Evaluating the impact of k-anonymization on the inference of interaction networks

Pedro Rijo, Alexandre P. Francisco, Mário J. Silva

INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Portugal

E-mail: {pedro.rijo,aplf,mario.gaspar.silva}@tecnico.ulisboa.pt

Revised and extended version of a paper selected from PSBD 2014

Abstract. We address the publication of a large academic information dataset while ensuring privacy. We evaluate anonymization techniques achieving the intended protection, while retaining the utility of the anonymized data. The published data can help to infer behaviors and study interaction patterns in an academic population. These could subsequently be used to improve the planning of campus life, such as defining cafeteria opening hours or assessing student performance. Moreover, the nature of academic data is such that many implicit social interaction networks can be derived from available datasets, either anonymized or not, raising the need for researching how anonymity can be assessed in this setting. Hence we quantify the impact of anonymization techniques over data utility and the impact of anonymization on behavioural patterns analysis.

Keywords. Privacy-preserving data publishing. Academic data publishing. Network inference.

1 Introduction

The continuous increase of stored data is causing raised interest due to the new possibilities it can open to organizations. Data mining techniques enable the extraction of interaction patterns that could be used for customization and adaptation of services to individuals. Industry leaders now provide e-commerce services that routinely profile clients based on their previous searches and purchases to recommend products. Similarly, academic data can provide interesting insights over education institutions, helping to increase efficiency. For instance, cafeteria attendance could be predicted based on faculty and student schedules, and cleaning schedules may be optimally adjusted to attendance fluctuations. Other aspects, such as academic success, can be analysed and improved with similar approaches. On the other hand, the availability of such amount of data about a large academic population could be harmful if compromised. Malicious hackers might infer personal traits and behaviors from online activity patterns, daily schedules, individual addresses and other personal data in academic sites, to launch a variety of attacks or exploit private information.

Privacy definitions in datasets can be tuned by the owner before publishing the data. To control privacy, Sweeney proposed that datasets should adhere to k-anonymity [17]. The notion of k-anonymity states that, for each record, there are at least $k - 1$ other records whose values for a set of special attributes, called Quasi-identifiers (QIDs), are equal. QIDs can be used for instance to link a record with external data. Typical methods for achieving k-anonymity include Datafly [16], Mondrian [13] and Incognito [12]. On the other hand, anonymization distorts the data, decreasing their utility. Typical data mining techniques are highly dependent on data quality. Network inference in particular is highly affected by anonymization, with macroscopic properties of inferred networks changing deeply when using anonymized data.

This work quantifies precision loss in network inference when underlying data is subject to k-anonymity techniques. We explore multiple approaches for achieving required privacy and analyse the decrease of data utility as the level of privacy is raised, in the context of the implementation of a semi-automatic system capable of answering queries over academic data and of retrieving queried data fields respecting privacy issues. We study data utility variation with the level of privacy, tuned by the k parameter in k-anonymity methods, and we compare different methods available to achieve k-anonymity.

The proposed approach has been implemented at Instituto Superior Técnico (IST)[1], the school of engineering of the Universidade de Lisboa, Portugal. The Academic Information System of IST FénixEdu[2] manages all the information about Students, Teachers, Researchers, Classes, Subjects, and Courses. The amount of information on human behavior and interactions that can be inferred from such large dataset covering a sizeable population (more than 60,000 people on record) makes it appealing for many types of analyses, both by internal teams and anyone else interested in learning about this population. However, since data on FénixEdu can expose private information about academic agents, some protection measures need to be applied before data release and publication.

In the next section, we review the issues of privacy preservation, contextualising the current state of art. We introduce key definitions and concepts, such as utility metrics, and overview anonymization methods for achieving required privacy. Next, we will describe the experiments with FénixEdu data and discuss the obtained results, which provide insights about the performance of our methodology for data anonymization. Finally, we will present our conclusions and final discussions, including directions for future work.

2 Privacy Preservation

Benjamin Fung *et al.* recently surveyed privacy-preserving data publishing [10]. They report that early work, by Dalenius provided a very stringent definition of privacy protection, in which privacy-protected data sets access to the published data should not enable the attacker to learn anything extra about any target victim compared to no access to the database, even when the attacker has background knowledge obtained from other sources [6]. Dwork has shown later that such stringent definition is impossible to achieve [8], but it remains a starting point for addressing privacy protection. Most recent literature on Privacy Preserving Data Publishing (PPDP) considers a more relaxed notion of privacy protection assuming that the attacker has limited background knowledge [10].

Fung *et al.* survey provides a classification for privacy models based on their attack principles, identifying four attack models: Record Linkage, Attribute Linkage, Table Linkage, and Probabilistic Attack. In this work, we focus on protecting data from record linkage attacks, which occur if an attacker is able to link an individual to a record in published data. In the record linkage attack model, we assume that an attacker may know a Quasi-identifier

[1] `http://tecnico.ulisboa.pt/`

[2] `http://fenixedu.org/`

(QID) of the victim. QIDs are attributes in private information that could be used for linking with external information. Such attributes not only include explicit identifiers, such as name, address, and phone numbers, but also attributes that in combination can uniquely identify individuals, such as birth date and gender [17]. A data table is considered to be privacy-preserving if it can effectively prevent the attacker from successfully performing these linkages.

As an example for record linkage attacks, suppose that some academic institution publishes student records for research purposes. An attacker may know that one individual, the victim, is present on that dataset. Even after de-identification of the records, if the attacker knows some of the attributes such as *age*, *locality* and *gender*, it may find a unique record containing such values, discovering available information for that victim. In this case we say that a record linkage attack occurred.

We can however take a step further in what concerns dealing with linkage attacks. Cuzzocrea et al. [3, 4] discussed and studied the problem of preserving the privacy beyond table entries linkage, considering the setting of general aggregations and proposing a sampling-based framework for dealing with such issues. In this setting we are usually limited to approximated answers when querying data, but we get some guarantees about what we can infer from aggregations such as averages or counting over multi-dimensional data.

Anonymization techniques rely usually on generalisation and suppression operations for privacy preservation. Generalisation operations are applied based on a Value Generalisation Hierarchy (VGH) that provides information on how to generalise each attribute.

Privacy can be achieved in many ways. For example, besides anonymization, obfuscation and/or perturbation techniques may be used. Obfuscation tries to protect privacy by suppressing identifiers. By itself, obfuscation does not meet privacy requirements, since other released information, QID, may be used for linkage even with suppression of identifiers as the name or Social Security Number [15]. Perturbation is a technique that introduces new records or changes the existing ones. This technique could be used for achieving privacy requirements but it would make the data synthetic, yielding records that do not correspond to real-world entities represented by the original data [17].

2.1 Utility Metrics

Anonymization faces the problem of also distorting the data, which will then become less precise and less useful than the original when used for data analysis. Previous research proposed metrics to assess the information loss due to anonymization. In this work, we assess data utility of our academic dataset using metrics proposed by LeFevre [13] and Sweeney [16].

LeFevre's metrics consider the size of each equivalence class E of an anonymized table RT for measuring data distortion. This means that an higher value represents a bigger distortion over original data. Intuitively, the discernibility metric C_{DM} assigns to each tuple t a penalty determined by the size of the equivalence class containing t, equivalent to

$$C_{DM} = \sum_{E \in \mathcal{E}} |E|^2 \tag{1}$$

where $\mathcal{E}$ is set of equivalence classes. As an alternative, a normalised average equivalence class size metric (C_{AVG}) may be used, although its value depends on k parameter,

$$C_{AVG} = \frac{|PT|/|\mathcal{E}|}{k} \tag{2}$$

where PT is the private table and $|PT|$ is the total number of records. Both metrics are defined for a table, or a set of records, and are dependent on the number of equivalence classes and on the number of records in the dataset. The usefulness of these metrics to compare values for different datasets is very low, specially the discernibility metric, which does not take into account the number of records.

