Personal Data Minimization Technology

Joshua Civile, Ben Starrenburg & Yoren Wierenga
0907482 0984798 0957598

Abstact: This study aims to evaluate the effectiveness of various privacy-preserving techniques in transforming the data set ADULT while preserving the utility of the data for the purpose of building a predictive model for the attribute 'Salary' based on other attributes in the data set. The study uses anonymization techniques such as attribute mapping, k-anonymity, and l-diversity to transform the data set to reduce the risk of re-identification and sensitive information disclosure. The effectiveness of these techniques is assessed through the measurement of prosecutor, journalist, and marketeer risk. The research questions focus on the impact of these techniques on the privacy preservation and utility of the data for building the predictive model. The study also compares the transformed data set with the original data set in terms of utility and privacy preservation. The results of the study will contribute to the field of data privacy and support the development of privacy-preserving techniques for sharing data sets in a manner that adheres to the principles of privacy and data protection.

30 JANUARI 2023

HOGESCHOOL ROTTERDAM

GROUP 3

TABLE OF CONTENTS

I. INTRODUCTION

Data privacy is a critical concern in the era of big data and advanced analytics. As a data controller, it is essential to ensure that any sharing of data sets adheres to the principles of privacy and data protection.
In this study, we aim to share a transformed version of the data set ADULT with the objective of deriving a predictive model for the attribute 'Salary' based on the other attributes in the data set. The usage purpose of the transformed data set is to build a model that can accurately predict the salary of individuals in the data set based on their other attributes.

To ensure that the data set is shared in a manner that is privacy-preserving, it is necessary to transform the data set to reduce the risk of re-identification of individuals and the disclosure of sensitive information. In this study, various methods where used for transforming data sets, including anonymization techniques such as attribute mapping, k-anonymity and l-diversity.
This study will evaluate the effectiveness of these techniques in transforming the data set ADULT while preserving the utility of the data for the purpose of building a predictive model. To accomplish this, a range of methods were utilized to assess the effectiveness of these techniques through the measurement of prosecutor, journalist, and marketeer risk.

The research will attempt to answer the following research questions:

- What are the various privacy-preserving techniques used for transforming the data set ADULT?
- How effective are these techniques in reducing the risk of re-identification and disclosure of sensitive information?
- How does the application of attribute mapping, k-anonymity, and l-diversity affect the utility of the data for building a predictive model for the attribute 'Salary'?
- What is the impact of using different methods for transforming the data set ADULT on the prosecutor, journalist, and marketeer risk?
- How does the transformed data set ADULT compare to the original data set in terms of utility and privacy preservation?

The remainder of the report is structured in the following manner: In Section 2, the theoretical background necessary to understand the experiments will be provided. In Section 3, experiments that were conducted to investigate the topic will be described. In Section 4, the results of these experiments will be presented. And in Section 5, the conclusions that can be drawn from this work will be discussed.

II. THEORETICAL BACKGROUND

In this section, we review the work previously done on the subject of personal data minimization. First, we discuss the attribute mapping theory. Then we present work about k-anonymity and related concepts. After, we present the 3 chosen data utility and disclosure measures used to check performance and the strength of anonymity of the data set. At last, we discuss the suppression limit in ARX in order to measure the trade-off between privacy protection and data utility.

Attribute mapping is a technique for anonymizing data that is used to protect the privacy of individuals in a dataset. It involves modifying the attributes of the individuals in the dataset in order to make it more difficult to re-identify them [1].

Attribute mapping can be performed on a variety of attributes, including demographic attributes, sensitive attributes, and quasi-identifiers.

Demographic attributes are characteristics that describe an individual, such as their age, gender, and occupation. Sensitive attributes are attributes that reveal sensitive information about an individual, such as their medical history or financial information. Quasi-identifiers are attributes that can be used to re-identify an individual in combination with other available information [2].

Attribute mapping can be performed on one or more of these attributes in order to preserve the privacy of individuals in a dataset [3]. For example, demographic attributes can be modified by adding random noise to the values or by generalizing the values to a higher level of aggregation (e.g., converting an individual's age from a specific value to a range of values). Sensitive attributes can be suppressed or replaced with synthetic values that preserve the distribution of the original values while obscuring the actual values. Quasi-identifiers can be modified by adding random noise to the values or by generalizing the values to a higher level of aggregation.

K-anonymity is a widely used concept in the field of data privacy and data management. It is a privacy model that requires that each individual in a dataset be indistinguishable from at least k-1 other individuals in the dataset with respect to a set of quasi-identifiers [4].The purpose of k-anonymity is to ensure that the disclosure of an individual's quasi-identifiers does not allow an attacker to re-identify the individual with a high degree of certainty [5].

The theory behind k-anonymity is that by making each individual indistinguishable from at least k-1 other individuals in the dataset [4], the privacy of each individual is protected. For example, if an attacker has access to a dataset that has been anonymized using k-anonymity with a value of k=3, the attacker would not be able to re-identify an individual with a high degree of certainty because the individual would be indistinguishable from at least two other individuals in the dataset.

K-anonymity is often combined with other privacy models, such as L-diversity and T-closeness, to provide a more comprehensive privacy protection. L-diversity requires that each group of individuals with the same quasi-identifiers have a minimum number of sensitive attribute values (L) to prevent the attacker from learning sensitive information about the individuals [6]. T-closeness requires that the distribution of sensitive attributes within each

group be similar to the distribution of sensitive attributes in the overall dataset, to prevent the attacker from learning sensitive information about the individuals.

Data utility measures

Classification accuracy

Classification accuracy is a measure of the accuracy of a classifier, which is a model that is used to predict the class or label of a set of data points [7]. It is defined as the number of correct classifications made by the classifier divided by the total number of classifications made.

Classification accuracy is often used as a measure of data utility when evaluating the performance of privacy-preserving data management systems, such as anonymization and data masking techniques. These systems are used to protect the privacy of individuals in a dataset by modifying the data in a way that makes it more difficult to re-identify the individuals.

Classification accuracy can be used to evaluate the utility of the modified data by comparing the accuracy of a classifier trained on the original data to the accuracy of a classifier trained on the modified data. If the accuracy of the classifier trained on the modified data is significantly lower than the accuracy of the classifier trained on the original data, then it is likely that the privacy-preserving data management system has reduced the utility of the data.

Non Uniform Entropy

Non-Uniform Entropy (NUE) is a data utility measure that is used in the context of data privacy. It provides a metric for quantifying the loss of information in a dataset due to data privacy protection measures, such as data generalization or suppression [8].

NUE is defined as the entropy of a dataset after it has been transformed to protect the privacy of its individuals. The entropy of a dataset represents the amount of uncertainty or randomness in the data. The NUE metric compares the entropy of the original dataset to the entropy of the transformed (i.e., privacy-protected) dataset and provides a measure of the loss of information that has occurred due to the privacy protection measures.

Granularity

Granularity refers to the level of detail and specificity in a dataset. In the context of data privacy, granularity is an important measure of data utility, as it affects the level of privacy protection that a dataset provides [8].

A high granularity in a dataset means that there is a high level of detail and specificity in the data, while low granularity means that the data is less specific and contains less detail. In terms of data privacy, a high granularity can increase the risk of re-identification, as it makes it easier to link personal information to specific individuals. On the other hand, low granularity can reduce the utility of the data, as it makes it less useful for analytical purposes.

