Data Anonymization Assures Statistical Properties and Privacy

Noelle Brown, Lu Cheng, Shanqing Gu, Alexandra Norman, and Lizzy Sterling

Abstract

With nearly all parts of our lives online and available as data, it is always challenging to protect data privacy ethically and legally when we transfer individual information across a boundary for public research purposes. Data anonymization is a process to make original data subjects no longer "directly or indirectly" identifiable. Different levels of anonymization should be considered for different scenarios, such as the power of re-identification and fragility of anonymization. Cautions should be taken for some anonymization techniques, like adding random variation to indirect identifiers, which can distort the data and impact the statistical analysis. By using openly available datasets, our team anonymized two datasets that include both categorical and numerical attributes without changing the statistics of the overall dataset. Two data anonymization tools (sdcMicro and ARX) were applied to help understand how much the data needs to be modified for true anonymity while still allowing for the analysis to yield results that were within the 95% confidence interval of the original data.

Key Words: anonymization, privacy, statistical properties, k-anonymity, i-diversity

I. INTRODUCTION

IN this data-driven world, it is always challenging to protect data privacy ethically and legally when we transfer individual information across a boundary for public research purposes. For example, Facebook recently suffered embarrassing data breach for 87 million users in this April. It is estimated that 63% of US inhabitants are likely to be uniquely identifiable only by quasi identifiers like birthday, sex, zip.

Data anonymization is a process to make original data subjects no longer "directly or indirectly" identifiable. For fully anonymized datasets, the data linkages to the original datasets can be irreversibly destroyed, but still reversibly in pseudonymized datasets. Different levels of anonymization should be considered for different scenarios, such as the power of re-identification and fragility of anonymization. Cautions should be taken for some anonymization techniques, like adding random variation to indirect identifiers, which can distort the data and impact the statistical analysis.

MSDS Program at Southern Methodist University, Dallas, TX 75275 USA
Noelle Brown (noelleb@smu.edu), Lu Cheng (lucheng@smu.edu), Shanqing Gu (shanqingg@smu.edu), Alexandra Norman (abnorman@smu.edu), and Lizzy Sterling (lsterling@smu.edu)

A. K-anonymization

K-anonymization is a standard for protecting private published data by data anonymization. The concept of k-anonymity was first introduced by Latanya Sweeney and Pierangela Samarati in a paper published in 1998. For k-anonymity to be achieved, there need to be at least k individuals in the dataset who share the set of attributes that might become identifying for each individual. A dataset is k-anonymous if every combination of identity-revealing attributes occurs in at least k different rows of the data set. Sensitive data will then be hidden in the crowd after k-anonymization. Generalization and Suppression are two commonly used methods to implement k-anonymization. Generalization is to replace individual attributes with a broader category, while suppression is to replace individual attributes with a (*) symbol.

Although k-anonymization can provide some useful guarantees, it has some limitations. It has no protection against the Background Knowledge and Homogeneity attacks. Homogeneity Attack leverages the case where all the values for a sensitive value within a set of k records are identical, while the Background Knowledge Attack leverages an association between one or more quasi-identifier attributes with the sensitive attribute to reduce the set of possible values for the sensitive attribute. Additionally, the dimensionality of the data must be sufficiently low. If the data is of high dimensionality, such as time series data, it becomes quite hard to ensure the same privacy guarantee as with low dimensional data.

B. I-diversity

One way to safeguard against these attacks in datasets that contain sensitive attributes, such as medical diagnoses or other

information that one would not want to be made public is to ensure that the data fulfills I-diversity. I-diversity is an extension of the k-Anonymity model that maintains diversity of sensitive fields. This model adds the promotion of diversity for sensitive values within that column with at least I distinct values for the sensitive field in each equivalence class. Since maintaining the anonymity of individual identity in k-anonymity may not protect the sensitive attributes, I-diversity prevents against these potential attacks on k-anonymized data with sensitive attributes.

While the k-Anonymity technique may be the most common for data anonymization, a challenge occurs when attempting to anonymize multiple attribute types. When applying k-anonymization to data with high dimensionality and scarcity, there is potential for a high information loss. In order to work around this challenge, one technique is to create a graph-based multifold model. In this model, variable associations are converted into uncertain form and sensitive attributes are made private through a fuzzing technique. The goal of this model is to preserve the sensitivity of sensitive attributes, association disclosures, and identity disclosure. The uncertain graph masks relationships between related attributes and allows for anonymization of multiple attribute types.

In this project, our team will first apply two data anonymization tools (sdcMicro and ARX) to a dataset to see which techniques best anonymize the data while preserving statistical significance within a 95% confidence interval. For our second analysis, we will create a classification model based on another dataset, anonymize this dataset, and see if the original model performs significantly better than the anonymized model. By analyzing both datasets as well as models, we will present a well-rounded report on anonymization tools and techniques.

II. ANALYSIS OF DATA ANONYMIZATION

In this section, we will explore the implementation of data anonymization on real datasets and investigate how data anonymization impacts the analysis results. We will begin by investigating the results on anonymizing the Broom County Employee Earnings dataset with sdcMicro and ARX and then analyze the results of a model created from an anonymized version of a College Score Card dataset.

A. Analysis of Broom County Government Employees Annual Earnings dataset

Part1: sdcMicro

sdcMicro is a flexible, new R package for anonymization of microdata and risk estimation. For categorical variables, sdcMicro uses deterministic procedures (recoding and local suppression) and probabilistic methods (swapping and PRAM). Recoding combines several categories into new, less revealing categories. PRAM stands for Post-RAndomization Method and ensures that each value of the categorical variable is altered to a different value according to a pre-prescribed PRAM matrix. For continuous variables, sdcMicro uses deterministic method micro-aggregation such that the dataset satisfies k-anonymity.

Briefly, sdcMicro can (1) check the impact on information loss (for example, applying k-anonymity on a high-dimensional dataset results in large information loss even for small values of k); (2) perform risk evaluation for direct and indirect identifiers; (3) measures confidence-based k-anonymity; (4) calculates system performance in terms of running time.

1. Data Exploration

The dataset from Broom County lists the annual employees' earnings from in 2009 to 2017[6], which consists of 24575 observations and 10 variables. While government employee salary information is publically available, it is desirable that salary information in the private sect be kept confidential. This dataset provides a good example of an attribute that individuals would not necessarily want to share with others, such as salary. No variables were dropped because of all missing values in this dataset.

For the anonymization steps, there are 4 categorical key variables (Earnings.Year, Department, Position.Title, Regular.or.Temporary, and Full.or.Part.Time), 3 numerical key variables (Regular.Earnings, Overtime.Earnings, Total.Earnings), and 2 identifying variables (Employee.Name, Union.Name) as confidential attributes that were then deleted.

