Data Anonymization Assures Statistical Properties and Privacy

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

Purpose

To anonymize data to allow for true privacy/ anonymity while yielding statistical results that are within a 95% confidence interval of the original data.

K-Anonymization & I-diversity

Common standards for data anonymization

k-Anonymity

- First introduced by Latanya Sweeny & Pierangela Samarati in 1998
- Need at least k individuals in the dataset who share a set of attributes that might be identifying
- Every combination of identity-revealing attributes needs to occur in at least k different rows of the data set
- Sensitive data will then be hidden in the crowd

Common Methods for k-Anonymization

Generalization

- Replace individual attributes with a broader category

Suppression

- Replace individual attributes with (*)
- Can replace all or some of the values in the column

i-Diversity

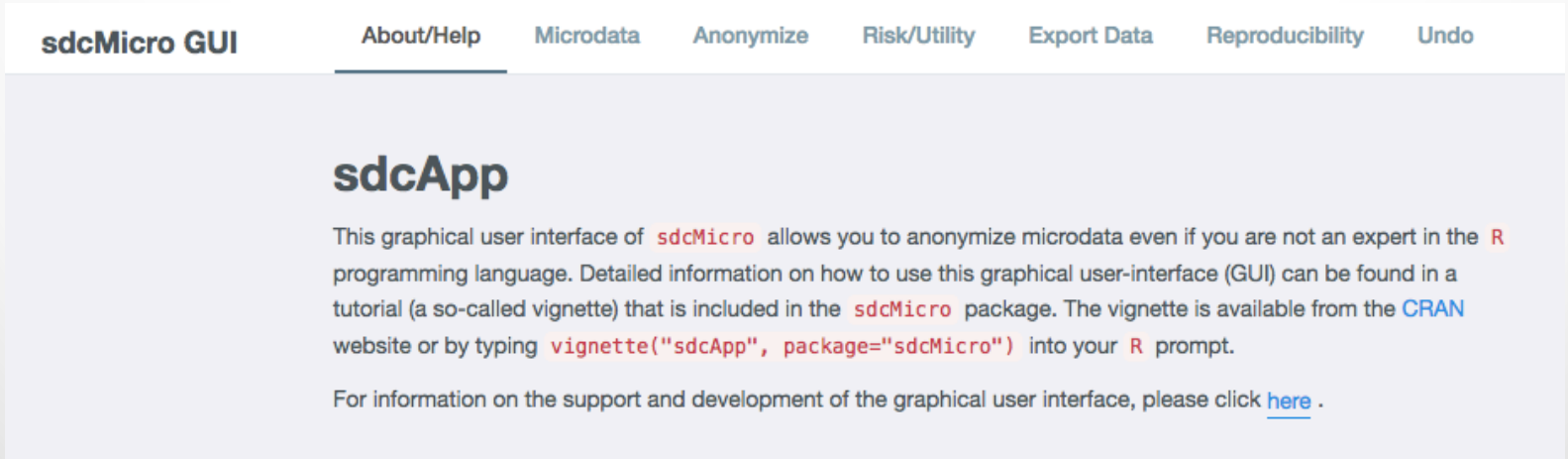
- Extension of the k-Anonymity model that maintains diversity of sensitive fields
- Adds promotion of diversity for sensitive values within that column – there must be at least i distinct values for the sensitive field in each equivalence class
- Maintaining anonymity of individual identity in k-anonymity may not protect the sensitive attributes
- Prevents against potential attacks on k-anonymized data with sensitive attributes

sdcMicro & ARX

Tools used

sdcMicro

- Flexible R package for anonymization of data and risk estimation
- Checks the impact on information loss
- Performs risk evaluation for direct and indirect identifiers
- Measures confidence-based k-anonymity
- Analyzes system performance in terms of running time



sdcMicro Anonymization Methods

Categorical Variables

- Deterministic Methods:
  - Recoding
  - Local suppression
- Probabilistic Methods:
  - Swapping
  - PRAM

Continuous Variables

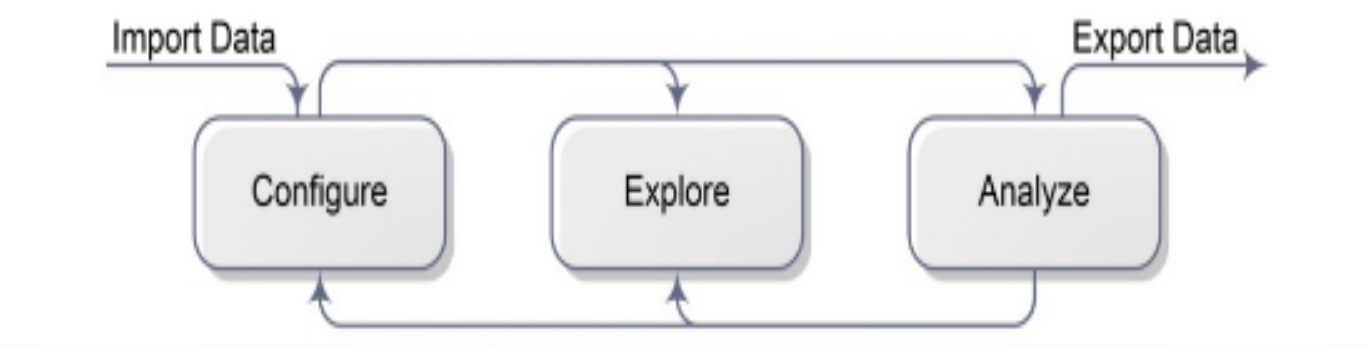
- Deterministic Method:
  - Micro-Aggregation
- Probabilistic Methods:
  - Adding Noise
  - Shuffling

ARX

- Comprehensive open source software for anonymizing sensitive personal data

ARX Workflow

- **Configuration**: Define transformation, privacy & coding models
- **Exploration**: Filter & analyze the solution space; organize transformations
- **Utility Analysis**: Compare & analyze input/output regarding utility
- **Risk Analysis**: Compare & analyze input/output regarding risk



