A Reviewer 1 Experimental Results

Requested Changes. "I also would like to make the following request. If the experimental results are convincing, I will recommend the paper for publication. Could the authors also evaluate the data synthetic methods using pairwise metrics, such as 'Pair' and 'Corr' in this paper. Could the authors also evaluate previous DP-focused tabular data synthetic methods such as DPSyn, PrivMRF and Private-PGM?"

A.1 Additional Metrics

		Adult				Airline			
	Method	**F1**	**CorAcc**	**Pair**	**Hist**	**F1**	**CorAcc**	**Pair**	**Hist**
	Real data	69.9$_{\pm 2}$	95.6	97.7$_{\pm 0}$	99.1$_{\pm 0}$	90.6$_{\pm 8}$	94.4	98.4$_{\pm 0}$	99.4$_{\pm 0}$
$\varepsilon = \infty$	CTGAN	59.5$_{\pm 6}$	72.7$_{\pm 3}$	**85.0**$_{\pm 1}$	91.2$_{\pm 1}$	87.2$_{\pm 3}$	74.2$_{\pm 1}$	**89.3**$_{\pm 1}$	**94.4**$_{\pm 1}$
	TVAE	63.2$_{\pm 2}$	75.1$_{\pm 3}$	84.5$_{\pm 1}$	**91.5**$_{\pm 1}$	85.8$_{\pm 5}$	70.3$_{\pm 3}$	82.8$_{\pm 3}$	90.3$_{\pm 2}$
	VAE	53.8$_{\pm 10}$	64.0$_{\pm 1}$	60.2$_{\pm 2}$	73.3$_{\pm 3}$	79.8$_{\pm 1}$	62.8$_{\pm 6}$	60.4$_{\pm 1}$	76.8$_{\pm 0}$
	GPT-2	**68.9**$_{\pm 0}$	**79.4**$_{\pm 1}$	83.7$_{\pm 0}$	90.7$_{\pm 1}$	**89.6**$_{\pm 5}$	**79.8**$_{\pm 5}$	85.5$_{\pm 2}$	90.8$_{\pm 1}$

Table 1: **Non-DP Benchmark.** The Real data baseline represents the optimal achievable performance, determined by evaluating metrics using real training data compared to real test data. **F1** is reported as averages of two ML Models (XGBoost and Logistic Regression). **Pair**, and **Hist** are reported as averages of two bin sizes (Bins 20 and 50). Each synthetic data generator is run five times and four synthetic datasets generated per run. Standard deviation reported after $\pm$. The best value per column is shown in **bold** while the second best value is underlined.