Sweeney defined a precision metric, *Prec*, which considers the "height" of the generalisation on the value generalisation hierarchy,

$$Prec(T) = 1 - \frac{\sum_{i=1}^{N_A} \sum_{j=1}^{|PT|} \frac{h}{|VGH_{A_i}|}}{|PT|\,|N_A|} \tag{3}$$

given that

$$f_1(\ldots f_h(t_{P_j}[A_i])\ldots) = t_{R_j}[A_i] \tag{4}$$

where $|PT|$ is the number of records in the private table PT being anonymized, N_A the number of attributes belonging to *QID* set, VGH_{A_i} is the VGH for attribute A_i in the *QID* set, f_1, ..., f_h are the generalizations on attribute A_i, $t_{P_j} \in PT$ and $t_{R_j} \in RT$ (released table) [17]. The higher the precision, the higher the utility of the data, meaning that the anonymized data is more similar to the original dataset. *Prec* ranges from 0 to 1.

2.2 Anonymization

Anonymization of a private relational dataset is the process of transforming the records in each private table into a released dataset in which none of the records in the released tables can be mapped to a single record in the corresponding private table.

The degree of anonymization of relational data can be measured through k-anonymity, a metric proposed by L.-Sweeney [17]. The notion of k-anonymity states that for each record there are at least $k-1$ other records whose values, for a set of special attributes, are equal. These special attributes with equal values correspond to quasi-identifiers (QID). In other words, for each of the records contained in the released table, the values of the tuple that comprise the quasi-identifier appear at least k times in the table. This is achieved through generalisation and suppression techniques. As a consequence, available algorithms for achieving k-anonymity produce as output a set of records where the value of each attribute may be turned into a class of values instead of the original value. For example, the age of an individual may be generalised from her actual age of 42 to the class of $[40 : 50]$.

In this study, we used implementations of three methods previously proposed for achieving privacy against record linkage attacks: Datafly [16], Mondrian [13], and Incognito [12]

2.2.1 Datafly

Datafly is a real-world implementation of *MinGen* a theoretical algorithm, also from the developer of Datafly, which provides k-anonymity protection with guaranteed minimal distortion. Datafly, on the other hand, uses a heuristic to make approximations, and so it does not always yield the optimum result, sometimes even distorting data. The user needs to start by identifying sensible attributes in the original private table *(PT)*. Then, by grouping attributes in the *PT*, the user defines the set of quasi-identifiers *(*QI_i*)*, assigning also a weight from 0 to 1 to each attribute QI_i representing its likelihood of being used in linking with external data. The user performing the anonymization needs also to specify the level of anonymization by specifying the value for parameter k. Finally, the user has to assign a preference value between 0 and 1 to each attribute in order to state which attribute should be preferentially subject to distortion.

The core Datafly algorithm has few steps. The first step constructs a frequency list, which contains distinct sequences of *QID* values in *PT*, along with the number of occurrences of each sequence. Each sequence in the frequency list represents one or more tuples in a table. The second step uses a heuristic to guide generalisation. The attribute having the highest number of distinct values in the frequency list is generalised. Generalisation continues until there remains k or fewer tuples having distinct sequences in the frequency list. The third step suppresses any sequences occurring less than k times. Complimentary suppression is performed in the fourth step so that the number of suppressed tuples satisfies the k requirement. The final step produces a table *MGT*, based on the frequency list, such that the values stored as sequences appear as tuples in *MGT* replicated in accordance to the original stored frequency.

One of the limitations of Datafly is that it makes crude decisions, generalising all values associated with an attribute and suppressing all values within a tuple. Another problem is related to the selection heuristic that selects the attribute with the highest number of distinct values as the one to generalise, leading to some unnecessary generalisations.

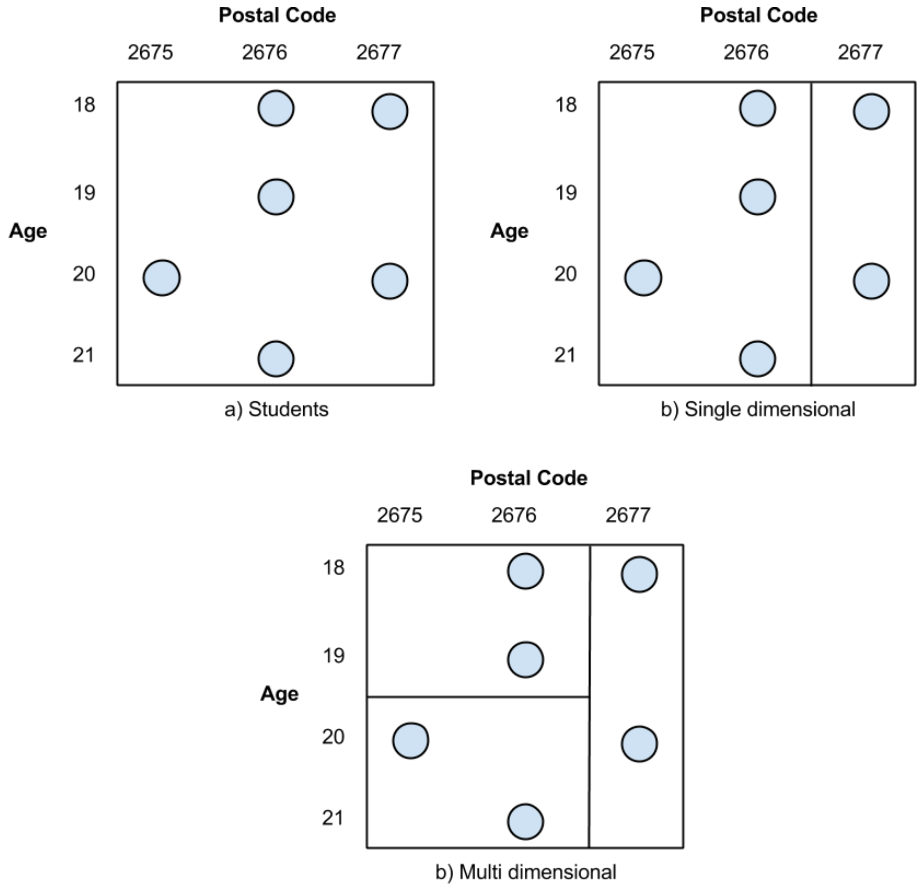
2.2.2 Mondrian Multidimensional k-Anonymity

Mondrian is a multi-dimensional approach to achieve k-anonymity, providing an additional degree of flexibility not seen in single-dimensional approaches. While in single-dimensional approaches a single *QID* attribute is chosen in each generalisation, in multi-dimensional approaches like Mondrian more than one attribute may be chosen (Fig. 1 illustrates the difference). This flexibility often leads to higher-quality anonymizations, in which released datasets typically contain more equivalence classes, providing more accurate information on *QID* attributes without violating k-anonymity. Mondrian operates through an attribute selection heuristic that determines which *QID* attribute an equivalence class will be partitioned on. The heuristic chooses the dimension with the widest (normalised) range of values. When multiple dimensions have the same width, it simply selects the first dimension that has an allowable cut. Once a dimension is chosen, the implementation performs partitioning independently of the corresponding Value Generalization Hierarchy (VGH). Partitioning is performed over the median value, such that any values less than or equal to the median resides in the left equivalence class and all other in the right equivalence class.

2.2.3 Incognito

For a given dataset there are multiple possible anonymizations satisfying the privacy definition imposed by k-anonymity, according to chosen quasi-identifiers and their VGHs. Incognito tries to choose the least generalised possible anonymization by exploring all possible anonymizations and ensuring soundness and completeness. On the other hand this leads to higher computational complexity and a considerable overhead in what concerns both time and space, in particular when dealing with larger datasets. Various implementations are possible [12]. Similarly to Datafly, Incognito can also perform suppression.

Consider a table with *Sex* and *Academic GPA* attributes as *QID*. Their respective VGH is represented in Fig. 2 and Fig. 3. Fig. 4 represents the possibilities Incognito tests to choose the least generalised anonymization. Incognito starts with original values for each of *QID* attributes and begins to apply generalisation operations to each of the attributes.



a) Students

b) Single dimensional

b) Multi dimensional

Figure 1: Spatial representation of students and partitioning (quasi-identifiers: postal code and age).