Below is the table of summary statistics for the numerical values in this dataset:

	Earnings Year	Regular Earnings	Overtime Earnings	Total Earnings
count	24575.000000	24575.000000	24575.000000	24575.000000
mean	2012.911129	32007.854281	1519.034378	33526.888659
std	2.593855	22771.776769	3490.212301	24261.086384
min	2009.000000	0.000000	-1427.100000	0.000000
25%	2011.000000	10162.145000	0.000000	10387.415000
50%	2013.000000	32180.020000	0.000000	33240.100000
75%	2015.000000	48060.540000	1093.755000	49785.995000
max	2017.000000	184639.580000	55890.400000	184639.580000

It is interesting to note that the minimum value for the attribute Overtime Earnings is a negative number, possibly suggesting a mistake or an outlier in the data.

2. Anonymization steps (computation time 1.66 minutes)

It may be useful to define a few terms used in the anonymization steps below. First, the "influence" method for performing micro-aggregation sorts observations by the most influential variable in each cluster. The "additive" method of adding noise adds noise completely at random to each variable depending on its size and standard deviation. The steps used to anonymize this dataset are listed below:

1: Microaggregation of numeric variable(s) "Regular.Earnings" (method="influence" and size=3)
2: Establishing 3-anonymity in key variables (with following order of importance: "Regular.or.Temporary", "Full.or.Part.Time", "Position.Title", "Department", "Earnings.Year")
3: Adding stochastic noise to variable(s) "Total.Earnings" (method="additive" and noise=15)
4. Adding stochastic noise to variable(s) "Regular.Earnings" (method="additive" and noise=15)
5. Adding stochastic noise to variable(s) "Overtime.Earnings" (method="additive" and noise=15)
6. Suppress values in variable "Earnings.Year" with individual risk above the threshold of 0.201
7. Suppress values in variable "Department" with individual risk above the threshold of 0.2
8. Suppress values in variable "Position.Title" with individual risk above the threshold of 0.104
9. Suppress values in variable "Regular.or.Temporary" with individual risk above the threshold of 0.04
10. Suppress values in variable "Full.or.Part.Time" with individual risk above the threshold of 0.04

3. Risk Measures

3.1 Risk measures for categorical key variables

For risk measures, 0 observations have a higher risk than the risk in the main part of the data, as compared to 6,138 observations in the original data. Based on the individual re-identification risk, we expect 301.77 re-identifications (1.23%) in the anonymized data set. In the original dataset we expected 6921 (28.16%) re-identifications. The detailed individual risks for anonymized and original data are shown in **Figure 1**.

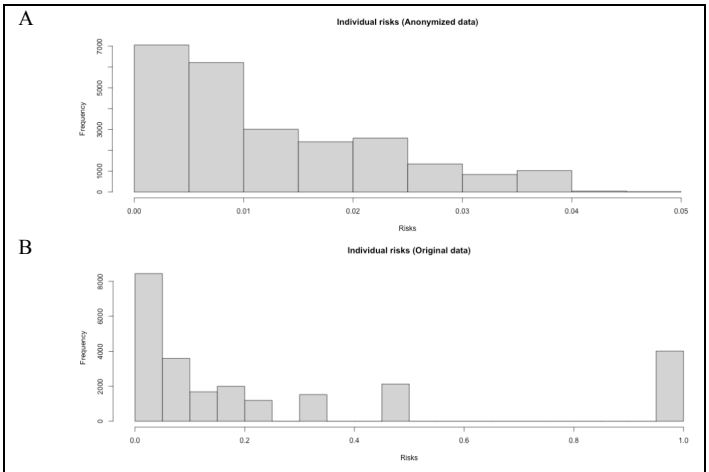


Figure 1: Distribution of individual re-identification risk levels

3.2 Information on k-anonymity

Below, the number of observations violating k-anonymity is shown for the original data and the modified dataset:

k-anonymity	Modified data	Original data
2-anonymity	0 (0.000%)	4010 (16.317%)
3-anonymity	0 (0.000%)	6138 (24.977%)
5-anonymity	0 (0.000%)	8857 (36.041%)

3.3 SUDA2 risk measure

The SUDA algorithm is used to search for Minimum Sample Uniques (MSU) in the data among the sample uniques to determine which sample uniques are also special uniques i.e., have subsets that are also unique. The Minimum Sample Uniques states that the pattern described by a set *M* appears in exactly one row of the dataset and every proper subset of *M* appears in multiple rows of the table. This finds the smallest combination of features that have a unique pattern for a specific record. The smaller the number of attributes contained in a unique pattern, the more risky the data is considered to be.

The table below shows the contribution of each categorical key variable to the SUDA scores. The contribution of a variable is the percentage of the total MSU's in the file that include this variable.

Variable	contribution
Earnings.Year	16.96
Department	8.95
Position.Title	91.32
Regular.or.Temporary	7.00
Full.or.Part.Time	4.71

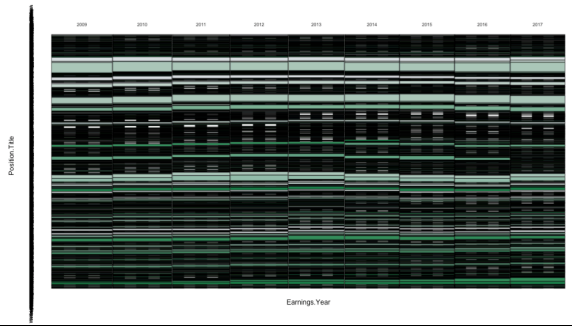
3.4 l-diversity risk measure

From the computation of distinct *l*-diversity of sensitive variable Total.Earnings, 11903 records violate l-diversity in at least one sensible variable. A dataset satisfies *l*-diversity if for every key *k* there are at least *l* different values for each of the sensitive variables. The statistics refer to the value of *l* for each record.

4. Visualizations of original and modified data

sdcMicro provides several different kinds of representations: (1) Graphically, univariate and bivariate mosaic plots can compare categorical variables before and after applying anonymization methods (for example, Position.Title and Earning.Year, as shown in **Figure 2**); (2) Tabularly, univariate and bivariate tabulations of categorical key variables can be used to compare the variables before and after applying anonymization methods; (3) Visually, display information loss based on the number of categories, the mean size of groups, and the size of the smallest category/group in the original and modified variables; (4) Observe the number of observations violating k-anonymity for a specified level of k.

A



B

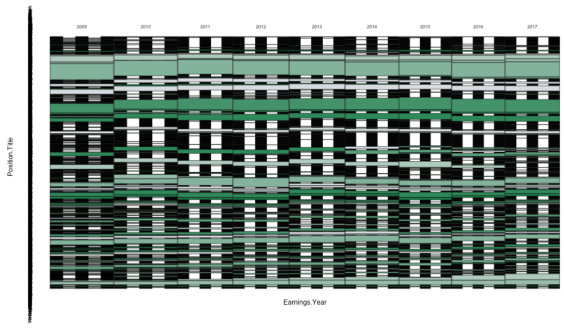


Figure 2: Mosaic plot for variables Position.Title and Earnings.Year before and after applying anonymization

5. Numerical risk measures

5.1 Compare summary statistics of numerical key variables before and after applying anonymization methods

	Correlation	Standard Deviation	Interquartile Range
Regular Earnings	0.989		
Original		22771.777	37898.395
Anonymized		23015.442	37687.618
Overtime Earnings	0.989		
Original		3490.212	1093.755
Anonymized		3528.986	1410.01
Total Earnings	0.989		
Original		24261.086	39398.58
Anonymized		24518.439	39555.426

5.2 Information on risk for numerical key variables

In the original data, the upper bound of the risk-interval is assumed to be 100%, and the disclosure risk in the anonymized dataset is currently between 0% and 17.18%. The larger the deviations from the original data, the lower the upper risk bound will be.

