Benchmarking Differentially Private Synthetic Data Generation Algorithms

Yuchao Tao[1,3] Ryan McKenna[1,2] Michael Hay[1,4] Ashwin Machanavajjhala[1,3] Gerome Miklau[1,2]

1 Tumult Labs, USA 2 University of Massachusetts Amherst, USA 3 Duke University, USA 4 Colgate University, USA

Yuchao Tao[1,3]

Ryan McKenna[1,2]

Michael Hay[1,4]

Ashwin Machanavajjhala[1,3]

Gerome Miklau[1,2]

Algorithms

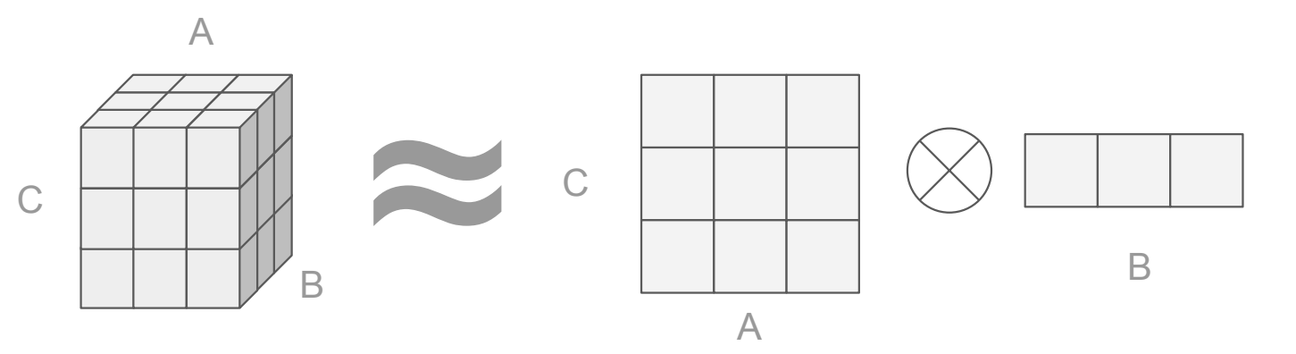
Inclusion Criteria

1. End-to-End DP
2. Tabular Data
3. Selected Publication Venue or Library
4. Publicly Available Source Code
5. No Public Data

Algorithm	Type
MST	Marginal
MWEMPGM	Marginal
PrivBayes	Marginal
DPGAN	GAN
DPCTGAN	GAN
PATEGAN	GAN
PATECTGAN	GAN
FEM	Workload
RAP	Workload
Kamino	Other
RON-GAUSS	Other