To balance privacy protection and data utility, data generalization techniques are often used to reduce the granularity of a dataset. For example, exact birthdates may be replaced with birth year ranges, or names and addresses may be removed. This reduces the risk of re-identification, while still maintaining some level of data utility.

Data disclosure measures

Prosecutor attack

The Prosecutor attack is a method used to assess the risk of re-identification or disclosure of sensitive information in a dataset. It is a type of attack in the field of data privacy that is used to evaluate the effectiveness of data protection measures.

The Prosecutor attack is defined as an attack in which an attacker tries to re-identify individuals in a dataset by linking the information in the dataset to background information [9]. This can result in the disclosure of sensitive information, such as names, addresses, and birthdates, which can then be used to identify individuals.

In a Prosecutor attack, the attacker starts with partial information about an individual and tries to link it to more specific information in the dataset. If the attacker is successful in linking the information, they can use the linked information to re-identify the individual [10].

Journalist attack

The journalist attack is a privacy threat that refers to the process of re-identifying individuals in a dataset by cross-referencing it with an "identification database" [9]. The goal of a journalist attack is to to re-identify individuals in a de-identified dataset.

In the context of data privacy, the journalist attack is a measure of data disclosure, as it tests the ability of a dataset to protect the privacy of individuals. The success of a journalist attack depends on the level of detail and specificity in the dataset, as well as the amount of publicly available information that can be used to re-identify individuals [10].

To prevent journalist attacks, data suppression and data generalization techniques are often used to reduce the granularity and specificity of a dataset. For example, exact birthdates may be replaced with birth year ranges, or names and addresses may be removed. These techniques can make it more difficult to re-identify individuals and, therefore, to disclose sensitive or private information.

Marketer attack

Market Basket Attack, also known as Marketer Attack, is a type of data disclosure attack that occurs in the context of data privacy. It is a privacy breach in which a malicious party attempts to infer sensitive information about individuals in a dataset based on the correlations between items in the dataset [9].

The Market Basket Attack is based on the assumption that items that are frequently purchased together in a market basket can reveal sensitive information about individuals. For example, the purchase of pregnancy test products may reveal sensitive information about an individual's pregnancy status [10].
The attacker can use this information to infer other sensitive information about the individual, such as their personal habits or medical conditions.

In the context of data privacy, the Market Basket Attack highlights the need for privacy-preserving techniques to be used when publishing or sharing data. For example, data generalization and suppression techniques can be used to reduce the granularity of data, or data can be randomized to reduce the correlation between items in the dataset [10].

Suppression limit ARX

ARX (Anonymization Research and Evaluation eXchange) is a privacy-preserving data management system that is used to anonymize large datasets. One of the key concepts in ARX is the suppression limit, which determines the minimum number of records in a dataset that must be suppressed in order to anonymize it [11].

The suppression limit is based on the theory of K-anonymity, which requires that the information of at least k individuals in a dataset be indistinguishable from each other in order to protect their privacy. The suppression limit of ARX is the minimum number of records that must be suppressed in order to achieve k-anonymity.

The suppression limit is an important measure of the trade-off between privacy protection and data utility. A higher suppression limit will result in a higher level of privacy protection, but will also reduce the utility of the data by reducing its size. On the other hand, a lower suppression limit will result in a higher level of data utility, but will also reduce the privacy protection of the data.

III. EXPERIMENTS

To run the experiments, ARX (Anonymization Research and Evaluation eXchange) has been used. The software has been obtained from arx.deidentifier.org/ and the version used is: 3.9.1. The experiments have been performed on a desktop running windows 10 21H2 with an AMD Ryzen 5 3600 hexa-core with 16gb ddr4 3600 MT/s.

The dataset used is called ADULT. Obtained from archive.ics.uci.edu/ml/datasets/Adult. The raw data set contains 48842 number of instances divided over 14 attributes. The raw dataset was reasonably cleaned well, but needed some additional preparation to be used in a usable manner.

Data preparation

At first all rows with zero values for non-numerical attributes were deleted. All rows containing the value "?" for the attributes were also deleted. Secondly a new explicit identifier was created with ID as attribute name, in order to differentiate between rows.

Next step was identifying the unnecessary attributes (UAT). Given the purpose of our data sharing, and usage for a salary prediction model, we have taking in consideration the distribution, skewness and correlation of the provided data to decide whether an attribute is necessary or not [12]. Table 1 below shows the UAT we have dropped with corresponding reason.

Table 1: Unnecessary attributes

UAT	Reason
Education_num	Index for attribute Education
Marital_statuse	Should not influence salary
Relationship	Should not influence salary
Race	Heavily skewed for white (Shouldn't influence salary)
Capital-gain	Heavily left skewed
Capital-loss	Heavily left skewed
Native-country	Heavily skewed towards USA
fnlwgt	No correlation

All transformations above resulted in a new modified data set. This data set has been used as basis for the attributes mapping necessary for ARX.

Attribute Mapping

In order to start the experiments, the attributes had to be mapped to a specific identifier. The attributes were split in two scenarios. Scenario a where external parties do not have access to the original data and scenario b where you and the external parties do have access to the original data.

Table 2: Attributes with identifier and reason

Attribute	Identifier	(abbr.)	Reason
ID	Explicit	EID	Directly identifies specific person
Age	Quasi	QID	Could be found in external data sources
Sex	Quasi	QID	Could be found in external data source
Education	Quasi	QID	Could be found in external data source
Occupation	Quasi	QID	Could be found in external data source
Workclass	Quasi/Non Sensitive	QID/NAT	Is Quasi Identifier for parties with access to original data source
Hours per week	Quasi/ Non Sensitive	QID/NAT	Is Quasi Identifier for parties with access to original data source
Salary	Sensitive	SAT	Sensitive, attribute trying to predict in model

Table 2 shows the remaining attributes from the modified data set, with their corresponding identifier. The attributes ‘Workclass’ and ‘Hours per week’ are the two attributes that have been assigned two different identifiers. As it is assumed that these two attributes are only Quasi identifying if and only if the external party has access to the original dataset. Table 3 Shows all attributes mapped with their identifier per scenario.

All the attributes and their related identifiers per scenario have been summarized in Table 3

Table 3: Attributes mapped per scenario

Attributes	A	B
ID		
Age		
Workclass		
Fnlwgt		
Education		
Education num		
Marital status		
Occupation		
Relationship		
Race		
Sex		
Capital gain		
Capital loss		
Hours per week		
Native country		
Salary		

LEGEND
EID
QID
NAT
SAT
UAT

Taxonomy Trees

For Every QID a related taxonomy tree has to be provided. In these experiments two different taxonomy trees have been created per scenario:

- Coarse Taxonomy Tree
  A rough and quick created tree with a simple structure and a few branches.

- Detailed Taxonomy tree
  A detailed optimum grouped attributes values differentiate over a larger amount of branches.

The approach used for the coarse taxonomy trees was that they have a maximum of 3 levels: Attribute value, one grouping layer and layer with 'any' as value. All coarse trees share the same base as the detailed taxonomy trees. They can be distinguished as the black rectangle indicates the Coarse tree.

Figure 1: Shows a combination of the coarse and detailed taxonomy tree for attribute age.

