Privacy-preserving Data Publishing For The Academic Domain

Pedro Rijo
pedro.rijo@ist.utl.pt

Instituto Superior Técnico, Lisboa, Portugal

October 2014

Abstract

This work addresses the release of a large academic information dataset. Since data is intended to be made accessible for a large audience, some measures need to be taken to protect individual privacy. It is shown how to build a system (FénixEdu-Priv) which is able to provide such data while considering privacy issues. FénixEdu-Priv is able to receive queries from external entities for accessing data and retrieve the desired data without exposing private information about academic agents. The resulting system enables an academic institution to easily provide information for external entities ensuring individual privacy. Furthermore, it allows to easily tune privacy levels on released data. This dissertation quantifies the impact of anonymization techniques over data utility and it also studies the impact of anonymization on behavioural patterns analysis. Released datasets will allow to better understand students and teachers, enabling the study of daily routines and improvement of the planning of many internal activities, such as cafeteria attendance, cleaning schedules or student performance. The dataset will also enable the study of interaction patterns on an academic population.
Keywords: Data Publishing, Privacy Protection, Privacy-Preserving Data Publishing, Social Network Analysis

1. Introduction

The continuous increase of stored data has raised interest on data analysis, due to the possibilities it can provide to organisations. Data mining techniques applied to such data enable the extraction of knowledge and interactions for customisation and adaptations of services to individuals. 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 techniques.

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 behaviours from online purchase 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. Privacy can be achieved through anonymization, for instance.

However, anonymization distorts data, decreasing its utility. Typical data mining results are highly dependent on data quality. Network inference[5] also suffers from anonymization, having the macroscopic properties from inferred networks changed when using anonymized data. This work explores multiple approaches for achieving the required privacy levels and analyses 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. It includes a study on data utility variation with the level of privacy in anonymization, and a comparison of the different methods available to achieve such anonymization. It also quantifies precision loss in network inference when underlying data is subject to anonymization techniques.

The proposed approach has been implemented at Instituto Superior Técnico (IST)[1], the school of engineering of the University of Lisbon, Portugal. The school's academic system, FénixEdu[2], manages all the internal information at IST. The amount of information on behaviours and human interactions

[1] http://tecnico.ulisboa.pt/
[2] http://fenixedu.org/

that can be inferred from such large dataset covering a sizable population (more than 60 000 people registered in the system) makes it appealing for many types of analyses, both within university and for anyone learning about this population. Furthermore, the plan to use FénixEdu in all University of Lisbon colleges, places FénixEdu-Priv into an higher relevance level, allowing access to information in 18 colleges and more than 50 000 people currently connected to the University.

This work intends to provide access to private data contained on FénixEdu system, while protecting individual privacy, as much of released data contains sensitive information about academic agents. Also, is intended to maximise data utility, i.e., minimise data distortion due to the application of privacy preserving techniques, for data analysis purposes and other studies from external entities.

In the remaining of this paper, I start by approaching privacy requirements and giving an overview of anonymization methods. Afterwards I approach the study of the effects of anonymization techniques and privacy definitions over data utility, showing how to provide valuable information without exposing individual privacy.

2. Background

Dalenius stated that, in privacy-protected datasets, 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 [4]. However, most literature on Privacy Preserving Data Publishing (PPDP) considers a more relaxed notion of privacy protection assuming that the attacker has limited background knowledge [4].

Fung provides a classification for privacy models based on its attack principles [4]. The classification distinguishes between 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 Quasi-identifier (QID) attributes 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 [14]. 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 from that victim. In this case we say that a record linkage attack occurred [4].

Anonymization techniques rely usually on generalisation and suppression operations for privacy preservation. Generalisation operations are applied based on a Value Generalisation Hierarchy (VGH) which 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 [10]. 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 in that records do not correspond to real-world entities represented by the original data [14].

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. To assess the information loss some metrics are required. In this work, we assess data utility of our academic dataset using metrics proposed by LeFevre [7] and Sweeney [13].

LeFevre's metrics consider the size of equivalence classes E of an anonymized table T 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 equivalence class containing t, equivalent to the following expression:

$$C_{DM} = \sum_{eqClassesE} |E|^2$$

As an alternative, a normalised average equivalence class size metric (C_{AVG}) may be used, although its value depends on k parameter:

$$C_{AVG} = \frac{totalRecords/totalEqClasses}{k}$$

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,

$$Prec(T) = 1 - \frac{\sum\sum \frac{h}{|DGH_{Ai}|}}{|PT| * |N_A|}$$

where DGH is the equivalent of VGH, $|PT|$ is the number of records of the Private Table being anonymized, and N_A the number of attributes belonging to the set of *QID*. The higher the precision, the higher the utility of the data, meaning that the anonymized data is more similar to the original dataset. *Prec* outputs values in the range 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 where 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. 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 [14]. These special atributes 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.