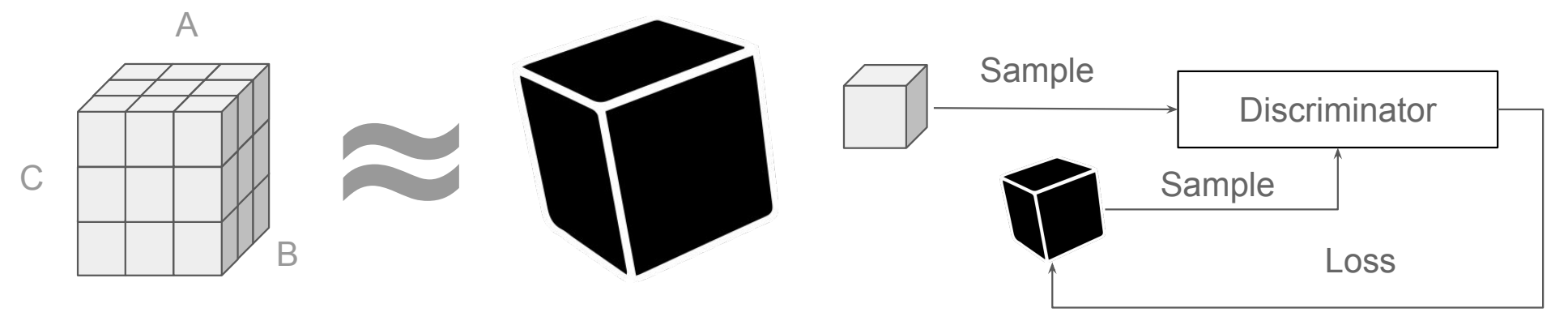
Marginal-based Algorithms

Approximate joint distribution by low-dimensional marginals



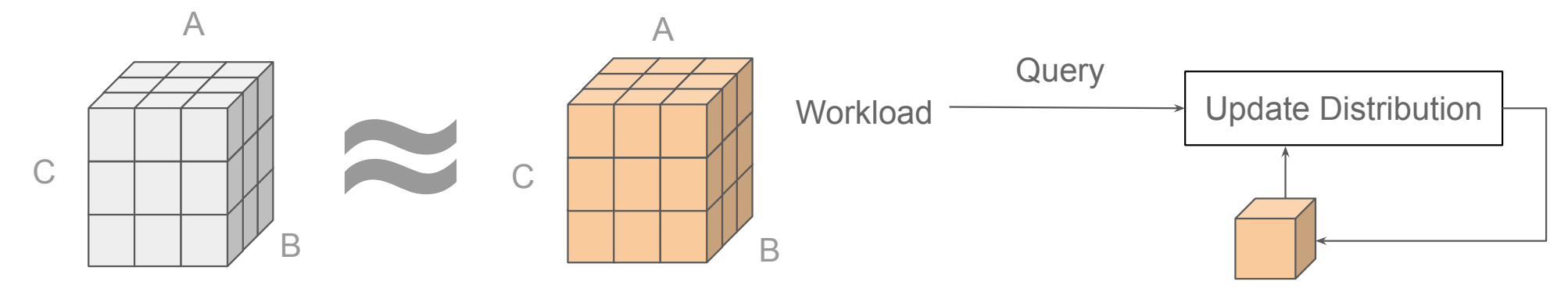
GAN-based Algorithms

Learn data by GAN

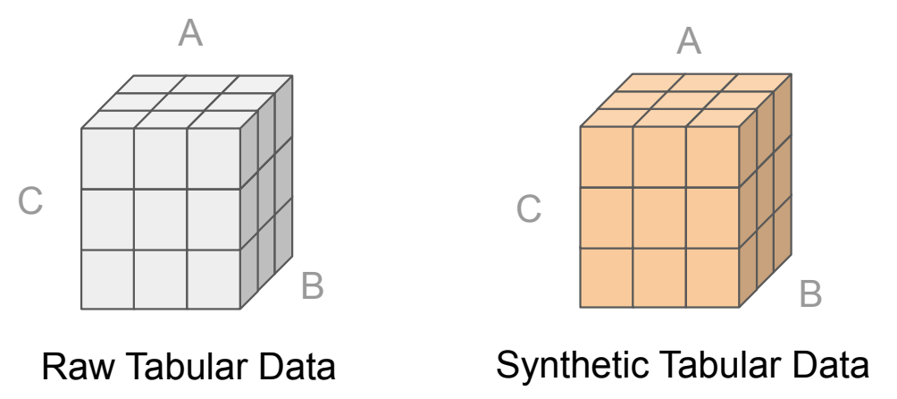


Workload-based Algorithms

Learn joint distribution by queries



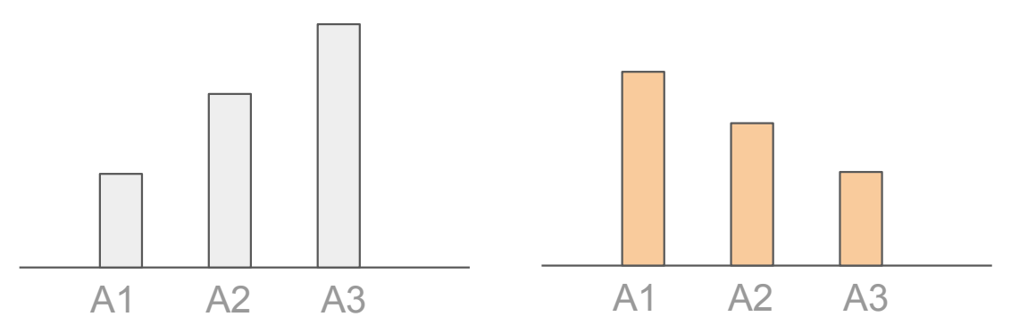
Metrics



Ind

Individual Attribute Distribution Similarity

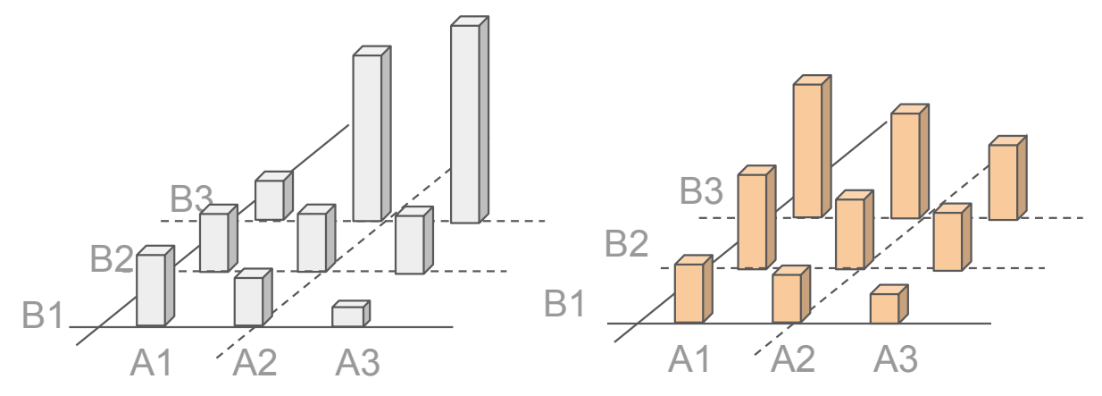
Score: 1-TVD



Pair

Pairwise Attribute Distribution Similarity