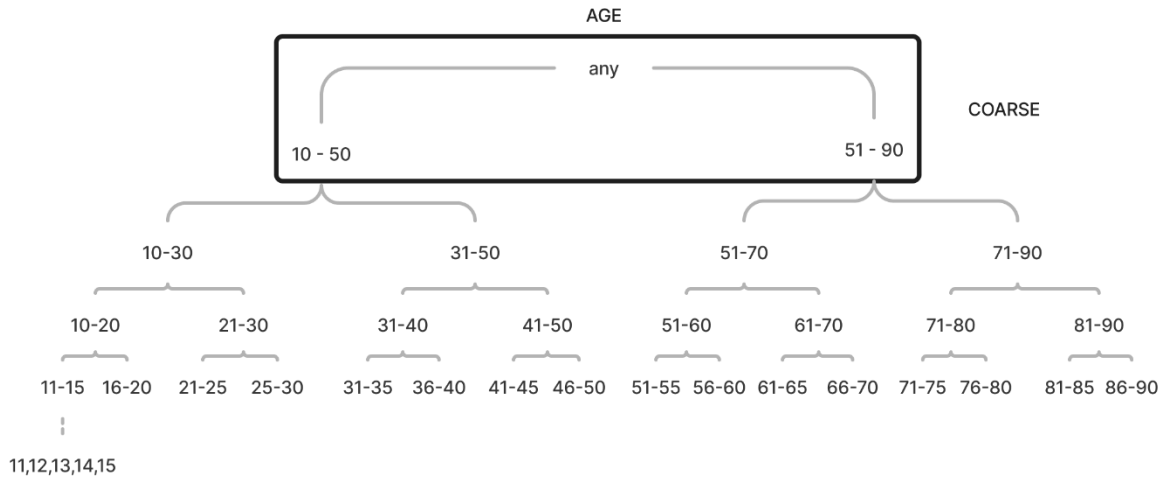


Figure 1:Combination of Coarse and Detailed Taxonomy tree for attribute age

This tree shows a clear distinction between the branches of the coarse and detailed tree. All detailed taxonomy trees have been created in mind to add at least a 1 and up to 3 extra branch layers where possible. All other Taxonomy trees can be found in Appendix I. In some cases, the coarse and detailed taxonomy trees are the same due to low differentiation between the attributes categories provided by the Adult data set.

Experiment and configuration

The goal is to anonymize the data per scenario with coarse and detailed taxonomy tree. The privacy model chosen for the QID is k-anonymity. The Privacy model chosen for the SAT (Salary) is distinct-l-diversity. For each combination 10 experiments have been performed with a different k value. The used k values are: 1,5,10,15,20,25,30,40,50,75 (notice that k =1 means

data without applying any anonymization models). As Salary is the only chosen SAT, and groups the value into two categories, '<=50K'and '>50K', the value of l in the l-diversity model can only be set to 2.

In every experiment ARX has been configured with a suppression limit of 10%. The experiment results had to comply to a couple of conditions:

- Be able to investigate the existence of bias against male versus female in the output.
- Be able to investigate the existence of bias against youth vs middle aged.
- Measure three utility measures with one being classification accuracy for 'salary'

In order to meet the different conditions a pre-set was added to the levels of the taxonomy trees. For the first condition the pre-set did not allow the tree passed the first level so the output will always contain the categories 'Male 'or 'Female'.

To meet the second condition, a pre-set was added to the taxonomy trees of age so that the output data would not show a group with an age difference greater dan 10 (i.e., the maximum level was set to 2 for detailed and 1 for coarse).

And the last condition was met by choosing Classification accuracy as utility measure in the 'Configure transformation' tab with salary as the target variable in the 'Attribute metadata' tab. The other utility measure results were extracted from 'Analyze utility' tab with subtabs 'Quality models' and 'Classification models'.

Because the experiments were performed with the preset to the taxonomy trees, the generalization state for the different attributes met the requirements. Therefore, the optimum generalization state provided by ARX in the 'Explore results' tab was used.

For example, the generalization state given by ARX with k=50 and l=2 is:

[Age, Education, Occupation, Sex] = [2, 1, 3, 0]

All other generalization states can be found in Appendix II

For each experiment the data has been logged for the following values:

- Value of k
- Value of l
- Utility Measures
    - Classification accuracy
    - Granularity
    - N-U Entropy
- Risk Measures (success rate)
    - Prosecutor
    - Journalist
    - Marketer
- Generalization state per QID

These values have been used in the visualization and analysis of the performances of the anonymization experiments as well as uniqueness and identifiability in chapter IV.

IV. RESULTS

In this chapter the results of the experiments done in Chapter III will be visualized and the graphs will be analysed. The graphs in Figure are grouped per utility and risk measure and each line represent the scenario a and b with coarse and detailed taxonomy trees. Secondly a utility privacy trade of has been visualized by choosing one utility measure and one risk measurement. Only the values from scenario a and b with detailed taxonomy trees have been used. These graphs are represented in Figure . At last, the Uniqueness and identifiability have been visualized using Python to draw histograms with percentage of the records uniqueness of the QID against all attributes. The dataset with k =10 and detailed taxonomy trees for each scenario were used, as well as the original input data. The graphs can be seen in Figure .

Each visualisation graph will be interpreted and explained in the sections below.

Trend Analysis

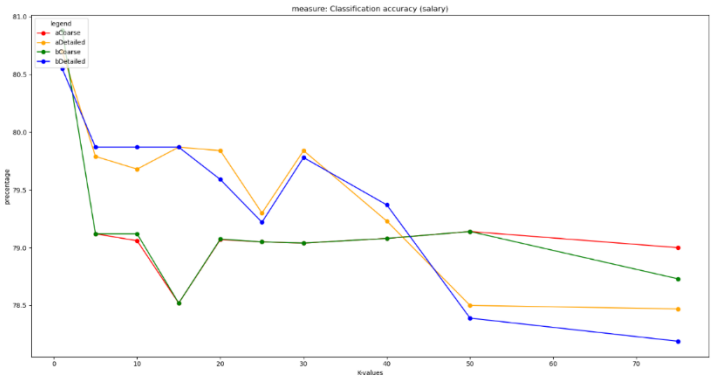
Figure can be separated into two parts. (1) where a, b and c refer to the utility measures and (2) where d, e and f refer to the risk measures.

the three chosen utility measures are plotted against the value of k seen in part (1). Figure 2a shows the classification accuracy plotted against the values of k. At first the scenario a and b used with a coarse taxonomy drop in accuracy significantly faster than the detailed taxonomy tree. Due to low level of branches in the coarse trees, the performance of the classification accuracy drops. With higher values of K, the performance of the classification accuracy drops off significantly for detailed taxonomy trees as they regularization tents to be in line with the coarse taxonomy trees.

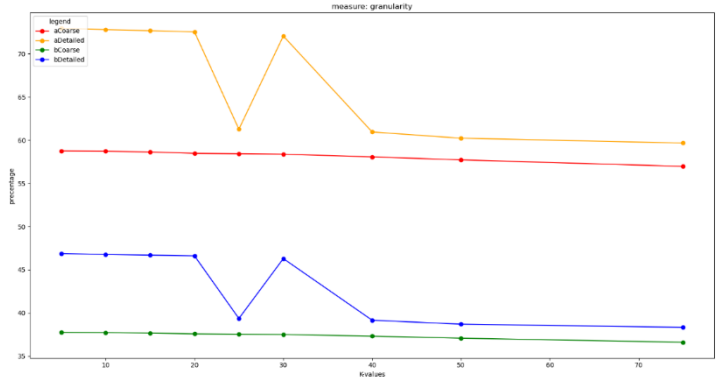
Figure 2b shows the granularity and N-U. Entropy plotted against the k values. As the k increases the granularity decreases this is in line with the fact that a low granularity means that the data is more generalized. As seen in the figure the granularity in scenario b is lower than a because of more generalisation of the attributes and thus a lower risk of re-identification.

