Concordia University

INSE 6180 Security and Privacy Implications of Data Mining(SUMMER 2020)

Analyzing Privacy Vulnerability of the Dataset, and Propose How to Prevent Attacks

DELIVERABLE 1 (D1)

Team members:

Peixing Ma (40080597)
Daniel Zakerifar (40054463)
Eyob Tessema (40059340)

July 11, 2021

1.1 Introduction

Recently, one of the subjects that many people has been studied on that is Publishing a data set with preserving privacy. It happend alot that some sensetive information of individuals compromised when the sencus financial data of the people, health profile data has been revealed. Privacy-Preserving Data Publishing (PPDP) is mainly effective at the time of modeling any database to table schema, where each row collectively contains information of a human being or a subject system[4].

Figure 1 shows data collection and publishing. Data publisher takes data from record owners (e.g., Arash and Cathy). For publishing data, the data publisher releases the data to a data miner or to the public, called the data recipient. In our work, a school collects data from students and publishes the records to an external center. So here the school is the data publisher, students are record owners, and the external center is the data recipient.In some data publishing scenarios, it is important that each published record corresponds to an existing individual in real life[3].

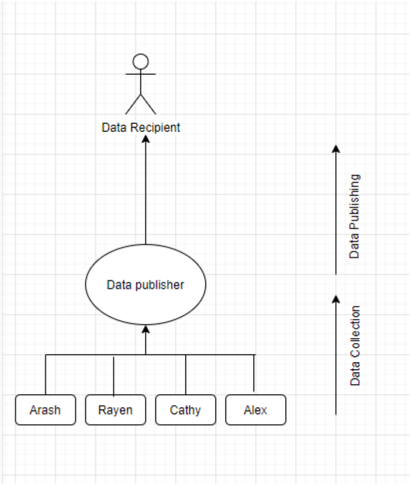


Figure 1.1: Data Collection and Data publishing

1.1.1 The Anonymization Approach

Distinct row: a row is a sequence of values with respect to corresponding attributes in a table. All rows with the same sequence of values are considered to be a distinct rows. In the most basic form of PPDP, the data publisher has a table of the form D(Explicit Identifier, Quasi Identifier, Sensitive Attributes,Non-Sensitive Attributes), where Explicit Identifier is a set of attributes, such as name and social security number (SSN), containing information that explicitly identifies record owners;
Quasi identifiers (QID) are attributes in a data set which are strongly related to and dependent on a person, and are used to de-anonymize individuals in the linkage attack. Example, sex, date of birth, etc.;
Sensitive Attributes consists of sensitive person-specific information such as disease, salary, and disability status;
Non-Sensitive Attributes contains all attributes that do not fall into the previous three categories.

Anonymization refers to the PPDP approach that seeks to hide the identity and/or the sensitive data of record owners, assuming that sensitive data must be retained for data analysis. Clearly, explicit identifiers of record owners must be removed. Even with all explicit identifiers being removed, a real-life privacy threat . For instance the date of birth, gender and postal code from a medical condition record to the voter registration list. Each of these attributes does not uniquely identify a record owner, but their combination, called the quasi- identifier , often singles out a unique or a small number of record. Usually, for the sake of protecting privacy, a data set is anonymized by removing some important information, such as name, of a person. However, this simple treatment does not guarantee the privacy. An adversary can easily find out the identity of a person by linking a combination of some seemingly unimportant attributes to another data set which also contains these attributes. This way of de-anonymization is called linkage attack[3]

1.1.2 ATTACK MODELS AND PRIVACY MODELS

For getting access to private data, attackers try to attack privacy models in diffrent ways. according to ways of attack on data, Privacy models can be differentiated in two separate classes.

1: Linkage attacks

1: Usually, for the sake of protecting privacy, a data set is anonymized by removing some important information, such as name, of a person. However, this simple treatment does not guarantee the privacy. An adversary can easily find out the identity of a person by linking a combination of some seemingly unimportant attributes to another data set which also contains these attributes. This way of de-anonymization is called linkage attack. A very famous example is the re-identification of Massachusetts Governor by linking the date of birth, gender and postal code from a medical condition record to the voter registration list.