In this study, we focus on three methods for achieving privacy against record linkage attacks:

- Datafly [13] was the first algorithm to generate datasets satisfying the k-anonymity definition. Datafly uses a heuristic to make approximations, and so, it does not always yield optimal result and it may even distort data more than required.

- Mondrian [7] takes a multi-dimensional approach to achieve k-anonymity that provides an additional degree of flexibility. Often, this flexibility leads to higher-quality anonymizations

- Incognito [8] selects, from the multiple possible anonymizations for any given table, the quasi-identifiers that satisfy the privacy definition imposed by k-anonymity and the least generalised possible anonymization in the given VGH.

Independently of the method, one can choose the k parameter for k-anonymity. This parameter must be adjusted for the intended level of privacy. The quality of the anonymization is also dependent of the VGH provided for each attribute, since generalisation operations are based upon them.

3. Implementation

3.1. Architecture of FénixEdu-Priv

We now focus on the approach taken to build FénixEdu-Priv, which will provide access to IST internal data, managed by FénixEdu while respecting privacy restrictions.

The interaction between the user requesting the data and FénixEdu-Priv, starts with the user, an external entity, submitting a query for the desired data fields (See Fig. 1). FénixEdu-Priv processes the query retrieving the original data available at FénixEdu. After receiving anonymization configurations, by the data owner, the data is passed through the UTD Anonymization Toolbox. The anonymized output is then released to the user in CSV (comma-separated values) file format.

3.2. FénixEdu

The first task of the anonymization process is to extract the data to be later released by FénixEdu-Priv. This task can be decomposed in two main steps, Data selection and Data cleaning, and can be achieved by adding new services as an extension to the base FénixEdu software.

3.3. Common Problems

Databases often have many data problems associated with noisy data, specially after data migration processes. When analysing such databases it is fundamental to preprocess data with some data cleaning techniques. Typical problems include missing values and domain inconsistency. Missing values can have their origin in database schema changes, or some incorrectness when manually introducing information. Domain inconsistencies, in its turn, are record attributes with values that do not make sense in such attribute.

3.3.1 Data selection

The first data processing step in FénixEdu-Priv, involves choosing which information will be made available. The selection process must be analysed considering eventual stakeholders and the future purposes of the information. Also, an academic data anonymization system, such as FénixEdu-Priv, must be built in such a way that adding or updating information can be made with no major effort.

For an academic dataset one can think of personal information about students and teachers, class and schedules details, and curricular track

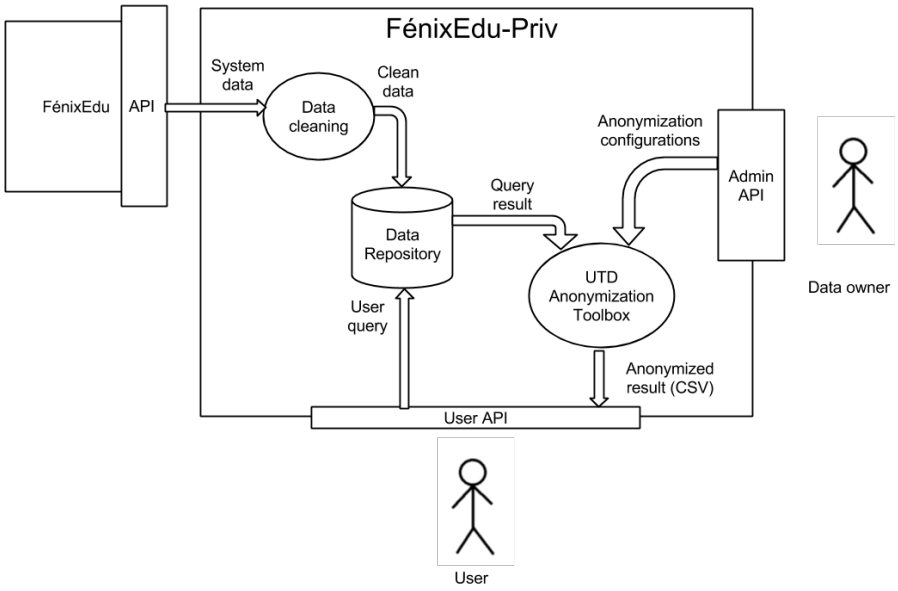

Figure 1: FénixEdu-Priv Architecture

records, both from students and teachers, to be a good start for available data. The full list of extracted fields from FénixEdu to FénixEdu-Priv can be consulted on `http://web.ist.utl.pt/ist167060/datastorm/doc.html`.