5.3 Information-loss criteria based on numerical key variables

IL1 is the sum of the absolute distances between the corresponding observations in the raw and anonymized datasets, which are standardized by the standard deviation of the variables in the original data.

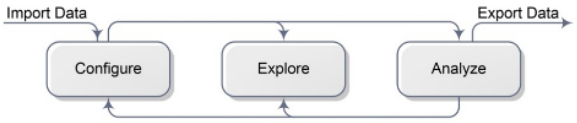
For the continuous variables in the dataset, the IL1s measure is defined as:

$$IL1s = \frac{1}{pn} \sum_{j=1}^{p} \sum_{i=1}^{n} \frac{|x_{ij} - z_{ij}|}{\sqrt{2} S_j}$$

where p is the number of continuous variables; n is the number of records in the dataset; $x_{ij}$and z_{ij}, respectively, are the values before and after anonymization for variable j and individual i; and S_j is the standard deviation of variable j in the original data. The difference in eigenvalues is a comparison of the robust eigenvalues of the data before and after anonymization. Measure IL1s is Inf and the differences of eigenvalues are Inf%.

Part 2: ARX

ARX is a comprehensive open source software for anonymizing sensitive personal data containing numerical and categorical variables. ARX is divided into four perspectives: configuration, exploration, utility analysis, and risk analysis. The workflow is shown below:



Configure: Define transformation, privacy, and coding models
Explore: Filter and analyze the solution space, organize transforamtions
Analyze: Compare and analyze input and output regrading risks and utility

I. Configuration

Configuration in ARX begins with determining which attributes need transformations by assigning them attribute types and how you are going to transform them. The supported attribute types are: Identifying attributes, Quasi-identifying attributes, Sensitive attributes, and Insensitive attributes. In this project, Employee.Name and Union.Name are considered Identifying attributes to an individual recognition. Other attributes are Quasi-identifying because they could identify an individual with attribute linkages. We had no sensitive attributes because there are no attributes that an individual would not want someone to know about them, such as a medical diagnosis. We also did not have any insensitive attributes. During this step we also need to make sure that the attribute is of the correct data type. The default is 'string', so we only needed to change Earnings.Year to type 'integer' and all of the Earnings categories (Regular, Overtime, and Total) to type 'Decimal'.

After choosing our attribute types, we need to delete any identifying attributes and transform any quasi-identifying attributes as shown in table below. There are four different types of generaliazation hierarchies: masking-based, interval-based, order-based, and date-based as well as the option for micro-aggregation. Generalization is typically used with ordinal data and non-interval data. Microaggregation, on the other hand, is typically used with numeric data where there is a wide range of values.

Attribute	Date Type	Date Transformation
Earnings Year	Integer	Transformation: Generalization; Level-0
Department	String	Transformation: Generalization; Level-0, -1
Position Title	String	Transformation: Generalization; Level-0, -1
Regular or Temp	String	Transformation: Generalization; Level-0 (R or T)
Full or Part time	String	Transformation: Generalization; Level-0 (F or P)
Regular Earnings	Decimal	Transformation: Micro-aggregation; Function: Arithmetic mean; Level-0, -1, -2, -3
Overtime Earnings	Decimal	Transformation: Micro-aggregation Function: Arithmetic mean; Level-0, -1
Total Earnings	Decimal	Transformation: Micro-aggregation; Function: Arithmetic mean; Level-0, -1, -2, -3

For our privacy model, we are using k-anonymity with 5 folds for quasi-identifiers (I-diversity would be considered if the sensitive attributes are still kept). For general settings and utility measures, the suppression limit and the enabled threshold for pre-computation are both set to 100%.

II. Exploration

During the anonymization process, ARX characterizes a solution space of potential transformations of the input dataset. For each solution candidate, it is determined if risk thresholds are met and data quality is quantified according to the given model. This perspective allows users to browse the result of this process and to select interesting transformations for further analysis.

The current subset solution space is displayed in **Figure 3**. Each node represents a single transformation, which is identified by the generalization levels that it specifies for the quasi-identifiers in the input dataset. Transformations are characterized by three different background colors: green denotes transformations that result in a privacy-preserving dataset, while orange denotes transformations that are optimal regarding to the specified utility measure. The solution space may also be visualized as a list or a set of titles (not shown). When not filtering the solution space, the optimal transformation is [0, 0, 2, 0, 0].

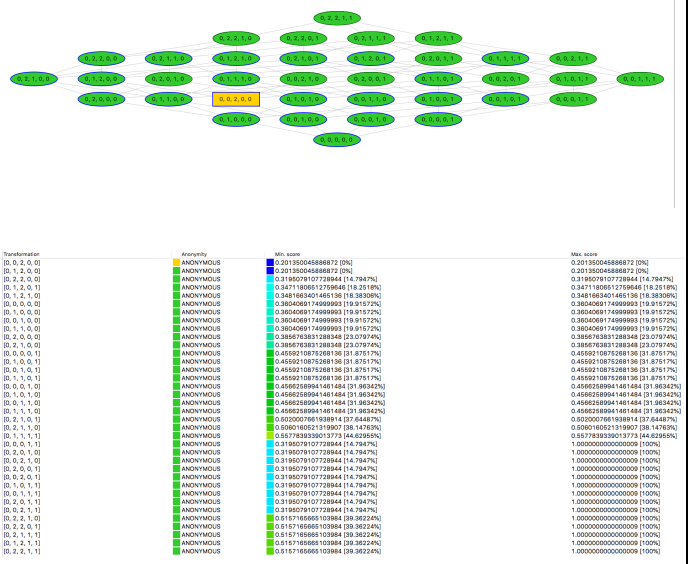


Figure 3. The solution space - each node represents a single transformation. Green denotes a transformation that results in a privacy-preserving dataset, while orange denotes the transformation that is optimal regarding to the specified utility measure.