Analysis of Broom County Government Employees' Annual Earnings Dataset: sdcMicro

24575 Observations

10 Variables

Dataset Information

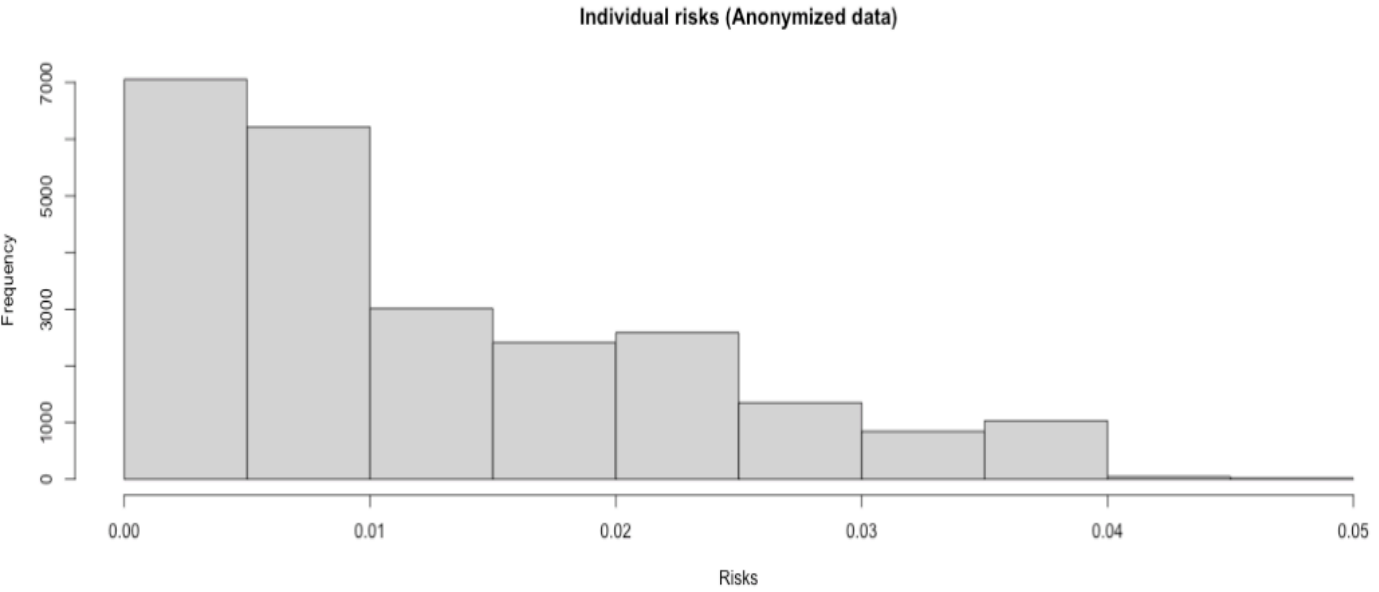
- 4 categorical key variables:
  - Earnings Year, Department, Position Title, Regular or Temporary, and Full or Part Time
- 3 numerical key variables:
  - Regular Earnings, Overtime Earnings, and Total Earnings
- Deleted two confidential/identifying attributes:
  - Employee Name and Union Name

	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

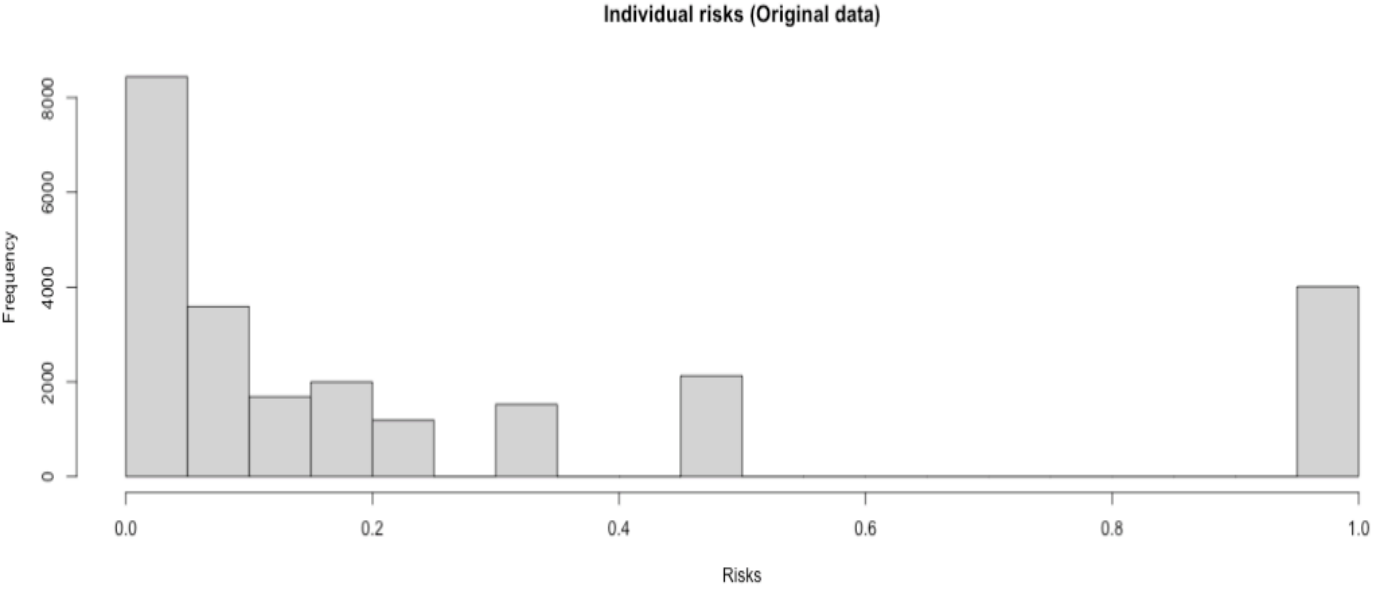
Anonymization Steps
(Computation Time: 1.66 Minutes)

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

Risk Measures



- **301.77** re-identifications (**1.23%**) in the anonymized data set



- **6921** (**28.16%**) re-identifications in the original

k-Anonymity & i-Diversity Risk Measure

Number of observations violating k-anonymity:

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%)

Number of observations violating i-diversity:

11903 records for Total Earnings

SUDA2 Risk Measure

- SUDA algorithm is used to search for Minimum Sample Uniques (MSU)
- Used to determine which are also special uniques (have subsets that are also unique)