3.3.2 Data cleaning

Data cleaning is one of most important and time-consuming processes [1]. It involves detecting and correcting corrupt or inaccurate records from a record set. This problem is caused by dirty data present in the database, which may be due to insertion errors, data migration processes, or even the result of data integration runs.

Data cleaning may be viewed as a sequence of phases:

1. *Data analysis:* Detection of inconsistent cases.
2. *Definition of a transformation workflow and mapping rules:* Definition of the sequence of operations to apply to dirty data.
3. *Verification:* Validation of the correctness and effectiveness of a transformation workflow and transformation definitions.
4. *Transformation:* Execution of defined transformation steps.
5. *Backflow of cleaned data:* Replace dirty data when processed data achieves an acceptable state for the established goals

The original dataset, collected from FénixEdu contained some problems, most of them related to missing values for some attributes. More complex cases were also present. For example, data about foreign students at IST, usually from mobility programs such as Erasmus[3], contain several issues, such as postal codes or phone numbers which were not conforming to the national (Portuguese) standards. Those records need special attention because they are subject to different conventions.

But even ignoring foreign student issues, many problems still arise with the data. One of the problems is that many attributes do not respect attribute domain rules. For instance, *Birth Year* had two problems: missing values, and domain inconsistencies. For this attribute some records presented values lesser that 1900 or greater than 2000. Such individuals are just too old or too young to be registered in this system. *Postal Code* had format problems. *Classrooms* have more complex missing values, since there were changes in the classroom names that happened along institution lifetime. These kind of problems were not corrected due to the complexity of the solution that would involve comparing older with actual blueprints. Besides domain problems, there were also technical issues. *Locality* for example had some values containing meta-characters due to encoding problems.

The chosen solution was to match problematic fields against a regular expression and make some adjustments. For dealing with invalid postal codes we matched every single postal code with an online database[4], removing those that were invalid.

[3] `http://www.erasmusprogramme.com/`
[4] `http://www.geonames.org/`

4. Results

In the previous sections I have described my approach to privacy preserving data publishing and presented current privacy preserving techniques, providing insights into the decisions behind the design of the proposed system.

In some cases the distortion suffered on original data applied by anonimization techniques can block useful conclusions in latter analysis. To assess precision loss I performed a set of experiences using real-world data from FénixEdu-Priv. I evaluated Datafly, Mondrian and Incognito methods, assigning different privacy levels through the assignment of different values to the k parameter and measuring data utility besides typical time performances.

4.1. Experimental Data and Setup

To assess the efficiency of FénixEdu-Priv two different approaches were taken: the first tries to provide a guideline for data administrators to chose the most suitable method and value to the k parameter according to each situation, comparing some combinations for a set of different queries; the second approach analyses the impact of anonymized data on latter studies, using as test case the inference of social networks from retrieved data by FénixEdu-Priv.

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

4.2. Administrator Guidelines

For this experiment, two different queries were tested, each one independently analysed. The first retrieves, for each student in the system, the Postal Code, in a total of 66,809 records. The other query retrieves, for each student, the respective grade for each enrolled subject, in a total of 321,203 records.

It is possible to distinguish two kind of attributes in this dataset for anonymization purposes: attributes with *logical domains*, for instance, dates can be naturally grouped in months, years, decades, and so on; and attributes with *non-logical domains*, for instance, person identifiers have not a natural hierarchy since we do not have any extra information about the referred person. The latter kind of attributes make it harder to find the optimal VGH that would maximise data utility.

The experiments were conducted with the three suitable methods provided by the UTD Anonymization Toolbox (Datafly, Mondrian, Incognito). For comparing such methods, metrics previously described for precision and data distortion were used.

The metric of precision cannot be applied to Mondrian since its implementation on the UTD Anonymization Toolbox does not respect the defined VGH for generalisation operations and there is not a practical method to find out the VGH used. This makes Mondrian unsuitable for logical domains since Mondrian can make arbitrary classes without taking into account the VGH. Nevertheless, Mondrian can be useful in non-logical domains since it helps to find a good VGH, making the query auditor work easier. In logical domains, Incognito is the best option to maintain logical generalisation. C_{DM} and C_{AVG} metrics present values hard to read but that can be used to compare among the three methods. Note that C_{AVG} depends on k.