By comparing these three methods, although all three ensure k-anonymity, given a private dataset and required parameters, namely the value k, one would say that Datafly provides in general non-optimal anonymizations by over-generalizing attributes, Incognito provides optimal anonymizations since it is a sound and optimal algorithm but at an higher computational cost, and Mondrian provides intermediate anonymizations in what concerns generalization optimality [13].

Note that, independently of the method, for k-anonymity one needs only to provide the k parameter. But, although this parameter can be adjusted for the intended level of privacy, the quality of the anonymization from over-generalization point of view depends also on the VGH used for each attribute, which may be provided by the user or found automatically depending on the chosen method and domain knowledge inherent to each attribute.

3 Experimental Analysis

Anonymization by generalisation decreases data utility. In some cases the distortion suffered by the original private data can prevent analysts from obtaining meaningful con-

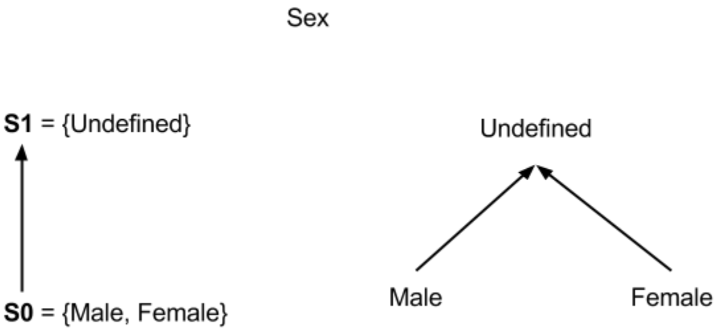


Figure 2: Sex VGH example for Incognito method.

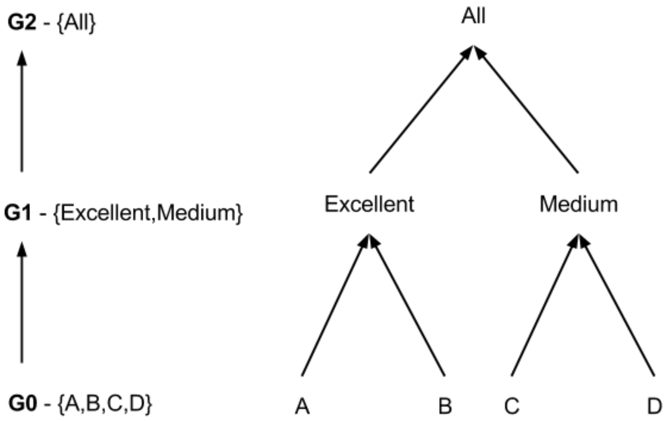


Figure 3: Academic grade VGH example for Incognito method.

clusions in latter analyses. To assess the precision loss of anonymization operations, we anonymized tables with query results from the FénixEdu academic system with Datafly, Mondrian and Incognito methods, assigning different values to the k parameter and measuring data utility.

We evaluated two datasets, resulting from two separate queries submitted to FénixEdu, to be analysed independently. The first, the *Postal Codes Dataset*, retrieves the postal code of the address of each person in the academic system. The second, the *Student Grades Dataset*, was obtained by retrieving, for each student, the set of all obtained grades, including fails and subjects in which the student was not evaluated.

In the first study, we analyse a common case of protecting an individual's address. In the second case, we simulate a situation where an attacker who is also a student can identify his set of grades. Once he identifies his own set of grades, the attacker can then attempt to identify each subject and other students. Consider that student A_1, who has identified his set of grades, has an unique grade in his academic curriculum (for instance, he only has one grade of 19 out of 20 in his set of grades). In this situation he may identify some colleague A_2, if A_2 his the only student with some other final grade (in the example, A_2 is

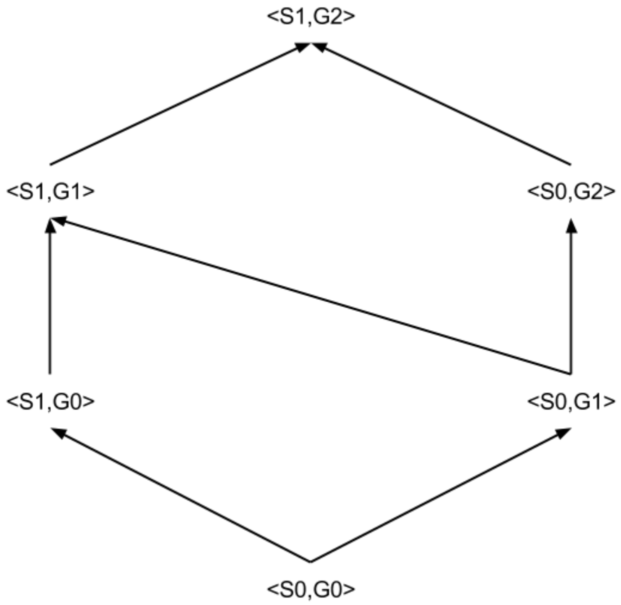


Figure 4: Generalisation options for Incognito method.

the only student with 20 out of 20). This kind of situation would compromise the dataset. The anonymization of some attributes will make the attacker unable to identify his set of grades, thus preventing record linkage attacks on the dataset.

We can distinguish two types of attributes in these datasets for anonymization purposes: *logical domains* and *non-logical domains*. In the first type, we have attributes like *birth date*, where the VGH naturally arises. It is straightforward that years may be grouped into decades, decades into centuries, and so on. In the second type of attributes we have for example person identifiers for which we do not have any logical hierarchy available. Hence, we can group any set of persons without violating any convention.

We anonymized both datasets with the three methods using the implementations in the UTD Anonymization Toolbox[3]. To compare the performance of the different anonymization methods, we used the precision and data distortion metrics introduced in Section 2.1. Although C_{DM} and C_{AVG} metrics present unbounded values that may be hard to interpret, they are still useful to compare the three methods. Note that, as discussed, C_{AVG} depends on k. Note also that the precision metric cannot be applied to Mondrian since its implementation in the UTD Anonymization Toolbox does not respect the defined VGH for generalisation operations and there is no practical method to find out the used VGH. This may make Mondrian unsuitable for logical numeric domains, such as ages, since it is more logical to group years by decade and Mondrian can make arbitrary classes without that concern. Nevertheless, Mondrian can be useful in other situations. Suppose that a person identifier is included in the *QID* set. Perhaps it is not relevant how to generalise individuals and Mondrian may find a good VGH for that case, making the work easier for the query

[3] http://cs.utdallas.edu/dspl/cgi-bin/toolbox/anonManual.pdf

Table 1: Data utility metrics for anonymization over postal codes dataset with multiple methods and k values. *Values are not exact since Mondrian does not respect provided VGH and there is no simple way to find the used VGH; values provided only for comparison.

Method	k	Prec	C_DM	C_AVG
Datafly	3	0.50	370,526,617	281.89
	5	0.50	370,526,617	169.14
	10	0.50	370,526,617	84.57
	30	0.25	1,782,649,453	247.44
	50	0.25	1,782,649,453	148.46
	100	0.25	1,782,649,453	74.23
Mondrian	3	0.38*	92,898,847	84.68
	5	0.38*	92,900,137	54.76
	10	0.38*	92,904,257	29.82
	30	0.37*	93,001,611	13.26
	50	0.37*	93,111,423	9.09
	100	0.37*	93,564,185	5.43
Incognito	3	0.50	370,526,617	281.89
	5	0.50	370,526,617	169.14
	10	0.50	370,526,617	84.57
	30	0.25	1,782,649,453	247.44
	50	0.25	1,782,649,453	148.46
	100	0.25	1,782,649,453	74.23

auditor. For logical domains, Incognito is the best option.

All reported experiments were conducted in an Intel(R) Core(TM) i5-2300 CPU @ 2.80GHz Quad-Core 64bits with 4GB RAM.

