

# Comparative Evaluation of Synthetic Dataset Generation Methods

Ashish Dandekar, Remmy A. M. Zen, Stéphane Bressan

December 12, 2017



# Open Data vs Data Privacy

- ▶ Open Data
  - ▶ Helps crowdsourcing the research problems
  - ▶ Helps in validating new solutions in the real-world setting
  - ▶ Helps increasing reproducibility of the results
- ▶ Data Privacy
  - ▶ Becomes an increasing concern with 'Data Deluge'
  - ▶ Risks the credentials of data owners with public release of the datasets

# Open Data vs Data Privacy

- ▶ Open Data
  - ▶ Helps crowdsourcing the research problems
  - ▶ Helps in validating new solutions in the real-world setting
  - ▶ Helps increasing reproducibility of the results
- ▶ Data Privacy
  - ▶ Becomes an increasing concern with 'Data Deluge'
  - ▶ Risks the credentials of data owners with public release of the datasets

Release realistic but synthetic Data

# Data Privacy and Data Utility

- ▶ Commonplace Techniques
  - ▶ Aggregating attribute categories
  - ▶ Recoding attribute values
  - ▶ Shuffling data across the records
  - ▶ Adding random noise

But what about the utility of the generated data?

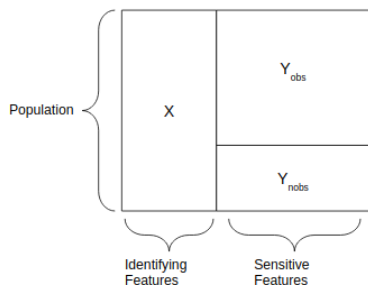


# Data Privacy and Data Utility

- ▶ Data Utility
  - ▶ Overall distribution should be preserved
  - ▶ Relationship among the features should be preserved
  - ▶ Statistical estimands of the features should be preserved

Research Question: How to maintain the utility of synthetically generated data while minimizing the risk of disclosure?

# Notation

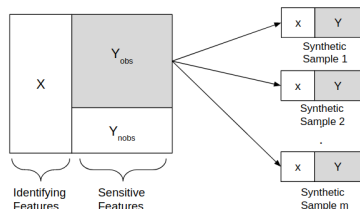


- ▶ Features are divided into two disjoint subsets:
  - ▶ Identifying features
  - ▶ Sensitive features
- ▶ Population is divided into two disjoint subsets:
  - ▶ Observed population
  - ▶ Unobserved population

# Multiple Imputation

- ▶ A three step procedure proposed by Rubin [6] to repopulate missing data
  1. Impute missing values in  $Y_{obs}$ ,  $m$  times using the knowledge of known values  $X$  and  $Y_{obs}$ .
  2. Perform analysis on each of  $m$  individual datasets
  3. Aggregate  $m$  results to get the final result

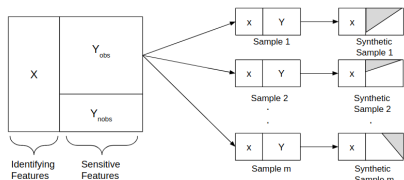
## Fully Synthetic Datasets [7]



- ▶ Consider for a fraction of records values of sensitive variables to be missing
- ▶ Impute missing values using known values
- ▶ Sample  $m$  datasets from population and release them publicly



# Partially Synthetic Datasets [4, 5]



- ▶ Instead of imputing all values of sensitive features, treat those values as missing which have high cost of disclosure
- ▶ For instance: HIV status of a person, annual income of a person above a certain threshold, etc.

# Data Synthesizers [3]

In order to synthetically generate different values, we use different data synthesizers:

- ▶ Linear Regression
- ▶ Decision Trees
- ▶ Random Forest
- ▶ Neural Networks

# Data Synthesizers [3]

## General Synthesis Procedure

We want to generate all features,  $Y_i$ , of a dataset except  $Y_0$

1. Generate values for  $Y_1$  using known  $Y_0$ . Let,  $Y_1^{(syn)}$  denotes synthetically generated values for  $Y_1$ .
2. Generate values for  $Y_2$  using  $Y_0$  and  $Y_1$ . Use  $(Y_0, Y_1)$  to train the data synthesizer and use  $(Y_0, Y_1^{(syn)})$  to generate  $Y_2^{(syn)}$ .
3. Generate values for  $Y_i$  for  $i = 3, 4, \dots$  using  $(Y_0, Y_1, \dots, Y_{i-1})$ .

# Dataset

Attribute Name	Variable Type
House Type	Categorical
Family Size	Ordinal
Sex	Categorical
Age	Ordinal
Marital Status	Categorical
Race	Categorical
Educational Status	Categorical
Employment Status	Categorical
Income	Ordinal
Birth Place	Categorical

- ▶ 1% random sample from US 2001 Census Data<sup>1</sup>
- ▶ Survey data of 316,277 heads of households
- ▶ We synthetically generate values for *Age* and *Income*

---

<sup>1</sup><https://usa.ipums.org/usa/>

# Metrics of Evaluation

## Utility Evaluation

**Distribution level** KL-divergence between histograms of original and synthetically generated values of a feature

**Statistical Estimators** Overlap between confidence intervals of estimators on original and synthetically generated feature data

We have used the estimators of mean and variance of the features [1] for evaluation.