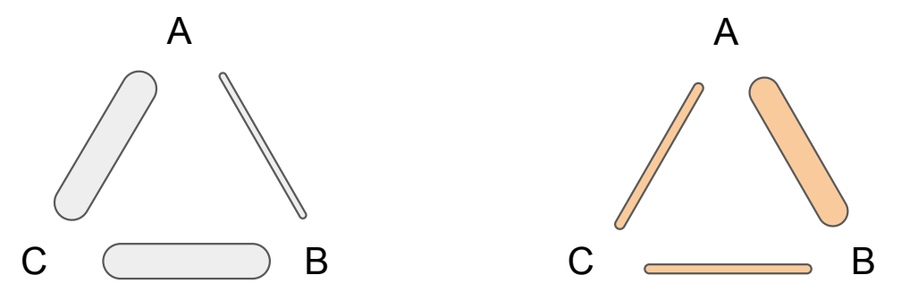
Score: 1-TVD



Corr

Pairwise Correlation Similarity

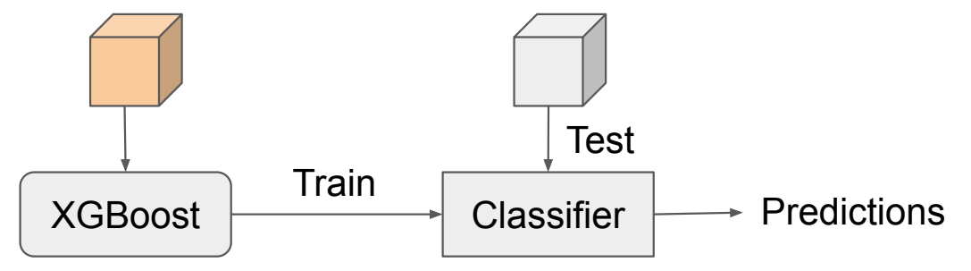
Edge Width: Correlation
Score: Matching Rate



F1

Classification Accuracy

Score: F1 (micro average)



Datasets

Name	Records	Cat.	Numeric	Label
Shopping	12330	9	10	Yes
Adult	32561	9	6	Yes
Bank	45211	13	8	Yes
Census	299285	29	12	Yes
Car	1728	7	0	Yes
Mushroom	8124	23	0	Yes
Scooter	27715	0	5	No

Epsilons

0.1 1.0 10

with delta = 1e-7

Findings

F1: No algorithm dominates.