2: probabilistic attack

Attackers have an intention to gain more information along with available background knowledge in the published record. This can be classified as a probabilistic attack if the difference between attackers prior and post beliefs are found. Figure 2 shows the different privacy models and attacks.

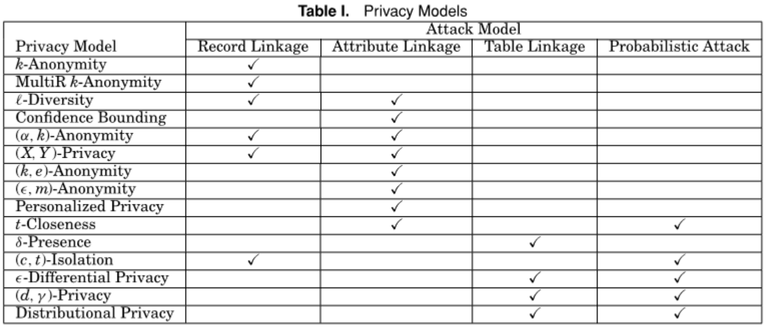


Table I. Privacy Models

	Attack Model			
Privacy Model	Record Linkage	Attribute Linkage	Table Linkage	Probabilistic Attack
k-Anonymity	✓			
MultiR k-Anonymity	✓			
ℓ-Diversity	✓	✓		
Confidence Bounding		✓		
(α, k)-Anonymity	✓	✓		
(X, Y)-Privacy	✓	✓		
(k, e)-Anonymity		✓		
(ϵ, m)-Anonymity		✓		
Personalized Privacy		✓		
t-Closeness		✓		✓
δ-Presence			✓	
(c, t)-Isolation	✓			✓
ϵ-Differential Privacy			✓	✓
(d, γ)-Privacy			✓	✓
Distributional Privacy			✓	✓

Figure 1.2: Privacy Models

Different types of attacks on privacy models are given below:[3]

Linkage Attacks

Record Linkage

In record linkage attackers try to match a value with values at a table and try to find out message or storage related to that value. In this attack, there is a probability to identify a victims record perfectly with help of additional information or knowledge.

Attribute Linkage

Attack that is performed with help of linking an attribute of record is considered as Attribute linkage. Through this process, an attacker may not get all the records but he/she might get some sensitive information regarding Victims' associated group.

Table Linkage

In table linkage attackers are able to find out all types of information stored in a table related to the victim. Both record linkage and attribute linkage are combined in table linkage

Probabilistic Linkage

This type of attack is not directly linked with the record, attribute or table of a record. Except these, probabilistic type attack mainly deals with attackers' belief on sensitive data of a record which is already published and analyzed by the attackers. This type of attack briefly can create a difference between prior belief and posterior beliefs of data to an attacker[4]. For each attack there is a privacy model. The following is the privacy model:

Privacy Models

k-Anonymity

To prevent record linkage through QID, we can use the notion of k-anonymity: Let T be a table with attributes A1, A2. . . .An, and QI = Ai...Aj where Ai...Aj in A1,...,An be the quasi-identifiers associated with it. And T[A1,..,An] represents a table T with attributes A1,..,An. Then, T is said to satisfy k-anonymity if and only if each sequence of values in T[QI] appears with at least k occurrences in T[QI].

In another words, the minimum group size on QID is at least k. A table satisfying this requirement is called k-anonymous.In a k-anonymous table, each record is indistinguishable from at least k 1 other records with respect to QID. Consequently, the probability of linking a victim to a specific record through QID is at most 1/k.

l-Diversity.

diversity principle, called l-diversity, to prevent attribute linkage. The l-diversity requires every qid group to contain at least l "well-represented" sensitive values.

