Technical Report: Differentially Private Event Logs with Case Attributes

Hannes Ueck[1], Robert Andrews[2], Moe T. Wynn[2], and Sander J. J. Leemans[1,3]

[1] RWTH Aachen, Germany
[2] Queensland University of Technology, Australia
[3] Fraunhofer FIT, Germany
hannes.ueck@rwth-aachen.de, s.leemans@bpm.rwth-aachen.de

1 Evaluation Results

This technical report details the results of the evaluation of the framework proposed in the main paper. Due to computational limitations, we were unable to obtain results for the combination of the BPIC13 event log and the DPGAN tabular data generation algorithm.

1.1 Stochastic process behaviour

We use three different measures to assess the similarity of the anonymised event logs process behaviour and the original event logs process behaviour. Figure 1 shows the Earth mover's conformance as proposed in [5]. We compute the Earth mover's conformance between the anonymised event logs and the original event log. An Earth mover's conformance of 1 signifies perfect conformance of the anonymised process behaviour to the original behaviour.

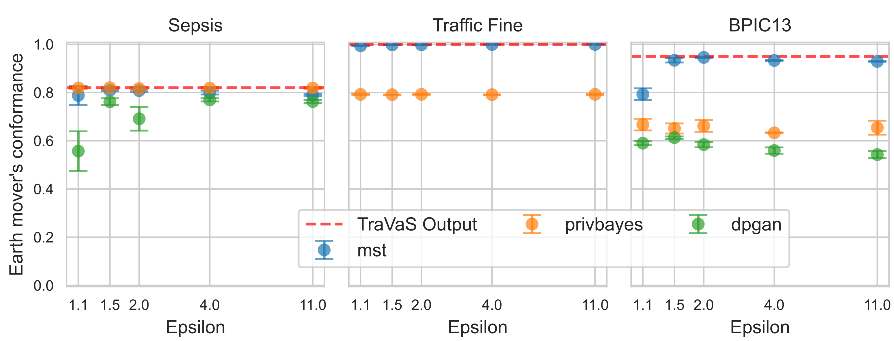

Fig. 1: Earth mover's conformance for values of ε.

Further, we assess how well the anonymised event logs can be used for process discovery and conformance checking tasks. We discover a process model from the anonymised event log using the inductive miner infrequent [3] with the noise threshold set to 20%. Then we replay the original event log on this discovered process model and compute the fitness and precision scores. A higher fitness score indicates that more traces from the original event log can be replayed on the private process model. For precision scores, a higher score indicates that the private process model does not allow for more behaviour than the original process model. Figure 2 shows the fitness scores for different values of ε across the different event logs and algorithms. Similarly, Figure 3 shows the precision scores.

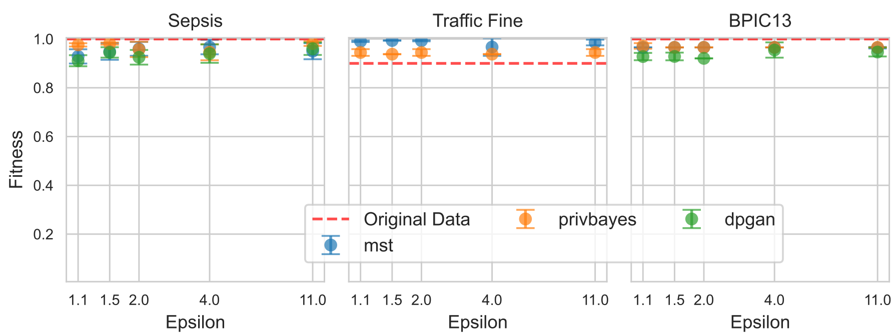

Fig. 2: Fitness for values of ε.

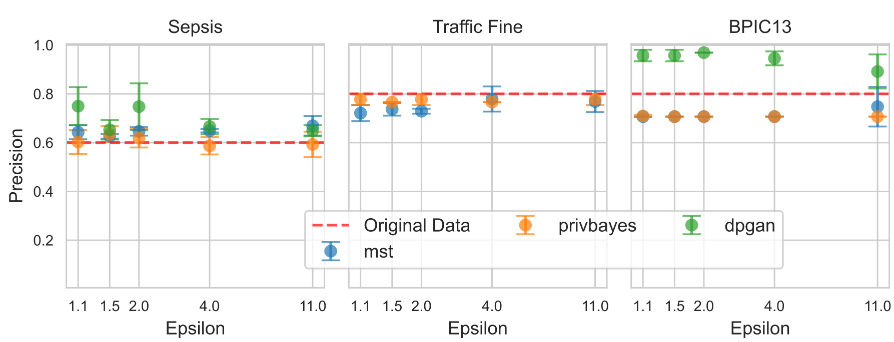

Fig. 3: Precision for values of ε.