3.1 Anonymization of the Personal Postal Codes Dataset

The Postal Codes Dataset contains the set of postal codes for each academic agent (Student, Researcher, Teacher) in a total of 66,809 records. The dataset is characterized by a logical hierarchy for the Postal Code attribute. Postal code assigns a region for each 4-digits group. This assignment is made such that a region with the code 2675 (Odivelas) is near 2676 (Amadora), for example. This property makes the VGH intuitive as each level groups one more digit, starting by the least significant. Portuguese postal codes contain 7 digits, but the 3 least significant digits were removed from the postal codes prior to the anonymization by the toolbox, since they do not represent a logical hierarchy. In Portugal such digits simply encode post carrier routes.

The results for the three methods under evaluation are shown in Table 1. It is possible to observe that both Datafly and Incognito achieved the same precision. A precision of 0.5 (50%) means that 2 digits were suppressed, and a precision of 25% means that only the most significant digit was kept. Mondrian gets a different precision, although presented values are not reliable due to its behaviour of not considering the defined VGH. Nevertheless we opt to present them as they provide a lower bound for Mondrian precision.

Anonymization decreases data utility with a significant impact for this dataset. The 2-digit prefix distribution of the dataset shown in Fig. 5 shows that some prefixes, such as

Table 2: CPU time taken by each anonymization method on the Personal Postal Codes dataset.

Method	**k**	**time (s)**
Datafly	3	91.16
	5	95.75
	10	96.32
	30	97.29
	50	100.44
	100	102.03
Mondrian	3	30.30
	5	30.48
	10	30.09
	30	30.17
	50	30.17
	100	29.69
Incognito	3	153.30
	5	127.85
	10	126.70
	30	158.47
	50	178.13
	100	179.28

43, 52, 94, 33, 98 and 72, have a low count, with values smaller than 30 people. Those postal codes refer regions with low representativity in IST population, likely caused by also having major academic institutions offering similar BSc and MSc degrees nearby. On the other edge of the distribution, the postal codes of areas near IST, namely within Lisboa and Setúbal (prefixes 26, 27 and 28), have high counts of enrolled students.

The regions with low count of students enrolled in IST reduce the utility of the anonymized Postal Codes Dataset, since the UTD Anonymization Toolbox implementation requires that every leaf of the VGH is at the same depth. If all regions contributed with a higher number of students, we could increase k to a higher value and still get high data utility. One possible workaround to avoid this loss of utility would be to cluster (manually or through an algorithm) the regions that are somehow related until all regions get a higher count.

To visualise the impact of anonymization, we created maps. Fig. 7, shows the distribution of home addresses of IST students in Portugal, while Fig. 8 presents the distribution of IST students in Portugal by home address using the anonymized data generated by Incognito with $k = 10$. The displayed coordinates of the anonymized postal codes are calculated as the centroid of all existent postal codes in each student equivalence class. Fig. 8 shows that students were clustered according to the equivalence class of their postal codes. It is also observable that some regions have more equivalence classes than others. This is simply due to the non-uniformity of the distribution of postal codes in Portugal, where 2-digit prefixes are not evenly distributed taking into account population density[4]. The two maps make evident how the skewed distribution of students by region strongly decreases data utility. A more detailed look over Lisbon metropolitan area, showing this effect is available

[4] see, for instance `http://en.wikipedia.org/wiki/Postal_codes_in_Portugal`

Table 3: Data utility metrics for anonymization of the Student Grades Dataset with multiple methods and k values. *Values are not exact since Mondrian does not respect the provided VGH and there is no simple way to find the used VGH; values presented only for comparison.

Method	k	Prec	C_DM	C_AVG
Datafly	3	1.00	16,480,722,265	13,383.46
	5	1.00	16,480,722,265	8,030.08
	10	1.00	16,480,722,265	4,015.04
	30	1.00	16,480,722,265	1,338.35
	50	1.00	16,480,722,265	803.01
	100	1.00	16,480,722,265	401.50
Mondrian	3	0.80*	28,403,897	10.81
	5	0.80*	28,465,783	6.98
	10	0.80*	28,847,609	4.17
	30	0.79*	32,933,353	2.31
	50	0.79*	39,615,933	1.96
	100	0.77*	61,595,543	1.71
Incognito	3	1.00	16,480,722,265	13,383.46
	5	1.00	16,480,722,265	8,030.08
	10	1.00	16,480,722,265	4,015.04
	30	1.00	16,480,722,265	1,338.35
	50	1.00	16,480,722,265	803.01
	100	1.00	16,480,722,265	401.50

in Fig. 9 and Fig. 10.

Regarding performance issues, Table 2 and Fig. 6 show the time taken by each method on the anonymization process for the 66,809 records of the Personal Postal Codes Dataset. Mondrian is the fastest method, and the k parameter does not influence the processing time. Both Datafly and Incognito require more processing time with increasing k. Incognito is more sensitive to the increase of k due to the need for processing every possible anonymization.

3.2 Anonymization of the Student Grades Dataset

The Student Grades Dataset contains the grades of each IST student for each subject, in a total of 321,278 records. Prior to anonymization, the grades were converted to the European ECTS scale[5]. Grades have values between A and E, and three other values: AP (Approved), RE (Reproved/Not-approved), and NA (Not available/Not evaluated). Grades were grouped in {A,B}, {C,D}, {E,AP}, and {RE,NA}. The VGH is illustrated in Fig. 11. Note that the UTD Anonymization Toolbox requires the mapping of each grade to a sequential integer.

Since we chose $QID = \{Subject, Grade\}$ we had also to define a VGH for subjects in addition to the VGH for grades. Since there are more than 4,000 subjects on the dataset and we do not have suitable and/or sound logical information about the subjects, there was no scope for a logical VGH. In addition, it becomes impractical to create an optimal VGH since we would need to try many combinations. The adopted solution was to create

[5] http://ec.europa.eu/education/tools/docs/ects-guide_en.pdf

Table 4: Data utility metrics for Incognito anonymization with varying k on the inferred interactions network

k	Prec	C_{DM}	C_{AVG}
3	0.68	2,156,660	2.30
5	0.66	2,411,474	1.96
10	0.65	3,213,374	1.67
30	0.62	7,414,960	1.58
50	0.61	12,571,942	1.64
100	0.60	24,999,132	1.68

two levels besides the original subject value similarly to what happened with grades VGH and as depicted in Fig. 12. An intermediate level was defined with four non-logical classes of equal size. The higher/top level, i.e., the root class, groups all subjects. Most likely in such cases, Mondrian would create a more efficient VGH with better results, but again at a prohibitive computational cost.

Table 3 presents the data utility metrics obtained by the three anonymization methods. Once again, Incognito achieves the best data utility. However, in this case, Datafly, which was predicted to overgeneralise, presents the same results as Incognito using both metrics. Mondrian on the other hand has provided less precision (note that, as before, precision is not exact when applied to Mondrian since its implementation on the UTD Anonymization Toolbox does not respect the defined VGH for generalisation operations, working only as a guideline). Despite that, Mondrian produces less distorted data according to remaining metrics. This happens because Mondrian created more equivalence classes, each containing fewer elements, leading to smaller values. Mondrian gets better results in those metrics due to the VGH definition provided to the *subject* attribute. While Incognito and Datafly generalise *subject* to the first level or to the root level (leading to an huge generalisation despite the level), leaving the *grade* attribute untouched, Mondrian generalises both attributes, distributing the generalisation distortion by them.

Given these two experimental datasets and their above analysis, we can conclude that, as predicted, Incognito is the best method for most situations. In situations where the VGH has not a logical hierarchy and it becomes hard to find an optimal, or good, hierarchy, it is appropriate to use Mondrian, since it defines its own VGH. The best k parameter for each situation depends on both the VGH for each attribute and on the record distribution for each equivalence class defined in the hierarchy. The anonymized data quality is highly dependent on the k parameter and the number of elements in the less populated class for the considered VGH level.