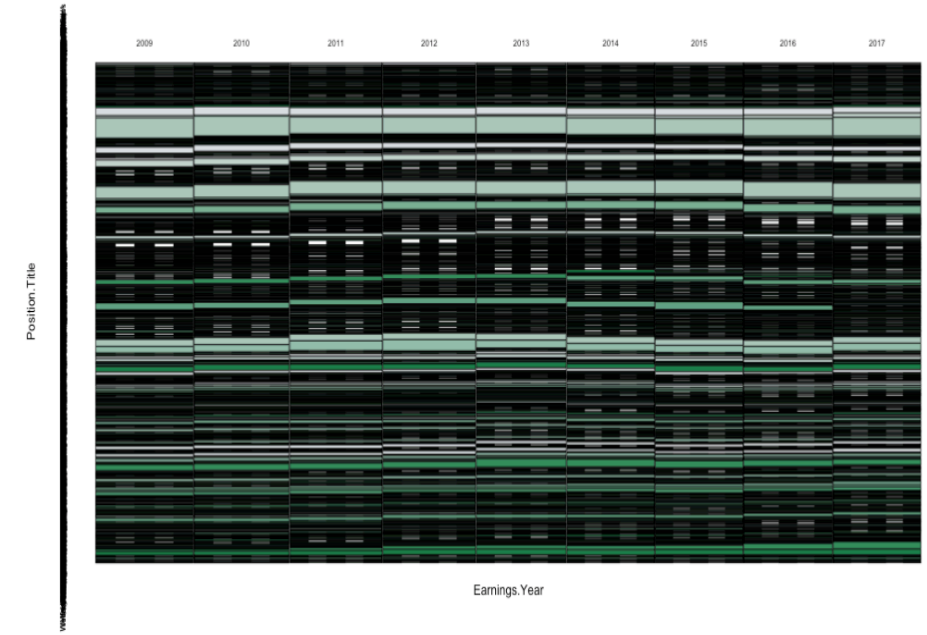
Contribution of each categorical key variable to the SUDA scores:

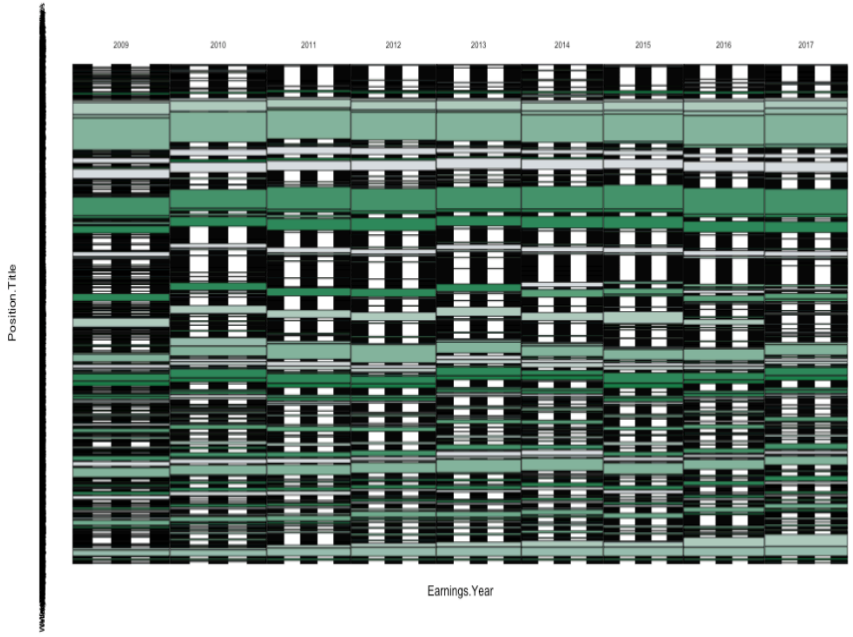
Variable	contribution
Earnings.Year	16.96
Department	8.95
Position.Title	91.32
Regular.or.Temporary	7.00
Full.or.Part.Time	4.71

The contribution of a variable is the percentage of the total MSUs in the file that include this variable

Visualizations

Mosaic plot for variables Position Title and Earning Year before (right) and after (left) applying anonymization





Risk Analysis

- Original data: upper bound of the risk-interval is assumed to be 100%
- Anonymized dataset: between **0%** and **17.18%**
- **Summary statistics before and after anonymization:**

	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

Analysis of Broom County Government Employees' Annual Earnings Dataset: ARX

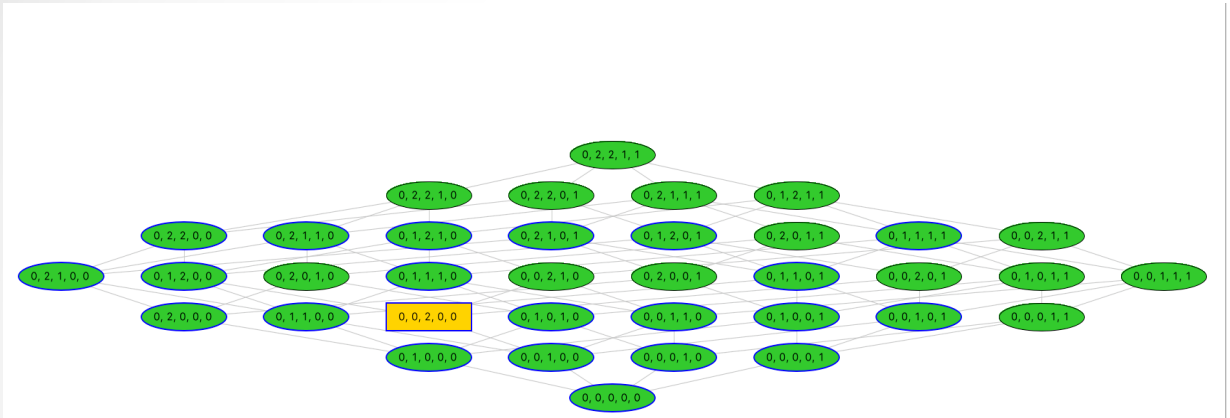
24575 Observations

10 Variables

Configuration

Attribute	Date Type	Date Transforamtion
Earning Year	Integer	Tranformation: Generalization; Level-0
Department	String	Tranformation: Generalization; Level-0, -1
Postion Title	String	Tranformation: Generalization; Level-0, -1
Regular or Temporary	String	Tranformation: Generalization; Level-0 (R or T)
Full or Part time	String	Tranformation: Generalization; Level-0 (F or P)
Regular Earnings	Decimal	Tranformation: Microaggregation;
		Function: Arithmetic mean; Level-0, -1, -2, -3
Overtime Earnings	Decimal	Tranformation: Microaggregation
		Function: Arithmetic mean; Level-0, -1
Total Earnings	Decimal	Tranformation: Microaggregation
		Function: Arithmetic mean; Level-0, -1, -2, -3