δ-Presence

To prevent table linkage,we can bound the probability of inferring the presence of any potential victim's record within a specified range $\delta = (\delta min,\delta max)$. [3]

1.2 Vulnerability analysis

Vulnerability means that the way that a data has the ability to compromise an individuals in a data set. In this section we want to analyse our date to find the most vulnerable part of our data. First we have to introduced our data. We got our data from [5]

1.2.1 Dataset information

This data approach student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). In [Cortez and Silva, 2008], the two datasets were modeled under binary/five-level classification and regression tasks. Important note: the target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade[5].

Attribute Information

Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets:
1 school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2 sex - student's sex (binary: 'F' - female or 'M' - male)
3 age - student's age (numeric: from 15 to 22)
4 address - student's home address type (binary: 'U' - urban or 'R' - rural)
5 famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6 Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â“ 5th to 9th grade, 3 â“ secondary education or 4 â“ higher education)
8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â“ 5th to 9th grade, 3 â“ secondary education or 4 â“ higher education)
9 Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), "at_home" or 'other')
10 Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11 reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12 guardian - student's guardian (nominal: 'mother', 'father' or 'other')
13 traveltime - home to school travel time (numeric: 1 less than 15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 greater 1 hour)
14 studytime - weekly study time (numeric: 1 - less than 2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - greater than 10 hours)
15 failures - number of past class failures (numeric: n if n is between 1 and 3, else 4)
16 schoolsup - extra educational support (binary: yes or no)
17 famsup - family educational support (binary: yes or no)
18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19 activities - extra-curricular activities (binary: yes or no)
20 nursery - attended nursery school (binary: yes or no)
21 higher - wants to take higher education (binary: yes or no)

22 internet - Internet access at home (binary: yes or no)
23 romantic - with a romantic relationship (binary: yes or no)
24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29 health - current health status (numeric: from 1 - very bad to 5 - very good)
30 absences - number of school absences (numeric: from 0 to 93)
these grades are related with the course subject, Math or Portuguese:
31 G1 - first period grade (numeric: from 0 to 20)
31 G2 - second period grade (numeric: from 0 to 20)
32 G3 - final grade (numeric: from 0 to 20, output target)

