Data Protection & Privacy

Project: Mondrian Multidimensional K-Anonimity

Authors:
Giacomo Garbarino (S4545532)
Manuel Parmiggiani (S4701853)

Basic Idea

During the analytic process, real data is often better to use instead of auto-generated samples, since it produces more realistic results.

The issue arises in terms of privacy whenever individual and private data is needed for the analysis purposes.

The aim of the anonymization process is to alter all of the information that can be used for the direct identification of the subjects while keeping untouched the information related to the analysis' domain of interest.

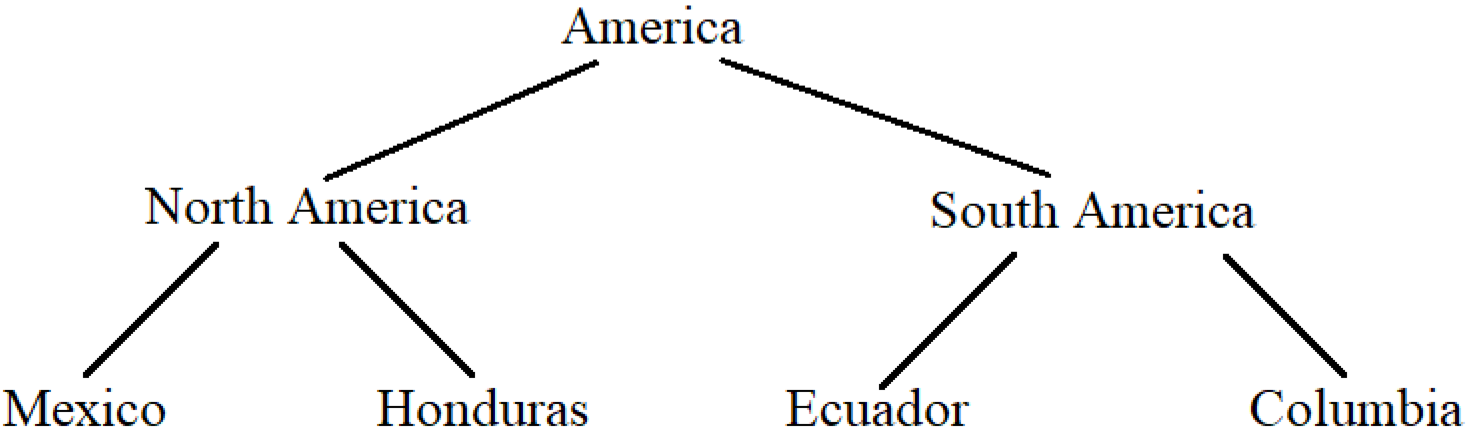
What we will do

The aim of the process is to identify and alter all of the columns that, merged with external informations, can lead to the direct identification of the owner of each record of the table, we will address those columns as "QI" (Quasi Identifiers).

How

For each QI column, we will replace some specific values with more generic ones, in order to make the dataset resistant to join attacks. We will rely on user-defined hierarchies inferred by each column's domain.

Hierarchy - Example in short



America

North America

South America

Mexico

Honduras

Ecuador

Columbia

The higher we go, the more generic the terms become

Does this solve the problem?

Generalization is a good way to make join attacks much less effective, but it doesn't solve the problem at all.

The solution that totally solves the issue consists in "hiding" the record associated to a certain user among other records that share the same values for the columns that can lead to the direct identification of the record's owner.

So, groups of records have to be processed simultaneously instead of single records.

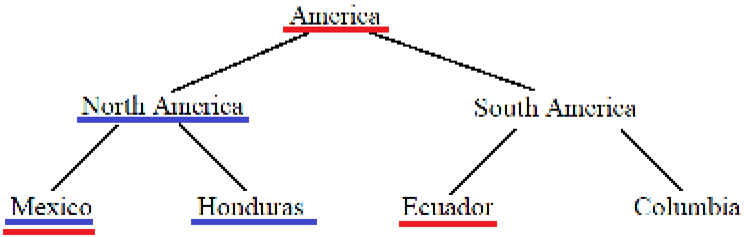
Final solution: K-Anonymization

A table is K-anonymous if every unique record occurs at least K times, so K is the minimum size of the equivalence class of each record.

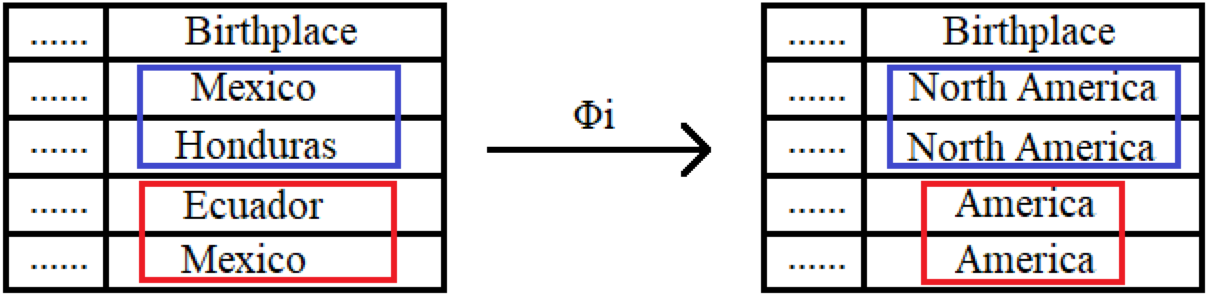
In our case, the final anonymized table has to be K-anonymous with respect to the QI columns only, since the rest of the data will have to be as much useful as possible for the analysis purpose, so it won't be altered.

