AUTUMN TERM 2022

LAB Assignment 2

Data anonymization tools

INTROSEC, Group 03
Lekhaz Adapa, lead3201
Jesper Blomqvist, jebl6563
Puja Poudel, pupo3339

Stockholms universitet

Table of contents

1. Background and Design

1.1 Background

1.1.1 Data anonymization tools

Data anonymization is a part of computer security; it falls under confidentiality of the so-called security triad, which consists of integrity, availability and confidentiality (Pfleeger, 2015). The need to keep data private and to control with whom data may be shared predates the modern computer era, but it has become an increasingly important aspect of information governance since computers with increased processing power and storage capacity, in tandem with a multitude of data generating and collecting devices, have been introduced on all levels in our society, which has lead to massive information stacks that may contain personal, sensitive information that can single out individuals on their own or be aggregated to achieve that goal. (Pfleeger, 2015)

Various laws and policies have been created and deployed to regulate this, but there are fundamental differences on how this is legislated and implemented in different countries. An important example is how the USA and EU differ in legislation regarding privately owned businesses and organizations handling personal data. In the US the ownership of the information falls to the entity that handles the data, whereas in the EU the data subject is considered to have ownership of the data (Pfleeger, 2015).

According to the General Data Protection Regulation (GDPR) in the European Union (EU), each data subject has the right to obtain confirmation from the controller whether or not his or her personal data is being processed. Data subjects can contact the controller at any time if they wish to exercise this right of conformation (GDPR, 2023).

Therefore, it may become important to know which rulesets a tool is compliant with and if not, what measures must be taken to be considered compliant.

The ability for users to examine and contribute to the tool is also of interest. Such transparency can help a potential user to make an educated choice when choosing a tool to handle sensitive data. It can also be used to identify and evaluate vulnerabilities by an adversary, so it is a double-edged sword.

The tool's performance and the capability for securing the integrity of data are important characteristics to consider when choosing and using any information system, as well as the need to consider who has access to the tool and the information it may contain.

The idea of privacy through anonymization can be viewed as strengthening confidentiality at the cost of availability and integrity. Transforming data in an anonymization process may restrict the ability to extract useful information from a dataset (Corporate Finance Institute, n.d). This balance between confidentiality, integrity and availability is a recurring theme in the field of computer and information security (Pfleeger, 2015).

For this assignment, we have decided to focus primarily on aspects related to how anonymization of data can be achieved rather than how to control access or to maintain the integrity of the data, but they are intertwined and interdependent of each other and at times it is not possible to disregard that interconnectedness.

1.1.3 Anonymization methods and techniques

1.1.3.1 K-anonymity

Dimakopoulos, Tsitsigkos and Nikolaos (n.d.) describes the K-anonymity concept as used in the field of data anonymization that prevents an individual from being re-identified after being released.The idea of K-anonymity is based on the idea that in a group of individuals with similar characters should not be able to stand out from the group. Common techniques for achieving k-anonymity are suppression and generalization techniques, where suppression aims at obfuscating the data values by removal or substitution of the data and generalization aims to abstract the data into general hierarchies, like age into age intervals, days into weeks or months, or geographical positioning into larger areas, in order to make identification of single individuals possible. (Dimakopoulos et al., n.d).

1.1.3.2 Km-anonymity

Km-anonymity is a weaker form of anonymity than of k-anonymity and works better with high dimensional data sets. In addition to taking into account a number n of quasi identifiers, the algorithm now limits the guarantee against adversaries who know only m of the n pseudoidentifiers. Some anonymization algorithms make sure that each combination of m quasi identifiers appears k times in the anonymized datasets, regardless of how many quasi identifiers there are. For datasets with set-valued attributes in Amnesia, m<<n. As with k anonymity suppression and generalization are key techniques behind this anonymization method. However, it differs from regular k-anonymity in the number of dimensions, or quasi identifiers, in the dataset and in the assumption that not all of these are known to an assailant. If the known number is significantly lower than the actual number it is assumed to still achieve anonymity (Dimakopoulos et al., n.d).

1.2 The test protocol

With this in mind, we set out to design a test protocol based on the general guidelines in the assignment instructions and repurposed some to suit our choice of tools. The protocol was complemented with additional features considered relevant for the tools' intended purposes.

- Availability - Platforms, measures how widely available the tool is.
- Transparency Open source, measures how much insight a user can gain and how much a user can contribute to the ongoing effort of maintaining the tool.
- Country of origin/jurisdictional/legislative compliance? For what parts of the world is the tool intended? (EU/US/etc. Our primary goal is to see whether the tool is usable within the EU, i.e compliant with the GDPR, without any extra measures needed)

- Methods and levels of anonymization.
- Data formats and data sources. Which formats and sources can be handled without extra steps of (pre)processing the original data?
- How easy is it for a user to install and use the system securely?
    - Are there parts that would be difficult for a naive user to complete, or complete securely?
    - Are there any kinds of delay involved that might make users too impatient to use the tool?
    - Is it possible to demonstrate where poor handling of the tool may contribute to a breach of anonymity?
- Purpose - What does each tool protect against
- Limitations - What might the tools not protect against?
- Documentation - How helpful is the documentation?
- Reliability - Is it possible to uncover the anonymised data if the user does not use the tool properly?
- What common misconceptions might exist about the tool?
- Compared to the other tools you are looking at?
- A short discussion on the cost/benefit ratio of the tool's benefits versus the potential loss of availability involved in installing, configuring and using the tool

1.3 Selected tools

We decided to investigate installations of three tools: ARX, Amnesia and G9 anonymizer. They were chosen based on top google search results for google on data anonymization tools and availability. Since some tools are commercially available products they were not readily available, even as trial or demo versions. We finally decided on the following three tools, of whom the first two were completely free to download and test and the third came with a 7 day trial period which did not hinder us in this assignment.

- **ARX** is an open source project developed by Fabian Passer and colleagues who are supported by the Technical University of Munich.
- **Amnesia** is based on ARX but developed separately by OpenAIRE, a non-profit founded in 2018.
- **G9 Anonymizer** is developed by Esito, a Norwegian company.

2. Experiment

We installed and evaluated each tool with a dataset according to the test protocol and then we compared the test protocols. Below we describe the actual processes of installing, configuring and performing the anonymization, followed by the test protocol and discussion.

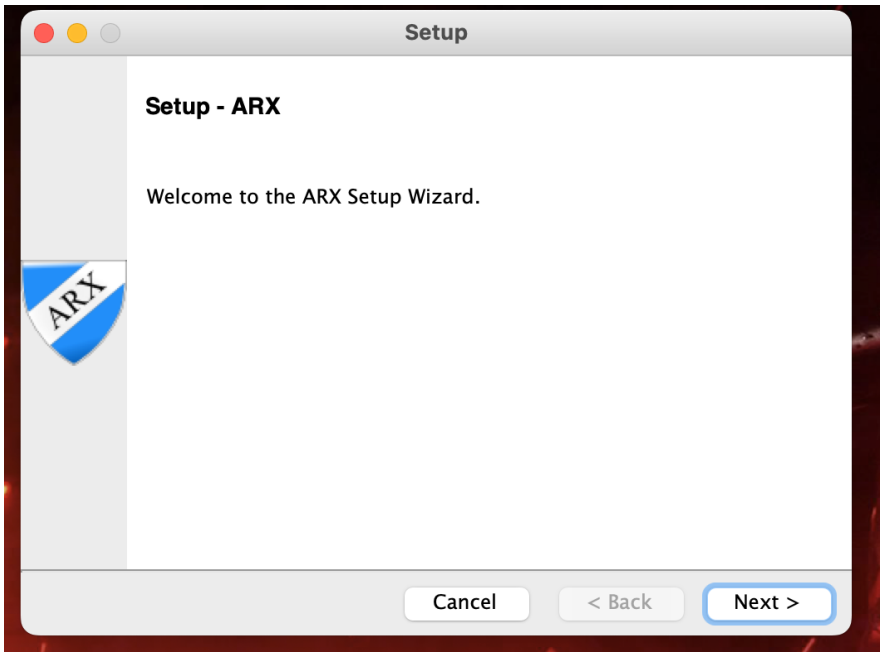
2.1 ARX Anonymization Tool

2.1.1 Test procedure

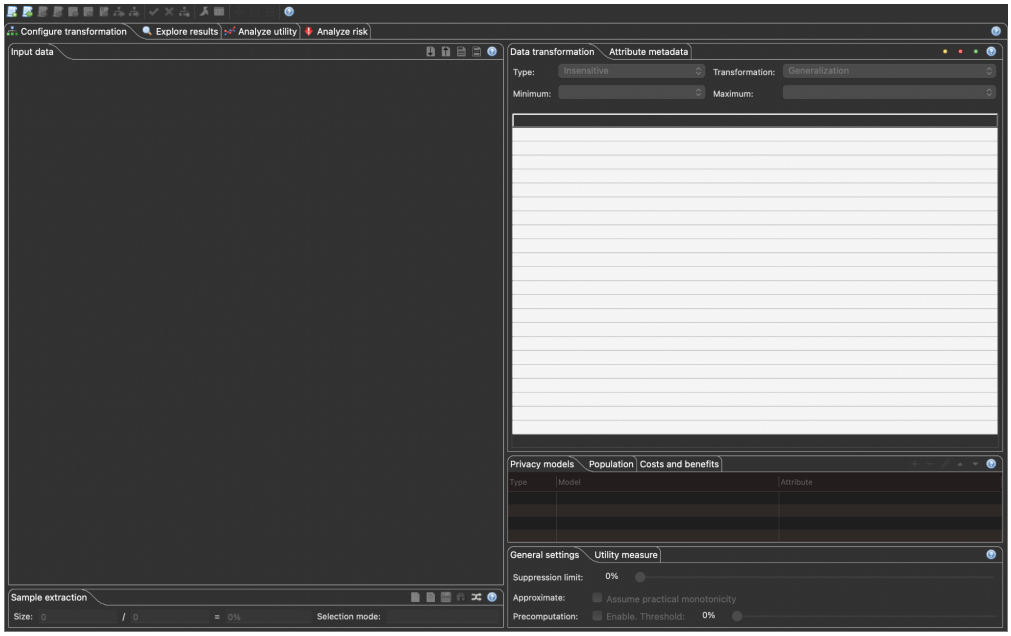
The following steps were taken to install and test ARX.