1.2.2 Data Analysis

First we have to visualized our data

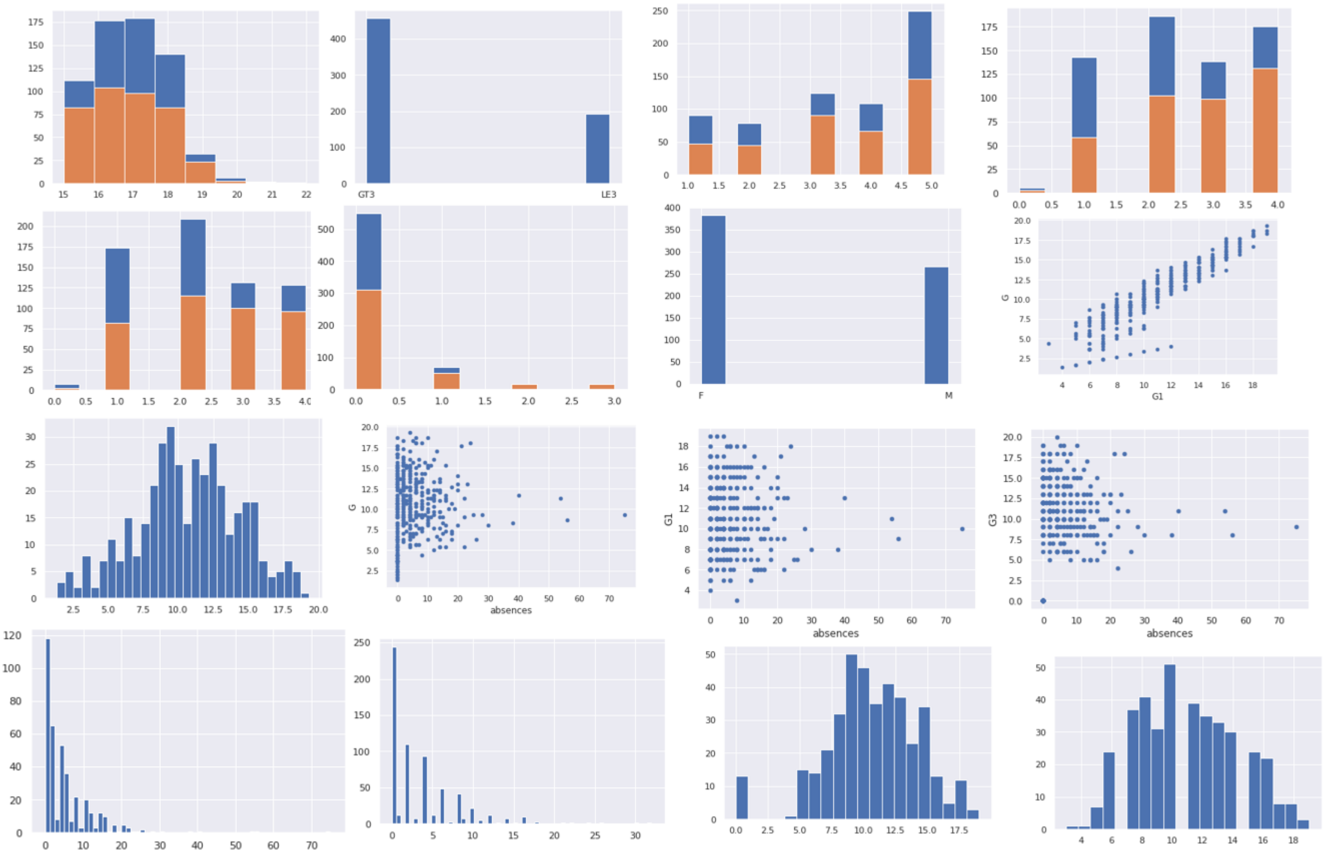
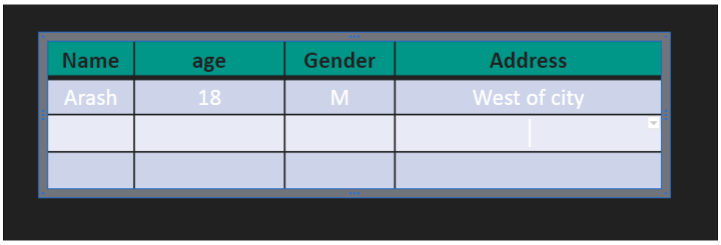


Figure 1.3: Data Visualisation

By looking at the last 7 pictures, we have found that this Data has been manipulated by two things. First is number of the absences. According to that, some even absences has been removed and added to the odd numbers. Also, they manipulated G1 by removing some grades and kept the average.

Finding Vulnerabilities

At the first look at the graph and data set,also by some investigate we can do the following attacks according to the vulnerabilities:
1: Background Knowledge Attack: By looking at the data, we can find that by having a some information about a person, we can compromise the person. More over we can find more information about the victim(the person who has been compromised). For instance, if some one knows about the number of absences it could be compromised. Specially students knows more each other that who came and who did not come to the class. Suppose one of the students wants to know about his/her friend grades: and he/she knows some of the QID identifier from his/her friend. So he/she tries to match that information with the table. In the worst scenario he/she could have the probability of finding the grade of the student.
2: Record linkage attack(A) Supposed that the school wants to publish to a data center. Suppose that the research center has access to the following table. By having these both table they can identify a person about grade.



Name	age	Gender	Address
Arash	18	M	West of city

Figure 1.4: Sample of a table for record linkage attack

3: Record linkage attack(B) In this data set there are a lot of cases that happened once. For example By querying number of students who are more than 19 years old, we only 5 records. With help of other attributes like sex and their parents' jobs, these students can be easily identified. Or number of individuals who are 22 is just one.