This principle is reaffirmed in figure 2c where the value of N-U. Entropy also decreases as the value of k increases. Entropy is a measure of uncertainty in a dataset, a low entropy value indicates that the uncertainty of a specific attribute in the dataset is low which means a high privacy protection. Scenario b shows a lower entropy level than scenario a. As scenario b uses more QID the privacy performance is greater which results in a lower usability.

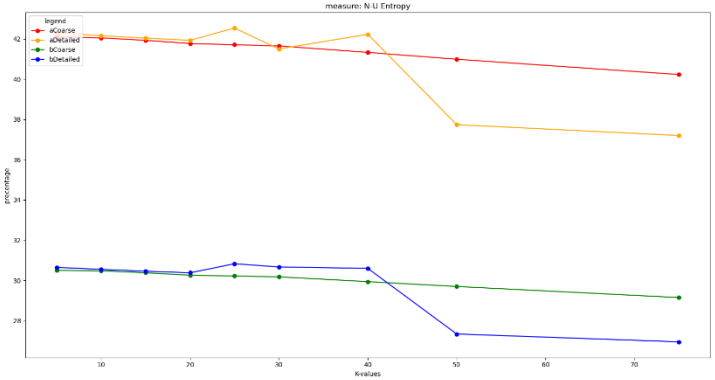
The risk measures of the output data can be seen in part (2). take Figure d, by increasing the value of k a higher privacy is yielded (i.e., The higher the value of K the lower the success rates). This is reflected the same for Figures 2e and 2f. The scenario used with a coarse taxonomy tree initially show a better privacy performance whereas the scenario used with a detailed taxonomy tree show an even better performance with higher k's. Note that the their seems to be no difference in the performance of the three risk measures since the three types of attack work in a fairly similar way.



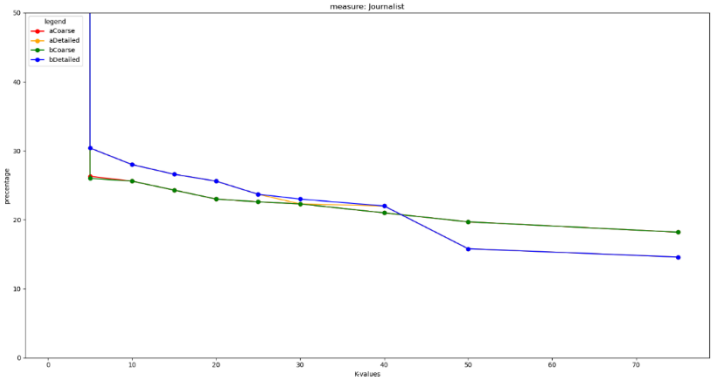
(a) Classification Accuracy



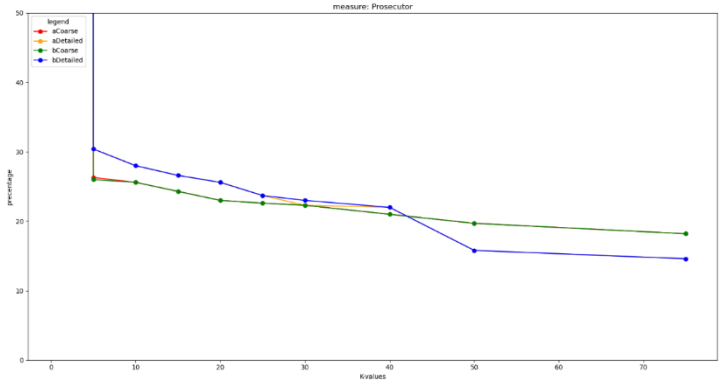
(b) Granularity



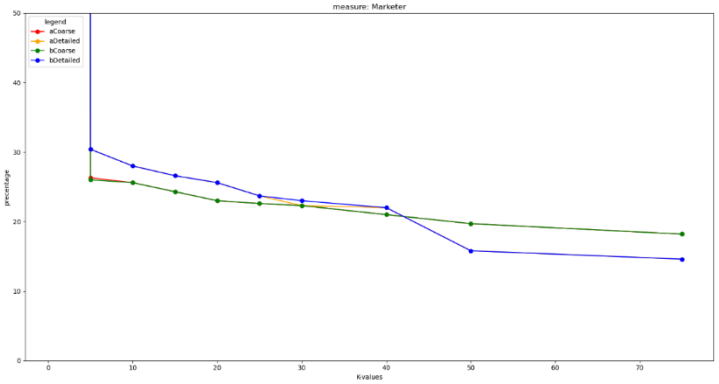
(c) N-U. Entropy



(d) Journalist



(e) Procecuter



(f) Marketer

Figure 2: Utility and risk Measurements for K

Utility-privacy trade-off analysis

The graphs in figure 3 show a clear positive correlation between increasing entropy and an increase in the prosecutor attack risk success rate in both scenario a and b. It is noteworthy to mention that the same trend is observed in both scenario A and B, which is in line with the observations made in the trend analysis where the entropy and success rates decreases as the value of k, privacy performance, increases