2.1.1.1 Setup/Installation process:

- After downloading the ARX Tool, we went through the setup process and this is how it looked like:

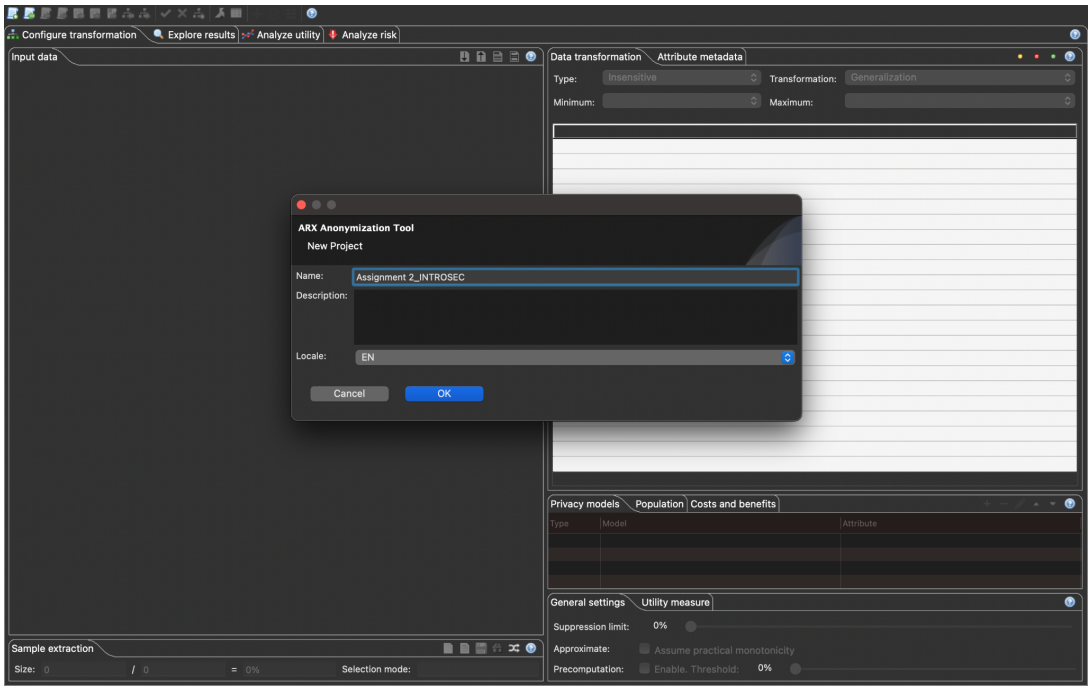


- As you can see the installation process is pretty straight forward and it's not much difficult for anyone.
- After the installation the ARX tool looks like this:



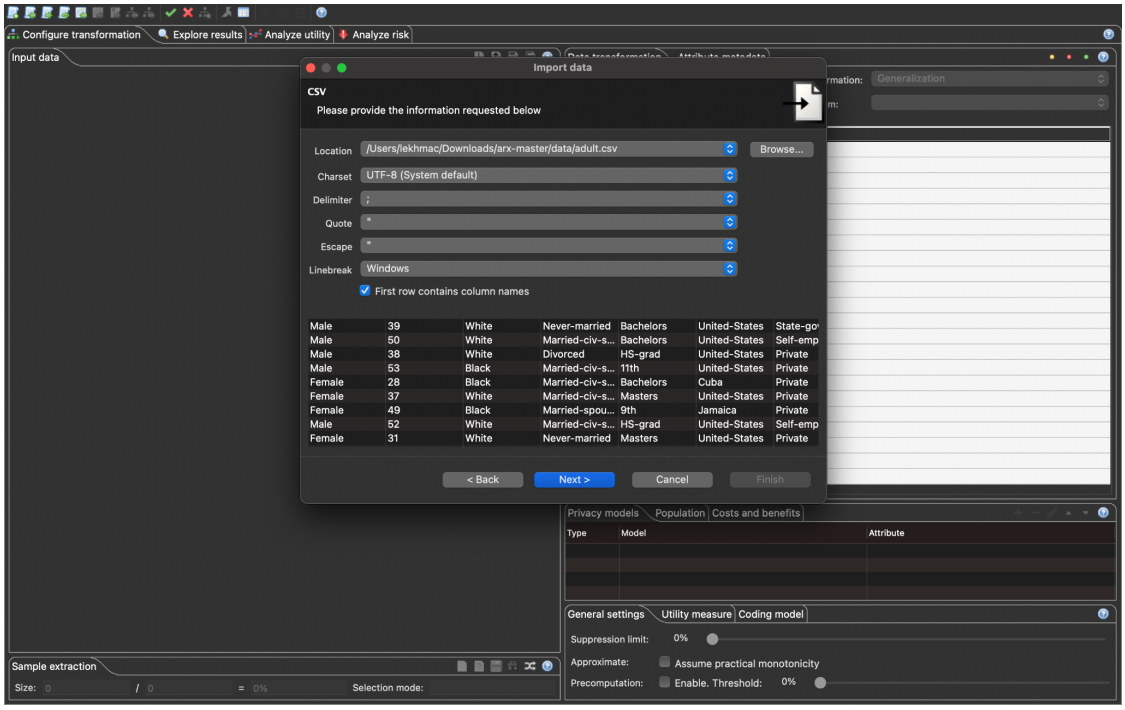
2.1.1.2 Performing data anonymization:

- ARX is a project based so we will first create a New Project.
  The screen looks like this:

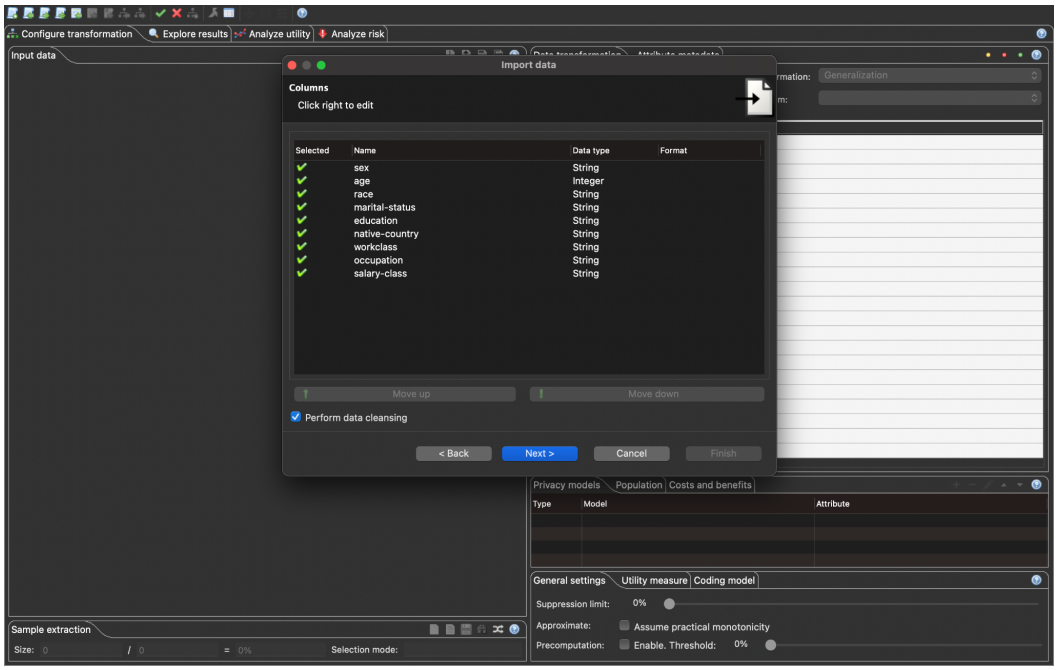


- And now we need to load a dataset. What we have noticed is that we can import only CSV, Excel(XLS, XLSX), and Database(JDBC).

- After loading the dataset we are able to see an interface like this:

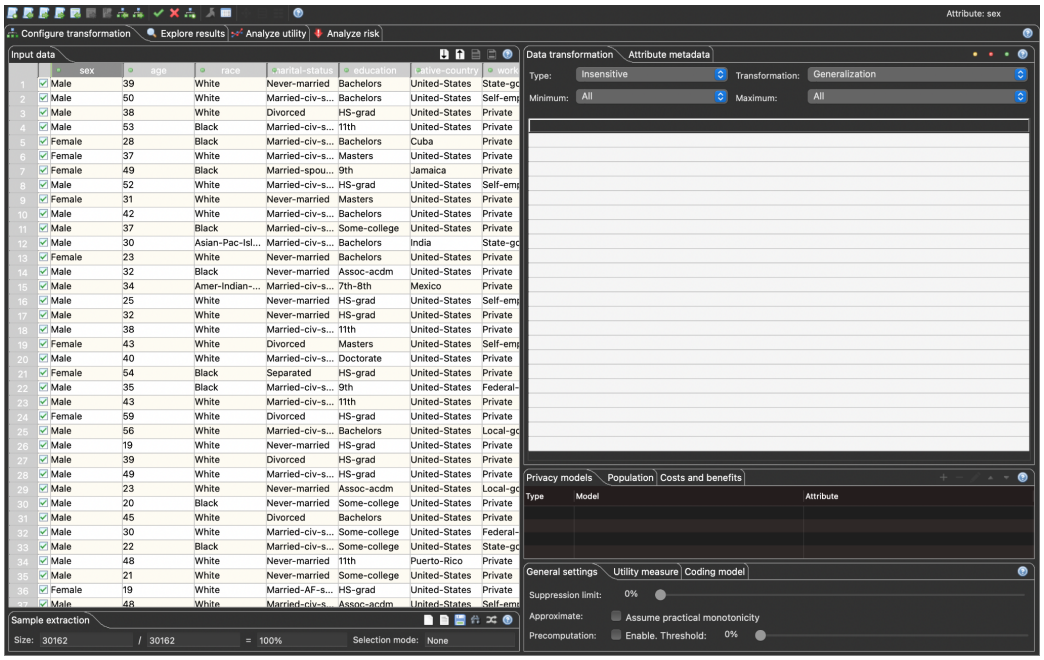


- And from the picture below we can exclude columns from the data and specify data types, and we can perform this later in the anonymization process.

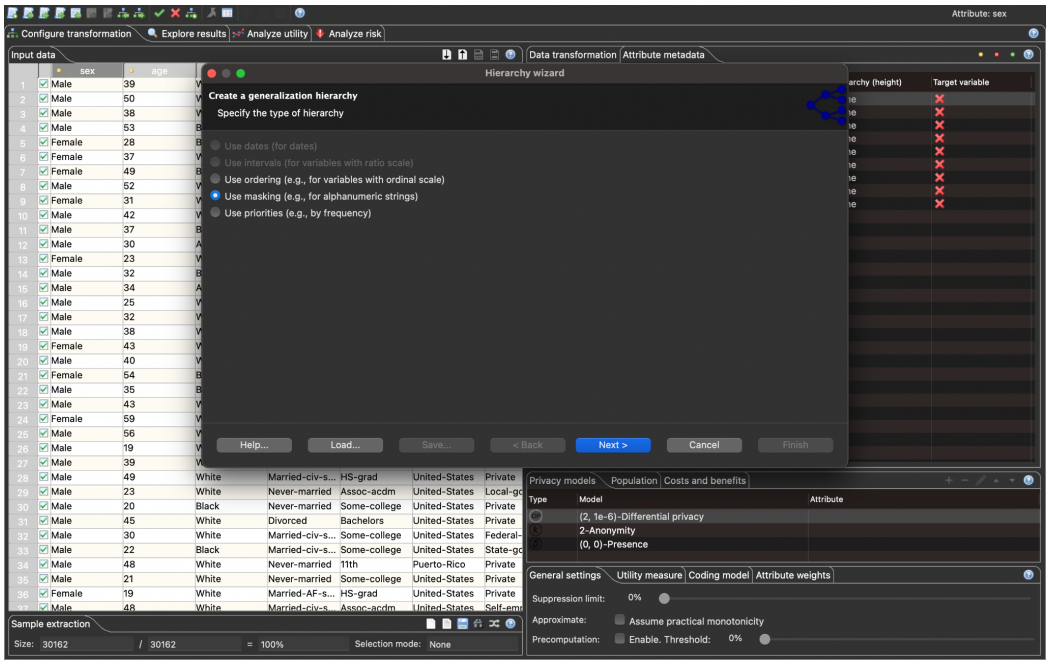


- Now we can see the imported data on the left hand side of the picture and at the right hand side we can see the attributes that are contained in the dataset. Here we can specify attribute type i.e. the attribute is a Direct Identifier or a Quasi Identifier and the data type i.e. string. We can create

and generalize the hierarchies. At the bottom we can specify privacy criteria and configure the anonymization process.