		Adult				Airline			
	Method	**F1**	**CorAcc**	**Pair**	**Hist**	**F1**	**CorAcc**	**Pair**	**Hist**
$\varepsilon = 1$, $\delta = 10^{-5}$	DP-GAN	33.5$_{\pm 20}$	39.1$_{\pm 3}$	41.2$_{\pm 4}$	63.7$_{\pm 3}$	40.2$_{\pm 24}$	37.5$_{\pm 5}$	22.2$_{\pm 13}$	44.7$_{\pm 12}$
	DP-CTGAN	**42.2**$_{\pm 20}$	49.3$_{\pm 3}$	59.2$_{\pm 2}$	75.7$_{\pm 2}$	67.1$_{\pm 8}$	31.6$_{\pm 3}$	62.2$_{\pm 2}$	78.7$_{\pm 2}$
	DP-VAE	0.0$_{\pm 0}$	45.2$_{\pm 1}$	40.3$_{\pm 1}$	61.8$_{\pm 2}$	26.5$_{\pm 28}$	42.4$_{\pm 1}$	20.6$_{\pm 0}$	41.8$_{\pm 1}$
	GPT-2								
	DP-Standard	27.8$_{\pm 15}$	**49.6**$_{\pm 2}$	68.4$_{\pm 1}$	85.7$_{\pm 2}$	60.5$_{\pm 7}$	49.9$_{\pm 2}$	77.0$_{\pm 2}$	90.3$_{\pm 3}$
	DP-2Stage-U	21.2$_{\pm 12}$	**49.6**$_{\pm 2}$	**76.1**$_{\pm 1}$	86.7$_{\pm 1}$	**68.5**$_{\pm 9}$	51.7$_{\pm 2}$	**80.8**$_{\pm 1}$	90.7$_{\pm 1}$
	DP-2Stage-O	30.4$_{\pm 17}$	49.5$_{\pm 2}$	72.3$_{\pm 1}$	**88.5**$_{\pm 1}$	55.2$_{\pm 18}$	**52.1**$_{\pm 2}$	80.1$_{\pm 1}$	**92.5**$_{\pm 1}$
$\varepsilon = 8$, $\delta = 10^{-5}$	DP-GAN	19.6$_{\pm 20}$	42.9$_{\pm 3}$	11.1$_{\pm 7}$	33.3$_{\pm 9}$	-	-	-	-
	DP-CTGAN	**46.5**$_{\pm 18}$	49.0$_{\pm 2}$	65.6$_{\pm 4}$	80.0$_{\pm 2}$	67.7$_{\pm 4}$	31.2$_{\pm 4}$	54.7$_{\pm 2}$	76.8$_{\pm 1}$
	DP-VAE	0.0$_{\pm 0}$	45.2$_{\pm 1}$	40.9$_{\pm 1}$	62.1$_{\pm 1}$	51.9$_{\pm 25}$	44.3$_{\pm 2}$	18.3$_{\pm 1}$	40.0$_{\pm 1}$
	GPT-2								
	DP-Standard	31.3$_{\pm 15}$	**52.0**$_{\pm 1}$	70.6$_{\pm 1}$	84.5$_{\pm 1}$	64.9$_{\pm 6}$	51.8$_{\pm 3}$	78.2$_{\pm 3}$	89.8$_{\pm 3}$
	DP-2Stage-U	22.4$_{\pm 15}$	51.4$_{\pm 2}$	**77.1**$_{\pm 1}$	86.9$_{\pm 1}$	**71.9**$_{\pm 9}$	**56.1**$_{\pm 2}$	80.9$_{\pm 1}$	90.4$_{\pm 1}$
	DP-2Stage-O	33.4$_{\pm 16}$	51.7$_{\pm 2}$	74.5$_{\pm 1}$	**87.9**$_{\pm 1}$	64.2$_{\pm 11}$	55.1$_{\pm 3}$	**81.1**$_{\pm 1}$	**92.3**$_{\pm 1}$

Table 2: **DP Benchmark.** The Real data baseline represents the optimal achievable performance, determined by evaluating metrics using real training data compared to real test data. **F1** is reported as averages of two ML Models (XGBoost and Logistic Regression). **Pair**, and **Hist** are reported as averages of two bin sizes (Bins 20 and 50). Each synthetic data generator is run five times and four synthetic datasets generated per run. Standard deviation reported after $\pm$. The best value per column for each ε is shown in **bold** while second best value is underlined.