MST						
ε	μ_{age}	$\mu_{infectionsuspected}$	$\mu_{hypotensie}$	$\mu_{infusion}$	$\mu_{oligurie}$	$\mu_{hypoxie}$
orig	70.08 ± (17.36)	0.81 ± (0.39)	0.05 ± (0.22)	0.76 ± (0.43)	0.02 ± (0.15)	0.02 ± (0.14)
11.0	66.77 ± (19.34)	0.67 ± (0.47)	0.03 ± (0.18)	0.62 ± (0.49)	0.02 ± (0.14)	0.02 ± (0.14)
4.0	65.25 ± (20.26)	0.67 ± (0.47)	0.03 ± (0.18)	0.61 ± (0.49)	0.04 ± (0.19)	0.02 ± (0.15)
2.0	55.64 ± (21.86)	0.67 ± (0.47)	0.05 ± (0.21)	0.62 ± (0.48)	0.04 ± (0.19)	0.04 ± (0.17)
1.5	55.66 ± (21.63)	0.70 ± (0.45)	0.05 ± (0.19)	0.62 ± (0.48)	0.05 ± (0.15)	0.02 ± (0.08)
1.1	55.27 ± (21.55)	0.51 ± (0.50)	0.50 ± (0.50)	0.50 ± (0.50)	0.51 ± (0.50)	0.40 ± (0.40)
PrivBayes						
ε	μ_{age}	$\mu_{infectionsuspected}$	$\mu_{hypotensie}$	$\mu_{infusion}$	$\mu_{oligurie}$	$\mu_{hypoxie}$
orig	70.08 ± (17.36)	0.81 ± (0.39)	0.05 ± (0.22)	0.76 ± (0.43)	0.02 ± (0.15)	0.02 ± (0.14)
11.0	63.44 ± (20.03)	0.64 ± (0.48)	0.28 ± (0.42)	0.59 ± (0.49)	0.25 ± (0.39)	0.04 ± (0.17)
4.0	61.86 ± (20.68)	0.65 ± (0.47)	0.12 ± (0.27)	0.59 ± (0.49)	0.09 ± (0.24)	0.03 ± (0.17)
2.0	63.54 ± (20.14)	0.61 ± (0.49)	0.09 ± (0.24)	0.58 ± (0.49)	0.15 ± (0.26)	0.17 ± (0.33)
1.5	62.63 ± (20.33)	0.66 ± (0.47)	0.24 ± (0.36)	0.58 ± (0.49)	0.16 ± (0.28)	0.08 ± (0.23)
1.1	59.60 ± (21.17)	0.61 ± (0.48)	0.26 ± (0.38)	0.58 ± (0.49)	0.23 ± (0.34)	0.08 ± (0.20)
DPGAN						
ε	μ_{age}	$\mu_{infectionsuspected}$	$\mu_{hypotensie}$	$\mu_{infusion}$	$\mu_{oligurie}$	$\mu_{hypoxie}$
orig	70.08 ± (17.36)	0.81 ± (0.39)	0.05 ± (0.22)	0.76 ± (0.43)	0.02 ± (0.15)	0.02 ± (0.14)
11.0	62.01 ± (24.28)	0.36 ± (0.48)	0.95 ± (0.22)	0.38 ± (0.49)	0.93 ± (0.24)	0.96 ± (0.20)
4.0	67.24 ± (19.96)	0.34 ± (0.47)	0.95 ± (0.21)	0.43 ± (0.49)	0.95 ± (0.20)	0.97 ± (0.18)
2.0	68.70 ± (20.72)	0.07 ± (0.24)	0.82 ± (0.34)	0.31 ± (0.46)	0.94 ± (0.22)	0.97 ± (0.15)
1.5	80.01 ± (13.24)	0.38 ± (0.49)	0.95 ± (0.19)	0.34 ± (0.46)	0.55 ± (0.49)	0.93 ± (0.23)
1.1	69.07 ± (25.51)	0.28 ± (0.44)	0.45 ± (0.49)	0.70 ± (0.45)	0.59 ± (0.49)	0.66 ± (0.47)

Table 1: Comparison of means and standard deviations for different privacy levels on the sepsis dataset using different tabular data generation algorithms

1.2 Descriptive statistics of case attributes

Table 1 shows the mean and standard deviation values of the case attributes for the Sepsis event log for all selected models and privacy budgets. The same is shown in Table 2 but for the Traffic fine event log. The BPIC13 event log is not shown because it contains only categorical attributes.