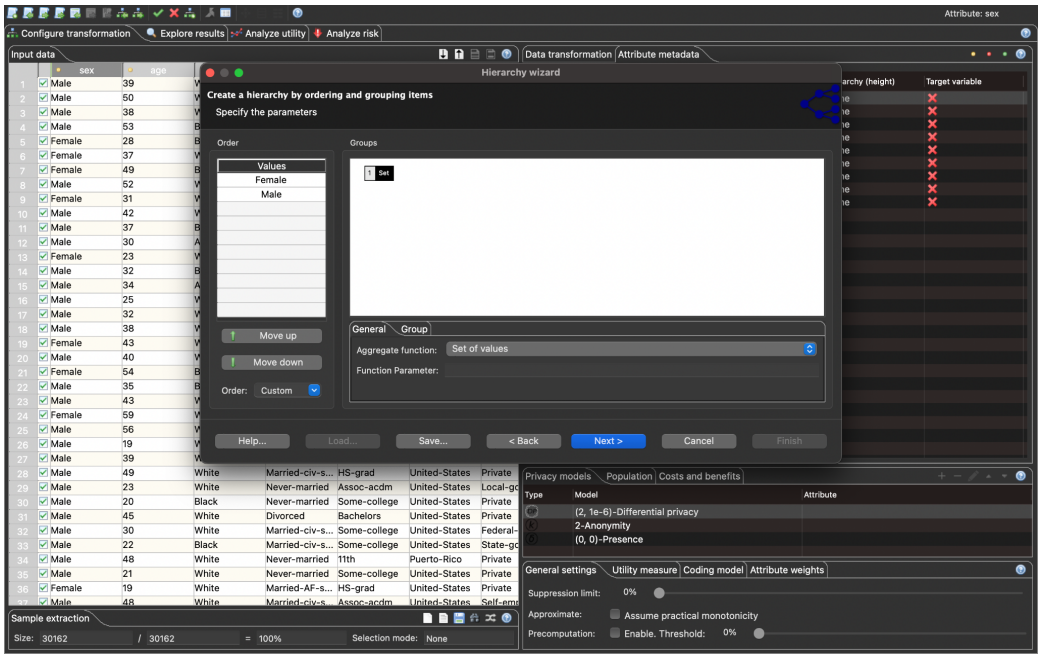


- Now we specify metadata about attributes. In our data everything we consider all the attributes are quasi identifiers. Now we start with the attribute “sex” and we perform a generalization hierarchy.

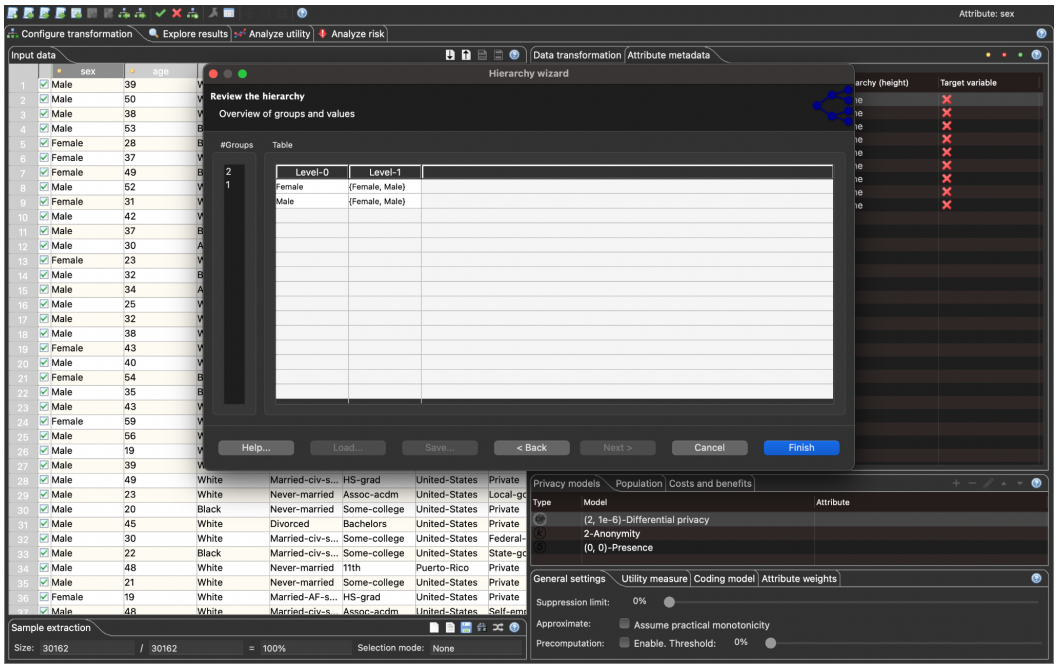


- As sex is a categorical attribute we use ordering. As shown in the above picture.
- On the left hand side we can create an ordered list of the attributes. On the right hand side we can combine consecutive values into groups. And as an example we will combine male and female

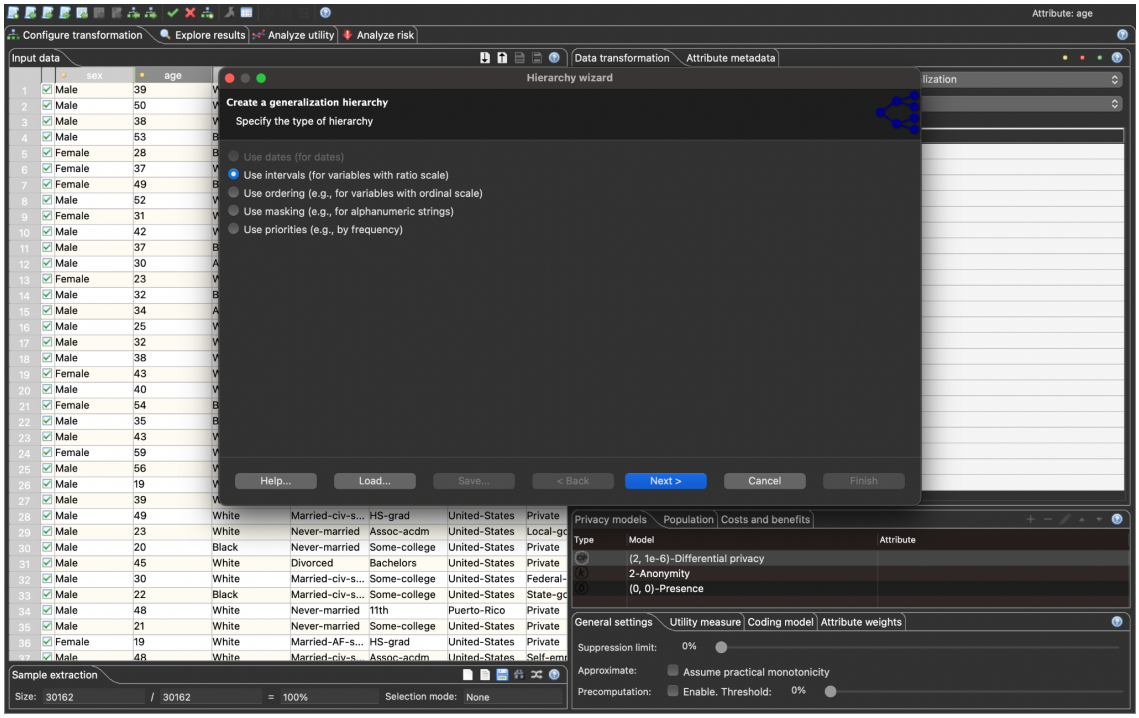
into one single group. To do that change that first click on the set and then change the value on size from 1 to 2 and click next.



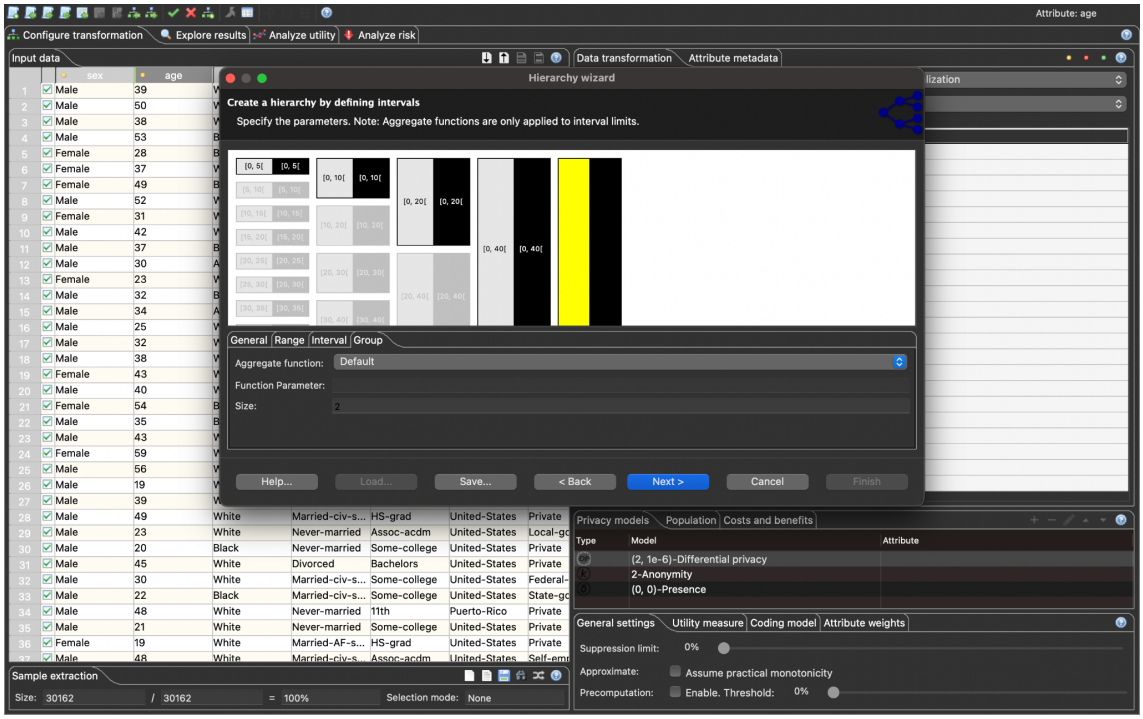
- Now you will see a window like this showing the groups and tables which are the combination of fame and male into one single group.



- Now we will do the same with the attribute “age”. And click generalization and we will choose intervals cause age is a numerical value.



Then we defined the range (we are top coding for the values that are >80). Following that we will choose the length of the basic interval and we have chosen 0-5. Finally we create a higher level of the hierarchy by merging intervals from lower level to higher levels.