Exploration

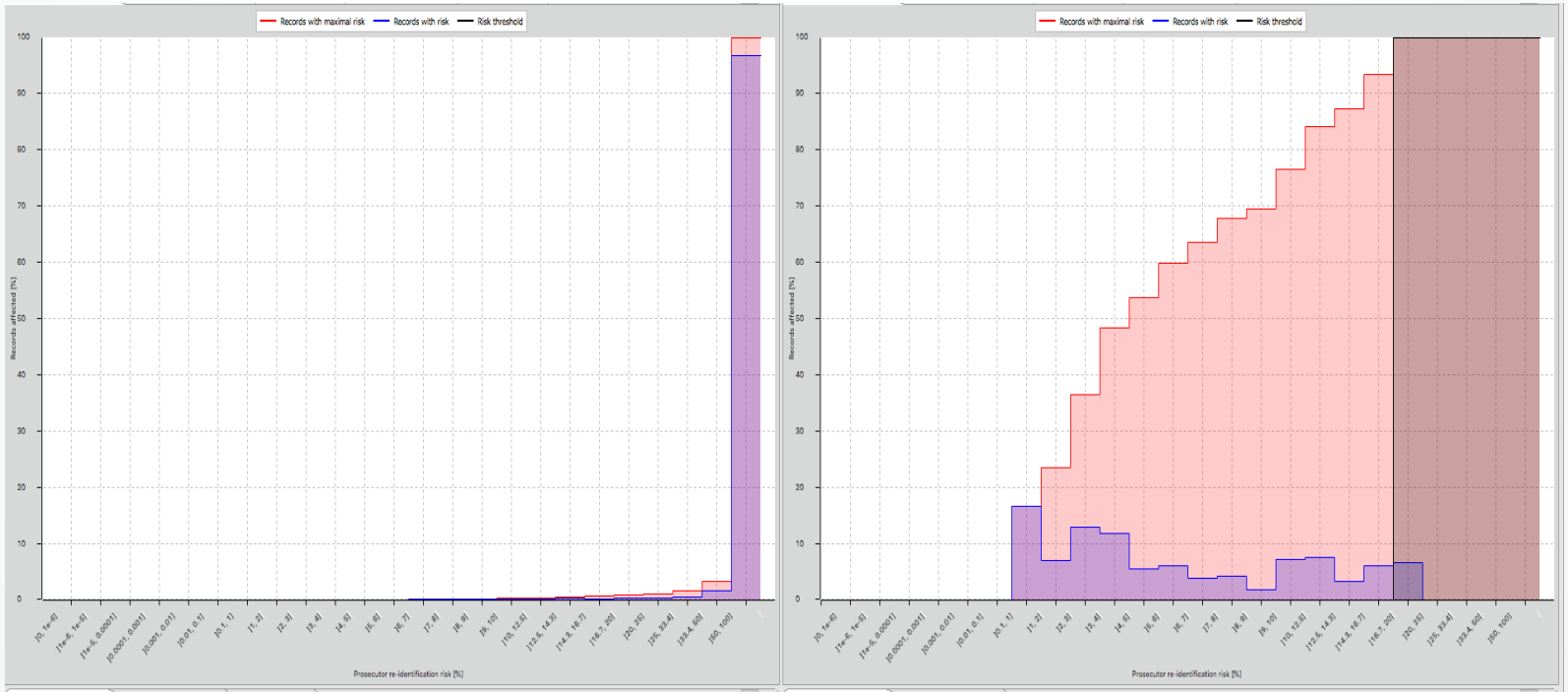


Transformation	Anonymity	Min. score	Max. score
[0, 0, 2, 0, 0]	ANONYMOUS	0.201350045886872 [0%]	0.201350045886872 [0%]
[0, 1, 2, 0, 0]	ANONYMOUS	0.201350045886872 [0%]	0.201350045886872 [0%]
[0, 2, 2, 0, 0]	ANONYMOUS	0.3195079107728944 [14.7947%]	0.3195079107728944 [14.7947%]
[0, 1, 2, 0, 1]	ANONYMOUS	0.34711806512759646 [18.2518%]	0.34711806512759646 [18.2518%]
[0, 1, 2, 1, 0]	ANONYMOUS	0.3481663401465136 [18.38306%]	0.3481663401465136 [18.38306%]
[0, 0, 0, 0, 0]	ANONYMOUS	0.3604069174999993 [19.91572%]	0.3604069174999993 [19.91572%]
[0, 1, 0, 0, 0]	ANONYMOUS	0.3604069174999993 [19.91572%]	0.3604069174999993 [19.91572%]
[0, 0, 1, 0, 0]	ANONYMOUS	0.3604069174999993 [19.91572%]	0.3604069174999993 [19.91572%]
[0, 1, 1, 0, 0]	ANONYMOUS	0.3604069174999993 [19.91572%]	0.3604069174999993 [19.91572%]
[0, 2, 0, 0, 0]	ANONYMOUS	0.3856763831288348 [23.07974%]	0.3856763831288348 [23.07974%]
[0, 2, 1, 0, 0]	ANONYMOUS	0.3856763831288348 [23.07974%]	0.3856763831288348 [23.07974%]
[0, 0, 0, 0, 1]	ANONYMOUS	0.4559210875268136 [31.87517%]	0.4559210875268136 [31.87517%]
[0, 1, 0, 0, 1]	ANONYMOUS	0.4559210875268136 [31.87517%]	0.4559210875268136 [31.87517%]
[0, 0, 1, 0, 1]	ANONYMOUS	0.4559210875268136 [31.87517%]	0.4559210875268136 [31.87517%]
[0, 1, 1, 0, 1]	ANONYMOUS	0.4559210875268136 [31.87517%]	0.4559210875268136 [31.87517%]
[0, 0, 0, 1, 0]	ANONYMOUS	0.4566258994146148 4 [31.96342%]	0.4566258994146148 4 [31.96342%]
[0, 1, 0, 1, 0]	ANONYMOUS	0.4566258994146148 4 [31.96342%]	0.4566258994146148 4 [31.96342%]
[0, 0, 1, 1, 0]	ANONYMOUS	0.4566258994146148 4 [31.96342%]	0.4566258994146148 4 [31.96342%]
[0, 1, 1, 1, 0]	ANONYMOUS	0.4566258994146148 4 [31.96342%]	0.4566258994146148 4 [31.96342%]
