information on quasi-identifier values has been lost. For k=2, the top 2 sample tables are k-anonymous.

exponentially grows with the number of quasi-identifiers and grows as a high-degree polynomial with the depth of the generalization hierarchy. If we had to check for all nodes whether they are k-anonymous (which is a slow operation), our algorithm would be too computationally demanding. Another computational bottleneck is the time required to compare the various outputs in the end based on the information loss metric. To address this difficulties, OLA adopts two tricks: minimizing the amount of nodes whose information loss is compared and *predictive tagging*. The first technique requires the information loss metric chosen to be monotonically increasing with each generalization step.

Firstly, OLA constructs a *lattice*, that is, a directed acyclic graph containing all the generalization nodes and in which edges go from one node to an immediate successor (i.e. a node produced by making a 1-step generalization on any attribute). It visits all the paths from the bottom, least generic node, to the top, most generic one. A k-minimal node is the k-anonymous node that is closest to the bottom node in one such path. In each path it conducts a "binary search" for the k-minimal node. This is accomplished by first selecting a node that is halfway between the bottom and the top one, then, if the node is k-anonymous, continue the search recursively in the bottom sub-lattice bounded above by this middle k-anonymous node. If the middle node was not k-anonymous, the algorithm continues the search recursively in the top sub-lattice. The search stops when a sub-lattice has only one vertex left, which will by necessity be a k-minimal node (in the worst case scenario the most generalized node will make all the records identical). Because the information loss is monotonic, we only need to keep track of k-minimal nodes, as all the nodes above in a path to the top, most general node would be discarded in the final step of the algorithm anyway. Additionally, when a new k-minimal node is found in this way, all the ones that had previously been found and that are descendants (more general) of the new one are discarded. As for predictive tagging, whenever a node is found to be k-anonymous, all the descendants are marked as such (certainly more generalization will not decrease the size of equivalence classes). Conversely, if a node is found to be not k-anonymous, we mark that on all the less generic ancestors.

The steps of the algorithm can be thus summarized:

1. Build the lattice of generalization nodes.
2. Find all the k-minimal nodes with a binary search in every lattice path, and predictively tag at every k-anonymity check.
3. Compute the information loss for the k-minimal nodes and output the one with the lowest value.

4.3.3 Checking nodes

I am convinced that because of the "jumps" from one node to the next at the binary search stage, when checking whether a node is k-anonymous, an implementation of OLA must simulate a release by applying the generalization rules to all the records in d, and then check the size of the resulting equivalence classes. For this reason, such a

check is $\Omega(|d|)$, i.e.: at least linear in the number of records. For the purpose of this project, the k-anonymity check has been implemented with a dictionary ADS recording how many entries have the same *signature*, or combination of identical quasi-identifier values. We iterate over all the input records, generalize them according to the node's specification and update the dictionary. Then, we verify whether within the bounds of tolerable suppression, it is possible to obtain equivalence classes of size at least k by looking at all the dictionary entries. This takes $O(|d|)$. The impact on memory of this operation is reasonable: the simulated release will take as much space as the input database and the supporting dictionary will probably take less space (and in the worst case, only a constant multiple of the memory required by the input, depending on the implementation of the dictionary and on how a quasi-identifiers signature is turned into a dictionary key). Furthermore, after the check has been completed, it is possible to remove both the simulated release and the dictionary from memory. In its original paper, this check has been implemented with a more inefficient sort of the records on the basis of quasi-identifier values, and then a linear iteration through those records to count the (adjacent) entries in the same equivalence class [18].

4.3.4 Alternative algorithms considered

Datafly [47] is a popular algorithm used in practice for providing anonymity in medical data. It uses a heuristic to decide which field to generalize: it selects the one with the most number of distinct values, and picks at random among ties. It is an extremely fast algorithm, with an $O(NlogN)$ overall complexity, where $N = |d|$. Having said that, it has been criticized for over-generalizing fields, and in fact its output is likely not globally optimal as it does not compute the value of information loss during operation. More precise statements on the qualitative difference between OLA and Datafly can be found in the paper proposing the original formulation of OLA [18]. The sub-optimality of Datafly has been critical in its rejection in favor of OLA.

Samarati [43] exploits the concept of the lattice like OLA. However, unlike OLA, it searches all nodes at one level of the lattice at a time, looking for the lowest level that contains a k-anonymous node. Then at that level, it chooses the best k-anonymous node according to an arbitrary criterion (that could be an information-loss metric). Samarati looks at different layers in a "binary search" fashion. If the lattice has height h, it will first look at level $h/2$ and then at level $h/4$ if it finds a k-anonymous node and at $3/4h$ otherwise. Note that nodes can be compared with respect to their information loss only at the final stage, within the same level, and so the solution that Samarati finds might not be globally optimal. Even though this algorithm is less wasteful (in terms of information loss) than Datafly, it makes it problematic to verify the impact of different metrics to the anonymization process.

Incognito [28] considers all the lattices built from all possible subsets of chosen quasi-identifiers. When it finds a k-anonymous node it performs predictive tagging within a lattice. Also, when it finds that a node is not k-anonymous, it marks as such the nodes in other lattices that contain a larger subset of quasi-identifiers. Incognito navigates each lattice bottom-up in a breadth-first fashion, in order to prune as many nodes as

possible. Then it compares all the k-anonymous nodes in the lattice with the full set of quasi-identifiers with respect to an information-loss metric. Incognito is thus globally optimal. Compared to OLA, Incognito has been shown to be significantly slower [18].

Mondrian [29] is a greedy multi-dimensional algorithm with a top-down approach (it refines attributes rather than generalizing them) and $O(NlogN)$ complexity. It is really fast and the quality of its output can be higher than that of optimal single-dimensional algorithms for numerical values, but it does not behave as well with categorical data, not possessing any mechanism to naturally group semantically close categories together. It is also worth reminding the reader that multi-dimensional recording suffers from the drawbacks described in Section 4.1.4.

4.4 Inverse OLA

OLA operates by building a lattice of all possible generalization nodes, finding all the ones that satisfy k-anonymity for a specific value of k and a maximum tolerated suppression, and among these nodes identify the one that gives the least information loss. This is useful if a data custodian has to abide by some strict and fixed anonymization requirements, and wants to find the best node given those circumstances. However, OLA does not adequately serve its purpose when the custodian has in mind a task which requires some bound information loss, and wants to find the best possible anonymization given those premises. Essentially, the process of choosing a node on the basis of anonymity, and only later information loss, would be reversed. It has been found that OLA can be modified to cover this different use case, thanks to the following facts:

- The information loss is monotonic, hence we know that the descendants of a node will by necessity have larger or equal loss.

- If a node is k-anonymous for $k = c$, all of its ancestors will have k at least c.

This modification of OLA has been called *Inverse OLA*. It takes as argument the maximum tolerated information loss for some metric (max_loss), and (in percentage of records) the maximum tolerated suppression. After having constructed a lattice, it identifies the acceptable nodes (with respect to information loss) through the same binary search across generalization strategies as OLA. It can also predictively tag nodes in the following way: if a node loses too much information, all the descendants can be discarded, if a node keeps enough information, all the ancestors will also be acceptable. When found, these "good" nodes are added to a set, like OLA's *k-minimal* set, but that instead of keeping all the least general nodes in a generalization strategy, keeps all the most general ones. After the search is complete, all the viable generalizations are compared on the basis of the minimum size of equivalence classes (with suppression), and the one with the largest value of k is output.

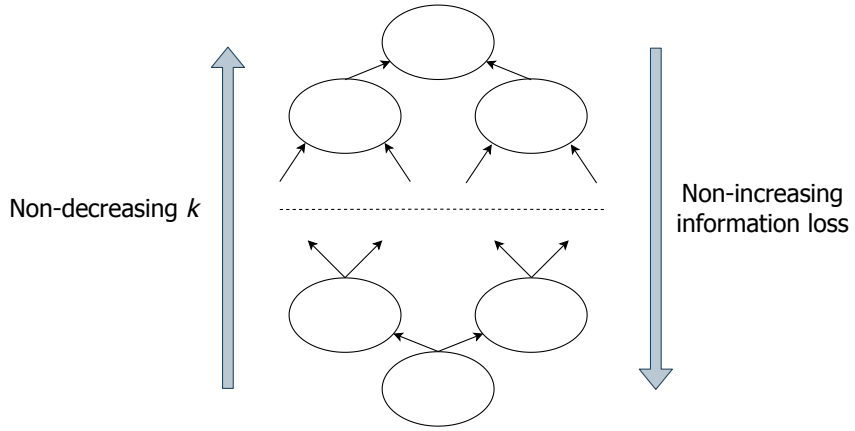


Figure 4.3: OLA goes up the lattice and finds the nodes with equivalence classes of size at least k. Among them, it goes down looking for the node with minimum information loss. On the contrary, Inverse OLA first finds the lower subset of nodes within a maximum acceptable information loss, then it goes up towards the nodes with highest k. This does not correspond exactly to the operation of the two algorithms, but is a useful abstraction to help intuition. Note that it is not guaranteed that among the k-minimal nodes the lowest information loss will be at the lowest height. We are only guaranteed that information loss is non-increasing on a path to the root. Similarly, the value of k is only non-decreasing on a path to the leaf.

4.5 ϵ-safe LA

4.5.1 The limits of k-anonymization

Regardless of the wide practical adoption of k-anonymity or of approaches inspired by it, and in addition to the attacks introduced in Section 2.4, it presents several weaknesses:

- As mentioned previously, deciding which collection of attributes are to be considered the quasi-identifiers is a complicated and often arbitrary decision, if we want to protect the data against many different potential attackers. Some of them might have knowledge which extends beyond what is readily available to the general public. For example they might be close to an individual (e.g. be their spouse) and use the extra information to restrict the range of possibilities and ultimately re-identify a k-anonymous release. Alternatively, they might work in the environment where the data was produced (such as a janitor finding a lab result in the trash).

- Extreme values can impact the output of a k-anonymity algorithm in a way that it facilitates undue inference. For instance, consider a de-identified database of HIV patients in a small town. If information about their income is recorded it can be revealing even if it is considered a quasi-identifier. If there is only one individual in the town with a gross income larger than 500,000 £ per year, in order to produce k-large equivalence classes, an algorithm based on generalization will generalize some values to a range large enough to encompass this large income.

From this an attacker can infer the presence of the wealthy citizen in the database.

- While it could be comforting to think about oneself as being "hidden in a crowd of k individuals" as defined by k-anonymity, such a statement is effectively irrelevant as it is not related to what its consequences are for privacy. Instead, differential privacy provides a rather precise definition of privacy to which some mechanism has to conform, i.e.: the variation of the output of neighboring databases is bounded and hence so are the consequences of the participation of an individual.