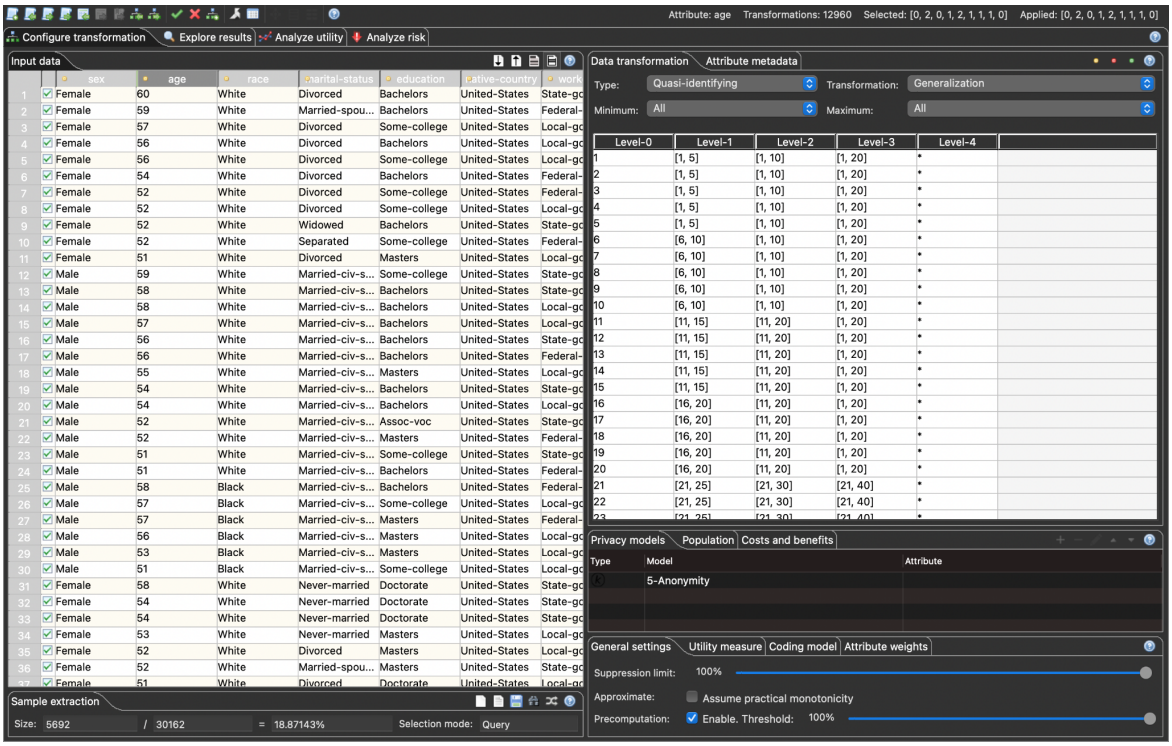
- And finally we can see the overview of the groups and values.

- So keep doing this operation for all the attributes, and as a result I will show you the final output of how they look, in the figure below. The outputs will consist of the remaining columns(race, marital-status, education, native-country, workclass, occupation, salary class) respectively.

2.1.1.4 Output

- Here's how the tool has anonymized the data. Below is the link to how we worked on the data.

https://youtu.be/azp5LNjTe20



- This is the output of anonymized data that we got after anonymization.

2.2 ARX Test protocol

Availability: Platforms Windows 64-bit, MacOS 64-bit, Linux/GTK 64-bit. Means the tool is available on all kinds of platforms which gives every user the ability to use the tool. We installed ARX on Windows 10/11 64 bit and worked on MacOS 64-bit.

Transparency and Documentation The source code is available on github. Documentation is available online (ARX, 2023).

Methods and levels of anonymization
ARX offers k anonymization through suppression and generalization techniques

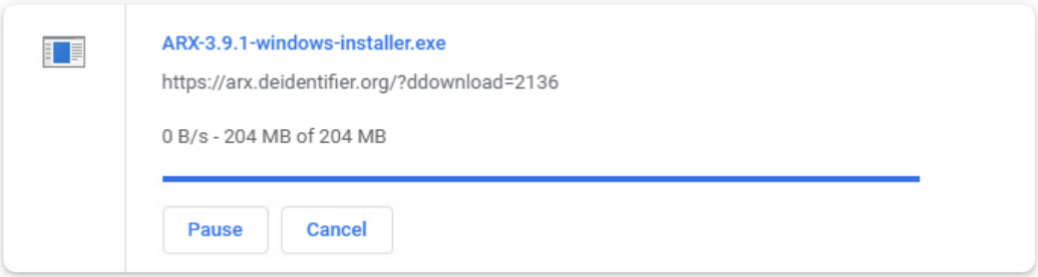
Datasets and supported data sources
ARX support a number of data formats and sources like CSV, XLS/XLSX and JDBC

Jurisdiction ARX is available in and compliant with EU regulations.

How easy is it for a user to install and use the system securely?

Are there parts that would be difficult for a user to complete, or complete securely?

- We encountered some issues downloading the file from the official ARX Website https://arx.deidentifier.org/downloads/ while installing this tool in the Windows Virtual Machine. Downloading took a little longer than expected. So we had to shut down our Virtual Machine in order to process the installation.



- In terms of installation, we haven't encountered any delays or issues, it's pretty straightforward. And there are no such parts that would be difficult for a naive user to complete or even complete securely.

Are there any kinds of delay involved that might make users too impatient to use the tool?

- We haven't found any kinds of delay regarding the data anonymization process but what we have observed is that in our example dataset that we have performed the Data Anonymization with required parameters i.e. the Privacy Model, Search Strategy and Transformation Model. In the Privacy Model we have many models and we have chosen k-anonymity as this is a well known privacy model aimed at protecting the datasets from re-identification in the prosecutor model.

- In search strategy we can select some kind of strategies that have been offered by the tool. We have five search strategies in this tool:
    - Optimal
    - Best-effort, binary
    - Best-effort, bottom up
    - Best-effort, top down

    - Best-effort, genetic.

- Next we have the Transformation model, we have two types of Transformation Models:
    - Global Transformation
    - Local Transformation
- In Global Transformation, this will result in full domain generalization, where each value of an attribute's domain is transformed to the same generalization level.
- While coming to Local Transformation, different generalization levels may be used for the same attribute value in different records. And we can also handle the number of iterations that we need to perform. As the number of iterations increases the time will be increased.
- Each strategy takes its own time to classify and analyze the data. Firstly, when we performed Data Anonymization choosing k-anonymity model with Optimal search and Global Transformation, we found that the solution space consists of 1792 transformations which have been classified and analyzed in approximately 0.079 seconds. This remains the same how many times we do also. Next we performed Optimal search and Local Transformation with 100 iterations, then we noticed a change in the number of transformations and time taken compared to global from 1792 to 50176 and 0,079 seconds to 1.407 seconds respectively.
- And this keeps changing when we change the value for the number of iterations. So as the value increases the time taken by the tool to anonymize data also changes.
- To justify the question: *Are there any kinds of delay involved that might make users too impatient to use the tool?* We have given the number of iterations as 1000000 and checked if this really affects the time, and we found that the time taken by the tool is just 9.518 seconds, which is really a decent time. So we conclude that we haven't found any delay involved that might make users too impatient to use tool.

Is it possible to demonstrate where poor handling of the tool may contribute to a breach of anonymity?

- We haven't found any poor handling of ARX tool that contribute to breach of anonymity while dealing with the smaller dataset, but we found that it goes under some breach of anonymity when we are working on large datasets. But these methods cannot be implemented in the area of Health Data Privacy.
- The dataset we worked on contains 30,000+ user details(which is usually not large). Here how it represents the risk analysis of the data before and after the anonymization.

- The left-hand side is the input dataset and to the right-hand side is the output dataset. These can be described as follow:
    - Records at risk: Records with risks above threshold.
    - Highest risk: Highest risk of a single record.
    - Success rate: Average records that can be re-identified.
- And we also thought if ARX can implement some future development is to improve its capabilities of processing high-dimensional data in two dimensions.

What might the tool not protect against? How helpful is the documentation ?

- Data subjects might be harmed if they are disclosed due to the possibility that they will be of interest to an attacker and this tool is a bit sensitive to the large datasets and Health Organizations. Obligations are also on the user to understand the nature of anonymization and reidentification processes.

What does the tool protect against

- Usually, this tool prevents the linking of individuals with sensitive attributes from a dataset, which are those attributes they do not want to be linked to.

What common misconceptions might exist about the tool?

- Misconception: ARX only works on text data.
  Reality: ARX can work on a wide variety of data types, including text, dates, numbers, and images.

- Misconception: ARX removes all identifying information from the data.
  Reality: ARX can remove and mask certain types of identifying information, but it may not be able to remove all identifying information from the data.

- Misconception: ARX is only useful for de-identifying data that is already in electronic form.
  Reality: ARX can de-identify the data that is already in the electronic form, but it is also useful to de-identify the paper documents by scanning them and then applying the de-identification process to the digital version.

- Misconceptions: ARX is foolproof and guarantees complete anonymity.
  Reality: ARX can be used to reduce the risk of re-identification, but it is not foolproof, there is always a risk that the data could potentially be re-identified.

Is it possible to uncover the anonymised data if the user does not use the tool properly?

- Even the data that is anonymized to 100%, they can be identified and hence if a user hasn't used the tool properly, yes, there is risk of uncovering the anonymized data. It is possible for anonymized data to be re-identified or de-anonymized if the anonymization process is not carried out correctly or if the anonymized data is linked to other datasets that contain additional identifying information.

A short discussion on the cost/benefit ratio of the tool's benefits versus the potential loss of availability involved in installing, configuring and using the tool.

Benefits of the ARX tool:

- Improved privacy and security: ARX helps to de-identify the sensitive data, reducing the risk of data breaches and unauthorized access to sensitive information.

- Compliance with regulations: ARX can help organizations meet their obligations under various privacy loss and regulations, such as the General Data Protection Regulation(GDPR) and the Health Insurance Portability and Accountability Act(HIPAA).

- Improved data quality: ARX can help to improve the quality of the data by removing or masking the irrelevant or sensitive information, making it easier to analyze and use.

Potential Costs of the ARX tool:

- Time and Resources: It may take time and resources to install, configure and use the tool, including training staff on how to use it effectively.

- Loss of availability: Using the ARX tool may require data to be taken offline or otherwise made unavailable for certain periods of time which could impact the organization's ability to use the data.

- Loss of data: The ARX can remove or mask certain types of data, which could potentially impact the organization's ability to use the data for certain purposes. It should be noted that the original dataset may be kept preserved

Overall, the cost/benefit ratio of using the ARX tool will depend on specific needs and goals of the organization, as well as potential privacy risks and costs. Investing in the ARX tool and other privacy-protection measures may be worthwhile for organizations that handle sensitive personal information and are at high risk of privacy breaches, while organizations that handle less sensitive personal information and have lower risks may be able to rely on other measures.

2.2 Amnesia

Amnesia was the second anonymization tool investigated.
Amnesia is developed and maintained within OpenAire, a non-profit partnership established in 2018 that strives to ensure open scholarly communication infrastructure to support European research (OpenAIRE, 2023)
Amnesia is based on ARX and like ARX, it is an open source project. It is available both as an online service and as a downloadable application to be used locally and it comes with an API for RESTful services used to expose various internal functions (Amnesia, 2022).
The developers strongly recommend that the online version should only be used for training and tutorial purposes, since it is not scaled for performance and the risk of breaking confidentiality and risking loss of integrity inherent in transmitting a data set to the web implementation.

2.2.1 Test procedure

The following steps were taken to install the Amnesia application and run evaluation according to our test protocol.

2.2.1.1 Installation process

Amnesia was installed on Windows 10/11 64 bit. The majority of tests were conducted on a local Windows 11 machine since the virtual machine provided for the assignment was inaccessible during the testing and it also had performance and storage issues.