[0, 2, 1, 0, 1]	ANONYMOUS	0.5020007661938914 [37.64487%]	0.5020007661938914 [37.64487%]
[0, 2, 1, 1, 0]	ANONYMOUS	0.5060160521319907 [38.14763%]	0.5060160521319907 [38.14763%]
[0, 1, 1, 1, 1]	ANONYMOUS	0.5577839339013773 [44.62955%]	0.5577839339013773 [44.62955%]
[0, 0, 0, 1, 1]	ANONYMOUS	0.3195079107728944 [14.7947%]	1.0000000000000009 [100%]
[0, 2, 0, 1, 0]	ANONYMOUS	0.3195079107728944 [14.7947%]	1.0000000000000009 [100%]
[0, 0, 2, 1, 0]	ANONYMOUS	0.3195079107728944 [14.7947%]	1.0000000000000009 [100%]
[0, 2, 0, 0, 1]	ANONYMOUS	0.3195079107728944 [14.7947%]	1.0000000000000009 [100%]
[0, 0, 2, 0, 1]	ANONYMOUS	0.3195079107728944 [14.7947%]	1.0000000000000009 [100%]
[0, 1, 0, 1, 1]	ANONYMOUS	0.3195079107728944 [14.7947%]	1.0000000000000009 [100%]
[0, 0, 1, 1, 1]	ANONYMOUS	0.3195079107728944 [14.7947%]	1.0000000000000009 [100%]
[0, 2, 0, 1, 1]	ANONYMOUS	0.3195079107728944 [14.7947%]	1.0000000000000009 [100%]
[0, 0, 2, 1, 1]	ANONYMOUS	0.3195079107728944 [14.7947%]	1.0000000000000009 [100%]
[0, 2, 2, 1, 0]	ANONYMOUS	0.5157165665103984 [39.36224%]	1.0000000000000009 [100%]
[0, 2, 2, 0, 1]	ANONYMOUS	0.5157165665103984 [39.36224%]	1.0000000000000009 [100%]
[0, 2, 1, 1, 1]	ANONYMOUS	0.5157165665103984 [39.36224%]	1.0000000000000009 [100%]
[0, 1, 2, 1, 1]	ANONYMOUS	0.5157165665103984 [39.36224%]	1.0000000000000009 [100%]
[0, 2, 2, 1, 1]	ANONYMOUS	0.5157165665103984 [39.36224%]	1.0000000000000009 [100%]

- Each node represents a single transformation
- **Green**: transformation that results in a privacy-preserving dataset
- **Orange**: optimal transformation for utility measure

Utility & Risk Analysis

Distribution of risks and re-identification risks for original (left panel) and anonymized data (right panel):



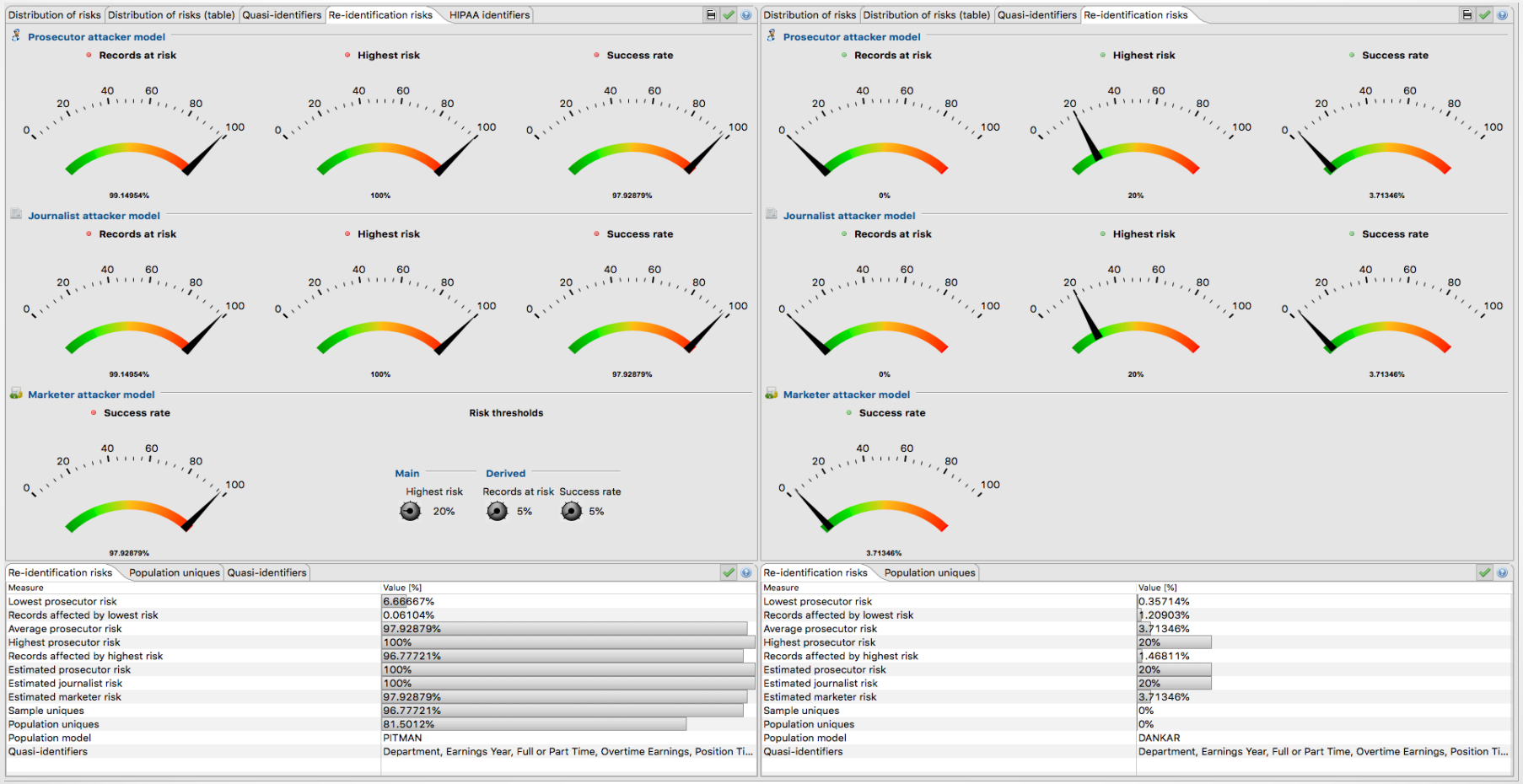
Re-identification risks | Population uniques | Quasi-identifiers