4.5.2 Differential Privacy under Sampling

An interesting approach suggested by Li, Qardaji and Su [31] attempts to bridge the gap between k-anonymity and differential privacy and introduces the concept of Differential Privacy under Sampling (DPS). A data access mechanism $\mathcal{M}$ is $(\beta, \epsilon, \delta)$-DPS if and only if $\beta > \delta$[2]and, after sampling tuples from a dataset with probability β, $\mathcal{M}$ is (ϵ, δ)-differentially private. DPS is a useful relaxation of differential privacy that allows to define an anonymization protocol based on an extension of standard k-anonymization algorithms. The relaxation exploits the additional uncertainty of the attacker caused by the sampling step. Conveniently, in many concrete contexts, it is customary to perform sampling prior to releasing a dataset. Unfortunately, DPS is not composable: it is not possible to apply in sequence P $(\beta_i, \epsilon_i, \delta_i)$-DPS algorithms, with $i \in [1, P]$, and find values of β, ϵ and δ for which DPS holds on the resulting output. This means it is not suitable for the interactive setting.

k-anonymization algorithms do not satisfy $(\beta, \epsilon, \delta)$-DPS because of the sensitivity to extreme values discussed above. In order to address this, we want the output not to be overly dependent on the input. To help achieve that, we define k-anonymization as a 2-steps process. Consider d the input tuples and $\mathcal{X}$ the set of possible output tuples. The first step outputs a function $g : d \to \mathcal{X}$. The second step applies such function to all the tuples in d and suppresses all the output tuples that do not have other $k-1$ doubles. Note that this model is powerful enough to describe all k-anonymization algorithms because of the flexibility that g can have, potentially just being a mapping between each input tuple to the desired output.

A first result of [31] is that if the procedure outputting g is independent of the input database, and all columns are treated as quasi-identifiers, then the output of k-anonymization satisfies DPS for some values of β, ϵ and δ. I claim that for this to hold, if we anonymize by generalization, either one of the following will be true:

- Call *conformable* all inputs consisting of the same quasi-identifiers, generalization hierarchy, and with databases that share the same schema. There will be some fixed generalization criterion according to which all conformable inputs will have the same generalization steps applied, regardless of the actual data

[2]Without this condition, $(\beta, 0, \delta)$-DPS holds trivially as, with probability (1-β), the record in which databases differ does not get sampled. Hence, if $\beta \leq \delta$, then $\mathbb{P}[\mathcal{M}(d^{(1)}) \in \mathcal{S}] \leq \mathbb{P}[\mathcal{M}(d^{(2)}) \in \mathcal{S}] + \delta$ for any $\mathcal{S} \subseteq \text{range}(\mathcal{M})$, any neighboring databases d^1, d^2, and any $\mathcal{M}$.

- Some element of randomness will affect the selection of the generalization steps to apply

As can be imagined, having g constant or random will likely result in a terrible trade-off between suppression and generalization, and represents a regression with respect to algorithms finding a good sequence of generalization steps through some heuristic, never mind OLA with its globally-optimal approach. Luckily, the same paper offers a useful relaxation of the above requisite. In fact, as long as the procedure outputting g is ϵ'-differentially private, then the output of k-anonymization is $(\beta, \epsilon, \delta)$-DPS, where

$$\epsilon \geq -ln(1-\beta) + \epsilon'$$
$$\delta = D(k, \beta, \epsilon - \epsilon')$$

and

$$D(k, \beta, \epsilon)^3 = \max_{n:\left\lfloor \frac{k}{\gamma} - 1 \right\rfloor \leq n \leq |d|} \sum_{j > \gamma n}^{n} F(j; n, \beta)$$
$$\gamma = \frac{e^\epsilon - 1 + \beta}{e^\epsilon}$$

where $F(j; n, \beta)$ denotes the probability mass function of the binomial distribution, and $|d|$ is the number of records in a given dataset.

Such an algorithm is said to achieve ϵ-safe k-anonymization. The four parameters, $k, \beta, \epsilon, \delta$, are related by the function D, with the first two controlling the quality of the anonymized data, and the latter two deciding the strength of the privacy guarantee. Due to the difficulty of producing an intuitive explanation for D, I refer the reader to the original proof in the appendix of [31]. The proof only works if all the fields are treated as quasi-identifiers.

In order to describe an extension of OLA, as defined in the previous section, that satisfies ϵ-safe k-anonymization, we need to consider many possible outputs, and use a differentially-private mechanism to select one of them.

4.5.3 The exponential mechanism

The exponential mechanism [35] allows to define a utility function u associated to a set of possible outputs O and select an output $o \in O$ in a differentially private manner that takes into account our preferences according to u. This is achieved by assigning larger probability to outputs with a larger utility.

Let Δu be the *sensitivity* of the utility score:

$$\Delta u = \max_{o \in O} \max_{d^{(1)}, d^{(2)}: \|d^{(1)} - d^{(2)}\|_1 \leq 1} |u(d^{(1)}, o) - u(d^{(2)}, o)|$$

[3]The full expression is very expensive to evaluate. However, empirically, the magnitude of the result of the summation decreases as $n \to |d|$. Hence, only the first few terms of the max operator need to be computed in order to estimate a reasonably trustworthy upper bound for δ.

Then, the exponential mechanism $\mathcal{M}_E(x, u, O)$ selects and outputs elements $o \in O$ for input x with probability proportional to $\exp(\frac{\epsilon u(x,o)}{2\Delta u})$. $\mathcal{M}_E(x, u, O)$ preserves $(\epsilon, 0)$-differential privacy.

4.5.4 Beyond OLA

It needs to be kept in mind that using the exponential mechanism there is still a (hopefully small) probability of producing an output that yields a very low utility, like the maximally-generalized output. What is more, we need to sample not just from k-anonymous nodes, as one might imagine, but from all possible nodes, as we need to sample from the same set of nodes for all conformable inputs. Hence, OLA needs the following radical modifications:

- It must calculate the utility of *each* node.
- No set of k-minimal nodes should be maintained because all nodes could be output in principle, whatever their utility.
- Because of the points above, it is not necessary for the input information loss to be monotonic anymore, which means suppression can be taken into account.
- An arbitrarily large suppression might need to be applied if the chosen node is not k-anonymous, therefore passing as an argument the maximum tolerable suppression is only useful to indicate our preference towards the nodes that only need some bounded suppression.
- All fields have to be treated as quasi-identifiers. Note that in this way the problem of the choice of quasi-identifiers is overcome at the cost of adding more dimensions. This could naturally also be done in a traditional k-anonymization framework.

It would be desirable to be able to recycle the information loss metric that OLA used to pick the best nodes, and have it be the utility for the exponential mechanism (or more precisely, we would be using the *multiplicative inverse* of the information loss). In this way we could express our preferences among the various candidate releases in a way that is consistent with OLA. Several difficulties are encountered. First of all, computing the sensitivity of an arbitrary utility function might be hard. If the sensitivity of the utility is too high, the probability mass will be comparable for all nodes and we might as well choose one uniformly at random. We are forced to use a banal utility function that does not have to be computed on the data, but rather on the features of a node, such as *Prec* [48]. Otherwise, calculating the utility for each node would make the execution time unreasonably long. But we would also like to discourage releases that in order to be k-anonymous require a lot of suppression. We can do so by incorporating a penalty for suppression in whichever utility function we use. Unfortunately, that would again make the computation of utility unfeasible. We could settle with a fixed penalty for all nodes that are not k-anonymous. But this penalty can only be small, if we do not want to make the sensitivity too large and if we want to avoid overwhelming the information loss term.

Regardless of the utility function of choice, let ϵ-safe LA be the algorithm that constructs a lattice, calculates the utility for all the nodes (using predictive tagging if available), and outputs a node chosen via the exponential mechanism.

A final note: in order to enjoy the benefits of ϵ-safe LA, we need to clearly state that we are sampling a generalization node, and that we are sampling records. An adversary, not knowing this, might perform an invalid inference. For example, for ϵ and δ suitably small, homogeneity attacks should not be a concern. The protection of ϵ-safe k-anonymity against homogeneity attacks comes indirectly from the uncertainty added by the records sampling stage, and by the suppression of records after the sampling of a node. Consider however the case of an adversary being aware that some specific individuals were involved in the data generation process, and believing that they will also participate to the data release. Then the attacker could try to exploit the homogeneity of attributes in an equivalence class to infer sensitive values. Regardless of the validity of the inference of an attacker, its consequences in the real world will be tangible. We cannot stop, of course, anyone from drawing wrong conclusions and acting upon them, but we can try to minimize this possibility.

Chapter 5

Experimental methodology

5.1 Information loss measures

One of the following three information loss metrics will be applied to OLA in order to select a lattice node to output. They capture different aspects of the data and the generalization process and will later be compared on an inference task. Note that the depth of the node in the lattice is not useful because it disregards the difference in depth of the generalization hierarchies for different quasi-identifiers. For example, it will treat the generalization of a quasi-identifier that can only be suppressed and that of a quasi-identifier that has many possible generalizations, equally, even though the latter will still convey information after one step.

Prec [48] assigns a penalty equal to the sum of the generalization steps applied to each quasi-identifier, in which each term is weighted by the depth of the generalization hierarchy for that quasi-identifier. In this way, for example, starting from the bottom node, Prec would prefer, among its successors, the one produced by the generalization of the node with the deepest hierarchy. In other words, it is assumed that the sensitivity of the utility for a node is smaller for attributes that can be generalized many times. Since this metric only uses the generalization signature of a node and not the release associated to that node, it is very fast to compute but less sophisticated than the following metrics.

DM*, as defined in [18] is the sum, for all equivalence classes in a release, of the square size of those equivalence classes. The rationale is that nodes in which the combination of all quasi-identifiers becomes less unique for some records should be penalized. Unlike the original Discernibility Metric (for example, [8]), from which DM* is inspired, it is a monotonic metric, as required by OLA to function. In Chapter 6 the difference between the DM* value on a release and on the original database is used.

Entropy-based information loss: take a database d (with $|d| = N$) made up of tuples $x^{(i)}$, and its associated release r with J quasi-identifiers. The posterior probability $\mathbb{P}[a|a']$ of a value a given its generalization a' is

$$\mathbb{P}[a|a'] = \frac{\sum_{i=1}^{N} I(x_j^{(i)} = a)}{\sum_{i=1}^{N} I(r_j^{(i)} = a')}$$

where I is an indicator function returning 1 if the condition passed as an argument holds and 0 otherwise. Then, this metric is so defined

$$-\sum_{i=1}^{N}\sum_{j=1}^{J} log_2 \mathbb{P}[x_j^{(i)}|r_j^{(i)}]$$

where each $x_j^{(i)}$ is the jth quasi-identifier value for the ith tuple in d, and $r_j^{(i)}$ is the corresponding value in r. Intuitively, this measure penalizes releases where for true original values, their probability given their generalization is lower.

Note that all of these metrics are applied before suppression. If they were applied after, their behavior would not necessarily abide by the monotonicity property needed by OLA to build the k-minimal set. This means that the amount of suppression will not be factored in by OLA during the decision of which k-anonymous node to output. That is the purpose of the *max_sup* (maximum suppression) parameter: up to the percentage specified, it is assumed that suppressing records will not harm utility. It is possible to incorporate the number of entries suppressed into the metric only at the last stage, when OLA chooses among all the nodes in the k-minimal set. This could improve the output utility, but we would be forced to give up the claim over the global optimality of OLA: it is theoretically possible for some nodes yielding lower information to have been excluded from the k-minimal set, by heuristic rather than by certain knowledge that they are not better than the k-anonymous nodes below in a path to the bottom least-general root.

5.2 Privacy risk

Finding a good estimate of the risk of a privacy breach is way harder than computing a measure of the information loss but equally important in the determination of the worth of an anonymization technique. Differential privacy has a very elegant way of expressing this risk in a way that is independent of an attacker's background knowledge. However, as we have seen, differential privacy is hard to achieve with truthful tabular releases and another practical measure needs to be identified. As noted in [36], re-identification metrics tend to give a sense of false security because they are based on assumptions on the limits of the background knowledge available to an adversary. In particular, all the tabular privacy metrics that we have seen so far assume the existence of equivalence classes, and that it is impossible for an adversary to match an individual to a set of records with more precision than to any of the records in its equivalence class. Subsequently, these metrics use a parameter to model the mitigation of the difference, for an adversary, between $\mathbb{P}[x_s = v|r]$ [1] (posterior probability of a record x having value v at the index s of a sensitive attribute, given a release r) and the prior probability

$\mathbb{P}[x_s = v]$. Again, these parameters are k for the minimum number of records hiding each other in an equivalence class, l for the number of "well-represented" values in an equivalence class, and t for the difference in the distribution of sensitive values between equivalence classes and the whole release.