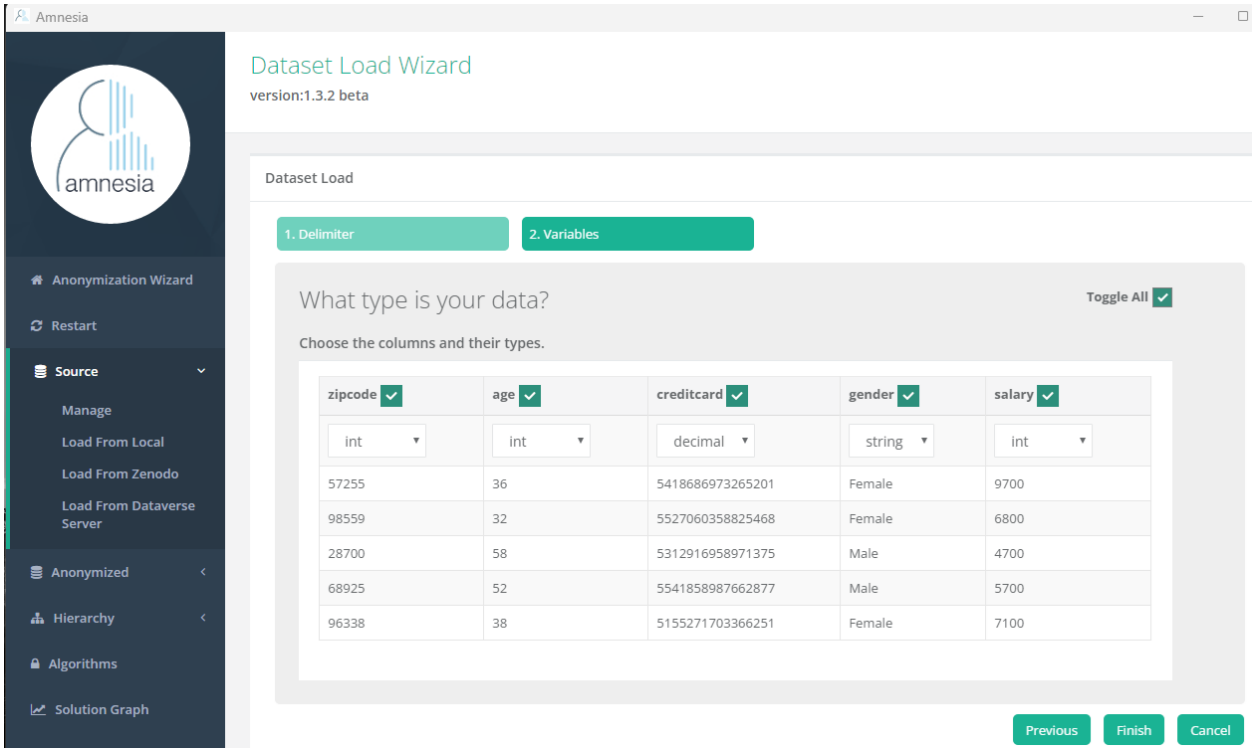
The installation was very straightforward utilizing a distributable msi package downloaded from https://amnesia.openaire.eu/download.html
Once downloaded, the installation process was completed by following the on screen Setup Wizard:

The only thing we could change was the installation folder.
Once installed, the user can proceed to load the dataset to be anonymized. Once loaded, the user needs to select the fields that are going to be processed and to provide additional information about the data contained in them, e.g. data types and relevance for identification.
In order to achieve k-anonymity, Amnesia utilizes suppression and generalization, that is removal and abstraction of data respectively.

2.2.1.2 Performing data anonymization

The data anonymization process begins with selecting a data set to process. We chose one of the test sets made available by OpenAIRE for tutorials called newData.txt which consists of 999 rows of comma separated values. For a more thorough comparison an independent, commonly used data set would have been preferable. In the end we decided there was not enough time allocated to this assignment to ensure that we would find and be able to use such a data set.

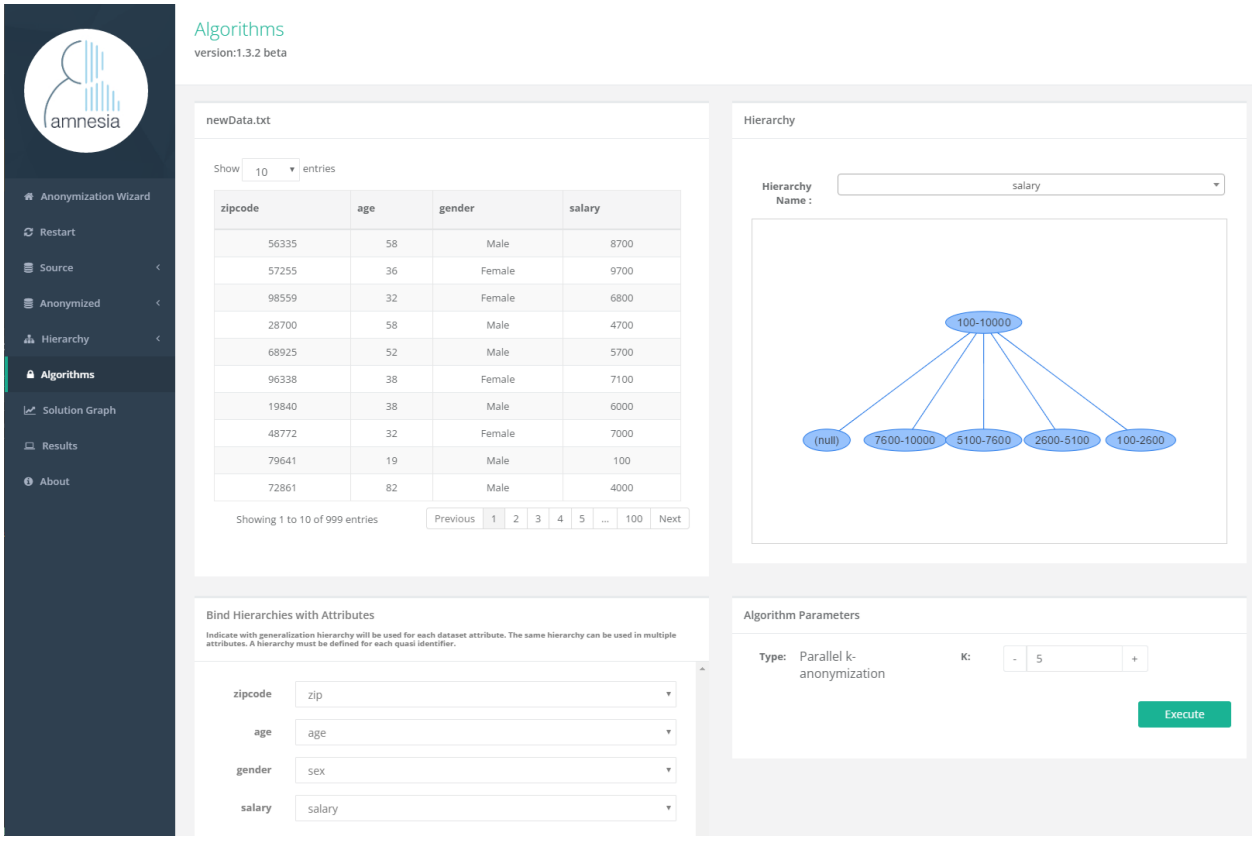
Just as with ARX, we need to perform some preprocessing of data types and to select which data we want to be processed. The tool aids us in this by suggesting data formats and likely candidates for processing. First we specify the delimiter used in the datasource and then we specify the data types for the variables:



The user is then presented with an overview of the dataset and can test for anonymity, which at this point is not meaningful, or proceed to hierarchies

In the hierarchy section the user can choose between autogenerating or loading previously defined hierarchies from a file. Some predefined hierarchies are included in the downloadable tutorial data set but we chose to autogenerate hierarchies based on the test data set.

Autogenerating a hierarchy consists of selecting the variable and choosing type. This depends on the variable's properties, e.g. if it is a string or integer or date.
Once hierarchies are created the user can proceed to algorithm selection where attributes in the data set are combined with the hierarchies created and the application suggests an anonymization method where the user can request the level of anonymity:

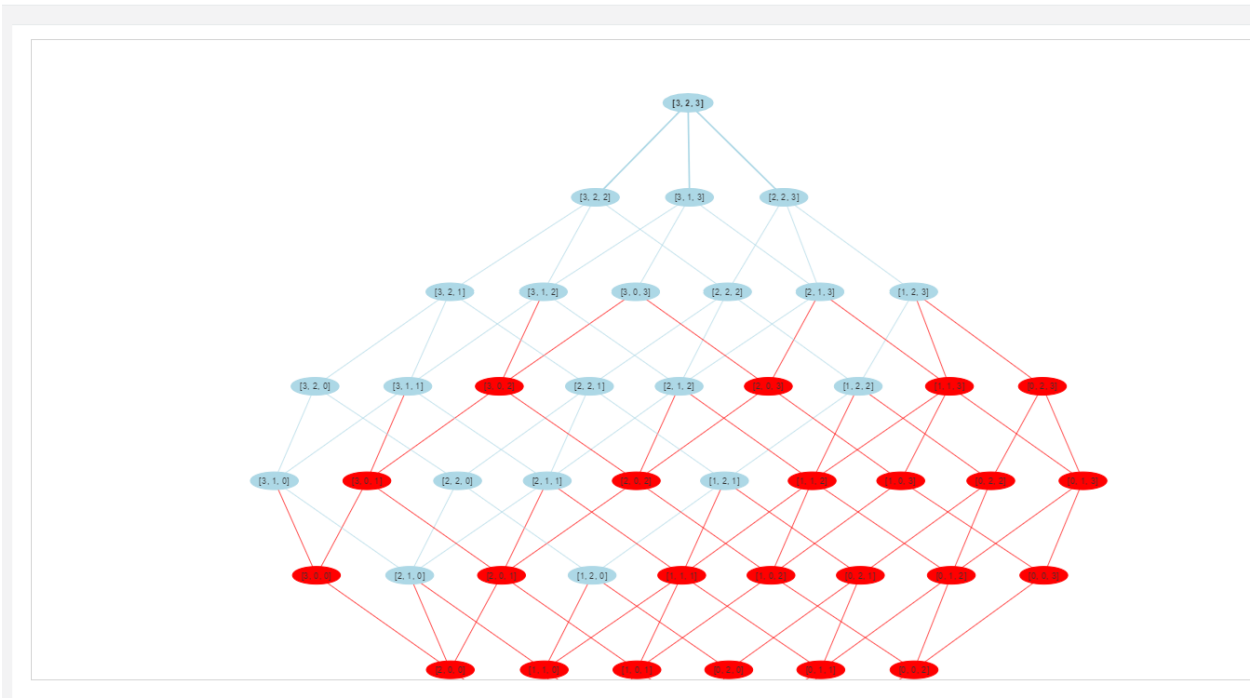


When all attributes are bound to hierarchies and the desired value of k is selected the algorithm can be executed on the data set by pressing the execute button,

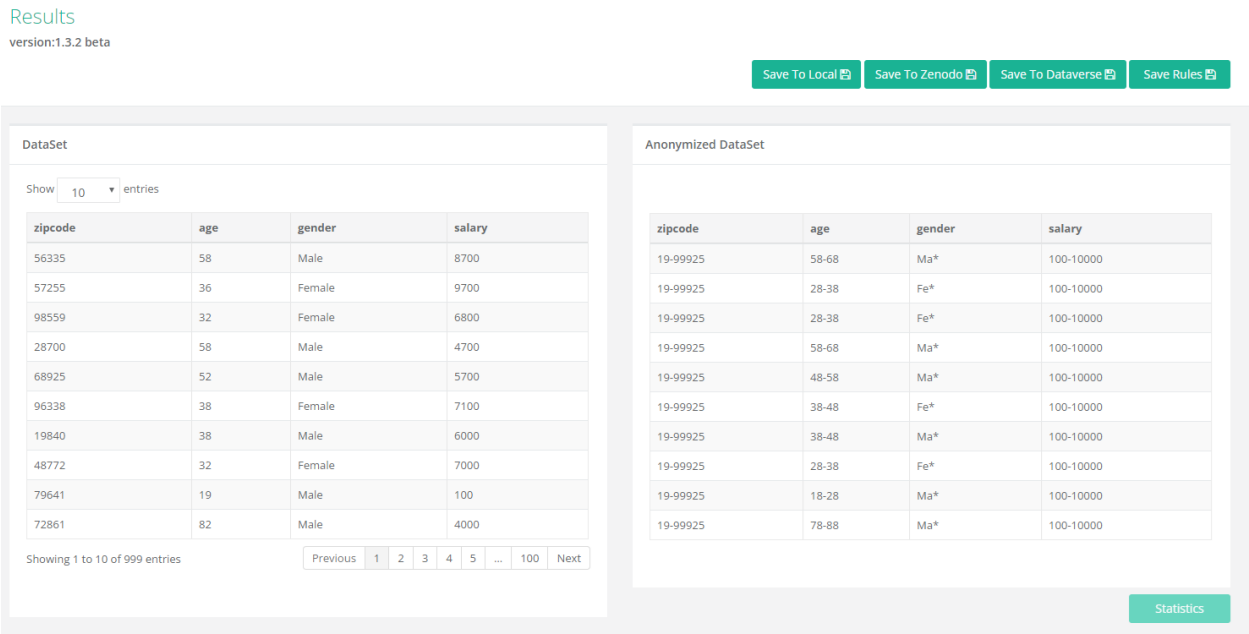
Just as with installation, much of the work is straightforward and the tutorial provides the user with information for each step but not so much about concepts such as generalization hierarchies and data types. It should be noted that the responsibility is on the user to understand the process, so if the user is illinformed, has misconceptions about critical concepts related to anonymization, there is a risk of not achieving the intended goal of confidentiality. We managed to anonymize data, but often at a higher level than intended. These tools are not meant to be used by a “naive” user with no prior understanding and time needs to be spent on familiarizing the concepts and going through the tool’s options. Luckily, there are good, basic tutorials and example files available.