MST			
ε	μ_{amount}	$\mu_{totalamount}$	μ_{points}
orig	71.42 ± (100.54)	23.97 ± (40.25)	0.08 ± (0.58)
10.0	115.18 ± (81.66)	54.21 ± (25.92)	0.87 ± (0.54)
3.0	130.36 ± (82.18)	54.29 ± (27.09)	0.87 ± (0.55)
1.0	130.54 ± (84.38)	33.97 ± (29.73)	0.47 ± (0.51)
0.5	84.90 ± (72.41)	33.98 ± (30.16)	0.47 ± (0.51)
0.1	71.44 ± (55.31)	30.66 ± (29.07)	0.25 ± (0.37)
PrivBayes			
ε	μ_{amount}	$\mu_{totalamount}$	μ_{points}
orig	71.42 ± (100.54)	23.97 ± (40.25)	0.08 ± (0.58)
10.0	71.15 ± (196.67)	42.50 ± (136.90)	0.14 ± (0.83)
3.0	68.15 ± (175.39)	43.67 ± (142.89)	0.13 ± (0.79)
1.0	71.11 ± (196.68)	42.49 ± (136.67)	0.14 ± (0.83)
0.5	68.11 ± (176.25)	43.81 ± (145.20)	0.13 ± (0.80)
0.1	71.13 ± (198.06)	46.18 ± (153.00)	0.16 ± (0.88)

Table 2: Comparison of means and standard deviations for different privacy levels on the traffic fine dataset using different tabular data generation algorithms

1.3 Correlation process behaviour - case attributes

Association measures to measure the correlation between the process behaviour and the values of the case attributes have been proposed by [4]. Figure 4 shows the correlation between the process behaviour and the case attributes for the Sepsis event log. The results for the BPIC13 and the Traffic fine event log are shown in Figure 5 and Figure 6. Unfortunately, we ran into a bug when computing the results for the anonymised event logs from PrivBayes.

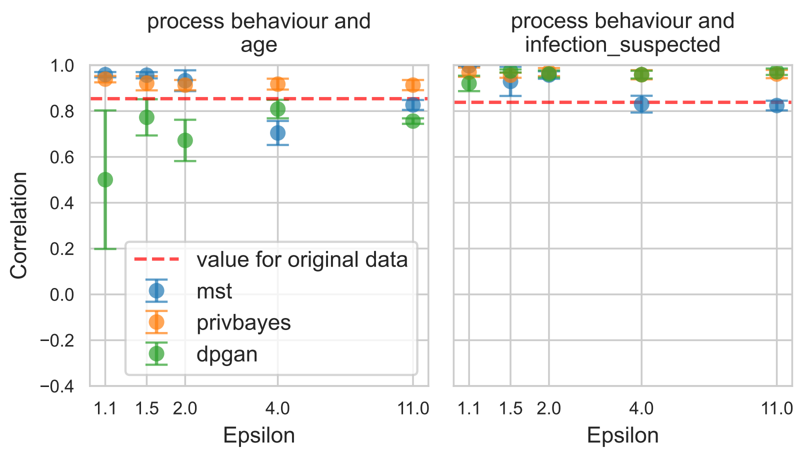

Fig. 4: Correlation values between process behaviour and case attributes of the Sepsis event log for values of ε.