It must also be stressed again that the assumption of the secrecy of some attributes can be misguided, or be broken after the data has been released. We would want to be able to adjust our metric on the basis of updated expectations on the availability of external information, and to assign to such availability a probability, rather than a binary value. In mathematical terms, letting H be the set of all the individuals in a release r with $|r| = R$, and calling $r_H^{(i)}$ the individual associated with record $r^{(i)}$, we desire

$$\forall h \in H, i \in [1, R] : \mathbb{P}[r_H^{(i)} = h] \leq m$$

for some appropriate constant m. We also want this to be valid for many different potential attackers, with access to different external information. Call Q the set of all possible *knowledge states*, where a knowledge state is a combination of sensitive columns that the adversary knows, perhaps because they have been leaked or have become available through some other source. Let P_Q be the probability distribution of an attacker being in different knowledge states. Then, the previous equation becomes

$$\forall h \in H, i \in [1, R] : \sum_{q \in Q} \mathbb{P}[r_H^{(i)} = h|q] P_Q(q) \leq m \tag{5.1}$$

$\mathbb{P}[r_H^{(i)} = h|q]$ can be calculated as the inverse of the number of records in r that are indistinguishable with respect to the quasi-identifiers and to the sensitive attributes in the knowledge state. If not all individuals from a population are included in r, it is possible to relax the calculation of the conditional probability and multiply it by the probability of inclusion in the release. Arguing that this is acceptable might be harder in case of a targeted suppression, where the individuals to include are not sampled uniformly at random. In any case, it is important to disclose the inclusion probability to the adversary, so that their estimate of the probability of guessing matches our model. As always, it is not possible and should not be our goal to stop an adversary from performing an invalid inference (we assume the adversary to be rational). I now define a release r to be m-concealing given a set of knowledge states and an associated probability distribution, if Formula 5.1 holds.

To better understand this definition, it is useful to think of m-concealing as a more general formulation of k-anonymity. Indeed, if we assume a set of columns to be known to the world with probability 1, and all the others to be unknown with probability 1, m-concealing corresponds exactly to k-anonymity, because all the knowledge states in which we know some of the secret columns, and/or we do not know some of the public columns, cannot occur, and are weighted by their zero probability in the summation. There will be only one term that in the summation receives a non-zero weight, and its

[1] With a slight overload of notation, r denotes the random variable corresponding to the output release being chosen.

weight will be one. We are now left with $\mathbb{P}[r_H^{(i)} = h|q]$, where q is the knowledge state in which we know the public columns (that we should start calling quasi-identifiers), and ignore the secret columns (the sensitive attributes). This single term corresponds precisely to the definition of k-anonymity, with $m = \frac{1}{k}$.

The assumptions that I have made are the following:

- An adversary can either have access to a full sensitive column or to no value at all for a sensitive attribute. This is of course a gross simplification, but it is required not to make the size of Q blow up and thus render the computation of this metric unfeasible. Hence, when modeling the probability of different knowledge states, a conservative approach must be followed. If one believes that it is possible that a single value for a sensitive attribute might be leaked with probability p, then the knowledge state of having access to the whole column should be assigned probability p.
- Knowing a sensitive column is statistically independent of knowing any other sensitive column.
- An adversary can only guess uniformly at random among the records that are indistinguishable with respect to the generalized quasi-identifiers and known sensitive attributes. An adversary cannot, on the other hand, use some background knowledge on the correlation between different attributes to improve the guessing probability. This is the sort of things that l-diversity and t-closeness address.

It is important to stress that saying that a release is m-concealing is *not* making a statement on the probability of an adversary discovering a sensitive value. For example, this condition holding does not contradict the possibility launching a successful homogeneity attack. Instead, it is meant as a modeling tool for the adversarial knowledge of sensitive attributes, and its implication on re-identification, with the goal of superseding k-anonymity's definition of equivalence classes, also used by l-sensitivity and t-closeness. These can in turn be built on a more solid foundation and then assert their specific additional requirements. Undoubtedly, this de-identification metric is going to make the claim on the pseudonymization of a dataset, more "reasonable", as required by the legal demands outlined in Section 2.1. Interestingly, m-concealing also satisfies the generalization property: if d is a dataset, and $r^{(1)}$ and $r^{(2)}$ are two generalizations of d such that $r^{(2)}$ is more general than $r^{(1)}$, if $r^{(1)}$ satisfies m-concealing, so thus $r^{(2)}$. To see that, consider the equivalence classes associated to a knowledge state, made up of records that are indistinguishable with respect to the columns in the knowledge state. These equivalence classes can only become larger by being merged through generalization. By increasing the pool of other indistinguishable records for some entries and some knowledge states, the value of m can only increase. As a consequence, it is possible to apply m-concealing to OLA and enjoy the benefits of predictive tagging. However, this has not been done in this project. Instead, m-concealing has been used to measure the degradation of the privacy of k-anonymous outputs given different assumptions on the background knowledge of an adversary.

For completeness, I will mention that there have been attempts at producing a more rigorous definition of privacy on tabular releases than those made available by the

combination of m-concealing and techniques inspired from k-anonymization. I am referring to *Bayes-optimal privacy* [32], based on the difference between the prior and posterior probability of inferring a sensitive attribute given a release. Given A the random variable whose distribution we are trying to model, and B, the random variable we are conditioning said distribution on, the posterior $\mathbb{P}[A|B]$ is computed by looking at all the equally likely disjoint random worlds [7] and dividing the frequency of the worlds satisfying $A \cap B$ by the number of those satisfying B. The resulting formula includes terms that to be computed need a knowledge of the joint distribution of quasi-identifier values and sensitive values in the *entire population*. These approaches suffer then from the following terrible drawbacks:

- If we assume that the adversary knows the true full distribution over attributes in the entire population, we do not know it anyway and cannot model it.
- If we assume that the adversary has some alternative inaccurate belief over what such distribution is, what is that belief? Many different adversaries will have many different beliefs about the distribution and picking one is an arbitrary choice.
- If we choose an approach similar to the one on which I have based m-concealing, and compute a sum in which terms are weighed by the probability of their associated knowledge state, we obtain a sum with an ungodly amount of terms. Each of these terms will reflect a possible belief about the joint distribution of column values. Even if all values were binary, the number of parameters needed to specify the full distribution would be exponential in the number of columns. Then, the number of knowledge states would be exponential in the number of parameters specifying the distribution.

5.3 Learning task

One of the interesting points to investigate is what the effect of anonymization is to concrete data analysis tasks, and how different information loss metrics are conducive to a better outcome. For this purpose, a dataset needs to be split into a training set, to fit a model, and a test set, for evaluation. The generalization/suppression algorithm is run exclusively on the training set. Records in the test set should not influence the selection of a generalization strategy. Furthermore, no suppression is applied to the test set, not to reduce it in size, which would make the validation of the model more fickle. As a consequence of the previous two points, it is very much possible that the test set will not be satisfying the criteria of privacy that our chosen anonymization algorithm enforces. The assumption is then that learning algorithms are to be trained on anonymized data, but run on non-anonymous data. However, in order to apply our trained model, the learning algorithm needs to operate on data with identical generalization. Hence, after a generalization node has been chosen, the test set gets generalized accordingly. It is of course possible to operate in a context in which the model is run on anonymized data, but that possibility is not explored because it is hard to investigate rigorously: the whole point of separating the data points into a training and a test set is that we want

to use the latter for evaluation, and fitting a model to it would be cheating. Similarly, here we aim to verify how well an anonymizing transformations fitted on training data generalizes to unseen data.

The UCI Adult dataset [6] has been chosen because it is an extremely well-known benchmark for machine learning classification models based on low-dimensional tabular data. No learning has been attempted on MIMIC-III at this stage, as existing benchmarks [22] rely on complete time-series. However, a set of good anonymization parameters has been identified on the basis of their performance on UCI adult and then applied to MIMIC-III, before measuring the resulting privacy gain. UCI Adult was extracted from the 1994 US Census database. It contains 48,842 entries (of which 16,281 are part of the test set), consisting of 14 attributes. One of the attributes has a binary value, signaling whether the income of the participant is above 50 thousand dollars per year. The task is, given the other attributes, to predict the value for *income*.

In the experiments that distinguish between sensitive attributes and quasi-identifiers, the quasi-identifiers have been considered to be the following:

age, work_class, education, marital_status, occupation, relationship, race, sex, native_country

All quasi-identifiers have been generalized according to the rules defined in Appendix B. The attribute *fnlwgt* contains information related to the calculation of the census and is irrelevant to the classification task, so it will not be included. The attribute *education-num* expresses the education level in years and is highly correlated with the categorical *education*, hence it will also be dropped. This means that the only attributes considered sensitive are:

capital_gain, capital_loss, working_hours_per_week

where the first two refer to investment sources, apart from wages/salary.

Experiments have been carried with logistic regression, a particularly apt baseline method for the classification of relatively low-dimensional data points. The implementation of the logistic regression classifier used in this work is based on [4]. In particular, numerical features are normalized by subtracting their mean and dividing by the standard deviation. Missing values (present in the training set) are filled with the value with the most occurrences in their column. Finally, categorical values are transformed to a *one-hot* encoding (a binary vector containing one entry per category and a value of 1 only on the correct category).

5.4 Computer environment

All experiments have been run on a machine equipped with a 2.3GHz Intel Core i5 processor and an 8-GigaBytes memory. The code for algorithms and experiments has been written in Python due to its prototyping speed and suitability for data science tasks, the latter being motivated by the vibrant landscape that produced the excellent library

Pandas [34], and a plethora of documentation. Python code has some drawbacks: it is dynamically typed and interpreted. As a result, these implementations are not very efficient if compared to their potential compiled counterparts. For instance, in the original paper [18], OLA is said to have significantly lower execution time than found in this project, regardless of more compute power and of a smarter k-anonymity check implementation. In most cases this extra slowness is not prohibitive for the purpose of experiments, and execution speed was not the prime object of investigation. Optimizations are naturally possible as efforts subsequent to this work. Some improvements might also include the parallelization of these algorithms, the capability of operating on SQL databases, and on data that is not entirely loaded in memory.