4.2.1 Personal Postal Code

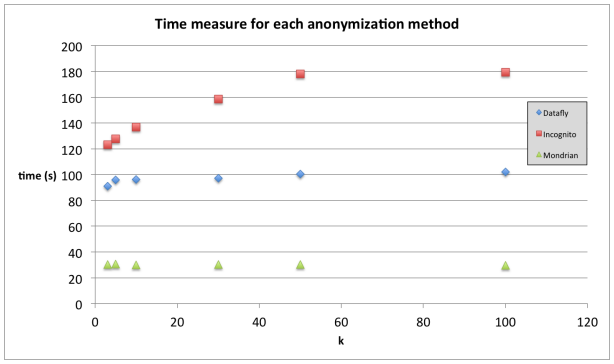

Figure 2: CPU time for each anonymization method on Postal Code dataset

The dataset of postal codes contains the set of postal codes for each student in a total of 66,809 records. The dataset is characterised by a logical hierarchy on the attribute. Postal code assigns a region for each 4-digits group, while the last 3 digits represent the mailman route. In this work we ignore those 3 digits since they do not contain any hierarchy. The assignment of the postal code is made such that regions with similar codes are close regions. For example, the code 2675 (Odivelas) is near 2676 (Amadora). This property makes the VGH intuitive as each level groups one more digit, starting by the least significant.

The results for the three suitable methods with multiple privacy levels over this dataset are shown in the Appendix A. 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 maintained. Mondrian gets a different precision, although not reliable due to its behaviour of

ignoring the provided VGH.

Anonymization decreases data utility with a significant impact for this dataset. Observing the dataset and the corresponding 2-digit prefix distribution available in the Appendix B we see that some prefixes, such as 43, 52, 94, 33, 98 and 72 have a low count, with values smaller than 30 people. We find that those postal codes refer to the areas of Porto, Bragança, Madeira, Coimbra, Azores and Évora. Porto, Bragança and Coimbra have low representativeness in IST since there are other major academic institutions in these areas, offering the same set of courses. Students of Madeira and Azores archipelagos are distributed in mainland Portugal. In addition to featuring an university, Évora is being severely affected by population ageing, with very few young people, therefore it has a very low number of students joining college. On the other edge of the distribution we have areas such as Lisboa and Setúbal (prefixes 26, 27 and 28) with high count of students due to the proximity with IST.

The existence of regions with low count of students joining IST reduces the utility of general dataset since the UTD Anonymization Toolbox implementation requires that every leaf of the VGH is at the same depth. One possible workaround would be to group manually some of the regions that are somehow related until all regions get a bigger count.

For better visualisation of the impact of anonymization, two maps were created and are shown in the Appendix C: the first one, present on Fig. 5, shows the distribution of IST students in Lisbon metropolitan area by its home address; the second map, shown in Fig. 6, presents the distribution of IST students through Lisbon metropolitan area by its home address using anonymized data. This data was anonymized using Incognito and a value of 10 for the k parameter. To infer the location of anonymized postal code it was calculated the centroid of all existent postal codes on the equivalence class of each student. It is possible to observe in Fig. 6 that students were clustered according the respective equivalence class of theirs postal code.

Regarding performance issues, Fig. 2 shows the time taken by each method on the anonymization process for handling the $66,809$ records. 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 has a bigger time requirement for the anonymization process, since it tests every possible anonymization.

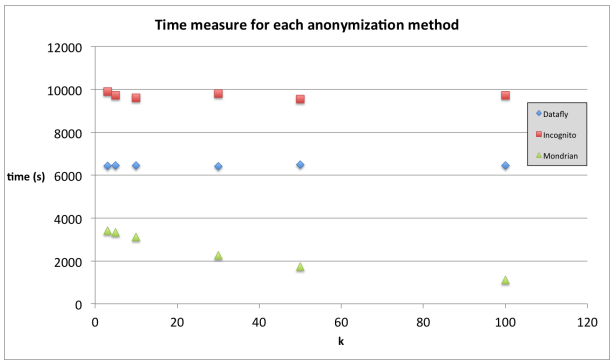

Figure 3: CPU time for each anonymization method on Student Grades dataset.

4.2.2 Student grades

This dataset contains the set of grades of each student at each subject they have enrolled in IST.

Grades were converted to the European scale, having values between A and E, and three other values: AP (Approved), RE (Reproved/Not-approved), and NA (Not available/Not evaluated). Grades were then grouped in {A,B}, {C,D}, {E,AP}, and {RE,NA}.

In this case, the *QID* attributes are *Subject* and *Grade* attributes, leading to the need of defining a VGH for subject in addition to the VGH for grades. Since there are more than $4,000$ subjects on the dataset and no extra information is available about the subjects, there is no scope for a logical VGH. It also becomes impractical to create an optimal VGH since we would need to try many combinations. The solution was to create two levels besides the original value: set of quartiles of ground values, corresponding to a total of 4 classes, each containing 1034 subjects; root class which groups all subjects. Most likely in such cases, where the definition of a VGH is simplified, Mondrian will create a more efficient VGH resulting in better results.