	Metric (group)				
Mechanism	GT	Ind	Pair	Corr	F1
MST	69%	95%	81%	52%	44%
MWEM-PGM	19%	0%	14%	29%	33%
PrivBayes	9%	0%	0%	19%	17%
Kamino	1%	0%	0%	0%	6%
FEM	0%	0%	0%	0%	0%
RAP	1%	0%	0%	5%	0%
PATECTGAN	4%	5%	5%	5%	0%
DPCTGAN	1%	0%	0%	5%	0%
RonGauss	0%	0%	0%	0%	0%
DPGAN	0%	0%	0%	0%	0%
PATEGAN	0%	0%	0%	0%	0%

Optimal Rate. For a combination of metric, dataset and epsilon, we count an mechanism as optimal if it achieves highest score in average. Here we report the optimal rate stratified by metrics.

F2: Marginal-based approaches are highly ranked

	Metric (group)				
Mechanism	GT	Ind	Pair	Corr	F1
MST	1.56	1.05	1.24	2.00	2.00
MWEM-PGM	2.88	2.76	2.62	3.86	2.17
PrivBayes	4.54	5.43	5.67	3.29	3.67
Kamino	5.26	4.27	4.93	7.87	3.67
FEM	4.91	4.30	4.35	5.95	5.06
RAP	5.94	5.83	5.39	7.17	5.27
PATECTGAN	6.17	6.45	5.90	4.90	7.65
DPCTGAN	6.56	6.84	6.68	5.16	7.75
RonGauss	7.35	7.06	7.11	7.61	7.61
DPGAN	8.46	9.06	9.44	6.78	8.60
PATEGAN	8.99	9.85	9.70	7.05	9.41

Average Rank. For a combination of metric, dataset and epsilon, we rank all the mechanisms by their average score. Here we report the average rank stratified by metrics.

F3: Many algorithms fail to preserve individual attribute distributions.

	Mechanism											
Dataset	Indep..	MST	PrivBa..	MWE..	RAP	Kamino	FEM	DPGAN	PATEC..	RonGa..	DPCT..	PATE..
Adult	0.98	0.98	0.74	0.95	0.70	0.85	0.76	0.59	0.57	0.59	0.59	0.46
Mushroom	0.99	0.99	0.97	0.95	0.88	0.78	0.78	0.70	0.68	0.68	0.67	0.58

Metric “Ind” at epsilon = 1.

F4: Marginal-based algorithms consistently obtain the highest correlation accuracy.

F5: Many algorithms fail to preserve correlations more accurately than independent.

	Mechanism											
Dataset	MST	MWE..	PrivBa..	Indep..	PATEC..	DPCT..	FEM	RonGa..	DPGAN	Kamino	PATE..	RAP
Adult	0.71	0.66	0.60	0.53	0.53	0.50	0.49	0.42	0.35	0.09	0.38	0.32
Mushroom	0.36	0.42	0.15	0.13	0.13	0.13	0.18	0.36	0.40	0.38	0.34	0.34

Metric “Corr” at epsilon = 1. Color indicates below (blue) or above (orange) the baseline, *independent*.