3.3 Anonymization Impact on Network Inference

As observed above, compliance with k-anonymity decreased data utility. Data utility is highly dependent on the k parameter and on the distribution of original records over equivalence classes. In this section, we overview the influence of such decrease on data utility in knowledge extraction tasks, using network inference as study case. In addition to comparing the impact of each anonymization method based on data utility, we study the effect of anonymization on the inference of interaction networks.

We use as inferred networks to analyse the impact of anonymization two two student-teacher networks that infer interactions through class enrolments.

Student-Class-Teacher The main network. Represents the interactions between students with enrolled classes and teachers with lectured classes. From that a tripartite graph is constructed.

Student-Student A network extracted from the previous one. Students are considered to interact with another student if they attended the same class.

Note that both networks are undirected. To infer those networks we used records of the type

$$T(Student, SubjectGrade, AttendedClass, Teacher, CUQ) \tag{5}$$

The table contains, for each student and attended class, the final grade of the student at such subject, as well as the lecturing teacher and the CUQ grade[6].

For this experiment, a sample of the population was selected. Since data was clustered by student and class, we decided to sample blocks of rows. This decision was guided by a locality principle: we assume that records close to each other tend to belong to the same student, or to students from the same academic period and same course. If records were selected in a completely random way, it could happen that each record belonged to a student from a different course or different academic period, making relationships almost nonexistent. Using this method, starting with 1,419,649 records, we selected a total of 142,000 records, about 10%.

For anonymization we defined the following set of quasi-identifier attributes:

$$QID = \{SubjectGrade,\ AttendedShift,\ CUQ\} \tag{6}$$

Similarly to the previous experiment, StudentGrade was converted to the European scale. Following the previous analysis, Incognito was chosen for anonymizing query results. Table 4 presents the utility metrics for the anonymized dataset and for several values of k. Network inference was made upon original and anonymized data for comparison.

We computed several measures for each network, namely the number of vertices, the number of edges, the min degree, the max degree, the average degree, the average distance, and the diameter. The degree of a vertex is just the number of its neighbors, i.e., the number of edges incident on it. The average distance and the diameter are computed from the pairwise shortest distances between all pairs of vertices. Since we are considering unweighted graphs, the shortest distance between two vertices is just the minimum number of hops between them. Then, the average distance is computed by taking the average over all reachable pairs of vertices and the diameter is the longest shortest path found in the graph. There are some issues that we must take into account when computing measures, such as unreachable vertices leading to infinite values both for the average distance and the diameter, which are beyond the scope of this paper. We only state that we relied on standard approaches, such as considering the effective diameter instead of the diameter. We relied on Webgraph framework[7] and related tools [8] to compute these measures efficiently [1]. We refer also the reader to the work by Boldi *et al.*, on efficiently approximating the neighborhood function, for more details about the computation of the average distance and the diameter [2].

[6] The IST Course Unit Quality (CUQ) System is aimed at following up the functioning of each course unit, by promoting responsible involvement of students and teachers in the teaching, learning and assessment process. More info at `http://quc.tecnico.ulisboa.pt/en/`

[7] `http://webgraph.di.unimi.it/`

[8] `http://law.di.unimi.it/software.php`

Table 5: *Student - Class - Teacher* network properties for different anonymity levels.

k	Avg Dist	Diameter	Nodes	Edges	Min Deg	Max Deg	Avg Deg
orig	5.46	5.97	31,784	300,410	1	203	9.45
3	4.41	5.36	20,492	285,486	1	168	13.93
5	4.15	4.77	15,839	269,540	1	165	17.02
10	3.92	4.58	11,119	245,520	1	154	22.08
30	3.49	3.97	6,380	200,510	1	134	31.43
50	3.34	3.84	5,215	180,082	1	153	34.53
100	3.23	3.80	4,388	157,594	1	215	35.91

Table 6: *Student - Student* network properties for different anonymity levels.

k	Avg Dist	Diameter	Nodes	Edges	Min Deg	Max Deg	Avg Deg
orig	2.95	3.62	2,401	155,633	1	287	64.82
3	2.36	2.78	2,401	256,637	2	388	106.89
5	2.26	2.69	2,401	293,659	2	414	122.31
10	2.12	2.47	2,401	387,463	4	553	161.38
30	1.92	1.94	2,401	711,073	6	933	296.16
50	1.84	1.90	2,401	1,001,895	16	1,225	417.28
100	1.74	1.87	2,401	1,492,979	24	1,601	621.82

This first network is a tripartite graph where *Student*, *Teacher* and *Class* are the existing entities. Table 5 shows how inferred network properties change by adjusting the anonymization level through the k parameter.

It is possible to observe how network properties change with increasing k. k should affect the generalization of all *QID*, but *AttendedClass* has been the only attribute which has suffered generalisation. In practice, when k increases, nodes corresponding to *AttendedClasses* are grouped together. As the number of nodes decreases, the number of edges also decreases. Where we could observe one edge from a *Student* to multiple *Classes* before, there is now only one to the cluster created by generalisation. However, having *Classes* grouped increases both the indegree and the outdegree for every *Class* node now. With the decrease on the number of edges, both the average distance and the graph diameter also decrease. Using $k = 10$ as reference, one can observe that graph properties, such as the average distance and the graph diameter, have values of 72% and 77% with respect to the original network, which are still somewhat similar to the original values. The number of nodes suffers big differences (the anonymized network has about 30% of the nodes in the original network), and the number of edges remains similar (82%). The difference on the number of nodes justifies the 43% difference registered for the average degree.

The *Student-Student* network is inferred from equal enrolments by students. Two students are considered to interact if both were enrolled in the same class. The evolution of the network properties is shown in Table 6. As for the first network, it was expected that the clustering of classes would lead to more connected students (the number of nodes does not decrease, since students are not affected by the generalisation). Having students more connected is visible in the average distance, the graph diameter, and the degrees of vertices, that decrease with increasing k. Once again, using $k = 10$ as a reference, we get an anonymized average distance equal to 72% of the real value and a diameter with 68% accuracy. But the number of edges grows up to more than twice the original value. The outcome is that, as a consequence, the same also occurs to the remaining degree properties.

Hence, we can conclude that k-anonymity significantly affects network inference, leading in general to huge structural differences between networks inferred from original data and from anonymized data. This is true even when we are just considering rather general network measures. As we point out in the next section, alternative approaches are needed for inferring networks while preserving privacy.

4 Conclusions and Future Work

Data publishing could be a powerful tool for solving real world problems, allowing for instance better resource planning and scheduling in academic environments. However, despite the potential of solutions based on data mining techniques, privacy concerns are growing everyday and cannot be no longer be ignored. In particular, releasing large data sets where personal interactions could be inferred without addressing privacy issues can grant an attacker sensitive information about his target victims.

In our academic systems, the data owners can test different parameters before choosing the best suited anonymization method. However, this capability alone is not considered sufficient, and in fact provides little confidence about the level of privacy of datasets that might be released to the public if the release of published datasets is not carefully controlled. Many datasets are requested for a variety of useful purposes, corresponding to queries covering multiple aspects of academic life. However, the large variety of data that could be included might be used to infer implicit user behavior and re-identification could become easily achievable. For instance, the class schedules, or the network access data could be used to infer implicitly the commutes of individuals, or joint class attendance could be used to derive interaction networks aiming the university population that would give too many entry-points for possible re-identification attacks.

Privacy-preserving data publishing emerged from the opportunity that data provided for mining could solve real-world problems. Since a trade-off between data utility and privacy is required, it is hard to achieve low values for data distortion. Anonymization through k-anonymity presents itself as a method for protecting privacy. Unfortunately, anonymization also decreases data utility, making data mining and knowledge discovery tasks harder. Besides the dependence on VGH definition, k-anonymity strongly depends on the existence of outlier classes, which can severely impact the quality of anonymization methods.