K-Anonymization

Categorical Attributes



Taking into account the given hierarchy for a certain column, the anonymization process will look for a common "ancestor" for the values of the records of the group for that column.



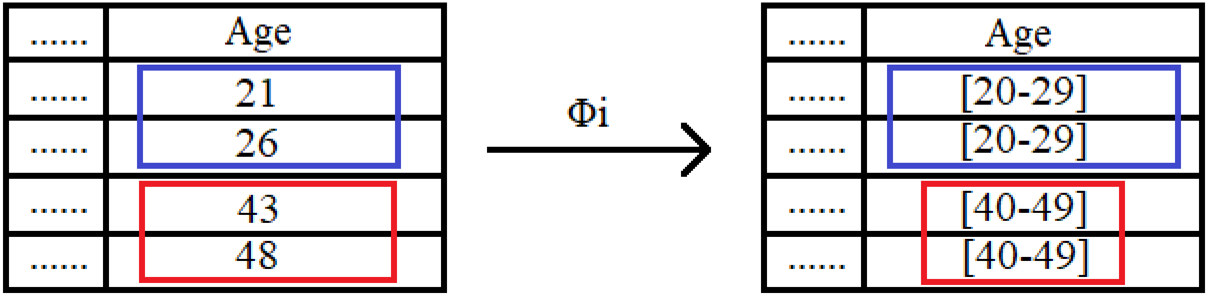
......	Birthplace
......	Mexico
......	Honduras
......	Ecuador
......	Mexico

......	Birthplace
......	North America
......	North America
......	America
......	America

Example with K=2

K-Anonymization

Numerical Attributes



......	Age
......	21
......	26
......	43
......	48

......	Age
......	[20-29]
......	[20-29]
......	[40-49]
......	[40-49]

K values are replaced with a range that may contain each of them

(Φi is the single-dimensional global recoding function that, applied to the values of one single column of the table, maps them into more generic ones)

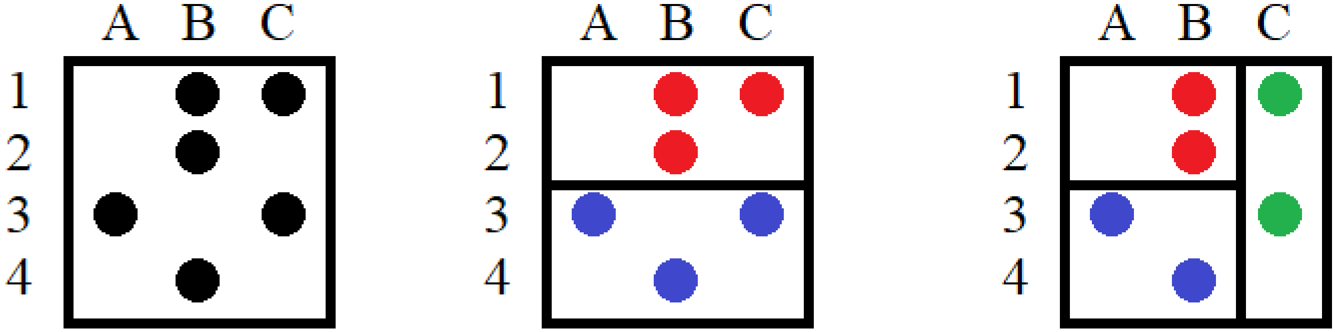
Example with K=2

Mondrian Multidimensional K-Anonymity

Grouping criterion

It is good to have groups consisting of records that have similar values for all the QI columns from the beginning since it simplifies the anonymization procedure.

In order to obtain this, the grouping is done following the multidimensional partitioning rule.



Single-Dimensional

Multidimensional

Mondrian Multidimensional K-Anonymity

Global recoding function

Since we are using the multidimensional rule for the grouping, we will also replace all the single-dimensional global recoding functions Φi with one multidimensional global recoding function Φ.

Applying Φ to all the QI columns of the dataset at the same time produces the same results of applying the Φi funtcions one to each column simultaneously.

2 important requirements

We must provide to the tool a file specifying the columns' types, the program cannot infer on its own what columns are sensitive data (SD) and what not, since this is established by us humans through a non-algorithmic reasonment.

ex : columns 'Age' and 'Annual Gain' are both numerical but the latter is a SD column while the former is not; without accessing external informations (in this case, the file we provide to the tool) the program cannot correct classify columns

We must provide to the tool a file specifying the hierarchy of the values for each categorical column, since those values often are related through non-algorithmic criteria.

ex : Canada is part of North America and North America is a fraction of the American Continent, we know this but the tool does not, we have to give to the tool this information (through a file)