Chapter 6

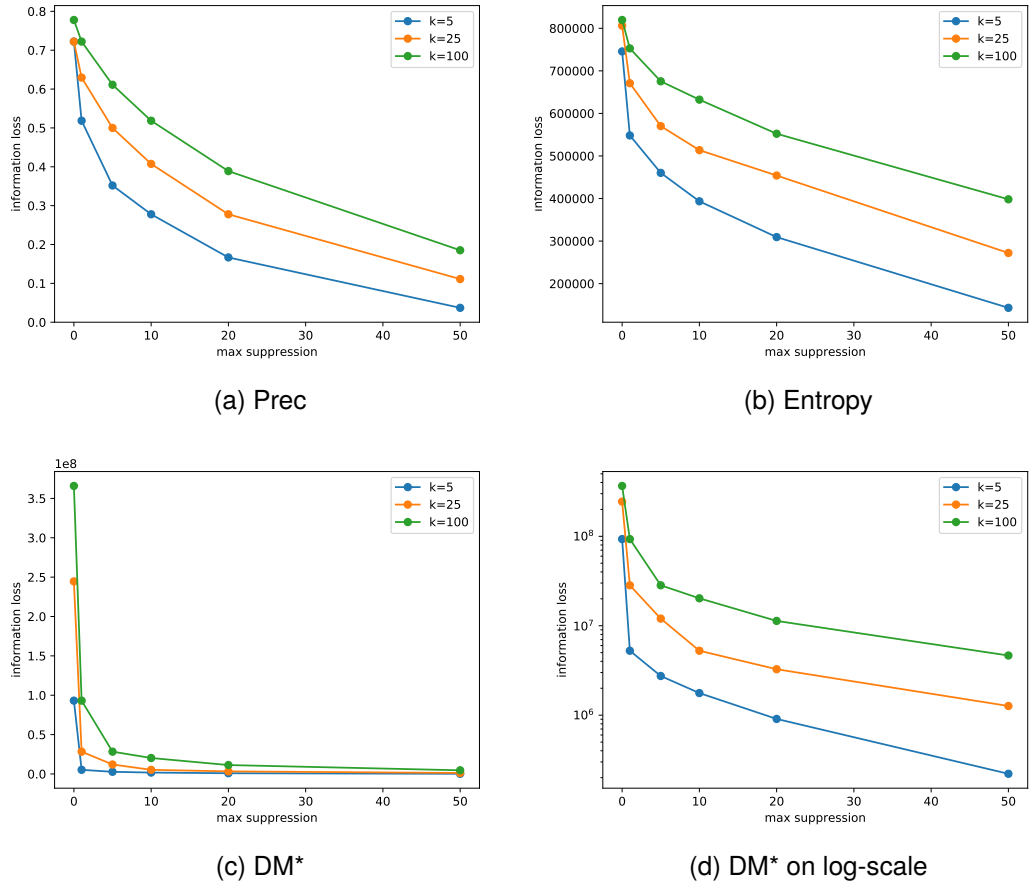
Experimental evaluation

6.1 Performance of OLA

6.1.1 Suppression versus information loss

The objective of this set of experiments was to explore the interaction between the suppression of records and the utility as represented by the k-anonymous node with lowest information loss, for different values of k and different loss measures. OLA has been run on the three information loss metrics listed in Section 5.1, with $k \in \{5, 25, 100\}$ and with $max\_sup \in \{0, 1, 5, 10, 20, 50\}$. The generalization rules applied can be reviewed in Appendix B. The results are shown in Figure 6.1, where Prec has been normalized so that the most general node will have loss equal to 1, and thus express information loss in a way that is independent of the number of quasi-identifiers (otherwise, the maximum value for information loss would be exactly equal to the number of quasi-identifiers). Furthermore, for DM* information loss has been calculated as the difference between the metric on the original dataset and on the chosen node. In other words, we are interested in the value of the metric on the non-anonymous dataset, and on how information is degraded by anonymization, not just in the absolute value of the metric on the output node. Entropy does not need this measure applied to it, because in its formula it already factors in the original dataset, and neither does Prec, whose root node has always a loss equal to zero. These adjustments make the results more easily comparable across different datasets, beyond just helping to select a node among various options.

DM* appears as the most extremely sensitive of the metrics, with an initial plunge of the information loss, and its curve then settling around $max\_sup = 5$. This is because it is calculated as a sum of squares of the size of equivalence classes. As a result, subtracting the value at the root node has only a negligible impact. For ease of comparison, the results of DM* are also plotted with a log scale, which makes the shape of its curve not too visually dissimilar from Entropy's. In general, all the metrics have a very large difference between the information loss with no suppression and that with $max\_sup = 1$, suggesting that even renouncing to that comparatively low number of entries could be



(a) Prec

(b) Entropy

(c) DM*

(d) DM* on log-scale

Figure 6.1: Relationship between maximum allowed suppression and information loss of the node chosen by OLA on UCI adult, with respect to three information loss metrics.

a worthwhile trade-off, for instance in the case of a small dataset in which a higher suppression is not tolerable. An explanation for this could be the removal of outliers, that only with large generalization could be hidden in appropriately-sized equivalence classes. Because of their rarity, the majority of them gets discarded even with low suppression, and the resulting gain in information loss becomes less pronounced with higher suppression. The difference between the first and second recorded max_sup points is especially marked in the case of a small k (k=5). Presumably, for large values of k, a 1% suppression is not sufficient to suppress all those equivalence classes that might be large, but still smaller than k, whereas this becomes feasible for a small k. Finally, it can be noted that for Entropy, the difference in loss for different values of k increases with max_sup, more regularly and more sharply than with Prec. The consequence of this "divergence" of the two curves entails that it will be easier for Prec to prefer a node with a larger k and a larger suppression. As an example, see how Prec is indifferent in the decision between the node for $k = 25$, $max\_sup = 20$, and for $k = 5$, $max\_sup = 10$). It is of course possible, in theory, that this judgement might be sensible. A larger value of k does not imply that most equivalence classes are that large, and the larger value of suppression might allow for a generalization in which most classes are on average smaller than for $k = 5$. However, it seems probable looking

k	max_sup	Prec	DM*	Entropy
25	0	0.802	**0.8227**	**0.8227**
25	10	0.8057	0.8155	**0.8426**
25	50	**0.8456**	0.8226	0.8432
100	0	0.802	**0.8215**	0.8052
100	10	0.8058	**0.8136**	**0.8136**
100	50	**0.8315**	0.8049	0.8255

Table 6.1: Classification accuracy for income on UCI Adult, after training the model on the anonymized training set, for different information loss metrics. The value in bold represents the best accuracy for some values of k and max. suppression

at Entropy's graph that this is not the case. Entropy, unlike Prec, does keep track of the size of equivalence classes, and appears to be making a different assessment of those two data points, clearly preferring $k = 5$. This is not a conclusive indictment of Prec: it might be the case that for a particular task, the relative depth of generalization is more important to consider than the size of equivalence classes. This needs to be specified at the outset, though, to pick a metric in a sensible manner.

6.1.2 Information loss and classification accuracy

In order to measure the quality of the decisions taken using different information loss metrics, the UCI Adult income prediction classification task has been performed before and after anonymization with different parameters. Results are shown in Table 6.1. On the original dataset, the classification accuracy of the model is 0.8519.

This comparison shows no definitive winner. DM* and Entropy tend to give better results with moderate suppression, with Entropy achieving the best performance, except in the case of $k = 100$ and no suppression. This changes when we allow up to 50% suppression. In that case Prec positively surprises, to the point that it manages to obtain the first and third best results across all runs. I will advance a hypothesis to explain this phenomenon: more suppression unlocks better nodes that could be output, but while Prec makes a choice that is agnostic of the actual generalized data, Entropy and DM* will heavily rely on it, and in particular they will degrade information on entire columns faster than Prec, with the promise of smaller equivalence classes. However, the promise is broken when suppression is applied. As a reminder, suppression is not taken into account by these metrics so as to make predictive tagging work, under the assumption that suppressing records is not going to harm utility. With a very large allowed suppression, this assumption becomes very misguided, and Prec manages to catch up. In general though, the amount of suppression does not seem to affect much accuracy, and in some cases it is actually correlated with a better performance of the classifier (as a result of less generic releases becoming viable). Similarly, and perhaps defying intuition, while there is a correlation between the increase of k and the decrease of accuracy, very often that is less significant for the purpose of classification than the impact of picking the right or wrong loss metric. In all cases the accuracy of the classifier stays above 80%. This is a very promising result showing that it is not impossible to increase the level of protection of individuals and still perform useful inference.

A less promising result can be read from Figure 6.2, showing how classification accuracy

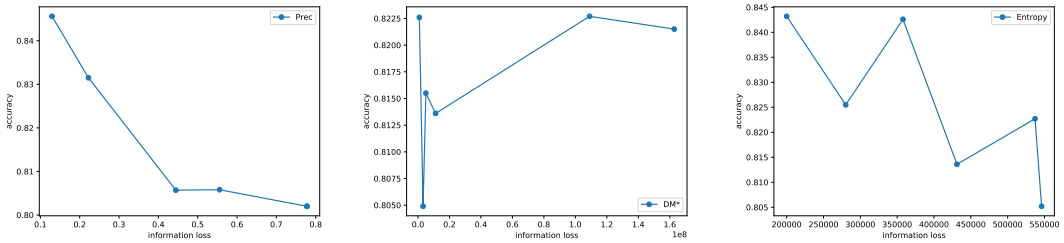


Figure 6.2: Relationship between information loss value for a node and resulting classification accuracy. From the left: Prec, DM* and Entropy.

varies with information loss. DM* looks completely unpredictable. Entropy seems to be exhibiting a clearer pattern, but there are not enough data points to make a rigorous statement, and anyway it still presents a decisively non-monotonic behavior. Prec is the only one with consistent repercussions on accuracy, which makes it the most useful metric for comparing utility across different sets of parameters, and possibly across datasets, and not just for selecting a node in a single run of OLA.

6.1.3 Evaluation on MIMIC-III

Chapter 3 highlighted that the release of MIMIC-III does not conform to k-anonymity if not for $k = 1$, even for a limited collection of fields. Experiments have been performed to apply the insights gained in this section so far to the anonymization of MIMIC-III. The algorithm Inverse OLA has been used, with values for the maximum information loss that have given acceptable results in Section 6.1.2, and three suppression settings: none, medium (no more than 10%), and large (no more than 50%). Only the metric Prec has been used, because previous experiments have shown that neither Entropy nor DM* seem to have the potential to to provide useful guidance on the information loss to require for a new dataset. Furthermore, Prec has a very simple interpretation and is the fastest to run. The objective was to successfully de-identify the union of patient demographics and a small sequence of procedures. The attributes for demographics are the same as those listed in Section 3.1, with the year of birth sampled from a normal distribution with mean 1980 and standard deviation set to 33. The sampling has been performed only once to construct the table, which has been reused for all runs. This table has then been joined with the first three procedures for a particular admission of each patient, to construct a final table consisting of nine fields. All rows with missing values were dropped, leaving 13,646 entries in total. The results are summarized in Table 6.2

For a value of Prec equal to 0.2, reasonable k-anonymity has been found to be unobtainable, regardless of suppression, given the dimensionality of the data. An information loss not larger than 0.5 has allowed to obtain a reasonable value of k (8) for medium suppression, and a very good value of k (228) for large suppression. If bringing the information loss up to 0.8 was considered acceptable, it would be possible to obtain an extremely large value of k (313) even with zero suppression.

Max. loss	Max. suppression	k
0.2	0	1
	10	1
	50	1
0.5	0	1
	10	8
	50	228
0.8	0	313
	10	1009
	50	7003

Table 6.2: Best value of k obtained using Inverse OLA, with Prec as the information loss metric. Results are reported for different maximum tolerated information loss and suppression on the joint view of patient demographics and three procedures.

A qualitative inspection of the anonymization produced by these experiments, however, reveals that in all cases in which k was larger than 1, the procedures have been maximally generalized, and thus essentially scrapped. So effectively the inclusion of procedures in the discourse was only apparent. The generalization hierarchy of procedures was defined as following: the original ICD9 procedure, code, its first digit, and complete suppression. In order to try and push Inverse OLA to sacrifice some demographic information to keep information about procedures, a modified version of Prec that allows passing a per-column weight has been used. This attempt has not produced any improvement. Even with very heavy weights for procedures, the best that could be accomplished was retaining the first code digit for one of them. The next step was trying to reduce the dimensionality of demographics, to only include gender and date of birth. With all the parameters combinations in Table 6.2, only in two cases the release did not suppress procedures completely, and had k>1. These are summarized in Table 6.3.