A.2 Additional Baselines

		Adult				Airline			
	Method	**F1**	**CorAcc**	**Pair**	**Hist**	**F1**	**CorAcc**	**Pair**	**Hist**
$\varepsilon = 1$, $\delta = 10^{-5}$	*Marginal*								
	AIM	**59.6**$_{\pm 6}$	**82.7**$_{\pm 1}$	**77.1**$_{\pm 9}$	88.4$_{\pm 5}$	**77.3**$_{\pm 5}$	**83.7**$_{\pm 2}$	46.7$_{\pm 3}$	68.1$_{\pm 2}$
	MST	39.6$_{\pm 19}$	65.1$_{\pm 1}$	74.6$_{\pm 10}$	87.0$_{\pm 5}$	72.2$_{\pm 6}$	59.9$_{\pm 2}$	46.3$_{\pm 3}$	68.2$_{\pm 2}$
	DNN								
	DP-GAN	33.5$_{\pm 20}$	39.1$_{\pm 3}$	41.2$_{\pm 4}$	63.7$_{\pm 3}$	40.2$_{\pm 24}$	37.5$_{\pm 5}$	22.2$_{\pm 13}$	44.7$_{\pm 12}$
	DP-CTGAN	42.2$_{\pm 20}$	49.3$_{\pm 3}$	59.2$_{\pm 2}$	75.7$_{\pm 2}$	67.1$_{\pm 8}$	31.6$_{\pm 3}$	62.2$_{\pm 2}$	78.7$_{\pm 2}$
	DP-VAE	0.0$_{\pm 0}$	45.2$_{\pm 1}$	40.3$_{\pm 1}$	61.8$_{\pm 2}$	26.5$_{\pm 28}$	42.4$_{\pm 1}$	20.6$_{\pm 0}$	41.8$_{\pm 1}$
	GPT-2								
	DP-Standard	27.8$_{\pm 15}$	49.6$_{\pm 2}$	68.4$_{\pm 1}$	85.7$_{\pm 2}$	60.5$_{\pm 7}$	49.9$_{\pm 2}$	77.0$_{\pm 2}$	90.3$_{\pm 3}$
	DP-2Stage-U	21.2$_{\pm 12}$	49.6$_{\pm 2}$	76.1$_{\pm 1}$	86.7$_{\pm 1}$	68.5$_{\pm 9}$	51.7$_{\pm 2}$	**80.8**$_{\pm 1}$	90.7$_{\pm 1}$
	DP-2Stage-O	30.4$_{\pm 17}$	49.5$_{\pm 2}$	72.3$_{\pm 1}$	**88.5**$_{\pm 1}$	55.2$_{\pm 18}$	52.1$_{\pm 2}$	80.1$_{\pm 1}$	**92.5**$_{\pm 1}$

Table 3: **DP Benchmark.** The Real data baseline represents the optimal achievable performance, determined by evaluating metrics using real training data compared to real test data. **F1** is reported as averages of two ML Models (XGBoost and Logistic Regression). **Pair**, and **Hist** are reported as averages of two bin sizes (Bins 20 and 50). Each synthetic data generator is run five times and four synthetic datasets generated per run. Standard deviation reported after $\pm$. The best value per column for each ε is shown in **bold** while second best value is underlined.