1.3 Solution

For providing more privacy, we have to anonymize more information of the individuals. At the first step we can classify the students with number of absences near each other. we classify the absences like the following: 1-3,4-5,6-8, 9-10, 11-15, 16-20 More than 20 It also has has enough L-Diversity. The rest of our work is how to anonymize the data.

1.3.1 Quasi-identifiers

According to the definition of the k-anonymity, we need to find out a set of quasi-identifier from the data set to calculate the value of k. Here is an algorithm to select quasi-identifiers [2]

1.3.2 Algorithm of selecting quasi-identifiers

1. Select a set of attributes, which are strongly related to or dependent on a person, based on your judgement.

2. Find the power set P of the set of selected attributes.

3. Construct tables with each element in P and count the number of distinct rows of each table.

4. Quasi-identifiers are the set of attributes which construct the table with maximum number of distinct rows. If there is a tie, select the set of attributes with smallest size.

1.3.3 Experiment

For our data set, we select a set of attributes $A =$ school, sex, age, Mjob, Fjob, guardian. Then we wrote a program get the the power set $P(A)$ of the set A and counted number of distinct rows of each table constructed by an element of $P(A)$. For example, school and sex, age, guardian are elements of $P(A)$. The result is in figure 1:

set	number	set	number
{school}	2	{sex}	2
{school, sex}	4	{age}	8
{school, age}	12	{sex, age}	14
{school, sex, age}	21	{Mjob}	5
{school, Mjob}	10	{sex, Mjob}	10
{school, sex, Mjob}	20	{age, Mjob}	29
{school, age, Mjob}	41	{sex, age, Mjob}	51
{school, sex, age, Mjob}	69	{Fjob}	5
{school, Fjob}	9	{sex, Fjob}	10
{school, sex, Fjob}	18	{age, Fjob}	28
{school, age, Fjob}	39	{sex, age, Fjob}	49
{school, sex, age, Fjob}	66	{Mjob, Fjob}	24
{school, Mjob, Fjob}	37	{sex, Mjob, Fjob}	45
{school, sex, Mjob, Fjob}	66	{age, Mjob, Fjob}	89
{school, age, Mjob, Fjob}	109	{sex, age, Mjob, Fjob}	137
{school, sex, age, Mjob, Fjob}	160	{guardian}	3
{school, guardian}	6	{sex, guardian}	6
{school, sex, guardian}	12	{age, guardian}	18
{school, age, guardian}	26	{sex, age, guardian}	31
{school, sex, age, guardian}	44	{Mjob, guardian}	15
{school, Mjob, guardian}	27	{sex, Mjob, guardian}	28
{school, sex, Mjob, guardian}	47	{age, Mjob, guardian}	58
{school, age, Mjob, guardian}	75	{sex, age, Mjob, guardian}	92
{school, sex, age, Mjob, guardian}	115	{Fjob, guardian}	14
{school, Fjob, guardian}	23	{sex, Fjob, guardian}	26
{school, sex, Fjob, guardian}	41	{age, Fjob, guardian}	57
{school, age, Fjob, guardian}	69	{sex, age, Fjob, guardian}	89
{school, sex, age, Fjob, guardian}	107	{Mjob, Fjob, guardian}	52
{school, Mjob, Fjob, guardian}	75	{sex, Mjob, Fjob, guardian}	88
{school, sex, Mjob, Fjob, guardian}	116	{age, Mjob, Fjob, guardian}	134
{school, age, Mjob, Fjob, guardian}	152	{sex, age, Mjob, Fjob, guardian}	189
{school, sex, age, Mjob, Fjob, guardian}	207		

Figure 1. result of selecting algorithm for quasi-identifiers

From the table above, we select the set of attributes which construct a table with maximum number of distinct rows, {school, sex, age, Mjob, Fjob, guardian}.