F6: Marginal-based approaches preserve the classification accuracy

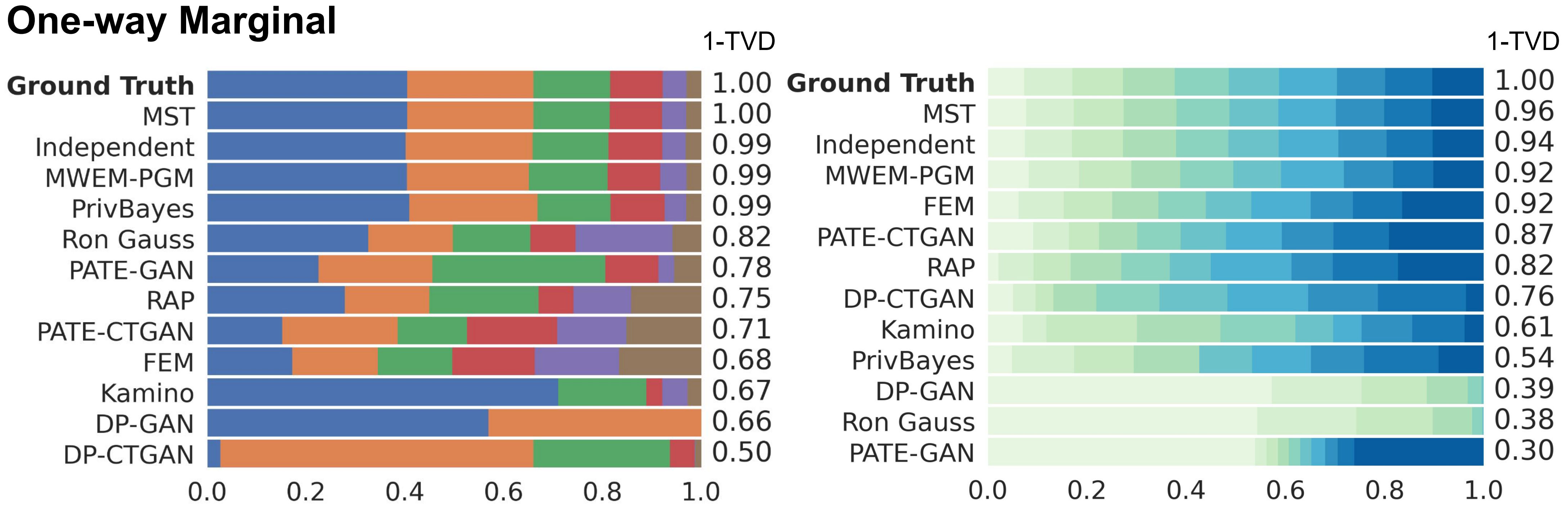
F7: GAN-based approaches fail to preserve the classification accuracy better than a simple majority classifier.

	Mechanism											
Dataset	MST	MWE..	FEM	PrivBa..	RonGa..	Kamino	RAP	DPGAN	Indep..	PATE..	PATEC..	DPCT..
Adult	0.63	0.74	0.44	0.66	0.39	0.66	0.55	0.33	0.45	0.35	0.43	0.39
Mushroom	0.98	0.97	0.90	0.77	0.77	0.76	0.70	0.69	0.50	0.50	0.43	0.36

Metric “F1” at epsilon = 1. Color indicates below (blue) or above (orange) the baseline, *independent*.

Qualitative Analysis: One-way Marginal

- A higher 1-TVD is better.
- Marginal-based algorithms (MST, MWEM-PGM and PrivBayes) accurately preserve the one-way marginal distributions.
- Many algorithms fail to preserve one-way marginal distributions accurately. For example, at the bottom, DP-CTGAN has 1-TVD 0.5 and visually it has a significant distortion of distribution.