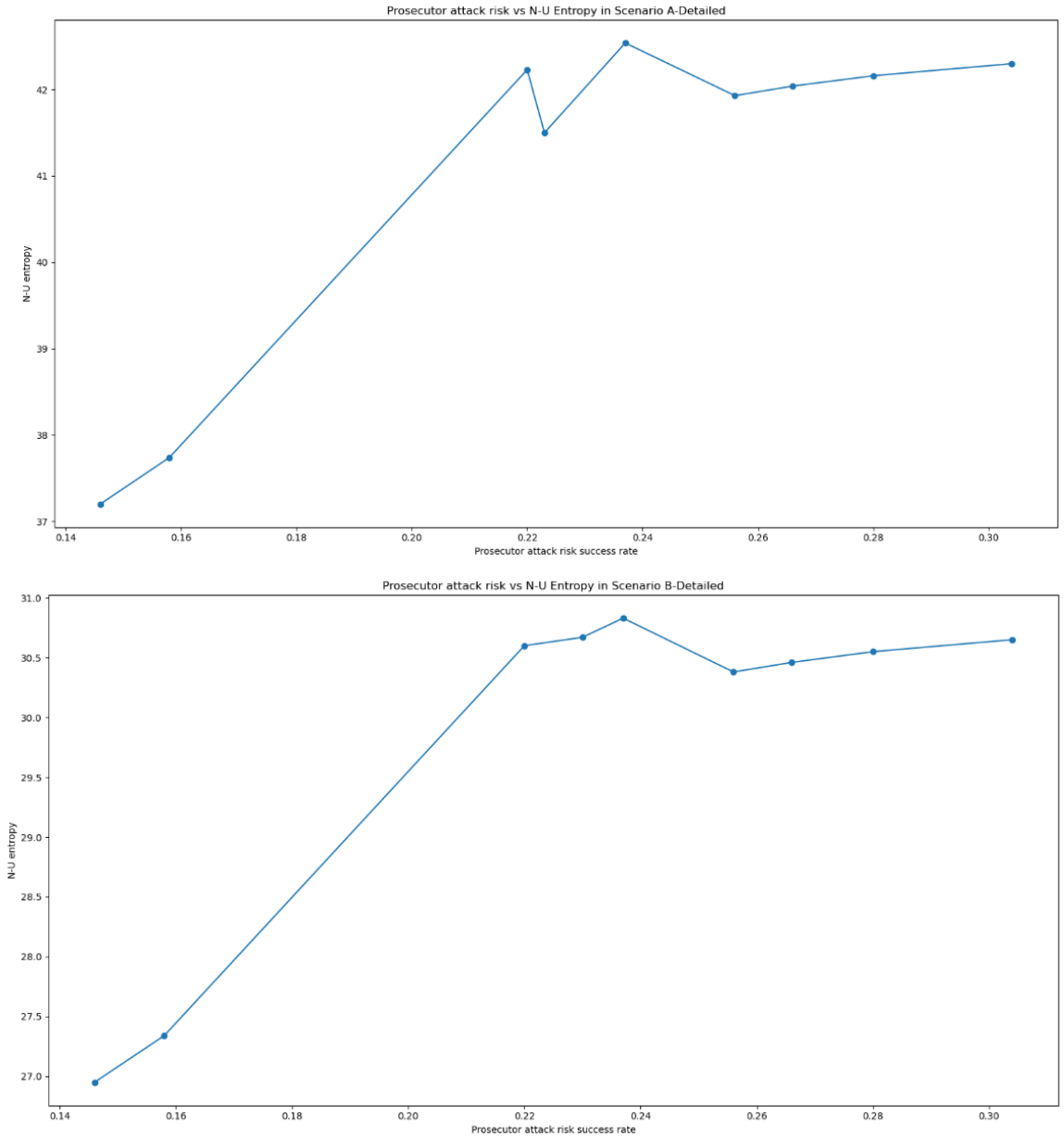


Figure 3: Utility-privacy trade of for N-U. Entropy vs. procescuter succes rate per scenario

Uniqueness and identifiability

The unique identifiability experiments have been carried out by creating a function that is call with different data and attributes. This function first selects the columns of the used quasi attributes, next it groups this data by row and calculates the size of each group. Secondly the size of each group is added to its respective row and a distinct select is done. Finally, the counts are plotted in a histogram, with uniqueness on the x-axis and the percentage of records on the y-axis, in the first subplot. After doing this with the quasi attributes, the same steps are followed, but with all attributes instead of the quasi attributes. Next the counts are plotted in the second subplot, also with uniqueness on the x-axis and the percentage or records on the y-axis. The full pseudocode can be seen in appendix III. The outputted figure can be seen in figure 4.

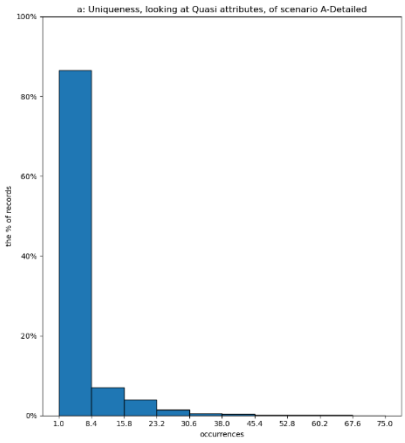
When analysing the outputted figure, there will be looked at the uniqueness of a data record in a microdata set with respect to a subset of attributes, because these records can be distinguished from others in the set. This refers to the degree to which values of the chosen attributes are unique within the dataset. If the combination of values in a subset of attributes is found in only a single record, then that record is considered unique. When a record is unique this might have consequences regarding privacy.

The consequences of having a unique record in a dataset is the risk of re-identification and attribution. This could possibly lead to multiple consequences. Your privacy is unwillingly taken away, it could lead to security thread, data could be used in legal liability for the data owners and trust could be lost.

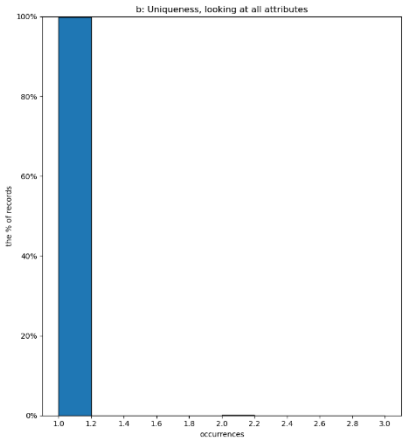
Figure 4a and 4b display the uniqueness in the original data. In the original data with all attributes almost every row is unique, this means re-identification would be very easy. By looking at the quasi attributes some records have a lot more identical occurrences in de dataset, meaning this already helped increasing the uniqueness. However, there are still a lot of unique occurrences left.

Figure 4c and 4d display the uniqueness in the output dataset of experiment a-detailed with a k-value of 10. When looking at all attributes, the graph looks a lot like figure 4b. Almost every row is still unique, however you can see that some records have multiple identical occurrences in the dataset. By looking at the quasi attributes the amount of unique records is lowered, it is also lower than in figure 4a. The biggest difference with figure 4a is how some records have a lot of identical occurrences, this is also why the x-axis is scaled a lot wider. This means that even though there are a lot of unique records, the records that are not unique are a lot more difficult to re-identify then before.

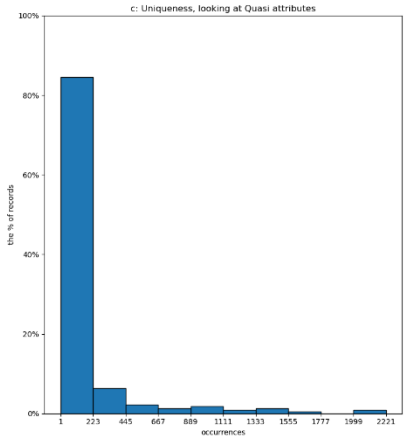
Figure 4e and 4f display the uniqueness in the output dataset of experiment b-detailed with a k-value of 10. In this experiment all attributes are of type quasi, meaning both figures display the same graph. There are no longer unique records which is a improvement compared to the figures 4a, 4b, 4c and 4d. Also even more rows have identical occurrences, meaning there is more privacy preserving in this experiment.