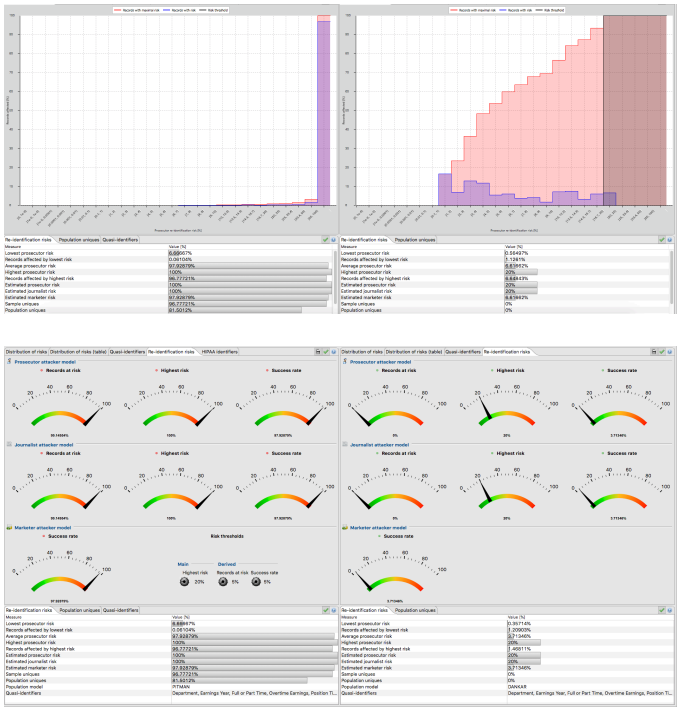


Figure 4. Distribution of risks and redidentification risks for original (left panel) and anonymized data (right panel). The distribution is displayed for both input and output data as a histogram or a table.

III. Utility analysis and risk analysis

For analyzing data from a utility perspective, ARX compares the transformed dataset to the original input dataset in these approaches: (1) comparing input and our output data; (2) giving summary statistics; (3) providing empirical distribution; (4) analyzing contingency; (5) displaying equivalence classes and records; (6) displaying properties of input and output data; (6) analyzing classification performance; (7) presenting data quality models; (8) providing local recording.

For a risk analysis perspective, various privacy risks can be analyzed including distribution of risks and re-identification risks as shown in **Figure 4**. There are three types of risks laid out in this analysis. The first is the risk of a Prosecutor attack. This attack is when the perpetrator targets a specific individual because they already know that that individual has information in the dataset. The second attack is the Journalist attack, where the attacker targets an individual but they may not possess background knowledge about that individual in the dataset. Finally, the Marketer attack is when the attacker attempts to re-identify a large number of individuals. This attack is only deemed successful if a large fraction of the records can be re-identified. We aim to minimize the risk of re-identification through these attacks.

Part 3: Statistical analysis

The statistical analysis will use one numerical attribute (Total Earnings) as an example to help address the question of how much the data needs to be modified for true anonymity while still allowing for the analysis to yield results that are within the 95% confidence interval of the original data.

The exported datasets were analyzed and summarized with Distributions, 95% Confidence Intervals, and p-values (**Figure 5**). The variable "total earnings" in the original dataset follows a bimodal-like distribution. When considering the large sample size (24,575), we will use two sample t-tests due to its robustness to non-normality scenario. The mean estimates with 95% confidence intervals are quite similar for the original and the anonymized data with both the sdcMicro and ARX anonymization processes. In addition, the p-values from the two-sample t-test results are 0.9803 (sdcMicro vs Original) and 0.0762 (ARX vs Original), indicating we cannot reject the null hypothesis that no difference exists between the true mean and the comparison value. We do not have evidence to suggest that the anonymized dataset is significantly different from the original dataset.

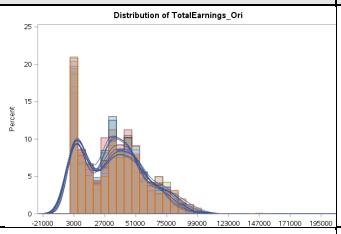
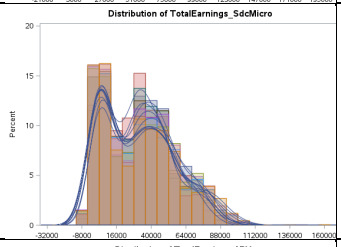
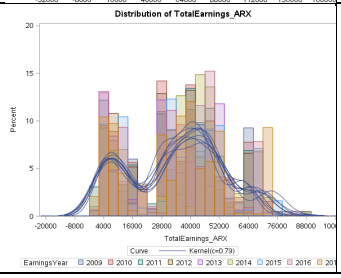
Dataset	Distribution of Total Earnings	95% Confidence Limits Assuming Normality	Two Sample t-test
Original		Mean estimate: 33535 95% CI: (32705, 34366)	Set original data as control group
sdc Micro		Mean estimate: 33289 95% CI: (32329, 34250)	p-value: 0.9803
ARX		Mean estimate: 34008 95% CI: (32705, 34366)	p-value: 0.0762

Figure 5. Statistical analysis for anonymized datasets with sdcMicro and ARX.

B. Analysis of College Score Card dataset

1. Introduction

The College Score Card dataset provides academic and management data from colleges in the United States in 2017. This dataset was chosen because there are several features that individuals would most likely want to be kept private, such as SAT score or family income level. There are several features in this dataset, but for our analysis we will only select 20 features that relate to college admission rate. Our goal is to build a classification model to predict the chance of being admitted to a particular school given other features of the school. Once this model is created, we will create another model based on an anonymized version of this dataset. The goal is to determine whether the models perform statistically different from one another. We will use Python and ARX as development tools in this analysis.

2. Data Description

The raw College Score Card data was cleaned to generate a data frame with 1,205 rows and 20 columns. A list of summary statistics and distributions of variables can be seen in **Figure 6**.

	count	mean	std	min	25%	50%	75%	max
UNITID	1205.0	188809.281328	57613.254233	100654.000000	153278.000000	186584.000000	215929.000000	487524.000000
CONTROL	1205.0	1.585892	0.504432	1.000000	1.000000	2.000000	2.000000	3.000000
ST_FIPS	1205.0	30.312863	15.103053	1.000000	18.000000	31.000000	42.000000	78.000000
REGION	1205.0	4.169295	1.946004	1.000000	3.000000	4.000000	5.000000	9.000000
LOCALE	1205.0	20.548548	9.685606	11.000000	12.000000	21.000000	32.000000	43.000000
CCBASIC	1205.0	18.819917	4.169969	-3.000000	18.000000	19.000000	22.000000	33.000000
CCSIZSET	1205.0	11.608299	3.286830	-2.000000	10.000000	11.000000	14.000000	17.000000
ADM_RATE	1205.0	0.662950	0.186794	0.050400	0.554300	0.680600	0.792800	1.000000
SAT_AVG	1205.0	1048.552697	127.737090	708.000000	969.000000	1034.000000	1109.000000	1507.000000
UGDS	1205.0	5850.506224	7672.533082	45.000000	1286.000000	2633.000000	6994.000000	54091.000000
COSTT4_A	1205.0	32340.424896	13114.782293	7312.000000	21132.000000	30207.000000	41573.000000	67572.000000
TUITIONFEE_IN	1205.0	20563.400000	12404.165436	580.000000	8903.000000	18400.000000	29950.000000	53000.000000
TUITIONFEE_OUT	1205.0	25245.264730	9514.377025	580.000000	17956.000000	24392.000000	30998.000000	53000.000000
AVGFACSAL	1205.0	7680.965975	2228.453530	1333.000000	6183.000000	7278.000000	8766.000000	21354.000000
PFTFAC	1205.0	0.689996	0.236361	0.040400	0.513100	0.685800	0.922500	1.000000
PCTFLOAN	1205.0	0.590684	0.167625	0.000000	0.478500	0.605200	0.712200	0.996300
DEP_INC_AVG	1205.0	72622.802115	20559.803886	18538.859601	58874.831081	71066.239344	86589.816723	147368.801210
IND_INC_AVG	1205.0	24048.715786	9247.266300	2318.081967	17902.392647	23633.770588	29824.676346	57794.477108
D_PCTPELL_PCTFLOAN	1205.0	6082.738589	7739.863127	48.000000	1374.000000	2798.000000	7374.000000	52671.000000