The results for the three suitable methods with multiple privacy levels over this dataset are shown in the Appendix D. Once again, Incognito achieves results that maximise the data utility. But, in this case, Datafly, which was predicted to overgeneralise, presents the same results as Incognito, using both metrics. Mondrian on the other hand, has achieved less precision (note that, as mentioned before, precision is not exact when applied to Mondrian). Despite that fact, Mondrian has presented less data distortion according the re-

maining metrics. This happens because Mondrian created more equivalence classes, and consequently with 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 on any of the levels), leaving the *grade* attribute untouched, Mondrian generalises both attributes, distributing the generalisation distortion by them. To increase achieved precision, it would be useful to define a more precise VGH, with more levels, since we can observe that every recorded has generalised that attribute one level, leading the remaining attributes with the original values.

Just as in the previous experiment, Fig. 3 shows the time taken by each method on the anonymization process for the dataset. Similarly to what happens in the postal code dataset, Incognito is the most time consuming method, while Mondrian keeps to be the fastest to producing the output. Despite what we observed before, Incognito has not been influenced by the k parameter, although such influence is visible in Mondrian, despite the fact that in the postal code dataset it was possible to observe the reverse: Incognito was influenced by k parameter; and Mondrian was not. Incognito approximately maintains its time requirements since precision also maintains the same value. This means that, for the first anonymization level, the dataset contains all equivalence classes with more than 100 individuals, and so, does not needs to perform more generalisation operations. On the other hand, Mondrian decreases time requirements with increasing k. This behaviour can be explained by the fact that, in order to maximise data utility, Mondrian keeps partitioning equivalence classes. For bigger values of k the maximum size of the equivalence classes is reached first, and for smaller values it takes more partitioning operations to achieve the desired result.

4.2.3 Method Comparison Conclusions

With these two experimental datasets it is possible to conclude that Incognito is the best method for most situations. Situations where the VGH has not a logical hierarchy and it becomes hard to find an optimal, or good, hierarchy can be more appropriate for Mondrian, which 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 on the less populated class of the VGH level. Regarding time requirements, Incognito has always bigger time requirements, while Mondrian has the fastest processing time. The available methods are dependent on the k parameter and on the data itself. The presented methods depend on the equivalence class distribution across the dataset, being, in some situations, affected.

Table 1: Metrics for anonymization with different values for k parameter

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

Despite these results, FénixEdu-Priv allows the data owner to test different parameters before choosing the most adequate anonymization method.

4.3. Anonymization Impact on Latter Studies

Besides the comparison of the impact of each anonymization method on data utility, I have approached the effect of anonymization on latter studies. Through the adherence of the dataset to k-anonymity, data has decreased its utility. In this section it is given an overview on the influence of such decrease on data utility in knowledge extraction tasks. I tested it with the inference of social network between teachers and students, based upon lectured and attended classes. Two networks were considered:

Student-Shift-Teacher The main network. Represents the interactions between students with enrolled shifts and teachers with lectured shifts. From that a tri-partite 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 shift.

Note that both networks are undirected. To infer those networks the submitted query returns records of the type:

T(Student, SubjectGrade, AttendedShift, Teacher, TeacherCUQ)

The query selected, for each student and attended shift, the final grade of the student at such subject, as well as the lecturing teacher and the CUQ grade[5].

For this experiment, a sample of the population was selected. Since data was grouped by student and shift, the best approach is to sample blocks of rows. This decision was due to locality principle. 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 different course or different academic period, making relationships almost nonexistent. From 1.419.649 records around 10% were selected in a total of 142.000 records.

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

QID={SubjectGrade, AttendedShift, Teacher-CUQ}

Similarly to the previous experiment, StudentGrade was converted to the European scale. Following the previous analysis in Section 4.2, Incognito was chosen for anonymizing query results. I present the metrics for anonymizing the dataset for various values of k in Table 1. Network inference was made upon original and anonymized data for comparison.

4.3.1 *Student-Shift-Teacher* network:

This first network is a tri-partite graph where *Student*, *Teacher* and $Shift$ are the entities. Table 2 shows how inferred network properties change with anonymization level, tuned through k parameter.