2.2.1.3 Output

Finally, after tampering with the various parameters during the preprocessing, hierarchy design and algorithm execution steps it is time to view the results. But first we are presented with a solutions graph where blue nodes are safe solutions given the parameters :



Our understanding is that the higher up in the graph, the more generalized the solution will be, and as a consequence of that there will also be more information loss.
At this step the user can examine different solutions and gauge the generalization level via a preview of the resulting anonymized data set and its constituent attributes, as well as getting information about information loss for the solution. When that is done and a solution is selected the user enters the final stage where the original and anonymized versions of the data set are presented, together with the ability to save it locally, online to Zenodo or Dataverse. The user can also save the rules for reuse.



As can be seen in this picture, the attributes' original values have been altered in order to achieve k anonymity, k=2.

Since we were not familiar with the data, it hard to say something about the result, but from the looks of it, our hierarchies are probably a little too broad and because of that we may have unnecessary integrity /information loss. Especially for the zipcode and salary attributes where all values seem to have been generalized into a single value.

2.2.2 Amnesia test protocol

Availability. Platforms supported for local applications are Linux 32/64 bit and Windows 32/64 bit. The tutorials show Amnesia being used on MacOS as well, but we could not find confirmation of that. Amnesia can therefore be said to be widely available to users.

Anonymization methods. Just like ARX, Amnesia is also utilizing k anonymization and km anonymization to produce anonymized datasets.

Transparency and documentation.
OpenAire provides documentation online with both documentation and tutorials, as well as an introductory seminar and a github repository for the project.
The documentation is good and easy to follow, however not many external sources or published papers are referenced when compared to ARX.

Datasets supported.
We did not find a complete list but samples were provided as CSV files (text files with data separated by various delimiters, not only comma separated values).

Amnesia is also capable of importing data from https://zenodo.org/ , https://dataverse.org/ and locally stored DICOM (an international standard for managing medical imaging information) images. These methods of importing data were not explored, however.

Country of origin/jurisdictional applicability? (EU/US/ETC)
Amnesia is developed within the EU for EU research. We have not found any documented compliance with other legislative regions.

Methods and levels of anonymization
Amnesia is primarily providing anonymity through suppression and generalization. Either through k anonymity or km anonymity.

Purpose - What does the tool protect against
The tool is intended to provide k level anonymization, thereby granting confidentiality for a given data set.

Limitations - What might the tools not protect against?
It is not a tool for authentication control or integrity preservation. In fact, the process of transforming a data set through an anonymization process is irreversible just because the data integrity is compromised. However, since the original data set is left intact and with multiple data sources a data set can be reidentified by combining other quasi identifiers that were left intact when considering only one data set.

Are there parts that would be difficult for a naive user to complete, or complete securely?
The installation and anonymization process is simple and straightforward. There are tutorials for new users but we do not think that neither the tool nor the task is suited for a “naive” user at all.
We do think it is noteworthy to keep in mind that data anonymization should not be undertaken lightly. A user needs to understand the concepts of data anonymization, be familiar with the data set and also understand how the tool works in order to select the right method and achieve the desired level of anonymity/confidentiality.
These tools may be fairly easy to use, but the users need to be aware of how they work, what they protect against and what they do not protect against and under which conditions these guarantees actually apply.

Are there any kinds of delay involved that might make users too impatient to use the tool?
The larger the data set, the higher the dimensionality, and the higher the value of k, the longer it will take to find an anonymization solution.

Is it possible to demonstrate where poor handling of the tool may contribute to a breach of anonymity?
Yes. Not identifying critical attributes as anonymization targets or against the recommendation of not exposing a dataset containing sensitive information by uploading it to the online version of the tool are two examples of when a possible breach of confidentiality can occur. Both of these can be said to be dependent on an ignorant user rather than on any security vulnerabilities. An example of handling

continuously updated datasets was also given. It is not recommended to reanonymize a dataset that has been appended with new data.

Compared to the other tools you are looking at?
The possibility to upload data in an non-secure manner is exclusive to this tool, but the other kind of user handling that may lead to breach of confidentiality is not tool specific but rather inherent problems related to the general field of information security. If you are not aware of concepts such as pseudo anonymization and reidentifcation by using multiple data sources containing shared data subjects it may lead to a breach of confidentiality regardless of tool being used.

Misconceptions

Pseudo anonymization vs anonymization.
Pseudo-anonymization is not anonymization, i.e 100% irreversible. It is merely a method to reduce linkability between a dataset and the identities of its constituent data subjects.
Pseudo anonymized datasets may be reverse engineered to reidentify individual records. e.g. by cross checking different databases and linking them based on shared quasi identifiers.

Example: medical records and public information from electoral registers can be combined and even though they may have been pseudo-anonymized individually, some combination of quasi identifiers (e.g. gender, zip code and date of birth) may exist that can be used to reidentify the individuals Not understanding the difference may lead to unintended consequences, including a breach of confidentiality.

Risks - Is it possible to uncover the anonymised data if the user does not use the tool properly?
We did not find any official checksums and the installer was not certified. This is a trust issue, security wise.

Using the online version for anonymization is not recommended, since it involves sending the original dataset to the server. This may result in a breach of confidentiality and therefore it is strongly recommended to use the downloadable version where the original dataset is processed locally and never leaves the premises.
The server cannot handle large datasets since anonymization is a computationally intensive process. This is not as much a security risk as a performance problem.
If not properly handled or based on misconceptions about the various methods and techniques, there is a risk of not achieving the desired level of confidentiality. With multiple data sources covering the same data subjects, it may be possible to re-identify individuals even though measures have been taken to properly anonymize each individual data set. However an adversary may find ways to gain access to and combine these datasets in such a way that re-identification of individual data subjects is possible. In addition to techniques and methods, there is also a need for regulatory frameworks to prohibit this.

Benefits of Amnesia

- Improved privacy and security: Amnesia helps to deidentify sensitive data, reducing the impact of data breaches and unauthorized access to sensitive information.
- Compliance with regulations: Amnesia can help organizations meet their obligations under various privacy loss and regulations, such as the GDPR in the EU
- Improved data quality: Amnesia can help with improving the quality of the data by removing or masking the irrelevant or sensitive information, making it easier to analyze and use.
  This, however, is a double edged sword since it can also restrict the usefulness of the data if applied too strictly.

Costs of Amnesia

- Time and Resources: It may take time and resources to install, configure and use the tool, including training staff on how to use it effectively. It is relatively easy to install and maintain, but the cost of training staff, not only in the use and maintenance of the tool, but also in the field of information security and various legislative aspects that may affect the use as well.

- Loss of availability: Using the tool directly on the original dataset may require data to be taken offline or otherwise made unavailable for certain periods of time which could impact the organization's ability to use the data. This could be remedied by a replication strategy for the organization's data so anonymization is always performed on a copy of the original dataset or during off hours.

- Loss of data: As a part of the anonymization process, the tool can suppress or generalize certain types of data, which could potentially restrict the organization's ability to use the data.

Just as with the other tools investigated, the cost/benefit ratio of using an anonymization tool will depend on specific needs and goals of the organization, external conditions such as legislation and policies, as well as potential privacy risks and costs. Investing in an anonymization tool and other privacy-protection measures may be worthwhile for organizations that handle sensitive personal information, while organizations that handle less sensitive personal information and have lower risks may be able to rely on other measures. This cannot be gauged by examining the tool alone, but must be done within a business specific context.

2.3 G9 Anonymizer

G9 Anonymizer is another software tool that is used to anonymize data(esito, 2022). It helps users to anonymize data by providing various methods for replacing or obscuring sensitive information. These methods may include masking, perturbation, and aggregation, which can be used to protect different types of personally identifiable information. The tool may also include features for testing and validating the anonymization process to ensure that it has been effective in protecting sensitive information. G9 Anonymizer is available as a standalone software application that can be installed and runs on a desktop or laptop computer. It is also available as a web-based application that can be accessed through a web

browser. It is used both in Windows and Linux OS. The tool provides full programmable anonymization logic using the Eclipse application. For our task, we installed the tool on Windows 64-bit.

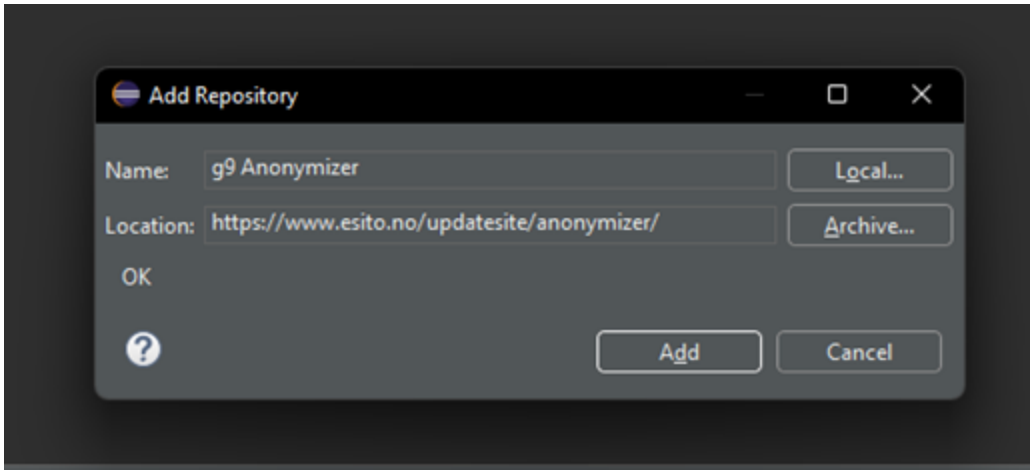
2.3.1 Test procedure

The following steps were taken to install and test G9 Anonymizer.