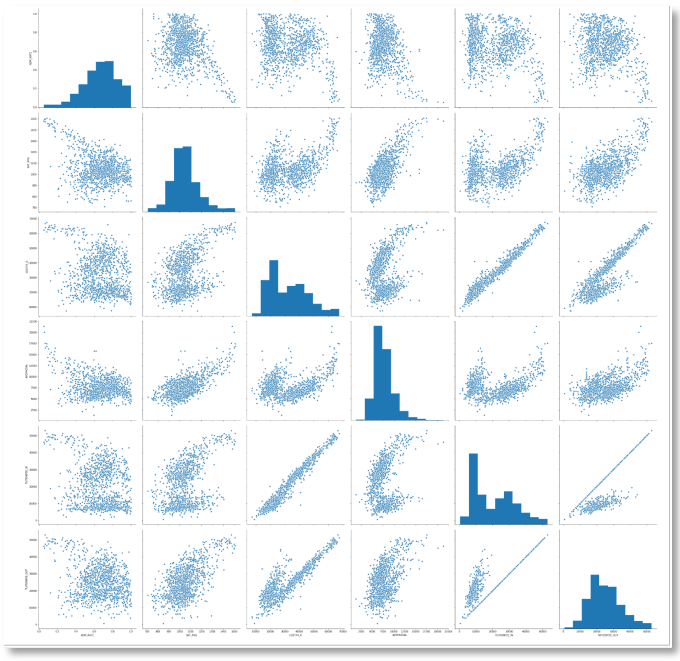


Figure 6. Summary statistics and variable distributions for the College Score Card dataset.

The goal of this model is to predict college admission rates given certain information such as SAT scores and income levels. The model will be evaluated based on its accuracy rate, which is a percentage of how well it predicts whether or not the student will be admitted as compared to the actual data. In order to do this, a binary attribute from the school admission rate variable is created as the **response variable** of our classification model. This variable is defined as follows:

1: College Admission Percentage > 50%
0: College Admission Percentage ≤ 50%

3. Classification Modeling – Original Data

The data will be divided into two sets, a training set and a testing set with split ratio of 80:20. Then, a random forest classification model is created using the training set to fit the model. The test set is then used for prediction and accuracy analysis. The model created from the original data has an accuracy percentage of 85.47%, indicating that around 85% of the predictions match with the actual data. F1 score is also a measure of test's accuracy. It considers both the precision p and the recall r of the test to to compute the score. The accuracy, F1 score and confusion matrix is shown below:

```
Random Forests Accuracy for original data
 0.8547717842323651
Random Forests F1 score for original data
 0.9148418491484185
Random Forests Confusion Matrix for original data
 [[ 18  25]
 [ 10 188]]
```

The calculations of sorted feature weights on this prediction are as follows:

	column name	weight
0	SAT_AVG	0.145219
1	COSTT4_A	0.112661
2	TUITIONFEE_OUT	0.106371
3	DEP_INC_AVG	0.090433
4	PCTFLOAN	0.081873
5	IND_INC_AVG	0.062084
6	TUITIONFEE_IN	0.058356
7	AVGFACSAL	0.054315
8	D_PCTPELL_PCTFLOAN	0.048110
9	UNITID	0.047623
10	UGDS	0.042687
11	PFTFAC	0.031059
12	LOCALE	0.027670
13	ST_FIPS	0.026179
14	CCBASIC	0.022727
15	REGION	0.021340
16	CCSIZSET	0.012943
17	CONTROL	0.008351

It is not surprising that SAT_AVG is the most important attribute in this model. When colleges look at applications, one of the main scores that they look at are standardized testing scores. This signifies that a potential student's SAT score has the largest weight in predicting whether or not they will be admitted to the university in this model.

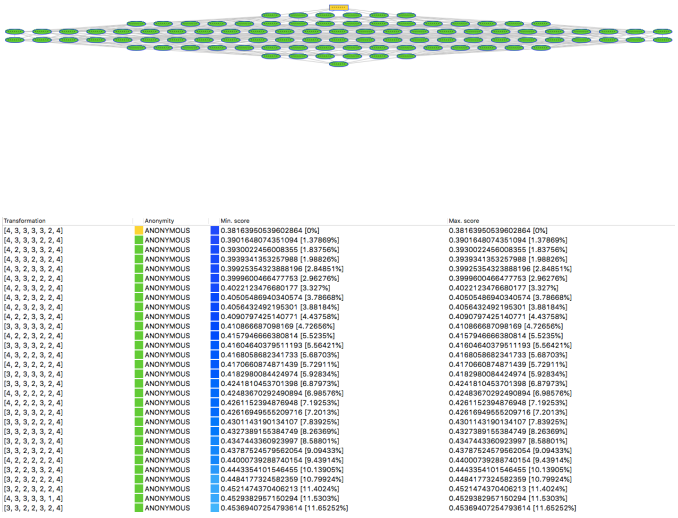
4. Data Anonymization

In this step, seven highly weighted continuous explanatory variables are anonymized using ARX. The first step is to create hierarchies. Since all attributes are continuous, we will use intervals to create a multi-level hierarchy for each attribute (shown in **Figure 7**). In this analysis, we set these seven attributes as quasi-identifiers while the remaining attributes were labeled "insensitive" and therefore did not require additional transformations. These seven attributes are SAT_AVG (average SAT score), COSTT4_A (average annual total cost of attendance), TUITIONFEE_IN (in state tuition and fees), TUITIONFEE_OUT (out of state tuition and fees), AVGFACSAL (average faculty salary), DEP_INC_AVG (average family income of dependent students), and IND_INC_AVG (average family income of independent students). Each of these variables could lead to identification when combined with other attributes, so they will be anonymized.

