Conceptualizing a Log Generator for Privacy-aware Event Logs

Kay Kaczmarek[1], Agnes Koschmider[2]

1 Introduction

Privacy-preserving process mining will gain importance in the future due to the mandatory implementation of protection guidelines and laws applicable when handling (personal) data. Mannhardt et al. [Ma19] pointed to two general privacy challenges that have to be addressed within process mining: technological privacy challenges (related to the design of privacy-by-design or privacy-by-default approaches) and organizational privacy challenges. Even if technical solutions exist to bridge these challenges, still the reluctance remains to share data and to trust those solutions. Process mining techniques uncover operational processes of companies and pose the risk to re-identify private data whose access has to be fully protected. Several privacy-preserving process mining techniques have been already suggested. For a comprehensive overview we refer to Elkoumy et al. [El21].

The intention of this paper is to suggest a log generator for privacy-aware event logs. The benefit of the log generator is to build synthetic event logs based on an original event log and to apply process mining on the synthetic log aiming to obtain the same results like with the original event log. Additionally, it should be allowed to sample the synthetic event log and to quantify the privacy risk for different event log sizes. From a technical perspective, we aim to use a training data set to learn constructing accurate and possibly generalized data sets. The trained model should allow to generate new data without retaining training data. Then the generated data is converted into an event log format. Statistical methods and conformance checking are used to evaluate the quality of the event log. Finally, the risk of re-identification is quantified using the approach of von Voigt et al. [Vo20].

2 Approach

Fig. 1 shows the process of synthetic, privacy-aware event log generation. First, the data must be prepared for the machine learning technique or for use as training data respectively. So far we have been using Recurrent Neural Networks with LSTM and Time-series Generative

[1] Kiel University, Kiel, Germany stu96465@mail.uni-kiel.de
[2] Kiel University, Group Prcoess Analytics, Kiel, Germany ak@informatik.uni-kiel.de

Adversarial Network. The synthetically generated data set can be transformed directly into an event log using PM4PY. The evaluation of the quality of the synthetically generated data is compared with the original data using several methods. We tried T-Distributed Stochastic Neighbor Embedding, Principal component analysis (PCA) and conformance checking. The re-identification risk is assessed in interaction with the quality evaluation.

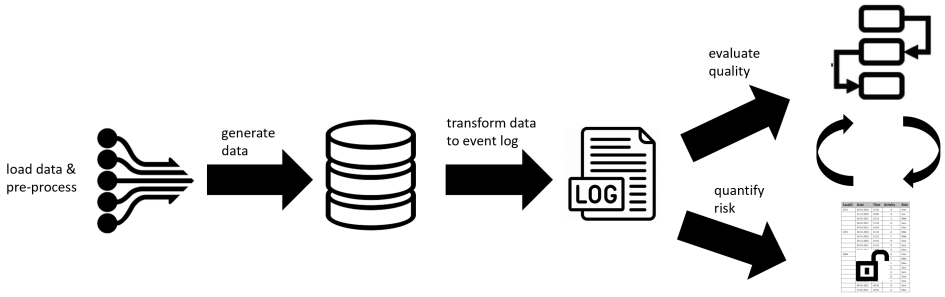

Fig. 1: Process discovery approach for location sensor event data.

3 Challenges

The implementation of a synthetic, privacy-aware event log generation has several challenges. It is difficult to efficiently learn synthetic sequential data using machine learning techniques. There is still a high manual effort required to generate the event log from diverse data types processed by the ML techniques. The generation of an event log is also limited how machine learning techniques learn and the input they expect (e.g., it is challenging to parse varying size of sequences).

Literatur

[El21] Elkoumy, G.; Fahrenkrog-Petersen, S. A.; Sani, M. F.; Koschmider, A.; Mannhardt, F.; von Voigt, S. N.; Rafiei, M.; von Waldthausen, L.: Privacy and Confidentiality in Process Mining - Threats and Research Challenges. CoRR abs/2106.00388/, 2021.

[Ma19] Mannhardt, F.; Koschmider, A.; Baracaldo, N.; Weidlich, M.; Michael, J.: Privacy-Preserving Process Mining - Differential Privacy for Event Logs. Bus. Inf. Syst. Eng. 61/5, S. 595–614, 2019.

[Vo20] von Voigt, S. N.; Fahrenkrog-Petersen, S. A.; Janssen, D.; Koschmider, A.; Tschorsch, F.; Mannhardt, F.; Landsiedel, O.; Weidlich, M.: Quantifying the Re-identification Risk of Event Logs for Process Mining - Empiricial Evaluation Paper. In: CAiSE. Bd. 12127. LNCS, Springer, S. 252–267, 2020.