Max. loss	Max. suppression	k	Information about procedures
0.5	10	12	First code digit for the first two, completely suppressed the third
0.5	50	104	First code digit for all three

Table 6.3: Best value of k obtained using Inverse OLA, Prec and given maximum tolerated information loss and suppression on the joint view of patient demographics and three procedures.

6.1.4 Protecting fields with m-concealing

If both demographics and procedures were necessary to a data analysis task, none of the k-anonymous releases in Table 6.2 would work well, regardless of the value of Prec. Additionally, the releases in Table 6.3, leave many features behind. We would then be forced to decide between two alternatives. We could release a version of the dataset that does not satisfy k-anonymity, or not consider procedures as quasi-identifiers. While these solutions are both undesirable, the latter is the lesser of two evils. However, there could be a better option: exploiting m-concealing to relax the privacy guarantee that considers procedures to be quasi-identifiers, and on the other hand strengthen the

assumption that they are not quasi-identifiers at all. To do so, a value has been assigned to the probability of a procedure being known by an adversary, and Inverse OLA has been run again with the same sets of parameters as before. Table 6.8 presents the result for different assumptions on such probability. In "Model A", the probability of knowing each of the columns has been set to 0.01, whereas in "Model B" it has been set to 0.05. In "Model C", it has been deemed more likely to know the first procedure (0.05) than the last two (0.01). Finally, "Model D" tried to be more ambitious and consider, besides the demographics, the first five procedures of a patient admission (all known with probability 0.01). A new table with these features has been constructed as before, containing 8,351 records. For the date of birth, the first 8,351 samples from the previous draw have been reused. The value reported is actually the reciprocal of of m as defined in Section 5.2, in order to facilitate comparison with k, and has to be roughly interpreted in a similar way, i.e. as the minimum number of entries that "hide" any particular record in the dataset. Because of the weighing, it is often not an integer, and it has been rounded to the first decimal. Additionally, only runs where $1/m$ was larger than 5 have been reported.

Table 6.4: Model A

Max info. loss	Max sup.	$1/m$
0.2	50	27.2
0.5	10	27.4
0.5	50	33.3
0.8	0	33
0.8	10	33.2
0.8	50	33.5

Table 6.5: Model B

Max info. loss	Max sup.	$1/m$
0.2	50	6.7
0.5	10	6.7
0.5	50	7
0.8	0	7
0.8	10	7
0.8	50	7

Table 6.6: Model C

Max info. loss	Max sup.	$1/m$
0.2	50	16.5
0.5	10	17.1
0.5	50	20.2
0.8	0	20
0.8	10	20.1
0.8	50	22.5

Table 6.7: Model D

Max info. loss	Max sup.	$1/m$
0.2	50	5.2
0.5	10	5.2
0.5	50	5.3
0.8	0	5.3
0.8	10	5.3
0.8	50	5.3

Table 6.8: Result of computing m-concealing on the output of Inverse OLA run on demographics, while modeling the probability of knowing procedures, for different models.

It is interesting to observe that the same combinations of maximum information loss and maximum suppression are the ones that lead to satisfactory releases, regardless of the model. Additionally, probably because procedures did not influence the output of Inverse OLA (the m-concealing value did not guide the choice of a node), the magnitude of $1/m$ does not vary much given a model. In conclusion, m-concealing seems to have successfully served its purpose. It allowed to discard the releases which could have created a risk given the potential for some sensitive attributes to be leaked. Conversely, it strengthened the confidence of the claim that the other releases had been

de-identified. Finally, do keep in mind that applying m-concealing directly to OLA (or Inverse OLA), instead of computing it on a pre-chosen release, would probably lead to an even more favorable outcome. This would mean substituting the k-anonymity check with an m-concealing check, for some m supplied as input.

6.2 Performance of ϵ-safe LA

6.2.1 Selection of parameters

Potential combinations of parameters and their effect on Differential Privacy under Sampling have been computed as described in Section 4.5.2. Because of the many parameters, a hierarchical approach has been adopted. Firstly, several lower bounds on ϵ have been correlated to, given different options for β, the required value for ϵ' (the parameter for the exponential mechanism). This has been done to understand the range of sensible values for ϵ'. However, not to decrease the probability of sampling useful nodes by an unreasonable amount, parameters requiring a negative ϵ' have not been taken into consideration. For β, the minimum value considered was 0.01, corresponding to a 1% probability of inclusion. This is considerably smaller than the maximum allowed suppression seen in the previous section, and has been chosen to widen the spectrum of values considered, given the lack of understanding on the behavior of ϵ' and its interaction with the other parameters. The maximum value for β was 0.9. As for ϵ, values ranged from 0.1 to 2, taking inspiration from non-extreme suggestions in [11]. The resulting ϵ' range was in $[0.01, 2]$[1], with $\epsilon' > \epsilon$ in all cases. Afterwards, the corresponding value of δ has been computed for different choices of k, and options that gave $\delta > 10^{-4}$ have been discarded, as prescribed in [42]. 50 was the smallest k found to satisfy this criterion for all simulated parameters. Note that, chosen in this order, k has no influence on ϵ, which intuitively represents the severity of a privacy breach, before the relaxation term δ is applied. Also note how a smaller β will lead to a certain decrease in the amount of records, but it is going to decrease the lower bound on ϵ, and by consequence make larger values of ϵ' viable, that in turn will increase the probability of sampling a node with a reasonable amount of suppression. Hence, it was initially unclear how β affects the number of output entries on average.

6.2.2 Comparison of penalties

The algorithm ϵ-safe LA has been executed on UCI Adult with several ϵ' between 0.01 and 2, with $k = 50$, and with different values p for the penalty factor. Two modes have been tried: in one the penalty factor was multiplied by the percentage of suppressed records, and in another it was multiplied by 1 for non-k-anonymous nodes and by zero otherwise. It is to be expected that smaller penalty factors will be more effective with the first mode, as the scale of the penalty to multiply the penalty factor by is larger. The penalty term thus calculated was added to the Normalized Prec value for

[1] A tighter upper bound would actually be slightly smaller, between 1.99 and 2.

each node, and the multiplicative inverse of the result was assigned as the utility of each node. Because the root node was never k-anonymous, its utility did not blow up with the inverse as a consequence of the information loss being 0. The sensitivity is calculated as the difference in utility between neighboring databases on the same node. This is exactly equal to the penalty factor in the case of k-anonymity-based penalty. For suppression-based penalty, the sensitivity is the penalty factor multiplied by the percentage of the database corresponding to one record. A fixed maximum suppression of 5% was applied to all runs, to increase the pool of penalty-free nodes. For $\beta = 0.75$, 10 samples of the dataset have been drawn and stored. Then, for each combination of parameters, the results reported are the average over the algorithm's output for each of the samples. To contain the running time of these experiments, a smaller subset of columns has been considered, giving rise to a lattice of 108 nodes in total. As a reminder, all columns need to be quasi-identifiers in the ϵ-safe framework.

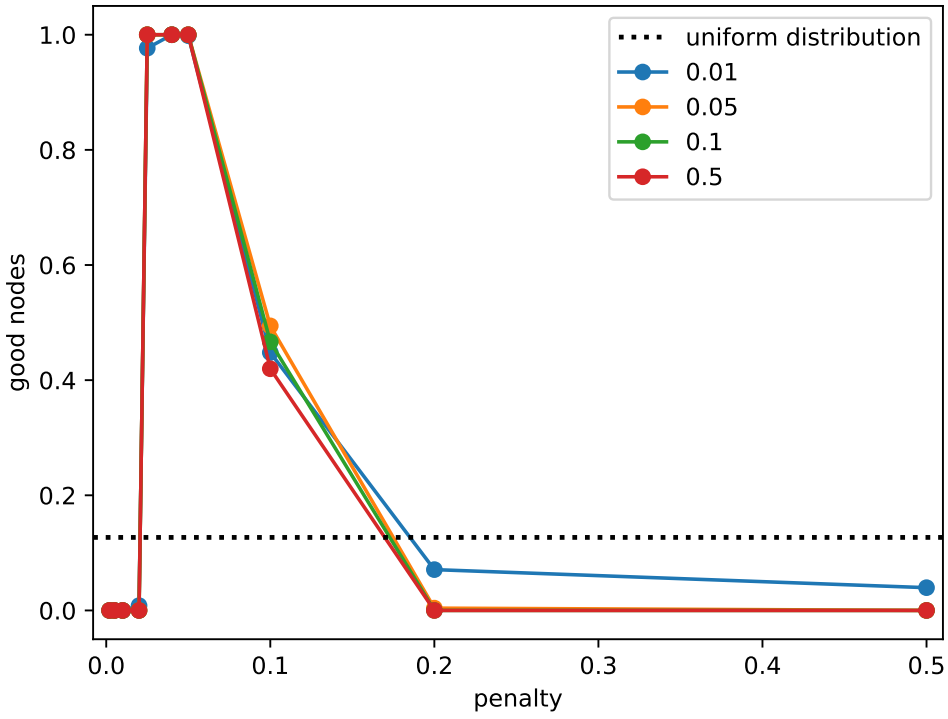


Figure 6.3: Relationship between penalty factor and total probability of outputting a "good" node for suppression-based penalty. Each line represents a different setting of ϵ', with the exception of the black dotted line representing the baseline.

For the purpose of measuring the quality of the output on average, "good" nodes have been defined to be those nodes whose Normalized Prec value is below 0.5, and suppression is below 5%. Then, the probability of sampling a "good" node is equal to the sum of the probability of sampling any of them. Figures 6.3 and 6.4 shows the way this probability varies with different penalty factors, for the two modes and for several settings of ϵ'. In both the graphs, a baseline has been marked with a dark dotted line, consisting of the probability of sampling a "good" node if the output were not chosen

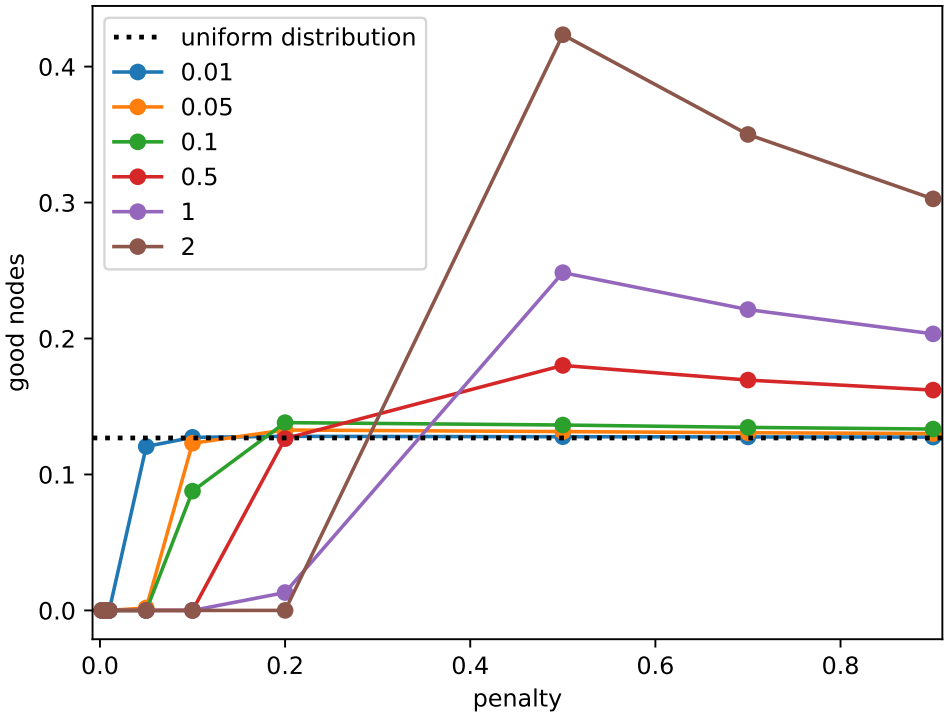


Figure 6.4: Relationship between penalty factor and total probability of outputting a "good" node for k-anonymity-based penalty. Each line represents a different setting of ϵ', with the exception of the black dotted line representing the baseline.

via the exponential mechanism, but uniformly at random in the lattice. Choosing the output uniformly at random is ϵ'-safe for (at worst) the same ϵ'.