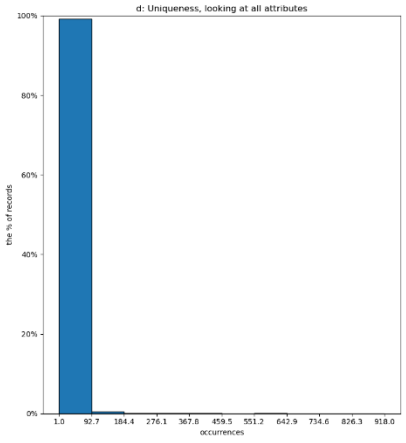
(a) original data only QID



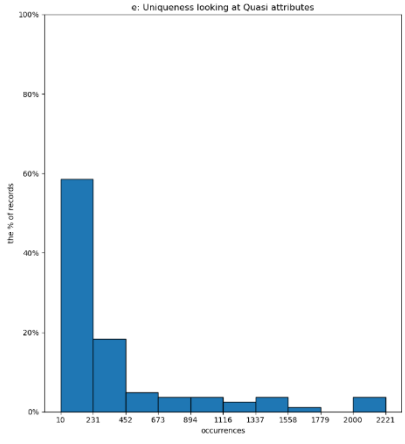
(b) original data all attributes



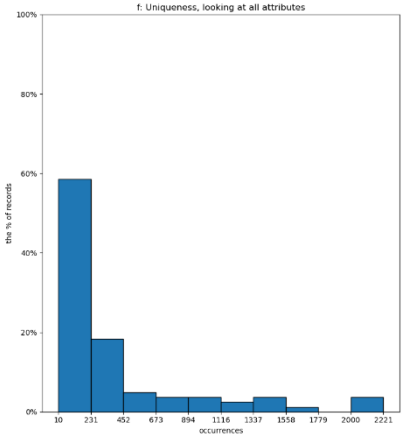
(c) Scenario a, only QID



(d) Scenario a, all attributes



(e) Scenario b, only QID



(f) Scenario b, all attributes

Figure 4: Uniqueness and identifiability

V. CONCLUSION

In these experiment couple of privacy preserving techniques have been applied. Firstly, the data has been cleaned and the attributes have been mapped to identifiers. Secondly the two main privacy models have been applied. K-anonymity for the quasi identifiers and l-diversity for the sensitive attributes. The level of privacy performance has been varied by increasing the k value.

The application of attribute mapping, k-anonymity, and l-diversity can affect the utility of the data for building a predictive model for the attribute 'Salary' by reducing the risk of re-identification of individuals and disclosure of sensitive information. Attribute mapping can be used to modify the values of attributes in the data set to protect sensitive information, while k-anonymity can be used to ensure that a minimum number of records in the data set have the same values for a set of attributes, making it difficult to identify an individual. L-diversity can be used to ensure that each group of similar records in the data set has a sufficient number of different values for a sensitive attribute, reducing the risk of disclosing sensitive information.

Using these models helped by reducing the amount of unique records of data. In scenario a there are still some unique records, however the records that are not unique are a lot more difficult to identify. In scenario b none of the records remained unique, they all records had at least ten duplicates as seen in figures 4e and 4f. However, the application of these techniques can also reduce the utility of the data, by reducing the amount of information available, to such an extent, that it can lead to a decrease in the accuracy of the predictive model.

The impact of using different methods for transforming the data set on the prosecutor, journalist and marketeer risk is that the accuracy, credibility and reliability of the data becomes less the higher the value of k gets. Using different methods results in data records becoming less unique, because of this it is more difficult extract and target an individual.

The relationship between utility measure and k value in k-anonymity can be seen as a trade-off between privacy protection and usability. High k values result in greater privacy protection but lower usability of the data and any analysis performed on it, while, vice versa , lower K values result in less privacy protection but higher accuracy of the data. Balancing these two factors is an important consideration in data protection.

VI. ACKNOWLEDGEMENT

The authors would like to extend their gratitude to the creators of the ADULT dataset, who have made it possible to conduct this research. The authors also wish to acknowledge the valuable support and guidance provided by our professor, Bargh, in the creation of this assignment. Their insights and expertise have been instrumental in shaping the direction and outcome of this study.

VII. REFERENCES

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

[2] A. S. L. R. a. Y. E. Amir. H, "M-Score: A Misuseability Weight Measure," *IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING,* pp. 419-428, May 2012.

[3] a. S. L. A. Majeed, "Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey," *IEEE Access,* pp. 8512-8545, 2021.

[4] Bayardo, J. Roberto and R. Agrawal, "Data privacy through optimal k-anonymization," *21st International conference on data engineering,* pp. 217-228, May 2005.

[5] V. S. lyengar, "Transforming Data to Satisfy Privacy Constraints," *international conference on Knowledge discovery and data mining,* pp. 279-288, July 2002.

[6] A. Machanavajjhala, J. Gehrke, D. Kifer and M. Venkitasubramaniam, "L-diversity: privacy beyond k-anonymity," *22nd International Conference on Data Engineering,* pp. 1-12, 2006.

[7] S. V. Stehman, "Selecting and interpreting measures of thematic classification accuracy," *Remote Sensing of Environment,* pp. 77-89, 1997.

[8] A. Amighi, B. S. Mortaza, S. Choenni, A. Latenko and R. Meijer, "On Protecting Microdata in Open Data Settings from a Data Utility Perspective," *The Fourteenth International Conference on Digital Society,* pp. 20-29, 2020.

[9] K. El Emam, "RE-DENTIFICATION RISKN DE-DENTIFIED DATABASES CONTAINING PERSONAL INFORMATION". USA Patent US 8,316,054 B2, 20 Novembre 2012.

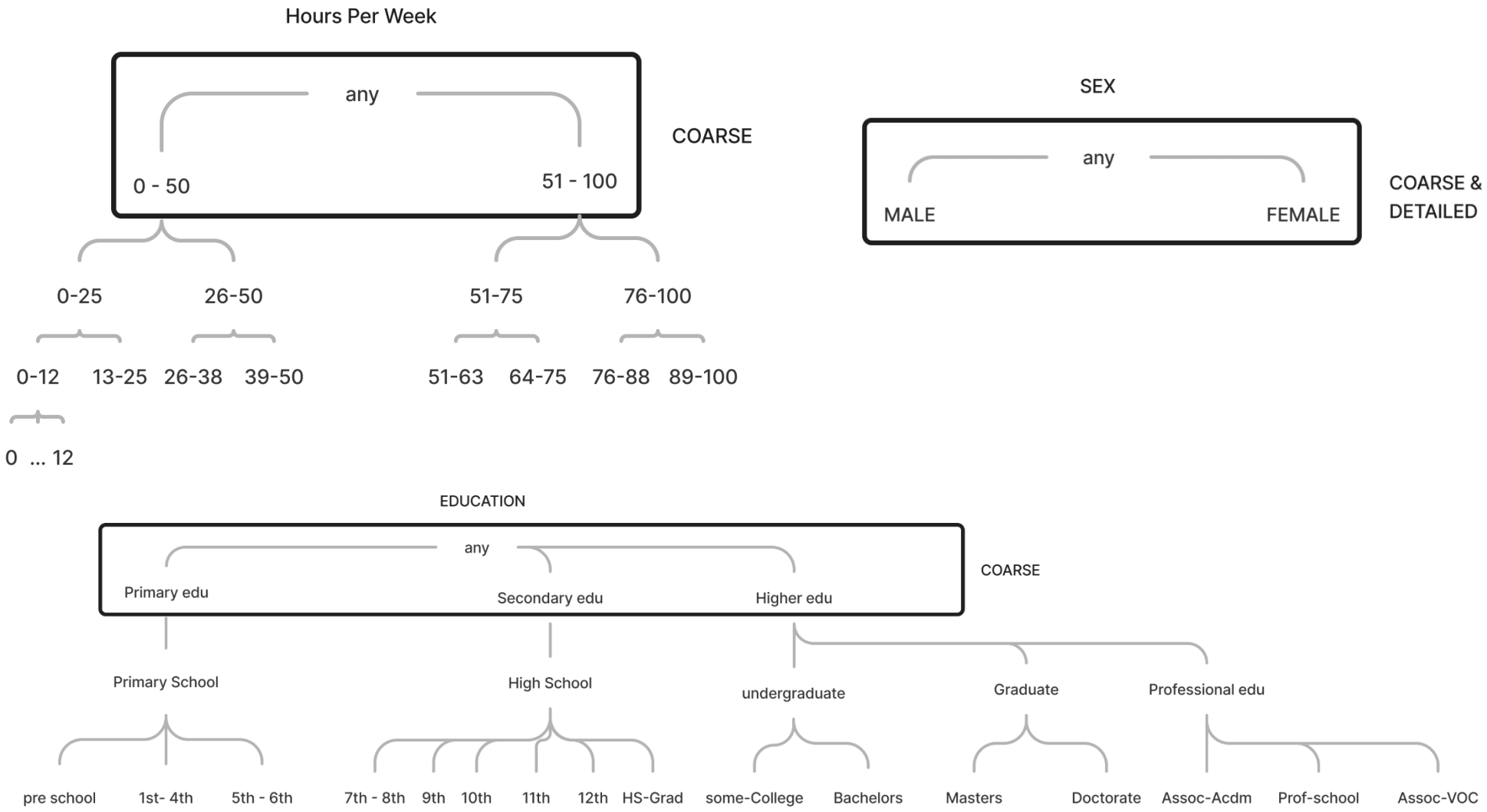
[10] W. Xia, Y. Liu, Z. Wan, Y. Vorobeychik, S. Nyemba, E. W. Clayton and B. A. Malin, "Enabling realistic health data re-identification risk assessment through adversarial modeling," *Journal of the American Medical Informatics Association,* pp. 744-752, April 2021.