Measure	Value [%]
Lowest prosecutor risk	6.66667%
Records affected by lowest risk	0.06104%
Average prosecutor risk	97.92879%
Highest prosecutor risk	100%
Records affected by highest risk	96.77721%
Estimated prosecutor risk	100%
Estimated journalist risk	100%
Estimated marketer risk	97.92879%
Sample uniques	96.77721%
Population uniques	81.5012%

Re-identification risks | Population uniques

Measure	Value [%]
Lowest prosecutor risk	0.56497%
Records affected by lowest risk	1.1261%
Average prosecutor risk	6.61662%
Highest prosecutor risk	20%
Records affected by highest risk	6.64843%
Estimated prosecutor risk	20%
Estimated journalist risk	20%
Estimated marketer risk	6.61662%
Sample uniques	0%
Population uniques	0%

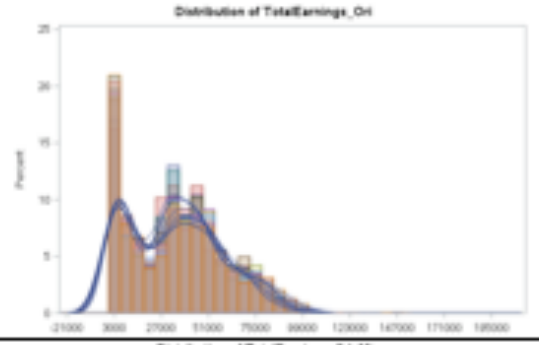
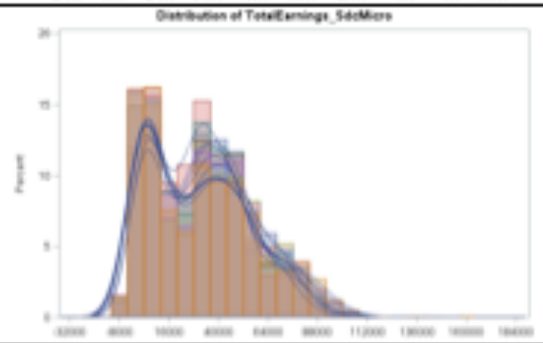
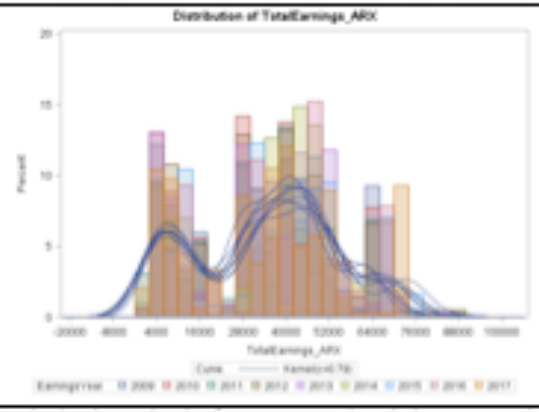
Risk Analysis Continued



Statistical Analysis

- The statistical analysis will use Total Earnings to see if the analysis yields results within a 95% CI of the original.
- P-values:
  - sdcMicro: **0.9803**
  - ARX: **0.0762**
- We do not have any evidence that the anonymized dataset is significantly different from the original dataset.