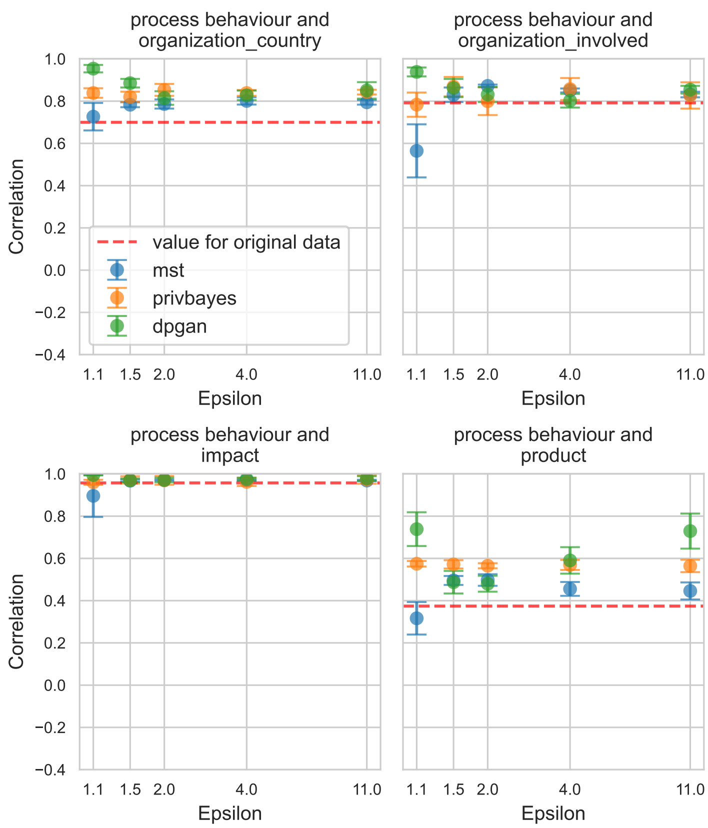

Fig. 5: Correlation values between process behaviour and case attributes of the BPIC13 event log for values of ε.

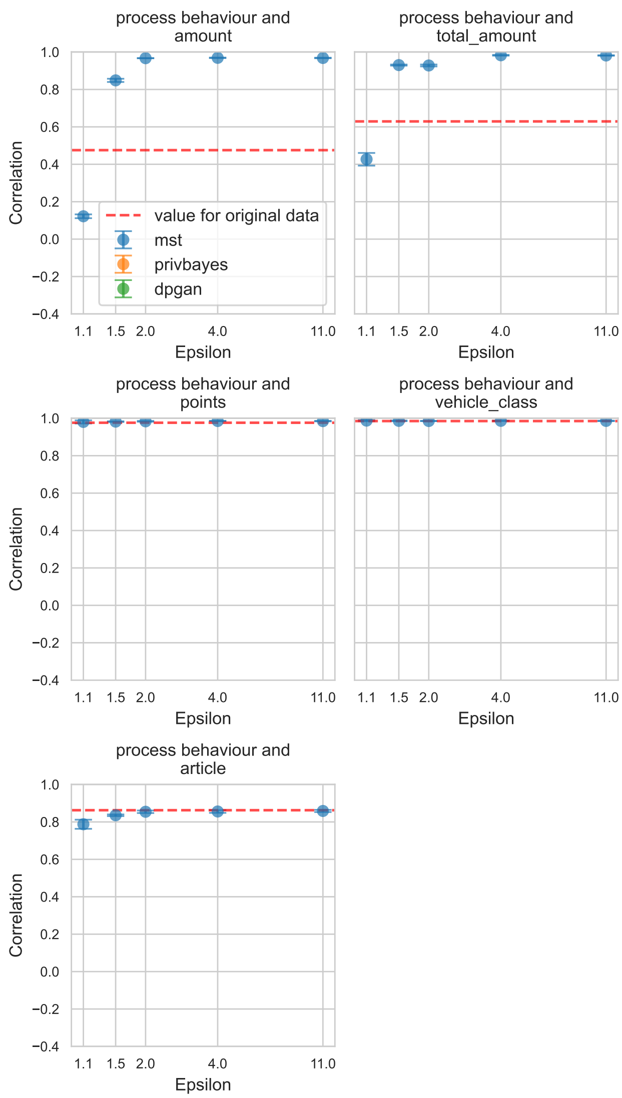

Fig. 6: Correlation values between process behaviour and case attributes of the traffic fine event log for values of ε.

1.4 Correlation between case attributes

For pairs of numerical case attributes, we use the Pearson correlation coefficient [1]. The point-biserial correlation coefficient is used to measure relationships between binary and numerical case attributes [2]. Figure 8 shows the results for the correlation between the numerical case attribute `age` and the other binary case attributes of the Sepsis event log. In Figure 7 results for the Traffic fine event log can be seen. The BPIC13 event log only includes categorical case attributes, thus this measure can not be computed.