2.3.1.1 Setup and installation process

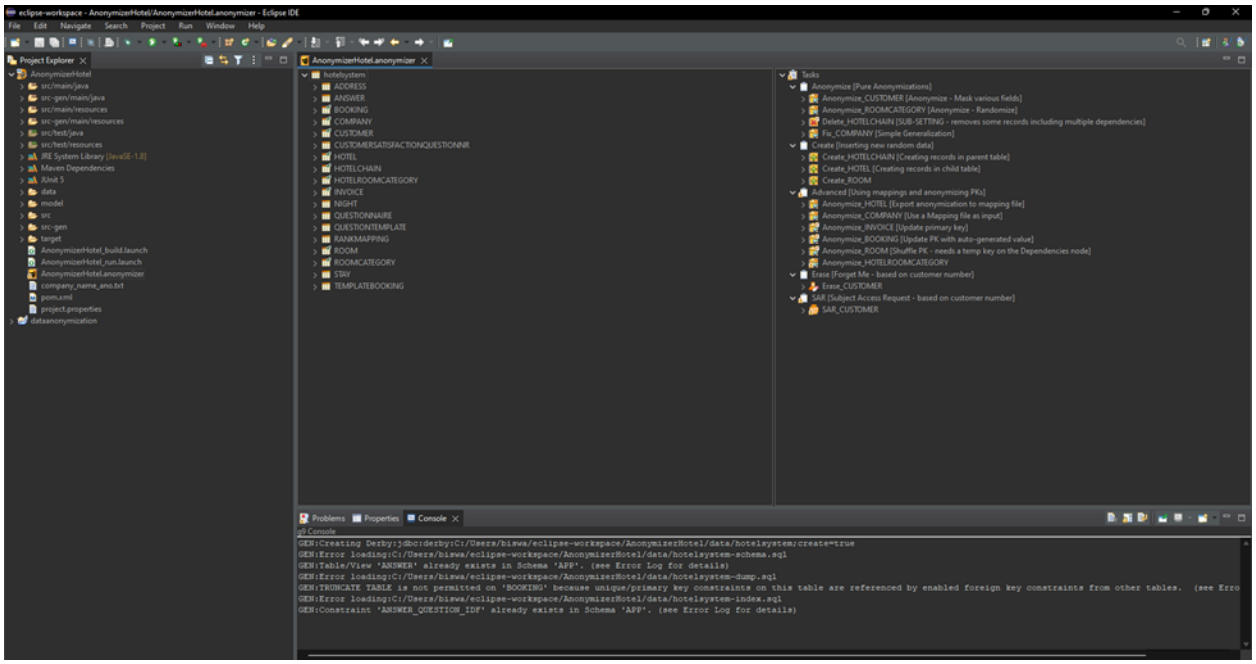
We downloaded the G9 anonymization tool from https://www.esito.no/en/products/anonymizer/. To use the tool, installing Eclipse plug-ins were required. Eclipse was installed from http://www.eclipse.org/downloads with the recommended IDE for Java Developers.

An additional plug-in that was required to use the G9 anonymizer was installed from Esito's update site by following the documentation from the website.



2.3.1.2 Performing data anonymization:

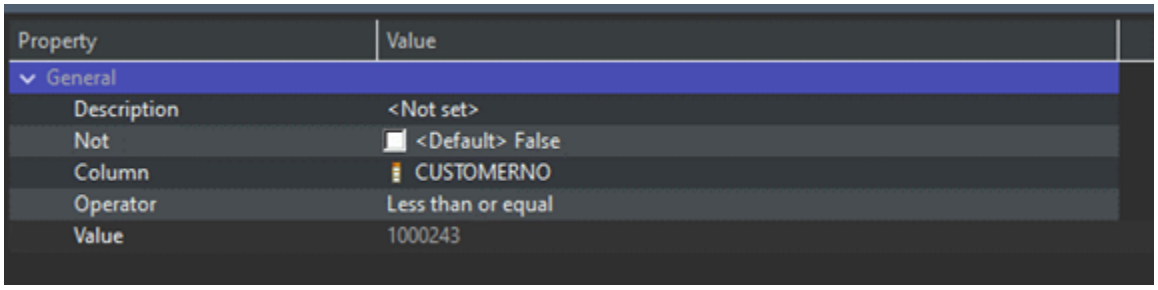
Following the installation of the necessary software and plugins, a new project was established to anonymize the data. To facilitate the subsequent steps in the project, the perspective and views of the working environment were adjusted and a shortcut was created (mosty, 2023). The project was then imported.



Before anonymizing the database, we will examine the data in the database table. The data in the database appears as shown in the picture below.

CREDITCARD	CUSTOMERNO	EMAIL	NAME	PASSWORD	PHONE	CUSTODIAN	LOCK_FLAG
<null>	1000234	robert.gibson...	Robert Gibson	<null>	+47 11 12 13 14	<null>	<null>
<null>	1000235	toby.baxter@i...	Toby Baxter	<null>	+47 11 12 13 14	<null>	<null>
<null>	1000236	harley.doyle@...	Harley Doyle	<null>	98765432	<null>	<null>
<null>	1000237	jordan.anders...	Jordan Anders...	<null>	+47 11 12 13 14	<null>	<null>
<null>	1000238	esito@crayon....	Esito	<null>	+47 11 12 13 14	<null>	<null>
<null>	1000239	louie.andrews...	Louie Andrews	<null>	+47 11 12 13 14	<null>	<null>
<null>	1000240	brecken.merc...	Brecken Mercer	<null>	+47 11 12 13 14	<null>	<null>
<null>	1000241	jens.barth@esi...	Jens Barth	<null>	+47 11 12 13 14	1000234	<null>
<null>	1000242	hudson.gould...	Hudson Gould	<null>	+47 11 12 13 14	<null>	<null>
<null>	1000243	<null>	<null>	<null>	<null>	<null>	<null>

The entire task has been explored and masking rules and conditions have been added to existing tasks. The properties of the data have been filtered with the WHERE condition and given properties.



Next, we began constructing the project by executing the task. Every time changes are saved to the anonymizer file, the generator starts and creates a comprehensive program that reflects all of the defined rules. The Console View displays the output of the generation process.

```
quit                - quits program
>trace
>run anonymize
>>> START 12:31:50(Task commit)
Reading internal config.properties
Starting Derby
Running example.anonymizer.anonymize.Anonymize
>>>   Anonymize
>>>     Anonymize_CUSTOMER
SELECT CUSTOMERNO, PHONE, NAME, EMAIL, CREDITCARD, PASSWORD, LOCK_FLAG, CUSTODIAN FROM
UPDATE APP.CUSTOMER SET PHONE = ?, NAME = ?, EMAIL = ?, CREDITCARD = ? WHERE CUSTOMERN
['+47 51289513','Evie Ellis','evie.ellis@gmail.com','4142834051119496',1000234]
['+47 87236119','Evelyn Johnston','evelyn.johnston@inmeta.no','4142834087046101',10002
['+47 77004074','Hannah Lewis','hannah.lewis@yahoo.com','4142834076954067',1000236]
['+47 53391573','Landon Hodges','landon.hodges@crayon.com','4142834053251552',1000237]
['+47 30209650','Dangelo Vega','dangelo.vega@crayon.com','4142834030059631',1000238]
['+47 44319116','Martha Hopkins','martha.hopkins@crayon.com','4142834044249103',100023
['+47 23238608','Sarah Harvey','sarah.harvey@crayon.com','4142834023108593',1000240]
['+47 81452779','Lucas Gonzales','lucas.gonzales@esito.no','4142834081432778',1000241]
['+47 34493845','Isabella Preston','isabella.preston@esito.no','4142834034353832',1000
['+47 69972016','Lilliana Saunders','lilliana.saunders@msn.com','4142834069791997',100
<<<     Anonymize_CUSTOMER (210ms)
>>>     Anonymize_ROOMCATEGORY
SELECT ID, INITIALPRICE, BEDTYPE, GUESTS, MAXDISCOUNT, ROOMQUALITY, LOCK_FLAG FROM APP
UPDATE APP.ROOMCATEGORY SET INITIALPRICE = ? WHERE ID = ?
[603.81,1]
[772.47,2]
[591.05,3]
[503.81,4]
[5251.20,5]
[668.55,6]
[1779.51,7]
[2693.10,8]
[2689.47,9]
[3394.02,10]
[1000.51,11]
[10253.18,12]
<<<     Anonymize_ROOMCATEGORY (92ms)
>>>     Delete_HOTELCHAIN
SELECT ID FROM APP.HOTELCHAIN WHERE ID = 0
SELECT ID FROM APP.HOTEL WHERE CHAIN_ID = 0
SELECT ID FROM APP.BOOKING WHERE HOTEL_ID = 15
SELECT SERIALNO, CUSTOMER_CUSTOMERNO FROM APP.STAY WHERE BOOKING_ID = 2006106
DELETE FROM APP.STAY WHERE BOOKING_ID = 2006106
SELECT SERIALNO, CUSTOMER_CUSTOMERNO FROM APP.STAY WHERE BOOKING_ID = 2007575
```

2.3.1.3 Output

Finally, we exited the script and checked the result. We can do this by simply refreshing the table if it is already open or by reopening the table.

CREDITCARD	CUSTOMERNO	EMAIL	NAME	PASSWORD	PHONE	CUSTODIAN	LOCK_FLAG
4142834044249...	1000239	martha.hopki...	Martha Hopkins	<null>	+47 44319116	<null>	<null>
4142834023108...	1000240	sarah.harvey@...	Sarah Harvey	<null>	+47 23238608	<null>	<null>
4142834081432...	1000241	lucas.gonzales...	Lucas Gonzales	<null>	+47 81452779	1000234	<null>
4142834034353...	1000242	isabella.presto...	Isabella Preston	<null>	+47 34493845	<null>	<null>
4142834069791...	1000243	lilliana.saunde...	Lilliana Saund...	<null>	+47 69972016	<null>	<null>
<null>	1000244	zaiden.mccon...	Zaiden Mccon...	<null>	+47 11 12 13 14	1000240	<null>
<null>	1000245	elian.harvey@...	Elian Harvey	<null>		<null>	<null>
<null>	1000246	pernille.groth...	Pernille Groth	<null>	+47 11 12 13 14	<null>	<null>
<null>	1000247	rory.mckinney...	Rory Mckinney	<null>	+47 11 12 13 14	1000234	<null>
<null>	1000248	franky.riggs@...	Franky Riggs	<null>	1234	<null>	<null>
<null>	1000303	chris.richard@...	Chris Richard	<null>	34535	<null>	<null>

Similarly, we can add other anonymization tasks by following similar steps as per requirements.

2.3.2 G9 Anonymizer test protocol

Availability. G9 is a standalone software application that runs both on Windows and Linux operating system. It is available both as tool that can be installed in our laptops or computers, and a web-based.

Anonymization methods. Just like the other two tools, G9 anonymizer also utilizes k anonymity for data anonymization.

Transparency and documentation. G9 anonymizer has high level of transperancy since it uses advanced anonymization techniques and provides details of the data to ensure that all of those data are anonymized correctly.
G9 anonymizer has comprehensive documentation that describes the features like data masking, encryption, validation, etc. and the ways to use them. It also provides guidelines on how to perform data anonymization.

Datasets supported CSV, JSON, and XML. In addition to these datasets, G9 anonymizer also supports relational databases such as MySQL, Oracle.

Country of origin/jurisdictional applicability? (EU/US/ETC)
G9 anonymizer is designed to provide data anonymization in accordance with applicable local, state, and federal laws, including European Union's GDPR (General Data Protection Regulation).

Methods and levels of anonymization
G9 uses various anonymization methods for data anonymization such as k-anonymity, pseudo-anonymization. To achieve a high level of data anonymization, G9 anonymizer uses a combination of pseudo anonymization and encryption techniques.

Purpose - What does the tool protect against

The tool provides protection against risks associated with unauthorized re-identification of individuals from anonymized data.