	Dataset		**XGB (F1)**	**XGB (AUC)**	**XGB (ACC)**	**LR (F1)**	**LR (AUC)**	**LR (ACC)**	**HIST (bin=50)**	**HIST (bin=20)**
$\varepsilon = 1$	Adult	AIM	**54.0**$_{\pm 4}$	**86.3**$_{\pm 1}$	**81.8**$_{\pm 0}$	**65.2**$_{\pm 0}$	**87.3**$_{\pm 0}$	**78.8**$_{\pm 1}$	83.5$_{\pm 0}$	**93.3**$_{\pm 1}$
$\delta = 10^{-5}$		MST	20.5$_{\pm 3}$	76.7$_{\pm 1}$	74.5$_{\pm 0}$	58.7$_{\pm 0}$	77.0$_{\pm 2}$	71.0$_{\pm 0}$	81.8$_{\pm 2}$	92.2$_{\pm 0}$
		DP-GAN	27.4$_{\pm 24}$	67.6$_{\pm 10}$	69.1$_{\pm 9}$	39.6$_{\pm 14}$	67.8$_{\pm 9}$	59.4$_{\pm 8}$	61.9$_{\pm 2}$	65.5$_{\pm 3}$
		DP-CTGAN	38.6$_{\pm 21}$	77.2$_{\pm 7}$	76.0$_{\pm 3}$	45.8$_{\pm 19}$	78.8$_{\pm 7}$	75.3$_{\pm 3}$	75.0$_{\pm 2}$	76.4$_{\pm 2}$
		DP-VAE	0.0$_{\pm 0}$	50.0$_{\pm 0}$	75.6$_{\pm 0}$	0.0$_{\pm 0}$	50.0$_{\pm 0}$	75.6$_{\pm 0}$	60.1$_{\pm 0}$	63.5$_{\pm 0}$
		DP-Standard	13.9$_{\pm 7}$	55.5$_{\pm 6}$	73.0$_{\pm 1}$	41.6$_{\pm 5}$	61.4$_{\pm 7}$	57.4$_{\pm 6}$	85.1$_{\pm 2}$	86.2$_{\pm 2}$
		DP-2Stage-U	10.3$_{\pm 5}$	48.4$_{\pm 4}$	74.5$_{\pm 1}$	32.1$_{\pm 6}$	49.3$_{\pm 8}$	49.4$_{\pm 4}$	86.3$_{\pm 1}$	87.1$_{\pm 1}$
		DP-2Stage-O	15.0$_{\pm 7}$	55.9$_{\pm 6}$	73.5$_{\pm 1}$	45.9$_{\pm 5}$	67.3$_{\pm 6}$	59.8$_{\pm 5}$	**88.2**$_{\pm 1}$	88.8$_{\pm 1}$
	Airline	AIM	**73.0**$_{\pm 4}$	**85.2**$_{\pm 3}$	**74.3**$_{\pm 4}$	**81.7**$_{\pm 0}$	**92.7**$_{\pm 0}$	**82.1**$_{\pm 0}$	65.9$_{\pm 0}$	70.2$_{\pm 0}$
		MST	69.2$_{\pm 8}$	80.4$_{\pm 6}$	74.0$_{\pm 5}$	75.2$_{\pm 0}$	86.3$_{\pm 0}$	76.4$_{\pm 0}$	66.1$_{\pm 0}$	70.3$_{\pm 0}$
		DP-GAN	38.3$_{\pm 25}$	65.1$_{\pm 14}$	60.3$_{\pm 6}$	42.1$_{\pm 24}$	62.7$_{\pm 12}$	59.2$_{\pm 6}$	43.4$_{\pm 12}$	46.0$_{\pm 12}$
		DP-CTGAN	64.1$_{\pm 10}$	74.0$_{\pm 8}$	65.9$_{\pm 6}$	70.1$_{\pm 5}$	79.4$_{\pm 7}$	70.1$_{\pm 5}$	78.5$_{\pm 2}$	79.0$_{\pm 1}$
		DP-VAE	1.0$_{\pm 3}$	54.4$_{\pm 11}$	56.7$_{\pm 1}$	52.0$_{\pm 14}$	61.4$_{\pm 14}$	58.0$_{\pm 8}$	41.5$_{\pm 0}$	42.2$_{\pm 0}$
		DP-Standard	55.8$_{\pm 3}$	62.1$_{\pm 7}$	60.0$_{\pm 6}$	65.1$_{\pm 6}$	68.4$_{\pm 10}$	64.9$_{\pm 8}$	89.6$_{\pm 4}$	91.1$_{\pm 3}$
		DP-2Stage-U	64.5$_{\pm 10}$	73.8$_{\pm 9}$	70.0$_{\pm 7}$	72.5$_{\pm 6}$	81.8$_{\pm 8}$	74.1$_{\pm 6}$	90.1$_{\pm 0}$	91.3$_{\pm 1}$
		DP-2Stage-O	53.5$_{\pm 16}$	61.8$_{\pm 16}$	59.2$_{\pm 13}$	56.9$_{\pm 20}$	63.3$_{\pm 22}$	60.8$_{\pm 18}$	**92.0**$_{\pm 1}$	**93.0**$_{\pm 1}$

Table 4: **DP Benchmarks.** For each dataset, the best value is highlighted in **bold** while second best is underlined. Results are averaged across five model runs with varying random seeds, with four synthetic datasets generated per run. LR refers to the Logistic Regression model, and XGB represents the XGBoost model.