It is clear that one of the main strengths of the suppression-based mode is its extremely small sensitivity, which allows the ratio of the probability of two nodes to be very large when they have a non-negligible difference in utility. Indeed, due to just how large the unnormalized probabilities output by the exponential mechanism have been found to be, it was necessary to use a software library that operates on arbitrarily-large integers, to avoid overflow problems. It can be seen that the effect of different ϵ' on the quality of the output nodes is very small when using suppression-based penalty. This result is interesting and counter-intuitive. What is probably happening is that, given any two nodes with "dissimilar" utility u_1 and u_2, the small sensitivity Δu makes the term $\epsilon' \frac{u_1 - u_2}{2\Delta u}$ very large, regardless of ϵ', because of the difference in magnitude between the sensitivity (of the order of 10^{-5}) and ϵ'. This term is then exponentiated, to compute the ratio between the probability of sampling either one of the two nodes. This ratio will then become huge. For nodes with closer utilities, ϵ' might have a larger impact, but it does not really matter which node we sample among the ones that have similar utility from the point of view of the metric being used, i.e. the probability of sampling "good" nodes. Instead, the quality of the output is extremely susceptible to the penalty factor. The graph for suppression-based penalty shows an initial, very sudden increase

of the percentage of "good" nodes way above the baseline, reaching almost 100% for a 0.04 penalty factor. The curve then steadily drops again and returns below the baseline when the penalty factor is set to 0.2. This is because the suppression-based penalty is initially useful not to choose nodes purely on the basis of a small value of Prec, but eventually overwhelms the criterion of utility, and nodes with an extremely low suppression are chosen even with a terrible information loss.

The other utility metric is way more reliant on ϵ', and even when that is set to 2, a fairly large value, the results are not as exciting, with a maximum probability of sampling a node with a good utility slightly over 0.4. It also must be considered that 2 is not an allowed value for ϵ' if, as in this case, β is 0.75 and we want ϵ to be smaller than 2. With the acceptable combinations of parameters computed in Section 6.2.1, the largest tolerable value of ϵ' (for $k = 50$) is approximately 1.99, with $\epsilon = 2$, and a much smaller β, namely 0.01. At any rate, 0.5 was the penalty factor which worked the best, beating the baseline, if marginally, in all cases.

6.2.3 Suppression versus probability of inclusion

Earlier, the question was asked of whether a larger β could increase the size of the output on average by allowing the penalty to contain suppression without varying ϵ. Consider the penalty based on whether a node was k-anonymous. Across all runs, the average suppression (the sum of the suppression of all nodes weighed by their probability) was lowest with the penalty factor equal to 0.9, and $\epsilon = 2$. This lowest average suppression was only less than 3.7% of records lower than the baseline, and only another 1% of records less than other points obtained with e.g. $\epsilon' = 0.01$. However, the lower value of ϵ' would allow, given the same ϵ, sensibly higher values of β. The baseline average suppression was low to begin with (about 1/10), but regardless, this observation seems to suggest that with this type of penalty, one needs to vary β by a large amount to obtain only a small gain in suppression.

As for the suppression-based penalty, as previously observed, it is essentially invariant to changes of ϵ', and thus also practically invariant to changes of β, whose effect, given fixed ϵ, can be absorbed at least to a large extent by smaller values of ϵ'. In essence, it seems that varying β should not be regarded as a tool for increasing the output quality, but only as a means to make ϵ as tight as possible, when releasing only a sampled subset of the data is known not to harm utility, at least up to a certain point. Figures describing the relationship between the penalty factor, ϵ' and suppression can be viewed in Appendix C.

6.3 Computational considerations

As already discussed, the analysis of the computational requirements of the algorithms used for these experiments was not a primary goal of investigation of this paper. Additionally, the running time of experiments is hard to interpret because of oscillations in the execution speed due to, among other reasons, competition of other applications

max_sup / ***k***	0	1	5	10	20	50
5	127	1267	2418	2058	1303	260
25	98	549	1308	2103	1869	773
100	55	215	635	1121	1657	1114

Table 6.9: Heat map representing the number of checked nodes when running OLA on UCI Adult, for different values of k and max_sup

on the machine used, and the amount of battery left. However, the following findings can be reported:

- **OLA on UCI Adult**: with the same settings as in Section 6.1.1, the k-minimal nodes where found after checking as many nodes as specified in Table 6.9, for different settings of k and max_sup. The table shows that for small k (5), medium amounts of suppression (5 and 10 percent) make predictive tagging the least effective. For larger k, the suppression that maximizes the number of checked nodes is shifted towards larger values, and on average the number of checked nodes decreases. Presumably, a more balanced split of k-anonymous versus non-k-anonymous nodes means more nodes must be checked. If Prec was used, the total running time was almost never more than 2 seconds per checked node. When looking at Entropy and DM*, both the checked nodes and the size of the k-minimal set are worth considering, because the computation of information loss is not trivial, and needs iterating over all records in a release associated to a node. In most cases, the running time for these two metrics was below 2 seconds times the number of checked nodes plus the nodes in the k-minimal set. No execution of OLA took more than 4 hours to run.
- **Inverse OLA**: as reported in Section 6.1.3, Inverse OLA was only run using Prec, and the statement in the previous bullet point applies here too: the running time was bounded from above by 2 times the number of checked nodes. Indeed, on average, the ratio between total running time and number of checked nodes was often smaller, and in some cases equal to one half or even lower. A very reasonable explanation can be formulated: even if the algorithm does not have to check all the nodes that were predictively tagged, navigating through all the generalization paths and visiting all their nodes can be lengthy because of the size of the lattice. The size of the lattice for UCI Adult was indeed larger than that for MIMIC-III, by a factor of 1.75.
- **m-concealing on MIMIC**: with the same settings as in Section 6.1.4, the computation of m-concealing did not ever require more than 4 seconds, for any of the combinations of fields inspected.
- **ϵ-safe LA**: during the experiments described in Section 6.2.2, after the utility of all nodes was computed, calculating the probabilities of all nodes and sampling one took under 3 seconds. The time required to compute the utility for a node was comparable to that of checking whether it was k-anonymous. Naturally, the suppression-based penalty was considerably slower as it needed to compute the suppression for all nodes.

Chapter 7

Conclusion

7.1 Discussion of experiments

The experiments conducted have highlighted how achieving k-anonymity exclusively by generalization is connected to an unnecessary loss of granularity. On the other hand, additionally suppressing records is going to markedly relax the generalization requirements, at the price of renouncing to outliers. In other words, a trade-off between the richness of the data in terms of variety and granularity has been identified. For the case in which both elements are precious, results have shown that the fastest gain in information loss based on granularity is obtained by going from no suppression to a small suppression, indicating how this might be a very valuable compromise. It must be noted however that the utility of a small suppression decreases for larger values of k. This also has to be taken into account when devising a release strategy. Also note how suppressing a small number of outliers might mean removing anomalies from the data and thus increasing its quality.

A comparison between different information loss metrics has lead to results whose interpretation is uncertain. It seems evident that this is also a case in which the choice must be guided by the judgement of the custodian, informed by the specific data analysis needs. Empirically, Entropy and DM* allowed for the best classification accuracy when moderate suppression was used. However, values of Prec can be explained intuitively, and it can be tweaked to give different weights to different fields when necessary (even though attempts of doing so in experiments have been inconclusive). Nothing excludes that different metrics might be used in conjunction, either by having both contribute to the assessment of nodes and "voting", or alternatively leading to different releases that might be manually inspected by the custodian before a final choice is made.

The power of OLA has been demonstrated. It was possible to obtain at least 80% classification accuracy on UCI Adult, after training on k-anonymous releases for k as large as 100. Even with an implementation of OLA that had not been optimized (as mentioned in Section 5.4), the running time was measured in no more than a few hours for a table consisting of tens of thousands of entries, which is a totally acceptable time for a custodian, given the assumption that releasing datasets is a sporadic activity. The

caveat is the unpredictability of such running time: depending on the parameters it varied from a few minutes up to 4 hours, on the same dataset. Additionally, it can be speculated that using deeper generalization hierarchies than in this paper, or using more quasi-identifiers, will make the required running time explode to reach days, weeks, and eventually make OLA's computation unfeasible. Even so, the generalization hierarchy used for these experiments was reasonable, and using more quasi-identifiers is very likely to be unfeasible for reasons deeper than computational: k-anonymity's curse of dimensionality. What is really important is that the algorithm scales well with the size of the input dataset. In this implementation, it scales linearly.

Inverse OLA has been introduced and shown to work well, covering the otherwise neglected use case of custodians wanting to protect the privacy of individuals as much as possible, but only after some minimum information loss target is achieved. Its main difficulty is closely intertwined with the difficulty of translating the value of an information loss metric across datasets and tasks. In particular, Entropy and DM* have been found not to be apt for the job, at least in the experimental scenario considered, because of the lack of correlation between their value and the success of income prediction on UCI Adult. Prec seems more suitable, but its value needs to be taken with caution, as it ignores the relative importance of fields in a determined task. For example, a seemingly appropriate value of Prec in Section 6.1.3 was deceptive, as it was associated with a complete suppression of information about procedures. More work needs to be done to identify an appropriate measure or set of measures that can allow Inverse OLA to be useful in practice.

In Section 6.1.3 it was also seen how the curse of dimensionality is aggravated by the size of the domain of a column's values, and by their distribution. Despite of the fact that the quasi-identifiers of MIMIC-III (basic demographics and three procedures) were nine, as many as those used for UCI Adult, it was so much harder to enlarge the equivalence classes for MIMIC-III without nullifying the values of procedures, because of the scarcity of entries mapped to the same procedure. The severity of this scarcity compounded when considering the three-elements time series. For this reason, in a k-anonymization setting, the procedures would traditionally be modeled to be unquestionably secret. This problem could be superseded by applying m-concealing, that as demonstrated manages to increase the confidence in the effectiveness of k-anonymization for the preservation of privacy, while readjusting the expectations on de-identification to a more realistic level. This holds provided it is possible to give a reasonable estimate of the probability of knowing sensitive columns, which cannot be taken for granted. In the case of a known breach of sensitive values, m-concealing can also be used to determine which individuals have been affected with high probability.