How easy is it for a user to install and use the system securely? E.g Are there parts that would be difficult for a naive user to complete, or complete securely?

We installed the Eclipse for Java developers' application for anonymization from the Eclipse website http://www.eclipse.org/downloads in our system rather than the virtual machine. We followed the user's manual for installing the application so it was quite easy. However, the installation of Eclipse was an additional step for G9, compared to ARX and Amnesia.
The tool provides the facility to ensure personal data from being disclosed to unauthorized users. However, for naïve users, the tool could be challenging to understand the types of information that need to be anonymized and the potential risk associated with anonymization, such as the risk of re-identification.

Are there any kinds of delays involved that might make users too impatient to use the tool?

While using the G9 anonymization tool some delays occurred because of the size and complexity of the data set being anonymized and the specific method being used (zdnet, 2021). The dataset to be anonymized was large, so it took a while to complete the process. Therefore, in such cases, the users might be impatient and leave the tasks halfway which would lead to incomplete or incorrect anonymization.

Is it possible to demonstrate where poor handling of the tool may contribute to a breach of anonymity?

If the G9 anonymizer is used improperly or if the user is not careful about how they use the tool, the data could be compromised. Additionally, failing to adequately test the anonymization process to ensure that it is effective in protecting sensitive information, could lead the anonymity to be compromised even if they use a G9 anonymizer.

Compared to the other tools you are looking at?

G9 is not very different from the other tools examined in that it is fairly straightforward to install, setup and use, and it also shares the common pitfalls of misconceptions and misuse that seem related to the field of anonymization and information processing, rather than a product of the tools.
It differs somewhat in the installation phase since it makes use of Eclipse, but we feel that despite this extra step, it is still fairly easy to install and set up.

What might the tools not protect against?

Re-identification of individuals: If the anonymized data is combined with other sources of information, it may be possible to re-identify individuals whose information was supposed to be anonymized (yourtechdiet, 2020).

Linkability: If different anonymized datasets contain common attributes or characteristics, it may be possible to link them together and potentially re-identify individuals.
Meta-data: Anonymization may not remove all meta-data from a dataset, and this information could potentially be used to identify individuals or link datasets together.
Insufficient protection: If the anonymization process is not applied correctly or thoroughly, sensitive information may be left in the dataset and could potentially be used to identify individuals.

What does each tool protect against?

G9 data anonymization tool might protect against:

- Identity theft: It can mask or remove personally identifying information such as name, social security number, and address.
- Discrimination: It can remove demographic information such as race, gender, and age that could be used to discriminate against individuals or groups.
- Privacy violations: It can remove sensitive information such as medical records, financial records, and personal preferences that could be used to violate an individual's privacy.
- Data breaches: It can prevent unauthorized access to data by encrypting it and requiring secure login credentials for access.

- Legal liabilities: It can ensure that data is handled by laws and regulations, such as the General Data Protection Regulation (GDPR) in the European Union.

How helpful is the documentation for different user profiles?

The G9 Anonymizer's documentation is quite helpful for users of all experience levels. For beginners, the documentation provides a comprehensive overview of the features and capabilities of the tool, as well as step-by-step instructions on how to use it. For more advanced users, the documentation includes detailed tutorials on advanced features such as customizing data anonymization policies and setting up data flows. Additionally, G9 Anonymizer provides databases anonymized for testing and development.

What common misconceptions might exist about the tool?

There are a few common misconceptions that may exist about G9 anonymizers:

- The G9 data anonymization tool is a complete privacy solution: While the G9 data anonymization tool can help protect sensitive personal information from being exposed, it is not a complete privacy solution (Utopia, 2021). Other measures such as secure data storage and access controls should also be implemented to ensure the highest level of protection.
- Data anonymization is a simple process: Data anonymization can be a complex process that requires careful planning and execution. It is important to consider the specific needs and requirements of the use case, as well as the risks and limitations of different anonymization methods.
- The G9 data anonymization tool is only suitable for certain types of data: The g9 data anonymization tool is capable of anonymizing a wide range of data types, including structured and unstructured data. It is suitable for use with a variety of data sources, including databases, spreadsheets, and text files.

Is it possible to uncover the anonymized data if the user does not use the tool properly?

Anonymized data could be re-identified or linked to specific individuals if the anonymization process is not applied properly or if the anonymized data is combined with other sources of information (gizmodo, 2019). This is known as a "breach of anonymity." For example: if the anonymized data is not properly secured, it may be possible for unauthorized individuals to access it and potentially re-identify individuals.

A short discussion on the cost/benefit ratio of the tool's benefits versus the potential loss of availability involved in installing, configuring, and using the tool.

The G9 data anonymization tool provides several benefits, including protecting sensitive personal data from unauthorized access and ensuring compliance with data privacy regulations. These benefits can potentially result in financial and reputational benefits for organizations that use the tool (mindtools, 2022). However, there is also a cost involved in installing, configuring, and using the tool. This includes the time and resources required to implement the tool, as well as any potential loss of availability that may occur during the installation and configuration process

3. Time Summary

Throughout the assignment we met regularly online to work together, discuss the tools, research areas, testing & results and finally writing the report. Between meetings we worked individually on the various tasks as well.

Member	Research & design	Installation & Setup	Tests	Report	Time Summary	Total
Jesper	**6**	**5**	**4**	**14**	**0,5**	**29,5**
Lekhaz	**5**	**4**	**4**	**15**	**0,5**	**28,5**
Puja	**6**	**8**	**3**	**13**	**0,3**	**30,3**
Total	**17**	**17**	**11**	**42**	**1,3**	**88,3**

Design: For the design part we took 4-5 hours to figure it out and come up with an outline on which type of data anonymization we should work with and we finally decided to work with three different kinds of data anonymization tools rather than for web anonymization.
We decided to work with three different kinds of data anonymization tools.We initially installed the tools on the virtual machine together, and discussed the test protocol. This took roughly 4 hours per student. Then we decided to assign responsibility for the tools among ourselves as one tool for each and later we discussed each of our work experience with the tool and pros and cons we faced while using the tool. For this part we took almost 12-14 hours per student with the tool and came up with an output that satisfies all the factors that are mentioned in the assignment instructions as well as a few added test parameters that we reckon is relevant for the topic. Firstly we have gone through each tool and looked up which tool supports what kind of platforms. Later we went through whether the tool is open source or not. Later on what kind of methods the tool can perform and support and up to what level does each tool is capable of. Next, we decided to work on the VM offered, but later thought that we are working with different kinds of tools so we decided to do this on our personal computers except one tool (G9 Anonymizer). We then went through the kinds of datasets that are supported by each tool so that we can work on single data so that we can compare them pretty easily. Next this consists of an installation process which is basically straight forward for both ARX tool and Amnesia while coming to G9 it's a bit more complicated as stated in the above documentation. After doing some work on each tool respectively we concluded that what the tool actually protects and how far does this protection goes and if it leads to any breaches in the anonymized data, the limitations and some policies of the tools. We also spent a great deal of time writing the report..

4. Reflections

This assignment has given us an opportunity to delve deeper into the aspect of privacy and, by extension, confidentiality. The specific tools may not have been so different but have given us the opportunity to explore the field of anonymization with its set of concepts and challenges.
The practical implementation of an anonymization tool is also dependent on other factors, relevant for information and computer security, such as legislature and policies as well as the balance between confidentiality, availability and integrity. Availability and integrity aspects need to be taken into consideration, since it is a tool and it does not operate in a vacuum. Since our assignment's primary focus was on the anonymization capabilities, we have not explored other information security aspects other than we found an obvious conflict and have tried to point those out in the results section.
In practice it has given us the opportunity to explore different mechanics for achieving anonymity, e.g suppression and generalization and by taking in the documentation given in webinars, research papers and commercial presentations we feel that we have become more knowledgeable, both in the academical fundamentals of the concepts and the more mundane implications of installing, using and maintaining software and datasets for various purposes in business related scenarios.
We have also reflected on factors such as transparency and the open source paradigm used for developing software and how that may benefit the field of data anonymization and also how that transparency could be misused and abused to create rather than eliminate vulnerabilities in the handling of personal information.

5.Reference list

Corporate Finance Institute. (n.d.). *Data Anonymization*. [online] Available at: https://corporatefinanceinstitute.com/resources/business-intelligence/data-anonymization/ [Accessed 5 Jan. 2023].

Dimakopoulos, M. T., Dimitris Tsitsigkos and Nikolaos. (n.d.). *Contact | Amnesia - Data anonymization made easy*. [online] amnesia.openaire.eu. Retrieved January 6, 2023, from https://amnesia.openaire.eu/features.html

Dimakopoulos, M.T., Dimitris Tsitsigkos and Nikolaos (n.d.). *Amnesia Anonymization Tool - Data anonymization made easy*. [online] amnesia.openaire.eu. Available at: https://amnesia.openaire.eu/index.html [Accessed 5 Jan. 2023].

F5.com. (2023). *The ARX system may experience poor NFS metadata performance*. [online] Available at: https://support.f5.com/csp/article/K10323 [Accessed 5 Jan. 2023].

General Data Protection Regulation (GDPR). (2013). *General Data Protection Regulation (GDPR) – Final text neatly arranged*. [online] Available at: https://gdpr-info.eu/chapter-3/ [Accessed 5 Jan. 2023].

Gizmodo. (n.d.). *Researchers Reveal That Anonymized Data Is Easy To Reverse Engineer*. [online] Available at: https://gizmodo.com/researchers-reveal-that-anonymized-data-is-easy-to-reve-1836629166 .

Haber, A.C., Sax, U. and Prasser, F. (2022). Open tools for quantitative anonymization of tabular phenotype data: literature review. *Briefings in Bioinformatics*, 23(6). doi:10.1093/bib/bbac440.

Iatropoulou, K. (n.d.). *OpenAIRE*. [online] OpenAIRE. Available at: https://www.openaire.eu/ [Accessed 5 Jan. 2023].

Matyszczyk, C. (2019). *Technology is making us more impatient, says study*. [online] ZDNet. Available at: https://www.zdnet.com/article/technology-is-making-us-more-impatient-says-study/.

Pfleeger, C.P., Pfleeger, S.L., Margulies, J. (2015). *Security in Computing (5th ed.)*. Prentice Hall.

Prasser, F., Eicher, J., Spengler, H., Bild, R. and Kuhn, K.A. (2020). *Flexible data anonymization using ARX - Current status and challenges ahead*. [online] GitHub. Available at: https://github.com/arx-deidentifier/arx [Accessed 5 Jan. 2023].

Prasser, F., Kohlmayer, F., Lautenschläger, R. and Kuhn, K.A. (2014). ARX - A Comprehensive Tool for Anonymizing Biomedical Data. *AMIA Annual Symposium Proceedings*, [online] 2014, pp.984–993. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4419984/ .

Utopia.Fans. (2021). *Safety Guide: What Is an Anonymizer gadget?* [online] Available at: https://utopia.fans/networks/safety-guide-what-is-an-anonymizer-gadget/ [Accessed 5 Jan. 2023].

www.esito.no. (n.d.). *g9 Anonymizer - Database masking and anonymization tool*. [online] Available at: https://www.esito.no/en/anonymizer/ Accessed 5 Jan. 2023].

www.mindtools.com. (n.d.). *MindTools | Home*. [online] Available at: https://www.mindtools.com/a7jgr0w/cost-benefit-analysis .

yourtechdietAdmin (n.d.). *Top 6 Data Anonymization Tools - 2020*. [online] YourTechDiet. Available at: https://yourtechdiet.com/blogs/6-best-data-anonymization-tools/ .