It would be interesting, as future work, to study the trade-offs of removing outliers to improve anonymization results. Are those outliers actually important? Or are they just untidy data? A new algorithm able to reason about the trade-offs of removing outliers, increasing data utility for the obtained anonymized result, and the inclusion of those additional records while decreasing data utility, could be subject to further research with high impact on the utility of large datasets published while addressing privacy-preserving concerns. The statistics community has been developing methods for anonymization, collectively known as *statistical disclosure control* that balance privacy and utility, which should be also researched in this context [11].

The achieved results provide evidence that k-anonymity may not properly handle sensitive data. This is more visible when studying interaction networks inferred from k-anonymized data. In fact, all privacy models offering a priori guaranteed privacy (differential privacy [9], t-closeness [14]) entail a great utility loss, because they place privacy first and utility second. Techniques studied for ensuring data privacy in the presence of data aggregations can also be useful in this context [3, 4]. However, k-anonymity deserves further attention, as there are other methods to achieve it besides generalisation, for example micro aggregation of the quasi-identifier attributes [7]. In addition to k-anonymity, newer approaches, such as the one used currently by official statisticians, can be called a posteriori disclosure risk protection[11, 18] and may provide more interesting results from the data utility point of view, because it makes privacy come second after utility. The idea is to use an anonymization method (e.g. noise addition, micro aggregation, generalisation, etc.) with some parametrisation that yields acceptable utility, then measure the extant disclosure risk when comparing the anonymized dataset with the original dataset (e.g. via record linkage) and, if the risk is too high, run again the anonymization method with more strict privacy parameters. This approach attempts to preserve as much utility as possible.

With respect to the study of interaction networks there are still some network structural properties that should be studied in this context, such as network clustering properties. For instance, assuming that interactions are inferred from sensible attributes, do clusters or communities reflect anonymization generalisation? If so, it could be still interesting to analyse clustering in networks built from anonymized data, without losing too much information.

In what concerns the study of interaction networks built from datasets with sensible private information, the research on social network anonymization has led also to relevant results [19]. Typical approaches do not only consider typical attributes, but they also take into account the network structure, which could be used in re-identification attacks, in particular through the use of different but related networks, making use of non-trivial data aggregations. Secure multiparty computation models, supporting privacy preserving distributed data mining protocols, may play an important role in this context as they ensure privacy preserving data aggregations obtained from different sources [5]. These alternative approaches suffer, as far as we know, from the lack of free available tools, making usability and comparisons harder. However, we believe that the combination of such approaches with the methods discussed in the present paper could improve the state-of-the-art on data anonymization.

Acknowledgements

A preliminary version of this work was presented at the ACM CIKM 2014 Workshop on Privacy and Security of Big Data (PSDB). We thank the anonymous reviewers for the provided comments and insights.

This work was partly supported by national funds through FCT – Fundação para a Ciência e Tecnologia, under projects PTDC/EIA-CCO/118533/2010, EXCL/EEI-ESS/0257/2012 and UID/CEC/50021/2013.

References

[1] P. Boldi and S. Vigna, The WebGraph framework I: Compression techniques. In *Proc. of the Thirteenth International World Wide Web Conference (WWW'2004)*, pp. 595–601. ACM, 2004.

[2] P. Boldi, M. Rosa, and S. Vigna. HyperANF: Approximating the neighbourhood function of very large graphs on a budget. In *Proceedings of the 20th international conference on World Wide Web*, pp. 625–634. ACM, 2011.

[3] A. Cuzzocrea, V. Russo amd D. Saccà. A robust sampling-based framework for privacy preserving olap. In *Data Warehousing and Knowledge Discovery*, pp. 97–114. Springer Berlin Heidelberg, 2008.

[4] A. Cuzzocrea and D. Saccà, D. Balancing accuracy and privacy of OLAP aggregations on data cubes. In *Proceedings of the ACM 13th international workshop on Data warehousing and OLAP*, pp. 93–98. ACM, 2010.

[5] A. Cuzzocrea and E. Bertino. Privacy preserving olap over distributed xml data: A theoretically-sound secure-multiparty-computation approach. *Journal of Computer and System Sciences*, 77(6):965–987, 2011.

[6] T. Dalenius. Towards a methodology for statistical disclosure control. Statistik Tidskrift 15:429-444, 1977

[7] J. Domingo-Ferrer and V. Torra. Ordinal, continuous and heterogeneous k-anonymity through microaggregation. In *Data Mining and Knowledge Discoveryi*, 11.2:195–212, 2005.

[8] C. Dwork. Differential privacy. In M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, editors, *Automata, Languages and Programming*, volume 4052 of *Lecture Notes in Computer Science*, pp. 1–12. Springer Berlin Heidelberg, 2006.

[9] C. Dwork. Differential privacy: A survey of results. In *Theory and applications of models of computation*, pp. 1–19. Springer Berlin Heidelberg, 2008.

[10] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv.*, 42(4):14:1–14:53, June 2010.

[11] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. S. Nordholt, K. Spicer, and P.-P. De Wolf. Statistical disclosure control. *John Wiley & Sons*, 2012.

[12] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data (SIGMOD '05)*, pp. 49–60. ACM, 2005.

[13] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In *Proceedings of the 22nd International Conference on Data Engineering (ICDE'06)*, pp. 25–25, 2006.

[14] Ninghui Li, Tiancheng Li, and S. Venkatasubramanian. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In *IEEE 23rd International Conference on Data Engineering (ICDE 2007)*, pp. 106–115. IEEE, 2007.

[15] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *IEEE Symposium on Security and Privacy (SP'2008)*, pp. 111–125. IEEE, 2008.

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

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

[18] L. Willenborg and T. De Waal. Statistical Disclosure Control in Practice. *Springer Science & Business Media*, 1996.

[19] Bin Zhou, Jian Pei, and WoShun Luk. A brief survey on anonymization techniques for privacy preserving publishing of social network data. In *ACM SIGKDD Explorations Newsletter*, 10.2:12–22. ACM, 2008.