As explained in Section 2.1.3, when anonymizing data, it is paramount for institutions to rely on precedents so as to strengthen the legitimacy of their legal position. While there is a significant amount of precedents for k-anonymity, which facilitates the selection of different widely-accepted values of k, it is more arduous to justify, for differential privacy, a value of ϵ, that anyway does not have any explicit intrinsic meaning [11]. In this sense, ϵ-safe k-anonymization could allow organizations to both operate in the realm of legal certainty, and at the same time provide the extra guarantee of differential privacy to the data subjects whose privacy is under threat. It is worth mentioning again

that the benefits of ϵ-safe LA are principally technical rather than legal: it accomplishes the admirable result of achieving differential privacy without falsifying records, and outputting a tabular release that can be manipulated with ease. It also does so, as shown in Section 6.2, while preserving a considerable amount of utility. This is a very important result, only overshadowed by the computational difficulties of calculating the utility of all nodes in a lattice.

7.2 Related work

The authors in [20] demonstrate that it is sometimes possible to exploit the specific needs of an application to improve its performance after anonymization. They evaluate a custom top-down k-anonymization method on UCI Adult, and show that the consequent loss of information can boost the accuracy of a classifier, as it makes the learned model more robust by means of de-noising of the training set.

Linking together k-anonymity and differential privacy has also led to the definition of (k, ϵ)-anonymity in [24]. Unlike $(\epsilon, \delta, \beta)$-DPS, the authors apply k-anonymity to some quasi-identifiers, and enforce ϵ-differential privacy on the remaining ones. They also leverage the clusters produced by k-anonymity to reduce the noise applied via differential privacy.

The only concrete application of differential privacy to healthcare that I am aware of was introduced in [51]. The authors attempt to solve the problem of identifying cohorts of patients to recruit for clinical trials. They use a modified version of the exponential mechanism to obtain, given an input dataset, a differentially private count of the eligible patients. This would allow a researcher to gauge the suitability of a site for a given trial. The traditional problems of setting and distributing a total privacy budget are not resolved by the authors.

Alternatives to the privacy-preserving techniques dealt with in this project could be organizational rather than technical. For example, there is a speculation [33] that it will become commonplace to train and test models against non-disclosed medical databases. Analysts could submit their models to the data holder, with the resulting simplifications with respect to data management, compliance with regulations and privacy-protection. In this scenario, tools such as Synthpop [38] (mentioned in Section 3) would become increasingly important to produce fictitious releases that allow at least to have a direct access to the shape, type and distribution of the data while designing models. This "flipped" approach could be facilitated if, as suggested by [41], the field of health informatics will be able to build prediction systems that are robust enough to be transferable across sites and contexts, once they have been trained on sufficient amounts of data. The approach that attempts to relate the environment in which the data is kept, and the way it is transferred, to the resulting privacy guarantees, is called *functional anonymization* and is introduced in [19].

7.3 Future work

Building on this paper, an interesting direction for future work regards the exploration of l-diversity or t-closeness algorithms based on m-concealing, and their implementation with a modification of OLA. The combination of mechanisms that guard against privacy breached through statistical inference, and of a metric which attempts to protect against external information or leaks, are bound to be extremely powerful. This should be verified experimentally, and the consequences for the quality of data should be investigated.

ϵ-safe LA is arguably the most fascinating product of this work. It has been shown to be a very promising technique that combines the strong definition of differential privacy and the convenience of a k-anonymization algorithm and release. However, more work is needed to understand the calibration of its many parameters and its connection to utility. In particular, setting the β probability of inclusion and the penalty factor has been shown to be both a subtle and crucial matter. Whether it is possible to use the idea behind Inverse OLA and allow for different possible values of k in the ϵ-safe framework should also be explored. Additionally, the problem of the cost of computing the utility for all nodes needs to be addressed. Some potential solutions might be based on a random choice of the nodes on which to precisely calculate utility, and an approximate estimation of their neighbors.

Fundamentally, this paper has discussed a variety of techniques that can be used for the de-identification of a limited number of features of Electronic Health Records. The second part of this project will need to address the issues regarding the anonymization of very high-dimensional time series, in combination with demographic information. This could be attempted using either a structured technique such as that described in [45], or by learning a representation of the patients that can meet some privacy target while allowing to perform some types of inference. It should then be attempted to evaluate anonymous results on a benchmark like [22], using deep learning to forecast e.g. the risk of mortality or the length of stay. As mentioned in Section 2.2, the precise filtering of the relevant records is paramount for the correctness of data analysis and it might collide with the anonymization techniques highlighted so far, in that the highly-granular features needed for filtering might be lost. This is another area that needs to be researched. Finally, the relationship between clinical concept extraction from text and anonymization needs to be looked into.

Bibliography

[1] Central Intelligence Agency - The World Factbook: United States of America. https://www.cia.gov/library/publications/the-world-factbook/geos/us.html. [Online, accessed: February 9th 2019].

[2] Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act privacy rule. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html. [Online, accessed: January 10th 2019].

[3] Regulation (EU) 2016/679 of the European Parliament and of the Council. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32016R0679. [Online, accessed: January 10th 2019].

[4] Animesh Agarwal. Logistic regression classifier on census income data. http://archive.is/gRGUC. [Online, accessed: February 24thth 2019].

[5] Charu C. Aggarwal. On k-anonymity and the curse of dimensionality. In *Proceedings of the 31st International Conference on Very Large Data Bases*, pages 901–909. VLDB Endowment, 2005.

[6] Arthur Asuncion and David Newman. UCI repository of machine learning databases, 1992, 2007.

[7] Fahiem Bacchus, Adam J. Grove, Joseph Y. Halpern, and Daphne Koller. From statistical knowledge bases to degrees of belief. *Artificial intelligence*, 87(1-2):75–143, 1996.

[8] Roberto J. Bayardo and Rakesh Agrawal. Data privacy through optimal k-anonymization. In *21st International conference on data engineering (ICDE'05)*, pages 217–228. IEEE, 2005.

[9] Inci M. Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K Jain, and Jiayu Zhou. Patient subtyping via time-aware LSTM networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 65–74. ACM, 2017.

[10] Raghavendra Chalapathy, Ehsan Zare Borzeshi, and Massimo Piccardi. Bidirectional LSTM-CRF for clinical concept extraction. *arXiv preprint arXiv:1611.08373*, 2016.

[11] Fida Kamal Dankar and Khaled El Emam. Practicing differential privacy in health care: A review. *Trans. Data Privacy*, 6(1):35–67, 2013.

[12] Thomas Desautels, Jacob Calvert, Jana Hoffman, Melissa Jay, Yaniv Kerem, Lisa Shieh, David Shimabukuro, Uli Chettipally, Mitchell D. Feldman, Chris Barton, et al. Prediction of sepsis in the intensive care unit with minimal Electronic Health Record data: a machine learning approach. *JMIR medical informatics*, 4(3), 2016.

[13] 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. ACM, 2003.

[14] Cynthia Dwork. An ad omnia approach to defining and achieving private data analysis. In *Privacy, Security, and Trust in KDD*, pages 1–13. Springer, 2008.

[15] Cynthia Dwork. Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*, pages 1–19. Springer, 2008.

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

[17] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[18] Khaled El Emam, Fida Kamal Dankar, Romeo Issa, Elizabeth Jonker, Daniel Amyot, Elise Cogo, Jean-Pierre Corriveau, Mark Walker, Sadrul Chowdhury, Regis Vaillancourt, et al. A globally optimal k-anonymity method for the de-identification of health data. *Journal of the American Medical Informatics Association*, 16(5):670–682, 2009.

[19] M.J. Elliot, C. Dibben, H. Gowans, E. Mackey, D. Lightfoot, K. O'Hara, and K. Purdam. Functional anonymisation: The crucial role of the data environment in determining the classification of data as (non-) personal. *CMIST work paper*, 2, 2015.

[20] Benjamin C.M. Fung, Ke Wang, and Philip S. Yu. Anonymizing classification data for privacy preservation. *IEEE Trans. Knowl. Data Eng.*, 19(5):711–725, 2007.

[21] Jeffrey E. Gotts and Michael A. Matthay. Sepsis: pathophysiology and clinical management. *Bmj*, 353:i1585, 2016.

[22] Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *arXiv preprint arXiv:1703.07771*, 2017.

[23] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[24] Naoise Holohan, Spiros Antonatos, Stefano Braghin, and Mac Aonghusa. (k, ϵ)-anonymity: k-anonymity with ϵ-differential privacy. *arXiv preprint arXiv:1710.01615*, 2017.

[25] Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, H. Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3, 2016.

[26] Daniel Kifer and Ashwin Machanavajjhala. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 193–204. ACM, 2011.

[27] Matthieu Komorowski, Leo A. Celi, Omar Badawi, Anthony C. Gordon, and A. Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720, 2018.

[28] Kristen LeFevre, David J. DeWitt, and Raghu Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 49–60. ACM, 2005.

[29] Kristen LeFevre, David J. DeWitt, and Raghu Ramakrishnan. Mondrian multidimensional k-anonymity. In *Proceedings of the 22nd International Conference on Data Engineering*, pages 25–25. IEEE, 2006.

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

[31] Ninghui Li, Wahbeh Qardaji, and Dong Su. On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In *Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security*, pages 32–33. ACM, 2012.

[32] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data*, 1(1), March 2007.

[33] David M. Maslove, Francois Lamontagne, John C. Marshall, and Daren K. Heyland. A path to precision in the ICU. *Critical Care*, 21(1):79, 2017.

[34] Wes McKinney. pandas: a foundational python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, 14, 2011.

[35] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science*, pages 94–103. IEEE, 2007.

[36] Arvind Narayanan and Edward W. Felten. No silver bullet: De-identification still doesn't work. *White Paper*, pages 1–8, 2014.

[37] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy*, pages 111–125. IEEE, 2008.

[38] Beata Nowok, Gillian M. Raab, Chris Dibben, et al. synthpop: Bespoke creation of synthetic data in R. *Journal of statistical software*, 74(11):1–26, 2016.

[39] Aaditya Prakash, Siyuan Zhao, Sadid A. Hasan, Vivek Datla, Kathy Lee, Ashequl Qadir, Joey Liu, and Oladimeji Farri. Condensed memory networks for clinical diagnostic inferencing. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[40] Niranjani Prasad, Li-Fang Cheng, Corey Chivers, Michael Draugelis, and Barbara E. Engelhardt. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. *arXiv preprint arXiv:1704.06300*, 2017.

[41] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1(1):18, 2018.

[42] Aaron Roth and Tim Roughgarden. Interactive privacy via the median mechanism. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 765–774. ACM, 2010.

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

[44] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604, 2018.

[45] Lidan Shou, Xuan Shang, Ke Chen, Gang Chen, and Chao Zhang. Supporting pattern-preserving anonymization for time-series data. *IEEE Transactions on Knowledge and Data Engineering*, 25(4):877–892, 2013.

[46] Huan Song, Deepta Rajan, Jayaraman J. Thiagarajan, and Andreas Spanias. Attend and diagnose: Clinical time series analysis using attention models. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[47] Latanya Sweeney. Datafly: A system for providing anonymity in medical data. In *Database Security XI*, pages 356–381. Springer, 1998.

[48] Latanya Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):571–588, 2002.

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

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[51] Staal A. Vinterbo, Anand D. Sarwate, and Aziz A. Boxwala. Protecting count queries in study design. *Journal of the American Medical Informatics Association*, 19(5):750–757, 04 2012.

[52] Yanshan Wang, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul Kingsbury, and Hongfang Liu. A comparison of word embeddings for the biomedical natural language processing. *Journal of biomedical informatics*, 87:12–20, 2018.

[53] Leon Willenborg and Ton De Waal. *Elements of statistical disclosure control*, volume 155, page 27. Springer Science & Business Media, 2012.