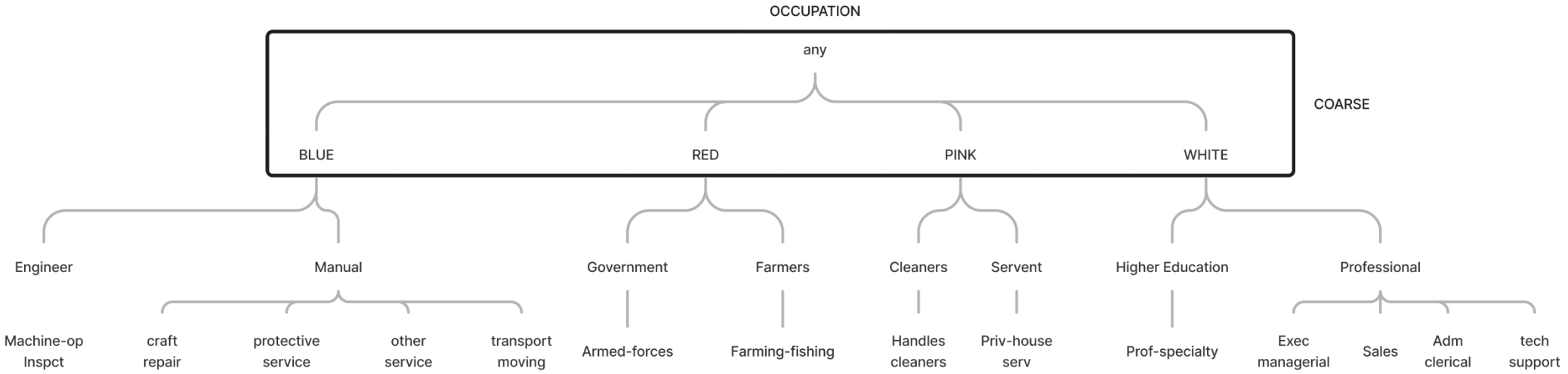
[11] ARX, "Utility analysis," ARX - Data Anonymization Tool, 2013. [Online]. Available: https://arx.deidentifier.org/anonymization-tool/analysis/. [Accessed 21 01 2023].

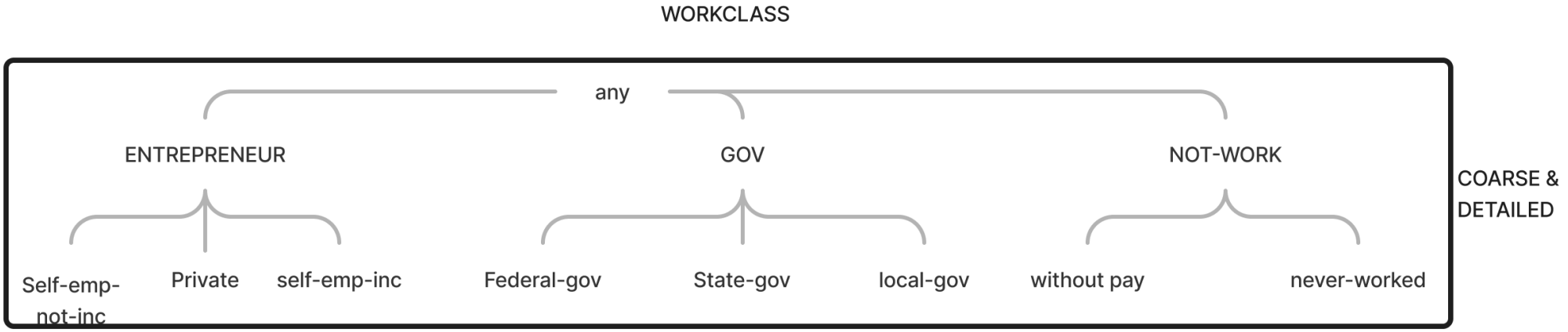
[12] K. Bhanot, "Will your income be more than $50K/yr? Machine Learning can tell," towardsdatascience, 1 April 2019. [Online]. Available: https://towardsdatascience.com/will-your-income-be-more-than-50k-yr-machine-learning-can-tell-92138745fa24. [Accessed 20 January 2023].

VIII. APPENDIX

APPENDIX I: Taxonomy Trees




OCCUPATION
any
COARSE
BLUE
RED
PINK
WHITE
Engineer
Manual
Government
Farmers
Cleaners
Servent
Higher Education
Professional
Machine-op Inspct
craft repair
protective service
other service
transport moving
Armed-forces
Farming-fishing
Handles cleaners
Priv-house serv
Prof-specialty
Exec managerial
Sales
Adm clerical
tech support


WORKCLASS
any
COARSE & DETAILED
ENTREPRENEUR
GOV
NOT-WORK
Self-emp-not-inc
Private
self-emp-inc
Federal-gov
State-gov
local-gov
without pay
never-worked

APPENDIX II: Generalization states

A:

Coarse

1. K1 – base data
2. K5
   - Transformation:
     i. Age: 1
     ii. Education: 2
     iii. Occupation: 1
     iv. Sex: 0
3. K10
   - Transformation:
     i. Age: 1
     ii. Education: 2
     iii. Occupation: 1
     iv. Sex: 0
4. K15
   - Transformation:
     i. Age: 1
     ii. Education: 2
     iii. Occupation: 1
     iv. Sex: 0
5. K20
   - Transformation:
     i. Age: 1
     ii. Education: 2
     iii. Occupation: 1
     iv. Sex: 0
6. K25
   - Transformation:
     i. Age: 1
     ii. Education: 2
     iii. Occupation: 1
     iv. Sex: 0
7. K30
   - Transformation:
     i. Age: 1
     ii. Education: 2
     iii. Occupation: 1
     iv. Sex: 0
8. K40
   - Transformation:
     i. Age: 1
     ii. Education: 2
     iii. Occupation: 1
     iv. Sex: 0
9. K50
   - Transformation:
     i. Age: 1
     ii. Education: 2
     iii. Occupation: 1
     iv. Sex: 0
10. K75
    - Transformation:
      i. Age: 1
      ii. Education: 2
      iii. Occupation: 1
      iv. Sex: 0