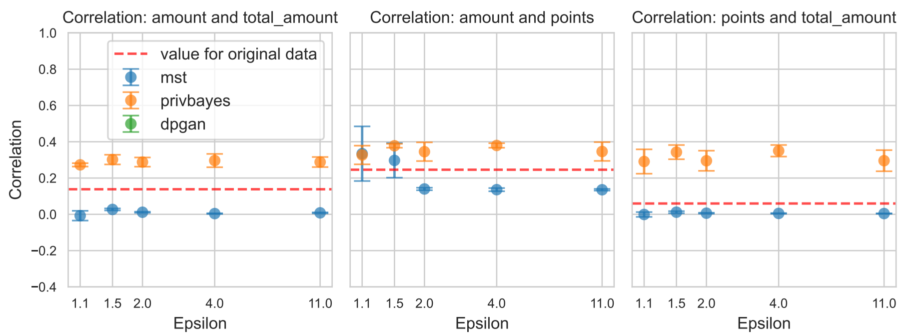

Fig. 7: Correlation values between case attributes of the Traffic fine event log for values of ε.

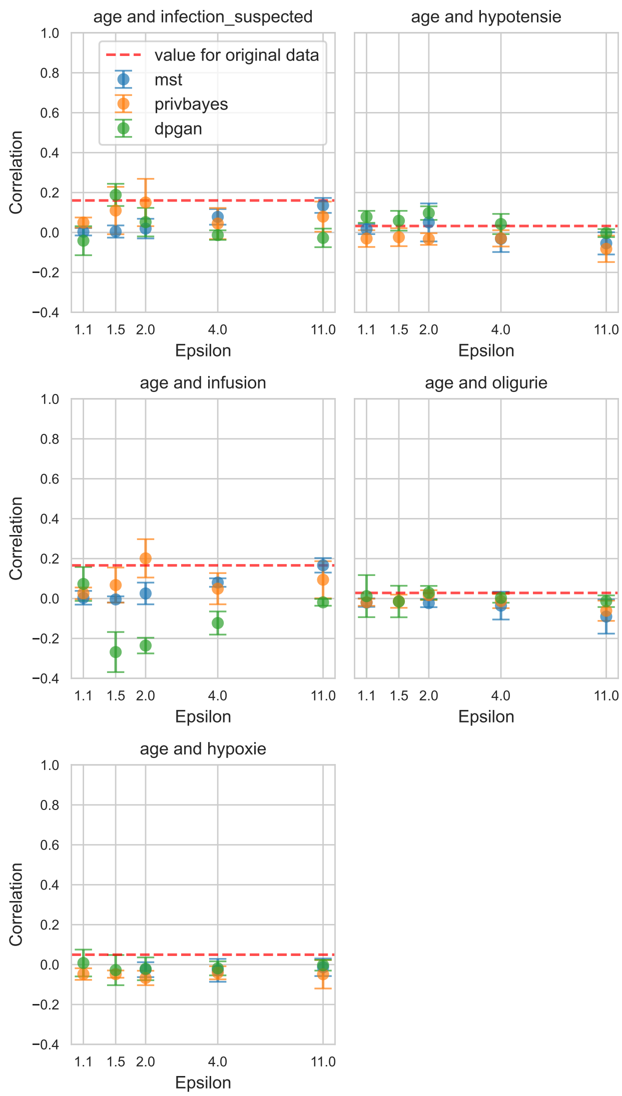

Fig. 8: Correlation values between case attributes of the Sepsis event log for values of ε.

References

1. Freedman, D., Pisani, R., Purves, R.: Statistics (international student edition). Pisani, R. Purves, 4th edn. WW Norton & Company, New York (2007)
2. Kornbrot, D.: Point biserial correlation. Wiley StatsRef: Statistics Reference Online (2014)
3. Leemans, S.J.J., Fahland, D., van der Aalst, W.M.P.: Discovering block-structured process models from event logs containing infrequent behaviour. In: Business Process Management Workshops. pp. 66–78. Springer International Publishing, Cham (2014)
4. Leemans, S.J., McGree, J.M., Polyvyanyy, A., ter Hofstede, A.H.: Statistical tests and association measures for business processes. IEEE Transactions on Knowledge and Data Engineering **35**(7), 7497–7511 (2023)
5. Leemans, S.J., et al.: Stochastic process mining: Earth movers' stochastic conformance. IS **102**, 101724 (2021)