Level-0	Level-1	Level-2	Level-3	Level-4	Level-5
18538.859601	[15000, 20000[	[15000, 30000[	[15000, 60000[	[15000, 120000[	[15000, 150000[
20781.158718	[20000, 30000[	[15000, 30000[	[15000, 60000[	[15000, 120000[	[15000, 150000[
22177.332762	[20000, 30000[	[15000, 30000[	[15000, 60000[	[15000, 120000[	[15000, 150000[
22189.883343	[20000, 30000[	[15000, 30000[	[15000, 60000[	[15000, 120000[	[15000, 150000[
22992.828029	[20000, 30000[	[15000, 30000[	[15000, 60000[	[15000, 120000[	[15000, 150000[
23318.498	[20000, 30000[	[15000, 30000[	[15000, 60000[	[15000, 120000[	[15000, 150000[
23941.73428	[20000, 30000[	[15000, 30000[	[15000, 60000[	[15000, 120000[	[15000, 150000[
24674.578431	[20000, 30000[	[15000, 30000[	[15000, 60000[	[15000, 120000[	[15000, 150000[
25904.696138	[20000, 30000[	[15000, 30000[	[15000, 60000[	[15000, 120000[	[15000, 150000[
25946.392744	[20000, 30000[	[15000, 30000[	[15000, 60000[	[15000, 120000[	[15000, 150000[
26708.371681	[20000, 30000[	[15000, 30000[	[15000, 60000[	[15000, 120000[	[15000, 150000[
27397.888889	[20000, 30000[	[15000, 30000[	[15000, 60000[	[15000, 120000[	[15000, 150000[
27477.531979	[20000, 30000[	[15000, 30000[	[15000, 60000[	[15000, 120000[	[15000, 150000[
28811.752022	[20000, 30000[	[15000, 30000[	[15000, 60000[	[15000, 120000[	[15000, 150000[
29332.537753	[20000, 30000[	[15000, 30000[	[15000, 60000[	[15000, 120000[	[15000, 150000[
29596.35443	[20000, 30000[	[15000, 30000[	[15000, 60000[	[15000, 120000[	[15000, 150000[
29883.227086	[20000, 30000[	[15000, 30000[	[15000, 60000[	[15000, 120000[	[15000, 150000[
29912.568293	[20000, 30000[	[15000, 30000[	[15000, 60000[	[15000, 120000[	[15000, 150000[
30484.471094	[30000, 40000[	[30000, 60000[	[15000, 60000[	[15000, 120000[	[15000, 150000[
30624.400389	[30000, 40000[	[30000, 60000[	[15000, 60000[	[15000, 120000[	[15000, 150000[
30740.237705	[30000, 40000[	[30000, 60000[	[15000, 60000[	[15000, 120000[	[15000, 150000[
30924.685535	[30000, 40000[	[30000, 60000[	[15000, 60000[	[15000, 120000[	[15000, 150000[
31052.215646	[30000, 40000[	[30000, 60000[	[15000, 60000[	[15000, 120000[	[15000, 150000[
31668.198368	[30000, 40000[	[30000, 60000[	[15000, 60000[	[15000, 120000[	[15000, 150000[

Figure 7. Hierarchies of Attributes

Then, we will set the privacy model and begin the anonymization process. Privacy models are used as criteria to help to choose the best hierarchy from all levels. This ensures that the anonymization process fulfills the requirement and results in a minimal loss of information. This analysis will be run using k-anonymity, k=5. A visualization of the solution space can be seen in **Figure 8**.



Transformation	Anonymity	Min. score	Max. score
[4, 3, 3, 3, 3, 2, 4]	ANONYMOUS	0.38163950539602864 [0%]	0.38163950539602864 [0%]
[4, 3, 3, 3, 2, 2, 4]	ANONYMOUS	0.3901648074351094 [1.37869%]	0.3901648074351094 [1.37869%]
[4, 2, 3, 3, 3, 2, 4]	ANONYMOUS	0.3930022456008355 [1.83756%]	0.3930022456008355 [1.83756%]
[4, 3, 3, 2, 3, 2, 4]	ANONYMOUS	0.3939341353257988 [1.98826%]	0.3939341353257988 [1.98826%]
[4, 3, 2, 3, 3, 2, 4]	ANONYMOUS	0.3992535432388196 [2.84851%]	0.3992535432388196 [2.84851%]
[4, 3, 3, 2, 2, 2, 4]	ANONYMOUS	0.3999600466477753 [2.96276%]	0.3999600466477753 [2.96276%]
[4, 2, 3, 3, 2, 2, 4]	ANONYMOUS	0.4022123476680177 [3.327%]	0.4022123476680177 [3.327%]
[4, 3, 2, 2, 3, 2, 4]	ANONYMOUS	0.4050548894034674 [3.78668%]	0.4050548894034674 [3.78668%]
[4, 2, 3, 2, 3, 2, 4]	ANONYMOUS	0.4056432492195301 [3.88184%]	0.4056432492195301 [3.88184%]
[4, 2, 2, 3, 3, 2, 4]	ANONYMOUS	0.4090797425140771 [4.43758%]	0.4090797425140771 [4.43758%]
[3, 3, 3, 3, 3, 2, 4]	ANONYMOUS	0.4108666870981691 [4.72656%]	0.4108666870981691 [4.72656%]
[4, 2, 2, 3, 2, 2, 4]	ANONYMOUS	0.4157946666380814 [5.5235%]	0.4157946666380814 [5.5235%]
[3, 3, 3, 3, 2, 2, 4]	ANONYMOUS	0.4160464037951193 [5.56421%]	0.4160464037951193 [5.56421%]
[4, 3, 2, 2, 3, 2, 4]	ANONYMOUS	0.4168058682341733 [5.68703%]	0.4168058682341733 [5.68703%]
[4, 2, 3, 2, 2, 2, 4]	ANONYMOUS	0.4170660874871439 [5.72911%]	0.4170660874871439 [5.72911%]
[3, 2, 3, 3, 3, 2, 4]	ANONYMOUS	0.4182980084424974 [5.92834%]	0.4182980084424974 [5.92834%]
[3, 3, 3, 2, 3, 2, 4]	ANONYMOUS	0.4241810453701398 [6.87973%]	0.4241810453701398 [6.87973%]
[4, 3, 2, 2, 2, 2, 4]	ANONYMOUS	0.4248367029249089 [6.98576%]	0.4248367029249089 [6.98576%]
[4, 2, 2, 2, 3, 2, 4]	ANONYMOUS	0.4261152394876948 [7.19253%]	0.4261152394876948 [7.19253%]
[3, 2, 3, 3, 2, 2, 4]	ANONYMOUS	0.4261694955209716 [7.2013%]	0.4261694955209716 [7.2013%]
[3, 3, 3, 2, 2, 2, 4]	ANONYMOUS	0.4301143190134107 [7.83925%]	0.4301143190134107 [7.83925%]
[3, 3, 2, 3, 3, 2, 4]	ANONYMOUS	0.4327389155384749 [8.26369%]	0.4327389155384749 [8.26369%]
[3, 2, 3, 3, 3, 2, 4]	ANONYMOUS	0.4347443360923997 [8.58801%]	0.4347443360923997 [8.58801%]
[3, 3, 2, 3, 2, 2, 4]	ANONYMOUS	0.4378752457962054 [9.09433%]	0.4378752457962054 [9.09433%]
[4, 2, 2, 2, 2, 2, 4]	ANONYMOUS	0.4400073928874015 4 [9.43914%]	0.4400073928874015 4 [9.43914%]
[3, 2, 2, 3, 3, 2, 4]	ANONYMOUS	0.4443354101546455 [10.13905%]	0.4443354101546455 [10.13905%]
[3, 2, 3, 2, 2, 2, 4]	ANONYMOUS	0.4484177324582359 [10.79924%]	0.4484177324582359 [10.79924%]
[3, 2, 2, 3, 2, 2, 4]	ANONYMOUS	0.4521474370406213 [11.4024%]	0.4521474370406213 [11.4024%]
[4, 3, 3, 3, 3, 1, 4]	ANONYMOUS	0.4529382957150294 [11.53039%]	0.4529382957150294 [11.53039%]
[3, 3, 2, 2, 3, 2, 4]	ANONYMOUS	0.4536940725479361 4 [11.65252%]	0.4536940725479361 4 [11.65252%]

Figure 8. The solution space - each node represents a single transformation. Green denotes a transformation that results in a privacy-preserving dataset, while orange denotes the transformation that is optimal regarding to the specified utility measure.