It is possible to observe how network properties change with increasing k. k affects the QID, but $AttendedShift$ has been the only attribute which has suffered generalisation. In practice, when k increases it starts grouping some of the nodes correspondent to $AttendedShifts$. With the decreasing of the number of nodes, it is not surprising that the number of edges also decreases since where before could be one edge from a *Student* for several $Shifts$, now there is only one to the cluster created by generalisation. But having $Shifts$ grouped increases both the indegree and the outdegree for every $Shift$ node now. With the decrease on the number of edges, average distance and graph diameter also decreases. Using $k = 10$ as reference, one can observe that graph properties, such average distance and graph diameter, have values 72% and 77% similar to the original network. Also, the number of nodes suffers big differences (anonymized network has about 30% nodes of original network), the number of edges keeps similar (82%). The difference on the number of nodes justifies the 43% difference registered on average degree.

4.3.2 *Student-Student* network:

The $Student - Student$ network is inferred from equal enrolments by students. Two students interact if they enrolled on the same shift. The evolution of network's properties is shown in Table 3.

Similar to the first network, it was expected that grouping *shifts* lead to more connected *students*. Obviously, the number of nodes does not decreases since *students* are not affected by generalisation. Having students more connected is visible in the average distance, graph diameter, and nodes degrees, that decrease with increasing k. Once again, using $k = 10$ as a reference, we get a anonymized average distance 72% of the real value and a diameter with 68% accuracy. But when dealing with the number of edges it grows up to more than twice the original value, occurring the same in the remaining degree properties as a consequence.

5. Conclusions

Data publishing may be a powerful tool for solving real world problems, allowing for instance to find a better solution for planning resources and scheduling. Even though the power provided by releasing such data could help significantly general population, it can also represent a fragility to individual privacy. In particular, releasing large datasets where personal interactions could be inferred without addressing privacy issues can make it possible to an attacker to obtain undesired information about his target victims. Privacy-preserving data publishing emerged from the opportunity that data provided for mining purposes could solve real world problems.

In FénixEdu-priv, the data owner can test different parameters before choosing the best suited anonymization method. However, this capability is not sufficient, and in fact gives little

[5]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/`

Table 2: *Student - Shift - Teacher* network properties for different anonymity levels

k	Avg Dist	Diameter	Nodes	Edges	Min Deg	Max Deg	Avg Deg
original	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 3: *Student - Student* network properties for different anonymity levels

k	Avg Dist	Diameter	Nodes	Edges	Min Deg	Max Deg	Avg Deg
original	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

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 could be requested for a variety of useful purposes, corresponding to queries covering multiple aspects of academic life. However, the academic system has a large variety of data that could be used to infer implicit user behaviours, making re-identification become easily achievable.

On the context of interaction network inference, anonymization plays a non-negligible role altering macroscopic properties of inferred network in comparison with the network inferred from original data. This is visible in characteristics, such as average distance or graph diameter. This make studies of this kind to produce non conclusive results.

5.1. Achievements

With this work several methods for achieving k-anonymity were studied, comparing both achieved data utility and time requirements. As mentioned before, there is not a best method for all scenarios, since the various methods act differently, dealing best with different domains. The presented results may provide useful guidelines for future applications of such techniques.

With the creation of FénixEdu-Priv it becomes easier to provide safe access to private information for each academic institution [11], specially in the 18 colleges adopting FénixEdu in the near future. With such amount of information available, an opportunity without precedents to study and improve this environment is opened.

The analysis on different techniques for protecting individual privacy and the comparison among them, namely in what concerns social network inference [12], provide useful guidelines for future interested on such topic.

5.2. Future Work

The achieved results evidence that k-anonymity may not be the best technique, conclusion highlighted when studying anonymized data against original one in the inference of interaction networks. In fact, all privacy models offering a priori privacy guaranteed (differential privacy [3], t-closeness [9]) entail a great utility loss, because they place privacy first and utility second. Even if other approaches besides k-anonymity may also be approached, k-anonymity still presents other methods that should be exploit: k-anonymity may be achieved through other methods besides generalisation, for example micro aggregation of the quasi-identifier attributes [2]. In addition to k-anonymity, newer approaches, such as posteriori disclosure risk protection[6, 15] may provide more interesting results from the data utility point of view, because make 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 reflect anonymization generalisation? If so, it could be still interesting to analyse clustering in networks built from anonymized data, without losing much information. Still regarding the study of interaction networks, that can be seen as equivalents to social networks, in anonymized data, a recent field may prove to be more adequate to this theme: Social Network Anonymization [16]. This area does not only considers typical attributes but it also takes into account the network structure, that could be used to re-identification attacks otherwise.

Unfortunately, those alternative approaches suffer from lack of free available tools, turning its application much more confined to specific professionals.

References

[1] A. Doan, A. Halevy, and Z. Ives. *Principles of data integration*. Elsevier, 2012.

[2] J. Domingo-Ferrer and V. Torra. Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery*, 11(2):195–212, 2005.