Statistical Analysis Continued

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

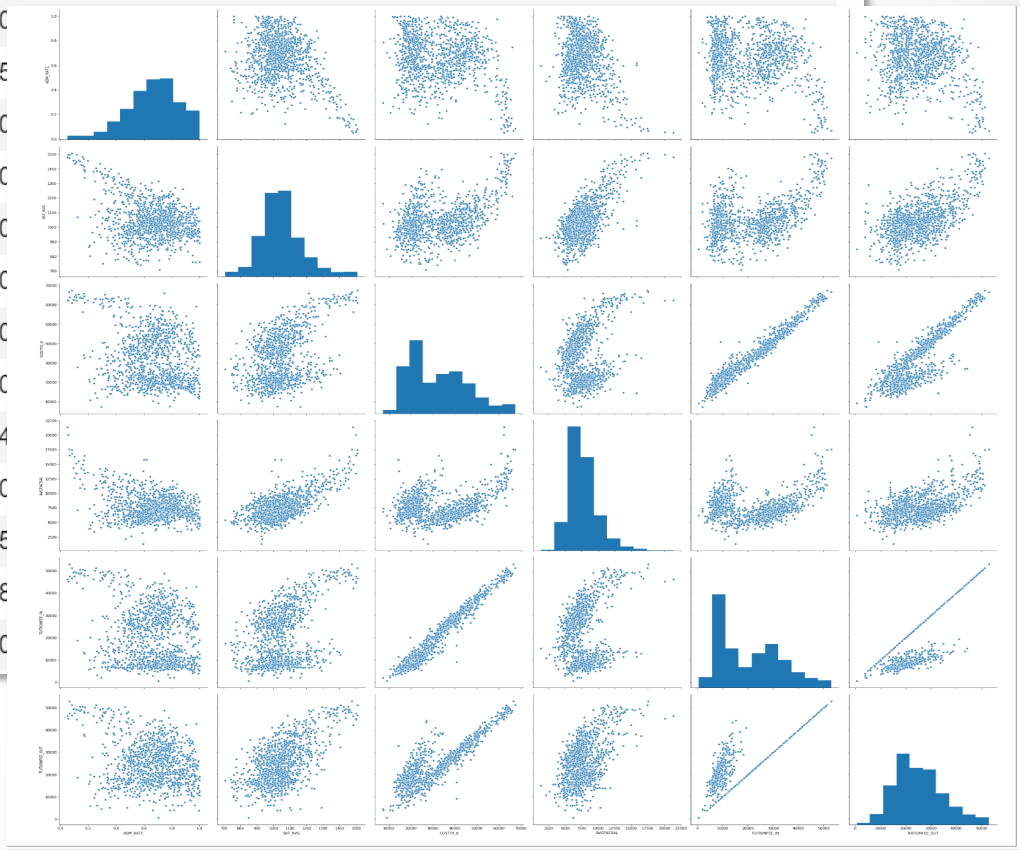
Analysis of College Score Card Dataset

1205 Observations

20 Variables (selected only those that related to college admission rate)

Dataset

	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	-2.00				
ADM_RATE	1205.0	0.662950	0.186794	0.05				
SAT_AVG	1205.0	1048.552697	127.737090	708.00				
UGDS	1205.0	5850.506224	7672.533082	45.00				
COSTT4_A	1205.0	32340.424896	13114.782293	7312.00				
TUITIONFEE_IN	1205.0	20563.400000	12404.165436	580.00				
TUITIONFEE_OUT	1205.0	25245.264730	9514.377025	580.00				
AVGFACSAL	1205.0	7680.965975	2228.453530	1333.00				
PFTFAC	1205.0	0.689996	0.236361	0.04				
PCTFLOAN	1205.0	0.590684	0.167625	0.00				
DEP_INC_AVG	1205.0	72622.802115	20559.803886	18538.85				
IND_INC_AVG	1205.0	24048.715786	9247.266300	2318.08				
D_PCTPELL_PCTFLOAN	1205.0	6082.738589	7739.863127	48.00				



Model

- Random forest model created using an 80/20 training/testing split – accuracy of 85.47%
- Response:

  1: Admission Rate > 50%

  0: Admission Rate ≤ 50%
- Accuracy, confusion matrix, & feature weights:

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

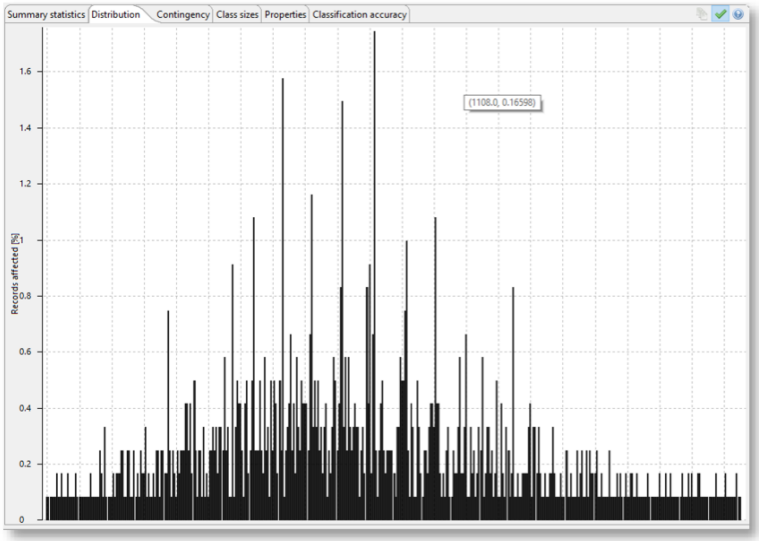
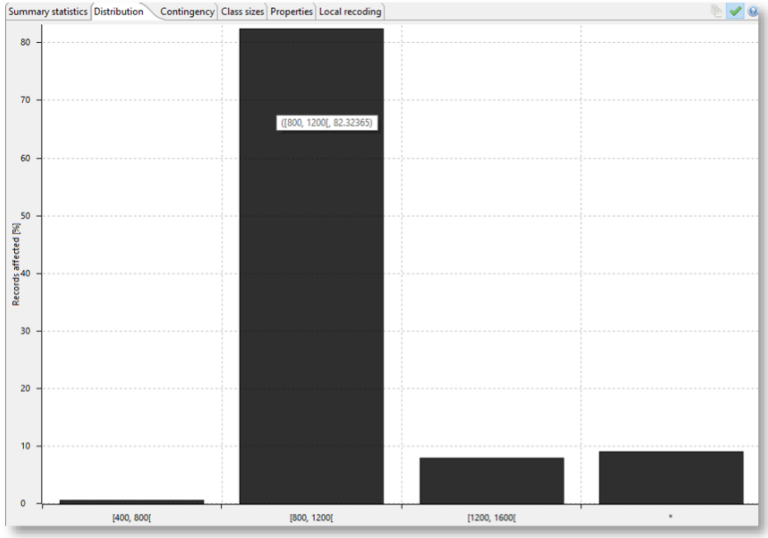
	column name	weight
0	SAT_AVG	0.145219
1	COSTT4_A	0.112661
2	TUITIONFEE_OUT	0.106371
3	DEP_INC_AVG	0.090433
4	PCTFLOAN	0.081873
5	IND_INC_AVG	0.062084
6	TUITIONFEE_IN	0.058356
7	AVGFACSAL	0.054315
8	D_PCTPELL_PCTFLOAN	0.048110
9	UNITID	0.047623
10	UGDS	0.042687
11	PFTFAC	0.031059
12	LOCALE	0.027670
13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