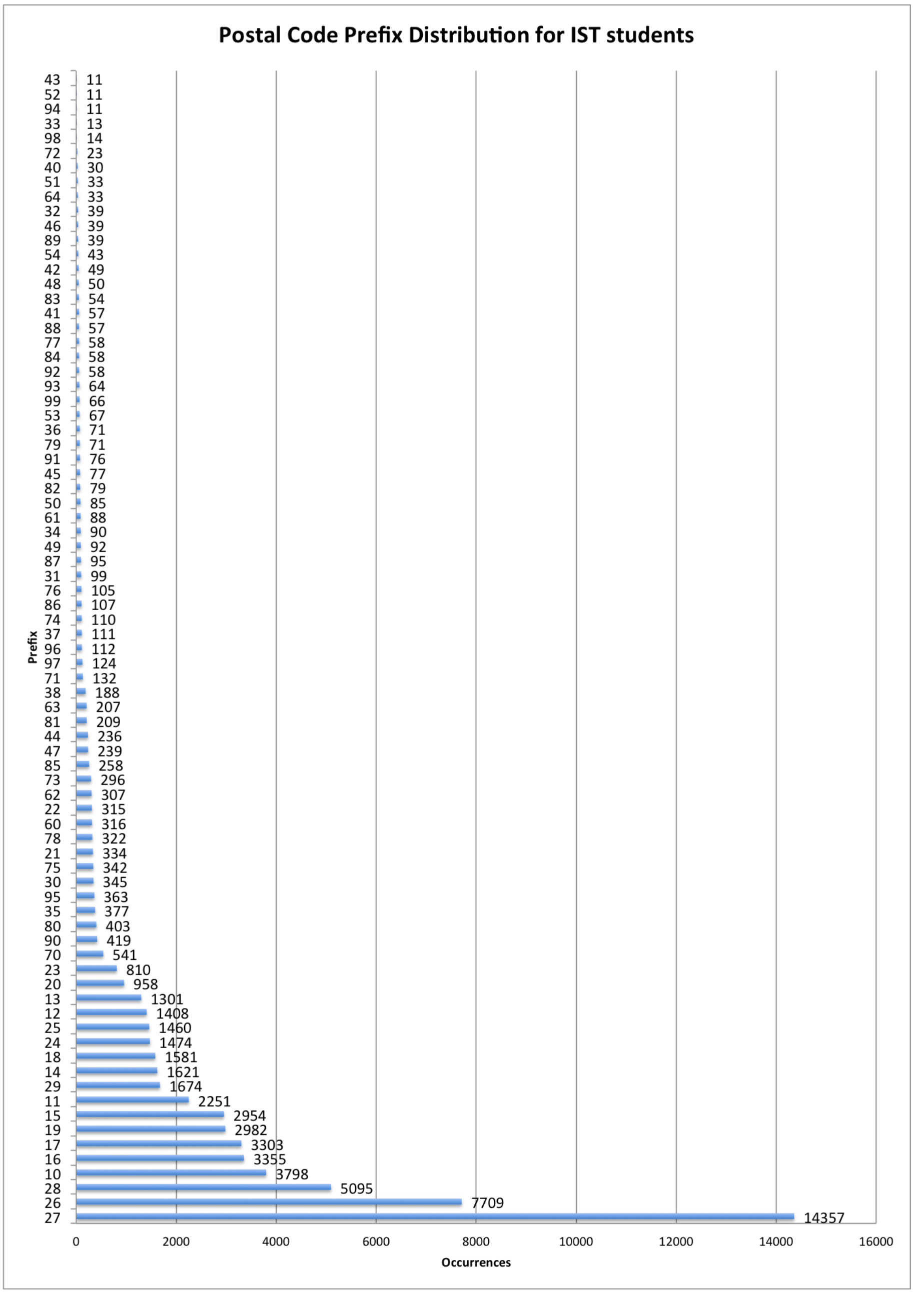


Figure 5: Postal Code prefix distribution taking into account the 2 most significant digits.

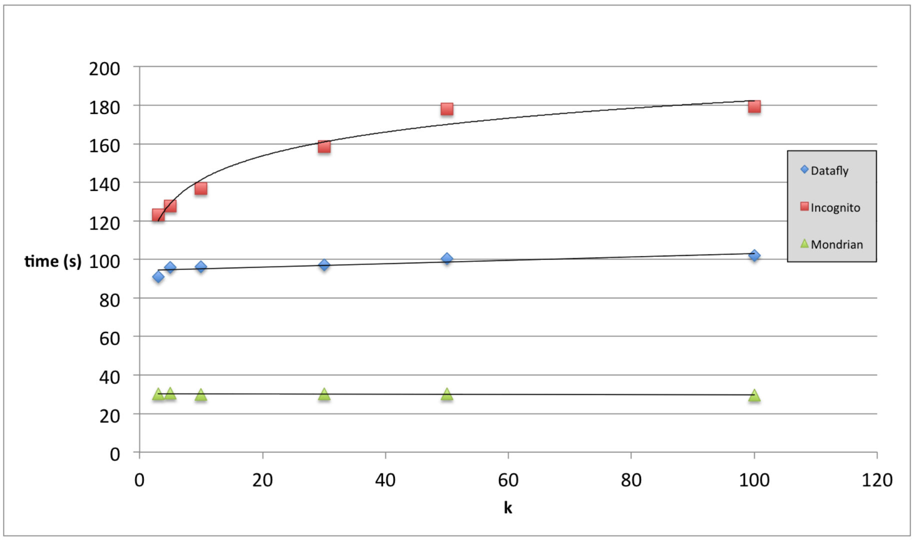


Figure 6: CPU time as a function of k taken by each anonymization method for the Personal Postal Codes dataset.

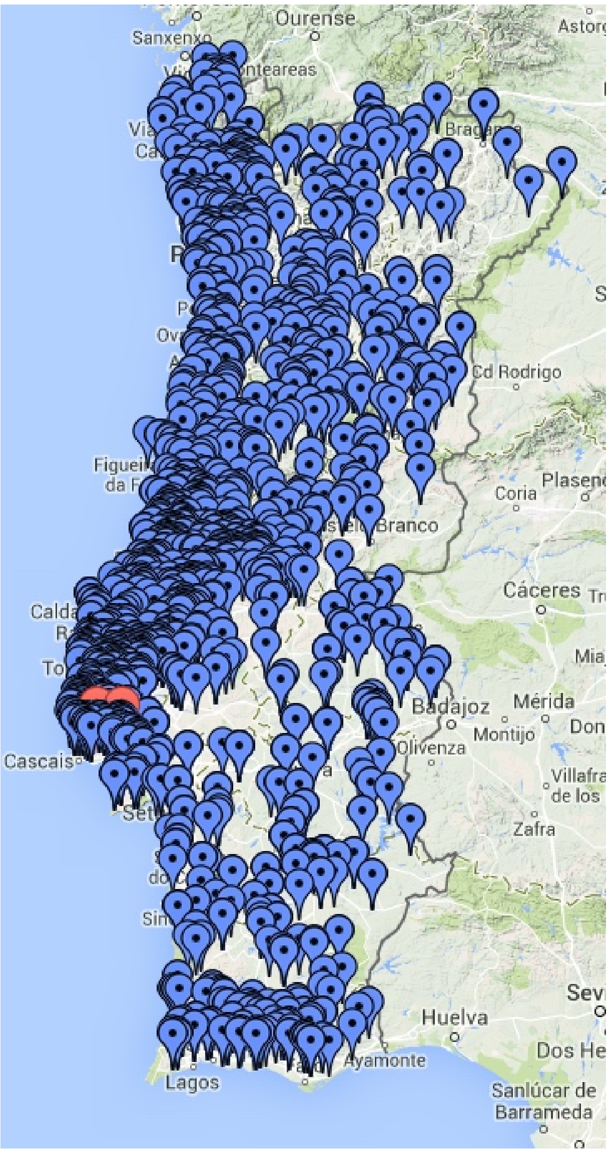


Figure 7: Distribution of IST students from continental Portugal by home address. Each blue dot represents a student.The red dots identify the location of the two IST campi.

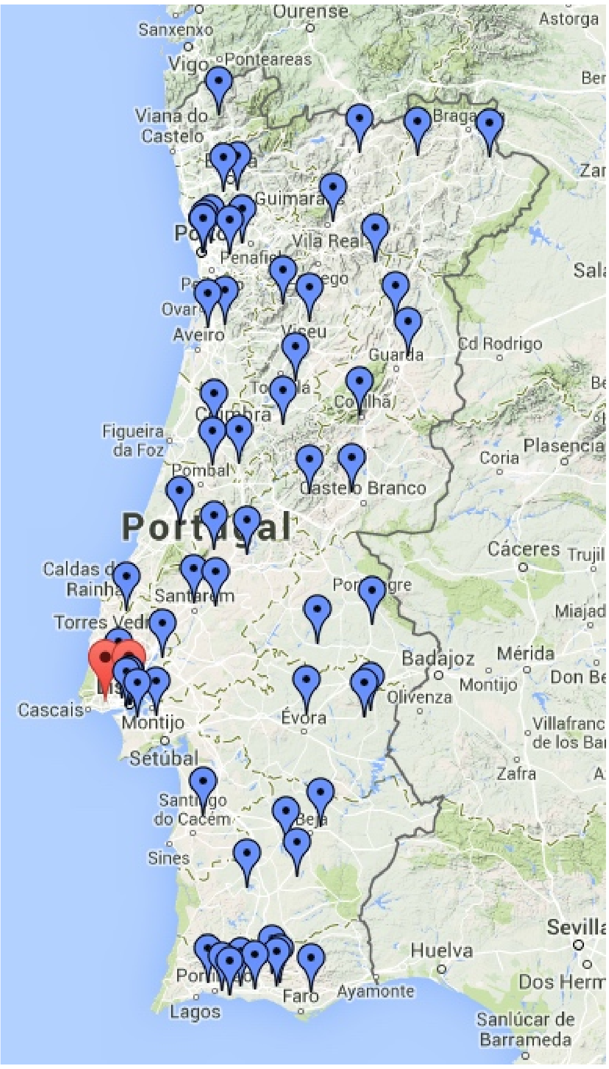


Figure 8: Distribution of Portuguese IST students by anonymized home address. Due to the anonymization process students were grouped, each blue dot represents a set of students. The red dots identify both IST campi. This data was anonymized using Incognito and a value of 10 for the k parameter.