Comparing the original data to the anonymized data, we can visually see the results of the interval anonymization (demonstrated in **Figure 9** for the attribute SAT_AVG).

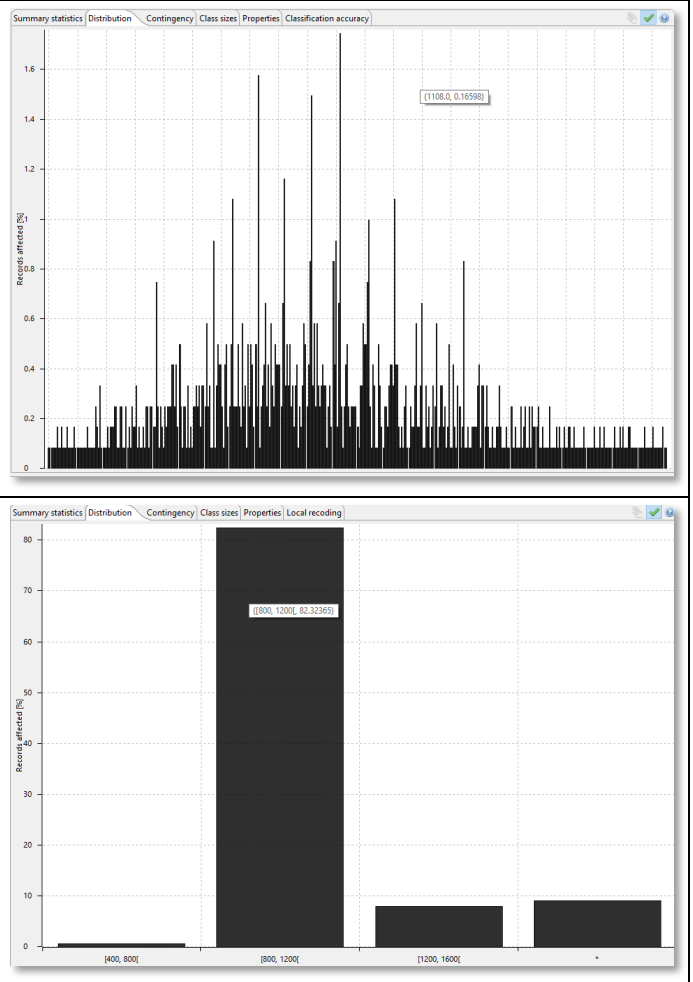


Figure 9. Distribution of the attribute SAT_AVG for the original data (above) and the anonymized data (below).

Checking the outliers, it is determined that the outlier rate is acceptable:

Summary statistics | Distribution | Contingency | Class sizes | Properties | Local recoding

Measure	Including outliers	Excluding outliers
Average class size	20.77586 (1.72414%)	19.21053 (1.75439%)
Maximal class size	110 (9.12863%)	106 (9.68037%)
Minimal class size	5 (0.41494%)	5 (0.45662%)
Number of classes	58	57
Number of records	1205	1095 (90.87137%)
Suppressed records	110 (9.12863%)	0

5. Utility Analysis and Risk Analysis

ARX provides utility and risk analysis for the anonymized data (**Figure 10**). This data's risk perspective shows that certain risk still does exist, although it is much less than the original data. Since the highest re-identification risk is only around 20%, we will proceed with our analysis with caution but note that there is risk involved with this dataset.

Figure 10. Distribution of risks and redidentification risks for original (left) and anonymized data (right).

6. Classification Remodeling – Anonymized Data

After the transformation, the newly anonymized data is exported from ARX. This will now be used to create a classification model with the same procedure as the original model (80:20 training/testing split). The variable types and a summary of the statistics of the anonymized data can be seen below:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1205 entries, 0 to 1204
Data columns (total 29 columns):
UNITID                1205 non-null int64
MAIN                  1205 non-null int64
CONTROL               1205 non-null int64
ST_FIPS               1205 non-null int64
REGION                1205 non-null int64
LOCALE                1205 non-null float64
CCBASIC               1205 non-null float64
CCSIZSET              1205 non-null float64
ADM_RATE              1205 non-null float64
SAT_AVG               1205 non-null object
UGDS                  1205 non-null float64
COSTT4_A              1205 non-null object
TUITIONFEE_IN         1205 non-null object
TUITIONFEE_OUT        1205 non-null object
AVGFACSAL             1205 non-null object
PFTFAC                1205 non-null float64
PCTFLOAN              1205 non-null float64
DEP_INC_AVG           297 non-null float64
IND_INC_AVG           297 non-null float64
OPENADMP              1205 non-null float64
D_PCTPELL_PCTFLOAN    1205 non-null float64
Binary_ADM_RATE       1205 non-null int64
SAT_AVG_NUM           1205 non-null int64
TUITIONFEE_IN_NUM     1205 non-null int64
TUITIONFEE_OUT_NUM    1205 non-null int64
AVGFACSAL_NUM         1205 non-null int64
COSTT4_A_NUM          1205 non-null int64
DEP_INC_AVG_NUM       1205 non-null int64
IND_INC_AVG_NUM       1205 non-null int64
dtypes: float64(11), int64(13), object(5)
memory usage: 273.1+ KB
```

	count	mean	std	min	25%	50%	75%	max
UNITID	1205.0	188809.281328	57613.254233	100654.000000	153278.000000	186584.000000	215929.000000	487524.000000
MAIN	1205.0	0.962656	0.189683	0.000000	1.000000	1.000000	1.000000	1.000000
CONTROL	1205.0	1.585892	0.504432	1.000000	1.000000	2.000000	2.000000	3.000000
ST_FIPS	1205.0	30.312863	15.103053	1.000000	18.000000	31.000000	42.000000	78.000000
REGION	1205.0	4.169295	1.946004	1.000000	3.000000	4.000000	5.000000	9.000000
LOCALE	1205.0	20.548548	9.685606	11.000000	12.000000	21.000000	32.000000	43.000000
CCBASIC	1205.0	18.819917	4.169969	-3.000000	18.000000	19.000000	22.000000	33.000000
CCSIZSET	1205.0	11.608299	3.286830	-2.000000	10.000000	11.000000	14.000000	17.000000
ADM_RATE	1205.0	0.662950	0.186794	0.050400	0.554300	0.680600	0.792800	1.000000
UGDS	1205.0	5850.506224	7672.533082	45.000000	1286.000000	2633.000000	6994.000000	54091.000000
PFTFAC	1205.0	0.689996	0.236361	0.040400	0.513100	0.685800	0.922500	1.000000
PCTFLOAN	1205.0	0.590684	0.167625	0.000000	0.478500	0.605200	0.712200	0.996300
DEP_INC_AVG	297.0	69545.031225	20357.536999	27397.888889	55778.547988	66744.586745	82170.054678	138870.366630
IND_INC_AVG	297.0	22870.538150	8870.485879	3294.250000	17194.739247	22697.640000	28976.715078	48226.638462
OPENADMP	1205.0	1.997510	0.049855	1.000000	2.000000	2.000000	2.000000	2.000000
D_PCTPELL_PCTFLOAN	1205.0	6082.738589	7739.863127	48.000000	1374.000000	2798.000000	7374.000000	52671.000000
Binary_ADM_RATE	1205.0	0.820747	0.383723	0.000000	1.000000	1.000000	1.000000	1.000000
SAT_AVG_NUM	1205.0	2.560996	0.976793	0.000000	3.000000	3.000000	3.000000	3.000000
TUITIONFEE_IN_NUM	1205.0	1.960996	1.210820	0.000000	1.000000	2.000000	3.000000	5.000000
TUITIONFEE_OUT_NUM	1205.0	1.453942	0.939487	0.000000	1.000000	1.000000	2.000000	3.000000
AVGFACSAL_NUM	1205.0	1.507884	0.657928	0.000000	1.000000	2.000000	2.000000	2.000000
COSTT4_A_NUM	1205.0	1.740249	1.237461	0.000000	1.000000	1.000000	2.000000	4.000000
DEP_INC_AVG_NUM	1205.0	1.783402	1.329598	0.000000	1.000000	1.000000	2.000000	5.000000
IND_INC_AVG_NUM	1205.0	0.940249	0.345471	0.000000	1.000000	1.000000	1.000000	2.000000

Once the model is fit and run, we find that the anonymized data results in a random forest model with an accuracy percentage of about 84%. This is also a promising percentage, as initially it seems as though the anonymized model is performing fairly well. The accuracy, F1 score and confusion matrix of the model created with the anonymized data are shown below:

```
Random Forests Accuracy for anonymized data
 0.8381742738589212
Random Forests F1 Score for anonymized data
 0.9037037037037037
Random Forests Confusion Matrix for anonymized data
 [[ 19  24]
 [ 15 183]]
```

7. Model Comparison

From the accuracy scores and confusion matrices, we can see that the original model is slightly better in terms of accuracy and classification than transformed data (accuracy of 85.477% from the original data vs. 83.817% from the anonymized data and F1 score of 0.9148 from the original data vs 0.9037 from the anonymized data). However, we need to run some statistical tests to determine if this difference is significant. The chart below lists the error rate and variance of each model:

	Model Name	Error Rate	Variance
0	Original Data Random Forests	0.145228	0.000103
1	Anonymized Data Random Forests	0.161826	0.000113

The 95% confidence interval of the difference in error rates of both models is (-0.0454, 0.0122). This provides evidence that the difference between original model and anonymized model is not statistically significant. Therefore, we do not have evi-

dence that the original model performs significantly better than the anonymized model when it comes to classifying the admission rate. Even after anonymizing 7 of the attributes, a random forest classification model can still predict the possibility of admission with an accuracy of about 84% and with an F1 score of about 0.9. No evidence supports the claim that the performance of the original model is significantly different than that of the anonymized model.

Additionally, when running a two-sample t-test to compare the predicted values from the original model and the anonymized model, we also find that there is no evidence that the predictions are significantly different (p-value 0.78). A 95% confidence interval for the difference between these predictions is between (-0.05, 0.0667). This is further evidence that the anonymized model can predict admission rates sufficiently well. This is promising for individuals or companies that need to predict or classify information while maintaining privacy and anonymity.

III. Conclusion

In both the dataset comparison (Broom County Government Employees Annual Earnings dataset) as well as the classification model comparison (College Score Card dataset), there was no statistical evidence that the original data or model was significantly different from the anonymized data or model. sdcMicro and ARX are both open-source tools that can be used to anonymize data to satisfy k-anonymity, i-diversity, or other anonymization standards. These tools make it accessible for individuals or companies to preserve individual privacy while allowing for statistical analyses to be run on the data. While this is not a comprehensive report on these tools, the results of our two datasets were promising.

In the future, it would be interesting to continue to use these tools on different datasets to anonymize data and determine whether there is any significant difference between the original data and the anonymized data. We could also use different standards for anonymization to add to the model being able to satisfy k-anonymity and i-diversity. Attempting to de-anonymize an already anonymized dataset would also be a good addition to this analysis. This would inform us of how difficult it really is to re-identify individuals based on our knowledge of anonymization techniques.

Overall, it is imperative to preserve individuals' privacy in data in order to respect those individuals and maintain a high standard for how data is viewed. With nearly every part of our lives made public in the age we live in, it is our duty as data scientists to preserve privacy as much as possible in order to prevent data breaches that reveal sensitive and potentially embarrassing or even harmful information about individuals. The techniques and tools analyzed in this report can aid those attempting to preserve persons' rights to privacy both in the real world and online.

IV. References

[1] "Broome County Annual Employee Earnings: Beginning 2009 | Open Data NY." State of New York, data.ny.gov/Government-Finance/Broome-County-Annual-Employee-Earnings-Beginning-2/jie5-3b37.

[2] "College Scorecard Data." School, collegescorecard.ed.gov/data/.

[3] Morland, William. "K - Anonymity: An Introduction." Engineering Privacy, 7 Apr. 2017, www.privitar.com/listing/k-anonymity-an-introduction.

[4] Nishara, N., and Reeta Pandey. "Enhancing Security in Public Clouds Using Data Anonymization Techniques." International Journal of Computer Applications, vol. 128, no. 1, 2015, pp. 33–36., doi:10.5120/ijca2015906428.

[5] "Privacy Models." ARX – Data Anonymization Tool, arx.deidentifier.org/overview/privacy-criteria/.

[6] Raskhodnikova, Sofya, and Adam Smith. "Lecture 2: K-Anonymity, Composition Attacks, and Differential Privacy." CSE 598A Algorithmic Challenges in Data Privacy. www.cse.psu.edu/~ads22/privacy598/lec-notes/Lec2-CSE598A.pdf.

[7] "Share and Discover Research." ResearchGate, www.researchgate.net/.

[8] Wang, Li E., and Xianxian Li. "A Graph-Based Multifold Model for Anonymizing Data with Attributes of Multiple Types." Egyptian Journal of Medical Human Genetics, Elsevier, 22 Sept. 2017, www.sciencedirect.com/science/article/pii/S016740481730192X

[9] Williams, Ryan, and Manuel Blum. K-Anonymity. 2007, www.cs.cmu.edu/~jblocki/Slides/K-Anonymity.pdf