Detailed

11. K1 – base data
12. K5:
    - Transformation:
      i. Age: 2
      ii. Education: 2
      iii. Occupation: 2
      iv. Sex: 0
13. K10:
    - Transformation:
      i. Age: 2
      ii. Education: 2
      iii. Occupation: 2
      iv. Sex: 0
14. K15:
    - Transformation:
      i. Age: 2
      ii. Education: 2
      iii. Occupation: 2
      iv. Sex: 0
15. K20:
    - Transformation:
      i. Age: 2
      ii. Education: 2
      iii. Occupation: 2
      iv. Sex: 0
16. K25:
    - Transformation:
      i. Age: 2
      ii. Education: 3
      iii. Occupation: 1
      iv. Sex: 0
17. K30:
    - Transformation:
      i. Age: 2
      ii. Education: 2
      iii. Occupation: 2
      iv. Sex: 0
18. K40:

- Transformation:
    i. Age: 2
    ii. Education: 3
    iii. Occupation: 1
    iv. Sex: 0

19. K50:
    - Transformation:
        i. Age: 2
        ii. Education: 1
        iii. Occupation: 3
        iv. Sex: 0
20. K75:
    - Transformation:
        i. Age: 2
        ii. Education: 1
        iii. Occupation: 3
        iv. Sex: 0

B:

Coarse

1. K1 – base data
2. K5
    - Transformation:
        i. Age: 1
        ii. Work class: 2
        iii. Education: 2
        iv. Occupation: 1
        v. Sex: 0
        vi. Hours per week: 2
3. K10
    - Anonymity: Anonymous
    - Transformation:
        i. Age: 1
        ii. Work class: 2
        iii. Education: 2
        iv. Occupation: 1
        v. Sex: 0
        vi. Hours per week: 2
4. K15
    - Anonymity: Anonymous
    - Transformation:
        i. Age: 1
        ii. Work class: 2
        iii. Education: 2
        iv. Occupation: 1
        v. Sex: 0
        vi. Hours per week: 2
5. K20
    - Anonymity: Anonymous
    - Transformation:
        i. Age: 1
        ii. Work class: 2
        iii. Education: 2
        iv. Occupation: 1
        v. Sex: 0
        vi. Hours per week: 2
6. K25
    - Anonymity: Anonymous
    - Transformation:
        i. Age: 1
        ii. Work class: 2
        iii. Education: 2
        iv. Occupation: 1
        v. Sex: 0
        vi. Hours per week: 2
7. K30
    - Anonymity: Anonymous
    - Transformation:
        i. Age: 1
        ii. Work class: 2
        iii. Education: 2
        iv. Occupation: 1
        v. Sex: 0
        vi. Hours per week: 2
8. K40
    - Anonymity: Anonymous
    - Transformation:
        i. Age: 1
        ii. Work class: 2
        iii. Education: 2
        iv. Occupation: 1
        v. Sex: 0
        vi. Hours per week: 2
9. K50
    - Anonymity: Anonymous
    - Transformation:
        i. Age: 1
        ii. Work class: 2
        iii. Education: 2
        iv. Occupation: 1
        v. Sex: 0
        vi. Hours per week: 2
10. K75
    - Anonymity: Anonymous
    - Transformation:
        i. Age: 1
        ii. Work class: 2
        iii. Education: 2
        iv. Occupation: 1
        v. Sex: 0
        vi. Hours per week: 2

Detailed

11. K1 – base data
12. K5:
    - Anonymity: Anonymous
    - Transformation:
        i. Age: 2
        ii. Work class: 2
        iii. Education: 2
        iv. Occupation: 2
        v. Sex: 0
        vi. Hours per week: 4
13. K10:

- Anonymity: Anonymous
- Transformation:
    i. Age: 2
    ii. Work class: 2
    iii. Education: 2
    iv. Occupation: 2
    v. Sex: 0
    vi. Hours per week: 4

14. K15:
    - Anonymity: Anonymous
    - Transformation:
        i. Age: 2
        ii. Work class: 2
        iii. Education: 2
        iv. Occupation: 2
        v. Sex: 0
        vi. Hours per week: 4
15. K20:
    - Anonymity: Anonymous
    - Transformation:
        i. Age: 2
        ii. Work class: 2
        iii. Education: 2
        iv. Occupation: 2
        v. Sex: 0
        vi. Hours per week: 4
16. K25:
    - Anonymity: Anonymous
    - Transformation:
        i. Age: 2
        ii. Work class: 2
        iii. Education: 3
        iv. Occupation: 1
        v. Sex: 0
        vi. Hours per week: 4
17. K30:
    - Anonymity: Anonymous
    - Transformation:
        i. Age: 2
        ii. Work class: 2
        iii. Education: 2
        iv. Occupation: 2
        v. Sex: 0
        vi. Hours per week: 4
18. K40:
    - Anonymity: Anonymous
    - Transformation:
        i. Age: 2
        ii. Work class: 2
        iii. Education: 3
        iv. Occupation: 1
        v. Sex: 0
        vi. Hours per week: 4
19. K50:
    - Anonymity: Anonymous
    - Transformation:
        i. Age: 2
        ii. Work class: 2
        iii. Education: 1
        iv. Occupation: 3
        v. Sex: 0
        vi. Hours per week: 4
    -
20. K75:
    - Anonymity: Anonymous
    - Transformation:
        i. Age: 2
        ii. Work class: 2
        iii. Education: 3
        iv. Occupation: 1
        v. Sex: 0
        vi. Hours per week: 4

APPENDIX III: Pseudo code algorithm task 6

```
function calculateUniqueness(data, attributes, QIDs, title.
subTitleQID, subtitleAll)
create figure
set figure title to “title”

extract allCounts by grouping data[QIDs] by QIDs with as_index as
False by  calculating the size of each group
create subplotQID in figure
create histogram in subplotQID using allCounts
set title of subplotQID to “subTitleQID”
set xlabel of subplotQID to “occurrences”
set ylabel of subplotQID to “the % of records”

extract allCounts by grouping data[attributes] by attributes with
as_index as False by  calculating the size of each group
create subplotAll in figure
create histogram in subplotAll using allCounts
set title of subplotAll to “subTitleAll”
set xlabel of subplotAll to “occurrences”
set ylabel of subplotAll to “the % of records”

call calculateUniqueness with originalData, originalAttributes,
QID_A, "Uniqueness in the original data", "a: Uniqueness, looking
at Quasi attributes, of scenario A-Detailed", "b: Uniqueness,
looking at all attributes" as inputs

call calculateUniqueness with data_k10a, allAttributes, QID_A,
"Uniqueness in experiment A-Detailed with k-10", "c: Uniqueness,
looking at Quasi attributes", "d: Uniqueness, looking at all
attributes" as inputs

call calculateUniqueness with data_k10b, allAttributes, QID_B,
"Uniqueness in experiment B-Detailed with k-10","e: Uniqueness
looking at Quasi attributes", "f: Uniqueness, looking at all
attributes" as inputs
```