[3] C. Dwork. Differential privacy: A survey of results. In *Theory and Applications of Models of Computation*, pages 1–19. Springer, 2008.

[4] 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.

[5] M. Gomez Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. page 1019, New York, New York, USA, 2010. ACM Press.

[6] 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.

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

[8] 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, pages 49–60, New York, NY, USA, 2005. ACM.

[9] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pages 106–115, April 2007.

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

[11] P. Rijo, A. P. Francisco, and M. J. Silva. Privacy-preserving data publishing in academic information systems. 2014.

[12] M. J. Silva, P. Rijo, and A. P. Francisco. Evaluating the impact of anonymization on large interaction network datasets. 2014.

[13] 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.

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

[15] L. Willenborg and T. De Waal. Statistical disclosure control in practice. *Lecture notes in statistics*, 155, 2001.

[16] B. Zhou, J. Pei, and W. Luk. A brief survey on anonymization techniques for privacy preserving publishing of social network data. *ACM SIGKDD Explorations Newsletter*, 10(2):12, Dec. 2008.

A. Data utility for anonymization over postal codes dataset.

	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

Table 4: Data utility 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.

B. Postal Code 2-digit Prefix Distribution for IST dataset

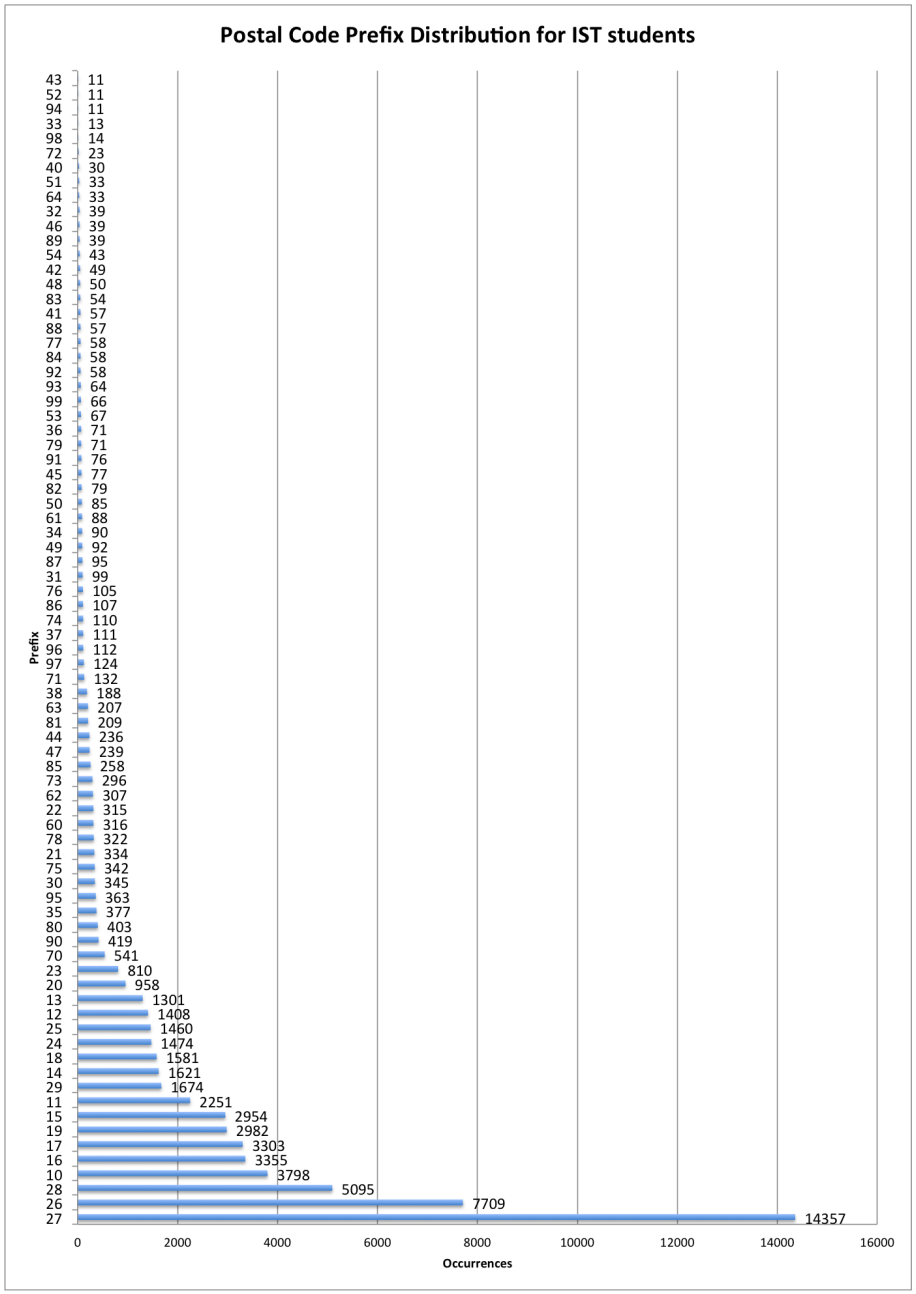

Figure 4: Postal Code Prefix Distribution

C. IST Portuguese students home address distribution using original and anonymized data

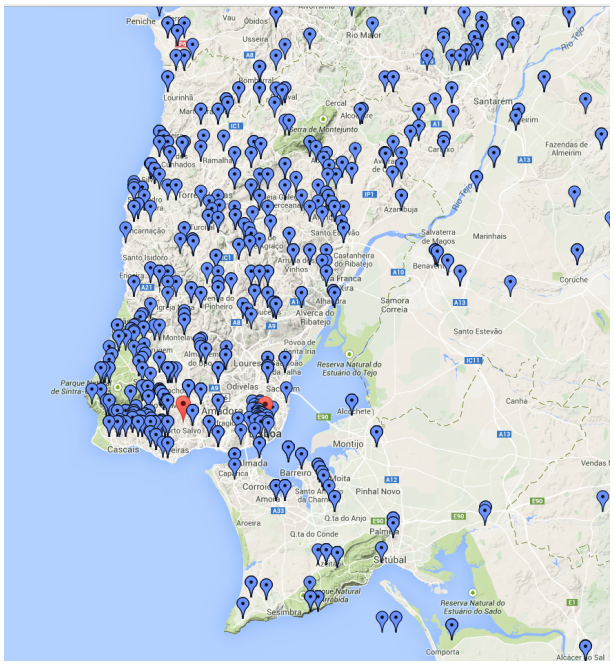

Figure 5: Distribution of Portuguese IST students by home address in Lisbon metropolitan area. The red dot signs IST Campus.

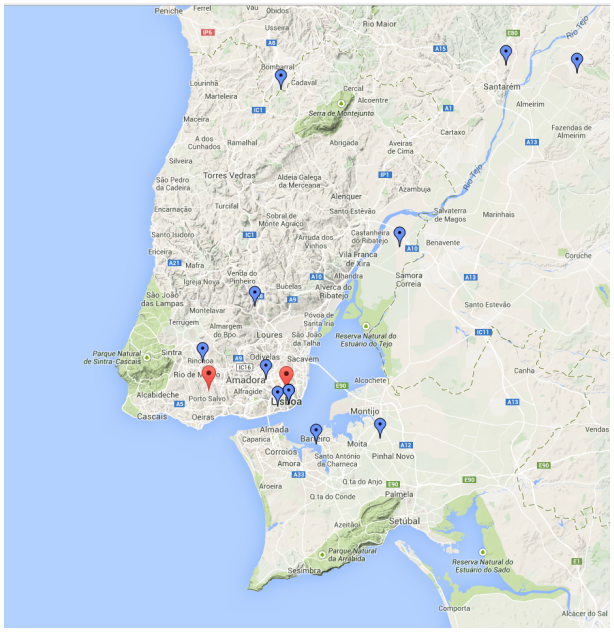

Figure 6: Distribution of Portuguese IST students by anonymized home address in Lisbon metropolitan area. Due to the anonymization process students were grouped, each blue dot representing a set of students. The red dot signs IST Campus.
This data was anonymized using Incognito and a value of 10 for the k parameter.

D. Data utility for anonymization over students grades dataset.

	k	Prec	C_{DM}	C_{AVG}
Datafly	3	0.75	4,928,090,203	3,568.92
	5	0.75	4,928,090,203	2,141.35
	10	0.75	4,928,090,203	1,070.68
	30	0.75	4,928,090,203	356.89
	50	0.75	4,928,090,203	214.14
	100	0.75	4,928,090,203	107.07
Mondrian	3	0.64*	28,403,897	10.81
	5	0.64*	28,465,783	6.98
	10	0.63*	28,847,609	4.17
	30	0.59*	32,933,353	2.31
	50	0.57*	39,615,933	1.96
	100	0.54*	61,595,543	1.71
Incognito	3	0.75	4,928,090,203	3,568.92
	5	0.75	4,928,090,203	2,141.35
	10	0.75	4,928,090,203	1,070.68
	30	0.75	4,928,090,203	356.89
	50	0.75	4,928,090,203	214.14
	100	0.75	4,928,090,203	107.07

Table 5: Data utility for anonymization of students grades 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 present only for comparison.