Anonymization

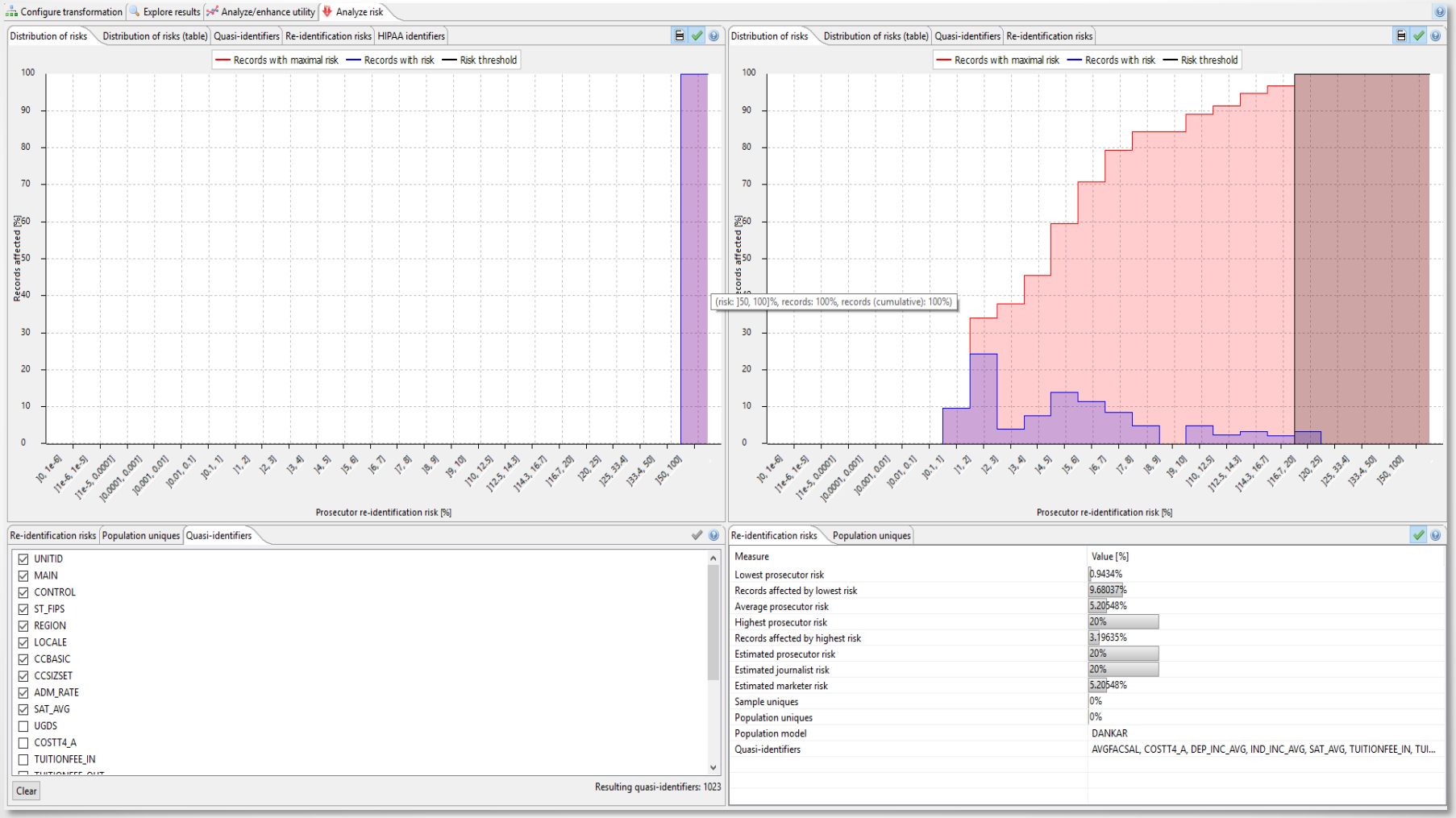
- Intervals used to create multi-level hierarchy for each attributes.
- 5-Anonymity used

- 7 attributes anonymized: SAT_AVG, COSTT4_A, TUITIONFEE_IN, TUITIONFEE_OUT, AVGFACSAL, DEP_INC_AVG, and IND_INC_AVG

Data Comparison

Attr	Original	Transformed
SAT_AVG		

Utility & Risk Analysis



Risk Analysis

Classification – Anonymized Data

	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	1.0		
CONTROL	1205.0	1.585892	0.504432	1.000000	1.000000	2.0		
ST_FIPS	1205.0	30.312863	15.103053	1.000000	18.000000	31.0		
REGION	1205.0	4.169295	1.946004	1.000000	3.000000	4.0		
LOCALE	1205.0	20.548548	9.685606	11.000000	12.000000	21.0		
CCBASIC	1205.0	18.819917	4.169969	-3.000000	18.000000	19.0		
CCSIZSET	1205.0	11.608299	3.286830	-2.000000	10.000000	11.0		
ADM_RATE	1205.0	0.662950	0.186794	0.050400	0.554300	0.6		
UGDS	1205.0	5850.506224	7672.533082	45.000000	1286.000000	2633.0		
PFTFAC	1205.0	0.689996	0.236361	0.040400	0.513100	0.6		
PCTFLOAN	1205.0	0.590684	0.167625	0.000000	0.478500	0.6		
DEP_INC_AVG	297.0	69545.031225	20357.536999	27397.888889	55778.547988	66744.5		
IND_INC_AVG	297.0	22870.538150	8870.485879	3294.250000	17194.739247	22697.6		
OPENADMP	1205.0	1.997510	0.049855	1.000000	2.000000	2.0		
D_PCTPELL_PCTFLOAN	1205.0	6082.738589	7739.863127	48.000000	1374.000000	2798.0		
Binary_ADM_RATE	1205.0	0.820747	0.383723	0.000000	1.000000	1.0		
SAT_AVG_NUM	1205.0	2.560996	0.976793	0.000000	3.000000	3.0		
TUITIONFEE_IN_NUM	1205.0	1.960996	1.210820	0.000000	1.000000	2.0		
TUITIONFEE_OUT_NUM	1205.0	1.453942	0.939487	0.000000	1.000000	1.0		
AVGFACSAL_NUM	1205.0	1.507884	0.657928	0.000000	1.000000	2.0		
COSTT4_A_NUM	1205.0	1.740249	1.237461	0.000000	1.000000	1.0		
DEP_INC_AVG_NUM	1205.0	1.783402	1.329598	0.000000	1.000000	1.0		
IND_INC_AVG_NUM	1205.0	0.940249	0.345471	0.000000	1.000000	1.0		

```
<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
```

Classification – Anonymized Data Continued

- Accuracy and confusion matrix of the new model:

```
Random Forests Accuracy for original data
 0.8381742738589212
Random Forests Confusion Matrix for original data
 [[ 19  24]
 [ 15 183]]
```

Model Comparison

- 95% confidence interval of the difference of error rate: (-0.0314, 0.0231)

	Model Name	Error Rate	Variance
0	Original Data Random Forests	0.132780	0.000096
1	Anonymated Data Random Forests	0.136929	0.000098

- 95% CI for difference in predicted admission rate: (-0.05, 0.0667), p-value: 0.78
- Conclusion: we do not have evidence that the models are significantly different.
- After anonymizing 7 of the attributes, the random forest classification model can still predict the possibility of admission fairly well.

Conclusion

- Both sdcMicro and ARX are good anonymization tools.
- In both of our datasets, there is no evidence that the anonymized dataset is significantly different than the original dataset.