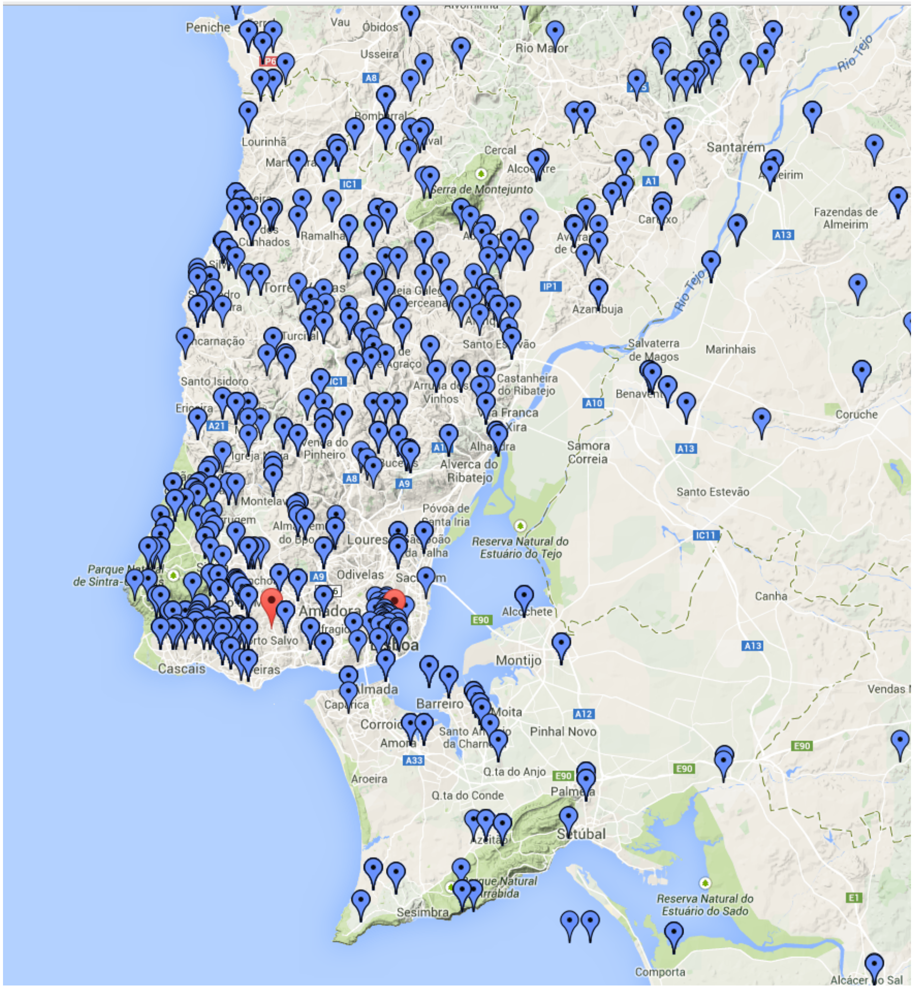


Figure 9: Distribution of Portuguese IST students by home address in Lisbon metropolitan area. The red dots identify both IST campi.

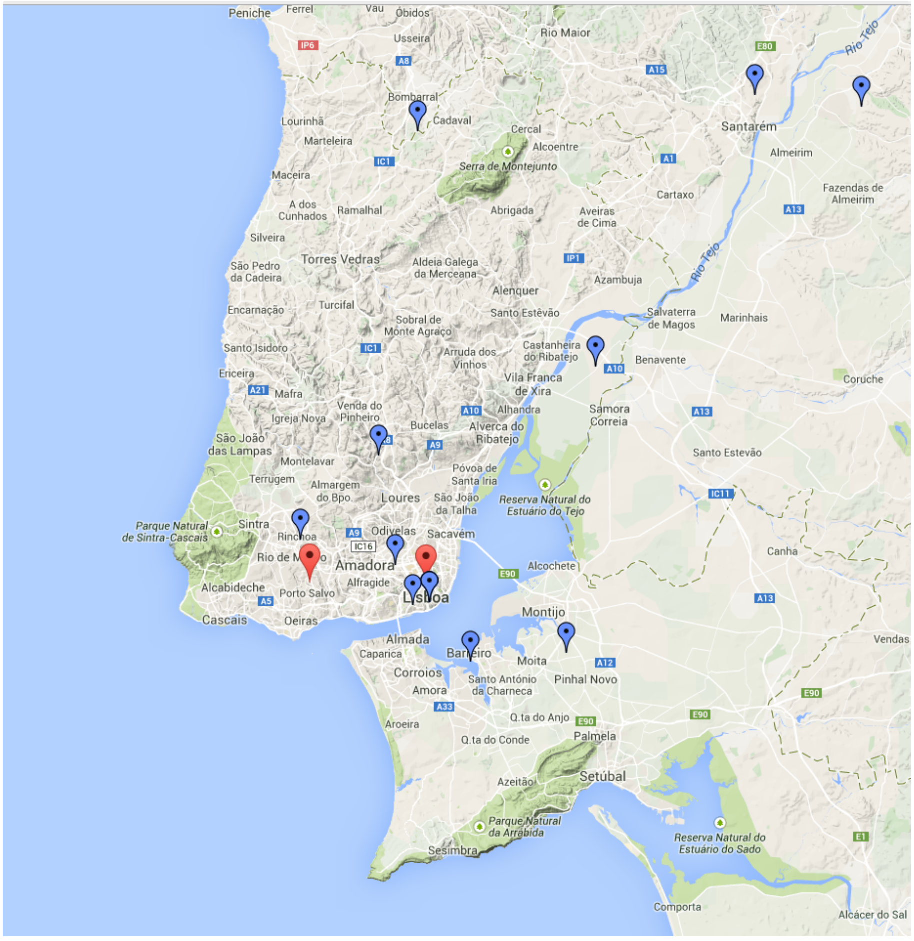


Figure 10: Distribution of Portuguese IST students by anonymized home address in Lisbon metropolitan area. Due to the anonymization process students were grouped, each blue dot representing a set of students. The red dots identify both IST campi. Data were anonymized using Incognito and a value of 10 for the k parameter.

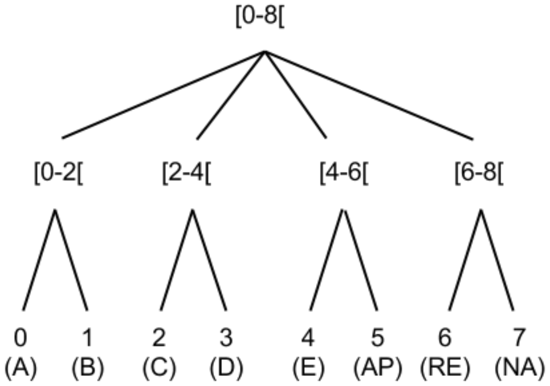


Figure 11: Value Generalization Hierarchy for the grade attribute in the Student Grades Dataset.

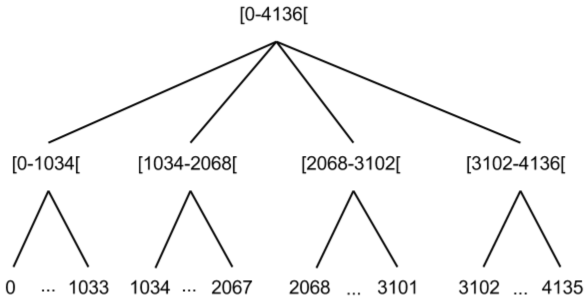


Figure 12: Value Generalisation Hierarchy for the subject attribute in the Student Grades Dataset.