Appendix A

Data management plan

This project was completed as part of the Master of Informatics degree program at the University of Edinburgh. In order to comply with the guidelines of the School of Informatics and in the presence of sensitive data being handled, a data management plan has been completed by filling a template through the tool available at https://dmponline.dcc.ac.uk/. This tool is principally aimed at producing statements over the collection of new data. This means that a large subset of the questions asked was not relevant, and the resulting data management plan is quite compact.

What data will be generated or reused in this research?
No original data will be generated in this research. We will be using the dataset MIMIC-III, from https://mimic.physionet.org/.

How will the data be documented to ensure it can be understood?
As no original data will be generated, we refer to the documentation at http://archive.is/vhCJh.

Where will the data be stored and backed-up?
The data will be stored on a private hard drive, and will not be backed up, as it is hosted at https://mimic.physionet.org/ and can easily be recovered from there.

How will you quality assure your data?
The data is provided with an MD5 checksum, which will be verified before the beginning of the research.

How will you manage any ethical and IPR issues?
Before obtaining the dataset, a compulsory training has been completed on ethics in health research involving human participants and on the legal framework under which the dataset was released (HIPAA). Additionally, an agreement has been signed that prevents us from attempting to re-identify the individuals in the dataset.

Which data do you plan to keep and for how long?
The entirety of the MIMIC-III dataset will be kept until June 2020.

Can you share your data? If not, how will it will be stored and preserved?
The data cannot be directly shared, and will be protected by keeping it in encrypted form on a private hard disk.

Appendix B

Generalization hierarchies

The following tables describe the generalization rules used in this work, with the quasi-identifier values that match the left-hand-side of an arrow being transformed to match the right-hand-side. Square brackets are used to indicate ranges, curly brackets indicate sets, and commas separate values. Null values are represented with a question mark. A * symbol matches any whole quasi-identifier value string. A character in italics matches any single character or digit of a quasi-identifier value string. Any other character matches the corresponding character in a quasi-identifier value string. Where an explanation is missing, a simple generalization was performed, by grouping together values listed between curly brackets.

B.1 UCI Adult

Country

Step	Rules
1	{Columbia, Peru, Ecuador, Trinadad&Tobago} → South America
	{United-States, Canada, Mexico} → North America
	{Puerto-Rico, Jamaica, Guatemala, El-Salvador, Cuba, Dominican-Republic, Haiti, Honduras, Nicaragua} → Central America
	{Hungary, Yugoslavia, Poland} → Eastern Europe
	{Greece, Portugal, Italy} → Southern Europe
	{Scotland, Germany, Ireland, England, Holand-Netherlands, France} → Western Europe
	{South Africa} → Southern Africa
	{China, Japan, Thailand, Cambodia, Philippines, Taiwan, India, Hong, Vietnam, Laos, Outlying-US(Guam-USVI-etc)} → East Asia
	{Iran} → Middle East

2	{South America, North America, Central America } → America {Eastern Europe, Western Europe, Southern Europe } → Europe {Southern Africa } → Africa {East Asia, Middle East } → Asia
3	{America, Europe, Africa, Asia} → World

Education

Step	Rules
1	{Some-college, Bachelors, Masters, Doctorate} → College {HS-grad, 9th, 10th, 11th, 12th} → High-School {1st-4th, Preschool} → Elementary-school {Prof-school, Assoc-acdm, Assoc-voc, ?} → Professional-Dev ? → ?
2	{College, High-School, Elementary-School, Professional-Dev, ?} → Education

Profession

Step	Rules
1	{Tech-support, Exec-managerial, Adm-clerical, Handlers-cleaners, Other-service, Prof-specialty, Priv-house-serv, Protective-serv, Armed-Forces} → Tertiary-Sector {Craft-repair, Machine-op,inspct, Sales} → Secondary-Sector {Farming-fishing, Transport-moving} → Primary-Sector ? → ?
2	{Tertiary-Sector, Secondary-Sector, Primary-Sector, ?} → Profession

Work class

Step	Rules
1	{Self-emp-not-inc, Self-emp-inc} → Self-employed {Federal-gov, Local-gov, State-gov} → Government {Without-pay, Never-worked} → Unemployed Private → Private ? → ?
2	{Self-employed, Government, Unemployed, Private, ?} → Workforce

Marital status

Step	Rules
1	{Married-civ-spouse, Married-AF-spouse, Married-spouse-absent} → Has-Spouse
	{Divorced, Separated, Widowed} → Had-Spouse
	{Never-married} → No-Spouse
	? → ?
2	{Has-Spouse, Had-Spouse, No-Spouse, ?} → Human

Age

Step	Rules	Explanation
1	[0,9] → 0:9	(merge range)
	[10,19] → 10:19	
	[20,29] → 20:29	
	[30,39] → 30:39	
	[40,49] → 40:49	
	[50,59] → 50:59	
	[60,69] → 60:69	
	[70,79] → 70:79	
	[80,89] → 80:89	
	[90,99] → 90:99	
	[100,109] → 100:109	
	[109,119] → 110:119	
2	{0:9, 10:19, 20:29, 30:39, 40:49} → 0:49	
	{50:59, 60:69, 70:79, 80:89, 90:99} → 50:99	
	{100:109, 110:119} → 100:119	

Race

Step	Rules	Explanation
1	* → Race	(suppression rule)

Sex

Step	Rules	Explanation
1	* → Sex	(suppression rule)

Relationship

Step	Rules	Explanation
1	* → Relationship	(suppression rule)

B.2 MIMIC-III

Language

Step	Rules	Explaination
1	* → Language	(suppression rule)

Religion

Step	Rules	Explanation
1	* → Language	(suppression rule)

Gender

Step	Rules	Explanation
1	* → Language	(suppression rule)

Procedure

Step	Rules	Explanation
1	*abcd* → *a*	(maintain first digit)
2	* → Procedure	(suppression rule)

Insurance

Step	Rules
1	{Private, Self Pay} → Not public {Medicare, Medicaid, Government} → Public ? → ?
2	{Not public, Public, ?} → Insurance

Marital status

Step	Rules
1	{MARRIED, LIFE PARTNER} → Has partner {SINGLE, DIVORCED, WIDOWED, SEPARATED} → No partner {?, UNKNOWN (DEFAULT) } → ?
2	{Has partner, No partner, ?} → Marital status

Date of birth

Step	Rules	Explanation
1	$[0, 1950] \rightarrow$ Silent generation $[1951, 1980] \rightarrow$ Baby boomer $[1981, 2000] \rightarrow$ Millennial $[2001, 3000] \rightarrow$ Generation Z	(merge range)
2	{Silent generation, Baby boomer, Millennial, Generation Z} → Date of birth	

Appendix C

Suppression with ϵ-safe LA

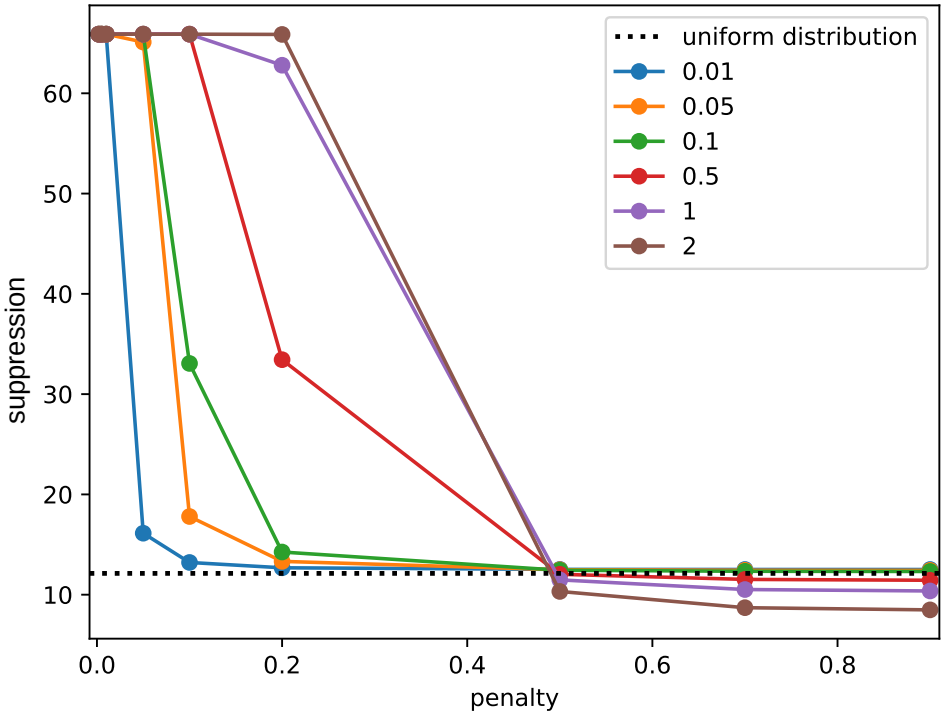


Figure C.1: Relationship between penalty factor and average suppression weighed by the probability of each node, for k-anonymity-based penalty. Each line represents a different setting of ϵ', with the exception of the black dotted line representing the baseline.

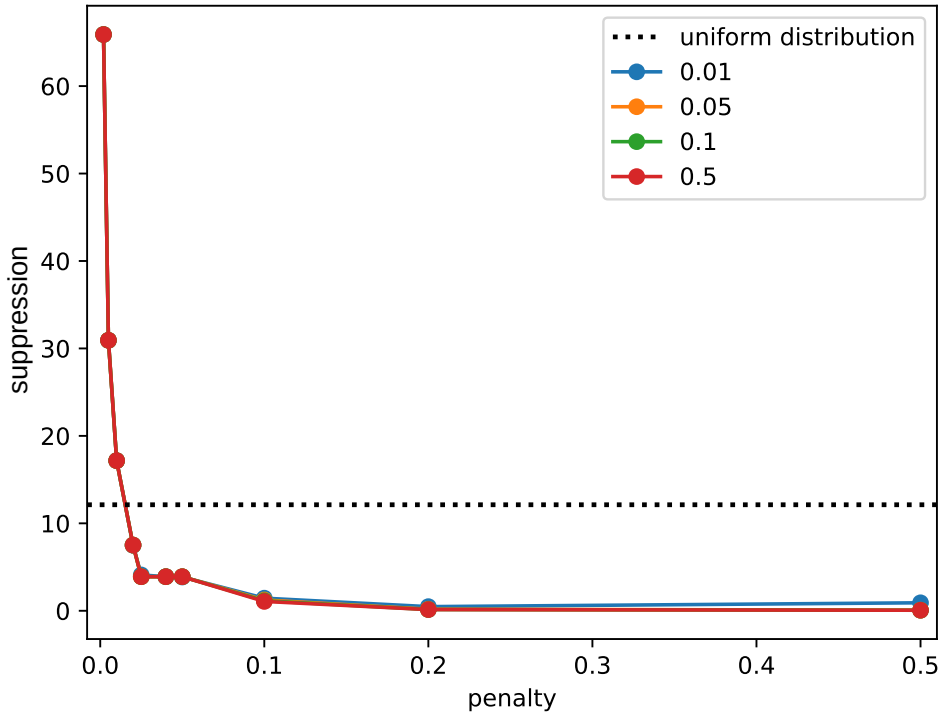


Figure C.2: Relationship between penalty factor and average suppression weighed by the probability of each node, for suppression-based penalty. Each line represents a different setting of ϵ', with the exception of the black dotted line representing the baseline.