G   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?

G.1   Additional Metrics

		Adult				Airline			
	Method	F1	CorAcc	Pair	Hist	F1	CorAcc	Pair	Hist
	Real data	$69.9_{\pm2}$	95.6	$97.7_{\pm0}$	$99.1_{\pm0}$	$90.6_{\pm8}$	94.4	$98.4_{\pm0}$	$99.4_{\pm0}$
$\varepsilon=\infty$	CTGAN	$59.5_{\pm6}$	$72.7_{\pm3}$	$\mathbf{85.0}_{\pm1}$	$\underline{91.2}_{\pm1}$	$\underline{87.2}_{\pm3}$	$74.2_{\pm1}$	$\mathbf{89.3}_{\pm1}$	$\mathbf{94.4}_{\pm1}$
	TVAE	$\underline{63.2}_{\pm2}$	$75.1_{\pm3}$	$\underline{84.5}_{\pm1}$	$\mathbf{91.5}_{\pm1}$	$85.8_{\pm5}$	$70.3_{\pm3}$	$82.8_{\pm3}$	$90.3_{\pm2}$
	VAE	$53.8_{\pm10}$	$64.0_{\pm1}$	$60.2_{\pm2}$	$73.3_{\pm3}$	$79.8_{\pm1}$	$62.8_{\pm6}$	$60.4_{\pm1}$	$76.8_{\pm0}$
	GPT-2	$\mathbf{68.9}_{\pm0}$	$\mathbf{79.4}_{\pm1}$	$83.7_{\pm0}$	$90.7_{\pm1}$	$\mathbf{89.6}_{\pm5}$	$\mathbf{79.8}_{\pm5}$	$\underline{85.5}_{\pm2}$	$\underline{90.8}_{\pm1}$

Table 8: **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
$\varepsilon=1,$	DP-GAN	$\underline{33.5}_{\pm20}$	$39.1_{\pm3}$	$41.2_{\pm4}$	$63.7_{\pm3}$	$40.2_{\pm24}$	$37.5_{\pm5}$	$22.2_{\pm13}$	$44.7_{\pm12}$
$\delta=10^{-5}$	DP-CTGAN	$\mathbf{42.2}_{\pm20}$	$49.3_{\pm3}$	$59.2_{\pm2}$	$75.7_{\pm2}$	$\underline{67.1}_{\pm8}$	$31.6_{\pm3}$	$62.2_{\pm2}$	$78.7_{\pm2}$
	DP-VAE	$0.0_{\pm0}$	$45.2_{\pm1}$	$40.3_{\pm1}$	$61.8_{\pm2}$	$26.5_{\pm28}$	$42.4_{\pm1}$	$20.6_{\pm0}$	$41.8_{\pm1}$
	GPT-2								
	DP-Standard	$27.8_{\pm15}$	$\mathbf{49.6}_{\pm2}$	$68.4_{\pm1}$	$85.7_{\pm2}$	$60.5_{\pm7}$	$49.9_{\pm2}$	$77.0_{\pm2}$	$90.3_{\pm3}$
	DP-2Stage-U	$21.2_{\pm12}$	$\mathbf{49.6}_{\pm2}$	$\mathbf{76.1}_{\pm1}$	$\underline{86.7}_{\pm1}$	$\mathbf{68.5}_{\pm9}$	$\underline{51.7}_{\pm2}$	$\mathbf{80.8}_{\pm1}$	$\underline{90.7}_{\pm1}$
	DP-2Stage-O	$30.4_{\pm17}$	$\underline{49.5}_{\pm2}$	$\underline{72.3}_{\pm1}$	$\mathbf{88.5}_{\pm1}$	$55.2_{\pm18}$	$\mathbf{52.1}_{\pm2}$	$\underline{80.1}_{\pm1}$	$\mathbf{92.5}_{\pm1}$
$\varepsilon=8,$	DP-GAN	$19.6_{\pm20}$	$42.9_{\pm3}$	$11.1_{\pm7}$	$33.3_{\pm9}$	-	-	-	-
$\delta=10^{-5}$	DP-CTGAN	$\mathbf{46.5}_{\pm18}$	$49.0_{\pm2}$	$65.6_{\pm4}$	$80.0_{\pm2}$	$\underline{67.7}_{\pm4}$	$31.2_{\pm4}$	$54.7_{\pm2}$	$76.8_{\pm1}$
	DP-VAE	$0.0_{\pm0}$	$45.2_{\pm1}$	$40.9_{\pm1}$	$62.1_{\pm1}$	$51.9_{\pm25}$	$44.3_{\pm2}$	$18.3_{\pm1}$	$40.0_{\pm1}$
	GPT-2								
	DP-Standard	$31.3_{\pm15}$	$\mathbf{52.0}_{\pm1}$	$70.6_{\pm1}$	$84.5_{\pm1}$	$64.9_{\pm6}$	$51.8_{\pm3}$	$78.2_{\pm3}$	$89.8_{\pm3}$
	DP-2Stage-U	$22.4_{\pm15}$	$51.4_{\pm2}$	$\mathbf{77.1}_{\pm1}$	$\underline{86.9}_{\pm1}$	$\mathbf{71.9}_{\pm9}$	$\mathbf{56.1}_{\pm2}$	$\underline{80.9}_{\pm1}$	$\underline{90.4}_{\pm1}$
	DP-2Stage-O	$\underline{33.4}_{\pm16}$	$\underline{51.7}_{\pm2}$	$\underline{74.5}_{\pm1}$	$\mathbf{87.9}_{\pm1}$	$64.2_{\pm11}$	$\underline{55.1}_{\pm3}$	$\mathbf{81.1}_{\pm1}$	$\mathbf{92.3}_{\pm1}$

Table 9: **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.