# Metrics of Evaluation

## Disclosure Risk Evaluation [2]

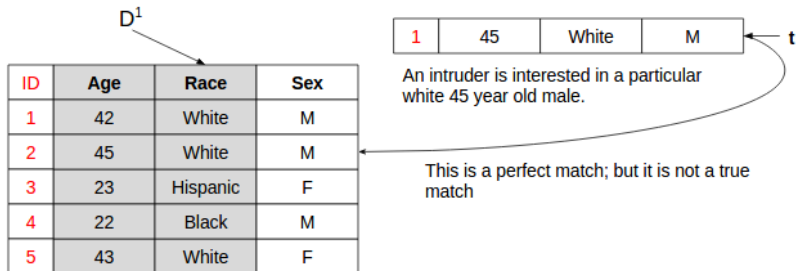
Suppose, an intruder has a vector of information  $\mathbf{t}$  about a certain user who is present in the released dataset. She calculates match probability of  $\mathbf{t}$  with every record  $j \in [1, \dots, s]$  in  $D = (D^1, \dots, D^m)$

$$Pr(J = j | D, \mathbf{t}) = \frac{1}{m} \sum_{i=1}^m \frac{1}{N_{(\mathbf{t}, i)}} \mathbb{I}(Y_j^i = \mathbf{t})$$

where  $N_{(\mathbf{t}, i)}$  is number of documents in  $D^i$  that match with  $\mathbf{t}$ . An intruder selects a record  $j$  with highest probability as the perfectly match.

# Metrics of Evaluation

## Disclosure Risk Evaluation [2]



An instance of the released dataset with Age and Race synthetically generated. (ID is not released)

# Metrics of Evaluation

## Disclosure Risk Evaluation [2]

We perform this experiment over different targets by an intruder.  
*True Match rate* and *False Match Rate* are taken to be the metrics for evaluation.



# Results: Partially Synthetic Data

## Utility Evaluation

Feature	Data Synthesizers	Original Sample Mean	Partially Synthetic Data		
			Synthetic Mean	Overlap	Norm KL Div.
Income	Linear Regression	27112.61	27117.99	0.98	0.54
	Decision Tree	27081.45	27078.93	0.98	<b>0.99</b>
	Random Forest	27107.04	27254.38	0.95	0.58
	Neural Network	27069.95	27370.99	0.81	<b>0.99</b>
Age	Linear Regression	49.83	24.69	0.50	0.55
	Decision Tree	49.83	49.83	0.99	<b>0.99</b>
	Random Forest	49.82	49.74	0.95	0.56
	Neural Network	49.87	49.78	0.90	<b>0.99</b>

We synthetically generate values for records with:

- ▶ Age < 26
- ▶ Income > 70000\$

# Results: Partially Synthetic Data

## Disclosure Risk Evaluation

Consider, an intruder who is interested in people who are born in US and earn more than \$250,000. We consider a tolerance of 2 when matching on the age of a person. *We assume that the intruder knows that the target is present in the publicly released dataset.*

<b>Data Synthesizers</b>	<b>True MR</b>	<b>False MR</b>
Linear Regression	0.06	0.82
Decision Tree	0.18	0.68
Random Forest	0.35	0.50
Neural Network	0.03	0.92

# Results: Partially Synthetic Data



## Efficiency Evaluation

<b>Data Synthesizers</b>	Linear Regression	Decision Tree	Random Forest	Neural Network
<b>Time (s)</b>	0.040	0.048	3.350	0.510

# Conclusion

- ▶ We present our preliminary results in comparative study of synthetic dataset generation techniques using different data synthesizers: namely linear regression, decision tree, random forest and neural network.
- ▶ The analysis shows that neural networks are competitively effective compared to other methods in terms of utility and privacy. They achieve this effectiveness at the cost of running time.

# References I

-  Jörg Drechsler, Stefan Bender, and Susanne Rässler.  
Comparing fully and partially synthetic datasets for statistical disclosure control in the german iab establishment panel.  
*Trans. Data Privacy*, 1(3):105–130, 2008.
-  Jörg Drechsler and Jerome P Reiter.  
Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data.  
*In International Conference on Privacy in Statistical Databases*, pages 227–238. Springer, 2008.

# References II



Jörg Drechsler and Jerome P Reiter.

An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets.

*Computational Statistics & Data Analysis*, 55(12):3232–3243, 2011.



Roderick JA Little.

Statistical analysis of masked data.

*Journal of Official statistics*, 9(2):407, 1993.



Jerome P Reiter.

Using cart to generate partially synthetic public use microdata.

*Journal of Official Statistics*, 21(3):441, 2005.



# References III



Donald B Rubin.

Basic ideas of multiple imputation for nonresponse.

*Survey Methodology*, 12(1):37–47, 1986.



Donald B Rubin.

Discussion statistical disclosure limitation.

*Journal of official Statistics*, 9(2):461, 1993.

# Partially Vs Fully Synthetic Dataset Generation [1]

- ▶ Fully synthetic dataset generation
  - ▶ Guarantees the least disclosure risk
  - ▶ Computationally expensive
  - ▶ Loses some utility due to full data synthesis
- ▶ Partially synthetic dataset generation
  - ▶ Provides good utility as many datapoints are retained without synthesis
  - ▶ Possesses more disclosure risk than fully synthetic datasets
  - ▶ Computationally inexpensive



# Results: Fully Synthetic Data

## Utility Evaluation

Feature	Data Synthesizers	Original Sample Mean	Fully Synthetic Data		
			Synthetic Mean	Overlap	Norm KL Div.
Income	Linear Regression	27112.61	27074.80	0.52	0.55
	Decision Tree	27081.45	27091.02	0.55	0.58
	Random Forest	27107.04	28720.93	0.54	0.64
	Neural Network	27185.26	26694.54	0.54	<b>0.99</b>
Age	Linear Regression	49.83	-192.21	0.50	0.56
	Decision Tree	49.83	49.83	0.56	0.56
	Random Forest	49.82	46.25	0.68	0.57
	Neural Network	49.76	54.32	0.75	<b>0.99</b>

## Disclosure Risk Evaluation

Fully synthetic datasets do not suffer from risk of disclosure since all values are synthetically generated.