PRETSA: Event Log Sanitization for Privacy-aware Process Discovery[1]

(Extended Abstract)

Stephan A. Fahrenkrog-Petersen
Han van der Aa · Matthias Weidlich

Information systems record data in the form of event logs, while executing business processes. Event logs can, therefore, be used for data-driven analysis of business processes. In recent years, various such analysis techniques have been proposed under the umbrella of process mining [1]. For instance, techniques for process discovery construct a model of a business process from an event log (see [2]), which can then be enriched with performance information for quantitative process analysis.

However, logs potentially contain sensitive information about individual employees involved in process execution. Due to legal frameworks such as the General Data Protection Regulation (GDPR), organizations are obliged to ensure a certain level of privacy and to protect the personal data of individuals [3]. In many scenarios, obfuscation of the event log is not sufficient to achieve such data protection. Rather, one has to rely on explicit approaches for data sanitization, which provide privacy guarantees through data transformation mechanisms. Data sanitization is typically lossy, meaning that the utility of the data for some analysis task is hampered. Therefore, it is necessary to develop techniques that achieve privacy, but preserve as much utility as possible for the analysis task at hand.

This work introduces a data sanitization technique suited for event logs used to discover performance-annotated process models. Specifically, we consider a trace linking attack on an event log with pseudonymized employee information that correlates events of the log with background knowledge on possible activity assignments during process execution. For this setting, we present PRETSA (PREfix-Tree based event log SAnitisation for t-closeness), a sanitization technique that guarantees privacy in terms of k-anonymity and t-closeness for the transformed event log. It thereby avoids disclosure of employee identities, their membership in the event log, and their characterization based on sensitive attributes, such as performance information. PRETSA takes up ideas on achieving k-anonymity for sequential data [4]. In essence, PRETSA constructs a prefix tree representation of an event log that is annotated with frequencies and attribute values. This tree is then step-wise transformed; subtrees are merged and relocated until the required privacy guarantees have been obtained. The resulting log transformations are comparatively fine granular. As a consequence, the log's utility for discovery of a performance-annotated process model is largely preserved.

Experiments with three real-world data sets demonstrate that sanitization with PRETSA yields

[1] This paper is in press as: Stephan A. Fahrenkrog-Petersen, Han van der Aa, Matthias Weidlich: PRETSA: Event Log Sanitization for Privacy-aware Process Discovery. 1st International Conference on Process Mining (ICPM), June 24–26, 2019, Aachen, Germany.

https://doi.org/10.1007/s00287-019-01203-z

Stephan A. Fahrenkrog-Petersen · Han van der Aa
Matthias Weidlich
Humboldt-Universität zu Berlin, Institut für Informatik, Berlin
E-Mail: {stephan.fahrenkrog-petersen, han.van.der.aa, matthias.weidlich}@hu-berlin.de

event logs of higher utility compared to methods that exploit frequency-based filtering, while providing the same privacy guarantees. In some cases, frequency-based filtering is not even able to provide any event log that fulfills the requested privacy guarantees. PRETSA, in turn, was always able to provide a sanitized event log with high utility. The latter is reflected in the amount of preserved process variants, the fitness of the resulting process model with the event log, and the deviation in the generated performance annotations compared to the annotations generated based on the original event log.

References

1. Van der Aalst WMP (2016) Process Mining – Data Science in Action. Springer, Berlin Heidelberg
2. Augusto A, Conforti R, Dumas M, La Rosa M, Maggi FM, Marrella A, Mecella M, Soo A (2019) Automated discovery of process models from event logs: Review and benchmark. IEEE T Knowl Data Eng 31(4):686–705
3. Mannhardt F, Petersen SA, Oliveira MF (2018) Privacy challenges for process mining in human-centered industrial environments. In: 2018 14th International Conference on Intelligent Environments (IE). IEEE, pp 64–71
4. Monreale A, Pedreschi D, Pensa RG, Pinelli F (2014) Anonymity preserving sequential pattern mining. Artif Intell Law 22(2):141–173