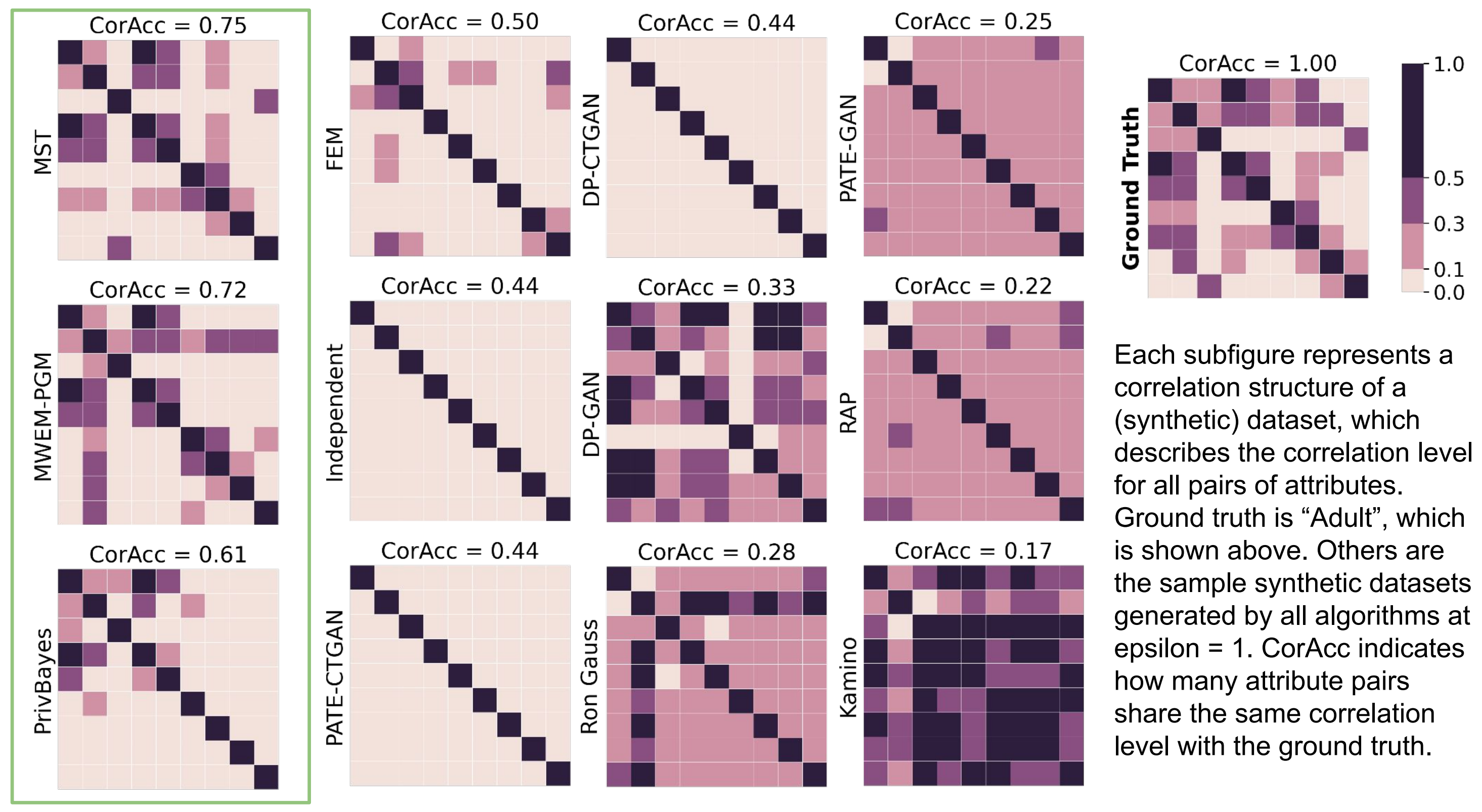
(a) Distributions of the *relationship* attribute: a categorical attribute with six possible values (shown at different colors).

(b) Distributions of the *age* attribute: a numerical attribute discretized by quantiles (which is why the Ground Truth appears uniform).

One-way marginal distributions of a single attribute from dataset *Adult* (marked as *Ground Truth*) and from sample synthetic datasets generated by all algorithms at epsilon = 1. The metric score is reported at the column 1-TVD, which measures the 1 - total variational distance of the distributions between the ground truth and the one of a synthetic data. Algorithm *Independent* is to assume all attributes are independent, add Laplace noise to all the one-way marginals and sample each attribute independently.

Qualitative Analysis: Correlation

- A higher CorAcc is better.
- Marginal-based algorithms (MST, MWEM-PGM and PrivBayes; first column) preserve the correlation structure accurately.
- Many algorithms fail to preserve the correlation structure more accurately than the baseline, *Independent*. For example, at the bottom right, Kamino has CorAcc 0.17 and visually it over-correlate many attribute pairs.



Each subfigure represents a correlation structure of a (synthetic) dataset, which describes the correlation level for all pairs of attributes. Ground truth is “Adult”, which is shown above. Others are the sample synthetic datasets generated by all algorithms at epsilon = 1. CorAcc indicates how many attribute pairs share the same correlation level with the ground truth.

1 Tumult Labs 2 University of Massachusetts Amherst 3 Duke University 4 Colgate University