1.4 Generalization algorithm

Now we have selected quasi-identifiers from all attributes in our data set. Since the combination of those quasi-identifiers provide too much information, we need to find a way to make value of them delivery less information. Here we introduced the generalization algorithm. [1]

1.4.1 Algorithm

1. Value generalization: remove one or more digits from the original value or a sequence of number. For example of generalizing a zip code,

$$95616 \longrightarrow 9561 \longrightarrow 956 \longrightarrow 95$$

2. Domain generalization: when values of an attribute is of category, then combine some values into one or introduce a higher level of description for these values. For examples,

$\{A, B, C, D\} \longrightarrow \{A, other\}$
$\{father, mother, grandfather, grandmother\} \longrightarrow \{parents, grandparents\}$
$\{LogAngeles, SanFrancisco, Minneapolis, ..., Miami\} \longrightarrow \{California, Minnesota, ..., Florida\}$

1.4.2 Experiment

We tried different ways of domain generalization for the quasi-identifiers, and found that the way of generalization which has the best performance is shown below.

1.5 Result and comparison

We calculated the number of distinct rows and their occurrence in the table for $T[QI]$ before and after the generalization algorithm. The details are in the figure 1.2 and figure 1.3.

In the table 1.2, it means that there are 129 distinct rows in the data set constructed with quasi-identifiers and each of them only occurs once. 37 distinct rows in the data set only occur twice in the data set and similar explanation for other fields. These values implies that there are 129 kinds of queries can be used to re-identify 129 individuals, more than half of the population of the data set, in a 100% probability. It means the data set is severely vulnerable from linkage attacks.

In the table 1.3, we can see that after applying generalization algorithm, every distinct row has more occurrences in the data set. Additionally, any distinct row has at least 5 occurrences in the data set. According to the definition of k-anonymity, it is 5-anonymity. It means that no matter what queries an adversary use, the result returned from the data base consisted of at least 5 rows. Then we can say the probability of any students in the data set being re-identified is at most $\frac{1}{5} = 20\%$, which is a relatively desirable result.

Number of distinct rows	occurrences	Number of distinct rows	occurrences
129	1	4	6
37	2	1	14
16	3	1	13
10	4	1	11
7	5	1	7

Table 1.1: Occurrences of distinct rows before applying generalization algorithm

Number of distinct rows	occurrences	Number of distinct rows	occurrences
2	12	1	41
2	6	1	29
1	38	1	15
1	21	1	34
1	11	1	9
1	5	1	19
1	124	1	13

Table 1.2: Occurrences of distinct rows after applying generalization algorithm

1.6 Remarks

K-anonymity with data generalization is a fast and handy mechanism that can be used to protect privacy of a data set from linkage attacks. However, it has disadvantages and limits.

First, we assume an adversary does not know about knowledge of attributes other than quasi-identifiers. If this precondition does not hold, k-anonymity does not guarantee privacy protection from linkage attack. Second, generalization algorithm can reduce the usability of the data set since it makes information less specific. For our data set, we eliminate two attributes, school and gender and make other quasi-identifier deliver less information. It makes it impossible for analyzers to analyze the relationship between a student's gender or school and his or her performance in class.

In conclusion, k-anonymity is useful to protect privacy of a data set against linkage attacks but we have to make trade-off between usability and privacy protection of the data set.

References

[1] L. Sweeney. *Achieving k-anonymity privacy protection using generalization and suppression.* International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 571-588.

[2] Omer, A.M. and Mohamad, M.M.B.. (2016). *Simple and effective method for selecting quasi-identifier.* 89. 512-517.

[3] BENJAMIN C. M. FUNG, KE WANG, RUI CHEN, PHILIP S. YU *Privacy-Preserving Data Publishing: A Survey of Recent Developments.*ACM Computing Surveys, Vol. 42, No. 4, Article 14, Publication date: June 2010.

[4] Tonny Shekha Kar, *A Study on Privacy Preserving Data Publishing with Differential Privacy*Ph.D thesis, University of Saskatchewan,2017.

[5] *https://archive.ics.uci.edu/ml/datasets/Student+Performance*