Quantifying the Re-identication Risk of Event Logs for Process Mining (Extended Abstract)[*]

Saskia Nuñez von Voigt[1], Stephan A. Fahrenkrog-Petersen[2], Dominik Janssen[3],
Agnes Koschmider[3], Florian Tschorsch[1], Felix Mannhardt[4,5], Olaf Landsiedel[3],
Matthias Weidlich[2]

Event logs for process mining often contain sensitive information that could be linked to individual process stakeholders by cross-correlating background information, e.g., in an emergency room process, certain events can indicate that a patient is in a certain condition. In general, case attributes can contain various kinds of sensitive data, revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, as well as financial or health information. Likewise, an event log can reveal information about the productivity [PLWF17] or the work schedule of hospital staff. Such kind of staff surveillance is a critical privacy threat. We therefore argue that, when publishing event logs, the risk of such re-identification attacks must be considered. The intention of this extended abstract is to raise awareness to the re-identification risk of event logs and to highlight the importance of privacy-preserving techniques in process mining [MKBW19]. We therefore provide measures to quantify this risk. Specifically, we provide an approach to express the uniqueness of data, which is derived from models that are commonly adopted by process mining techniques. The higher the uniqueness of an event log, the higher an adversary's chances to identify the target. In our case, a targeted re-identification is assumed, i.e., an adversary has information about specific individuals, which includes a subset of the attribute values. Given this background information, the adversary's goal is to reveal sensitive information, e.g., a diagnosis.

Our approach therefore explores the number of cases that are uniquely identifiable by the set of case attributes or the set of event attributes. We use this information to derive a measure of uniqueness for an event log, which serves as a basis for estimating how likely a case can be re-identified. We evaluate a number of so-called projections that can be considered as a kind of data minimization, effectively reducing the potential risk of re-identifying an individual in an event log. Projections refer to a subset of attributes in the event log. For instance, one projection might contain the sequence of all executed activities with their timestamps, while another projection only contains the case attributes. It has been shown that even sparse projections of event logs hold privacy risks [FaAW19].

[*] The original article is accepted for the 32nd International Conference on Advanced Information Systems Engineering (CAiSE 2020), Grenoble, France, June 8-12, 2020.

[1] Technische Universität Berlin, Germany, firstname.lastname@tu-berlin.de

[2] Humboldt-Universität zu Berlin, Germany, firstname.lastname@hu-berlin.de

[3] Kiel University, Germany, doj|ak|ol@informatik.uni-kiel.de

[4] SINTEF Digital, Trondheim, Norway, felix.mannhardt@sintef.no

[5] NTNU Norwegian University of Science and Technology, Trondheim, Norway

Therefore, in our evaluation, we will consider the re-identification risk for various projections.

To demonstrate the importance of uniqueness considerations for event logs, we conducted a large-scale study with 12 publicly available event logs from the 4TU.Centre for Research Data repository[1]. We categorized the records and assessed the uniqueness where cases refer to a natural person. Our results suggest that an adversary can potentially re-identify up to all of the cases, depending on prior knowledge. We show that an adversary needs only a few attributes of a trace to successfully mount such an attack.

In conclusion, generalization of attributes certainly helps to reduce the re-identification risk [ZBBC17]. Our results, however, show that combining several attributes, such as case attributes and activities, still yields many unique cases. In combination with lowering the resolution of values, e.g., publishing only the year of birth instead of the full birthday, we are able to reduce the re-identification risk. Such generalization techniques can also be applied to timestamps, activities, or case attributes. Along the lines of the data minimization principle, i.e., limiting the amount of personal data, omitting data is simply the most profound way to reduce the risk, which we clearly see when taking our projections into account. Consequently, the projections can be used to reduce the re-identification risk.

This paper shows that we as a community have to act more carefully, though, when releasing event logs, while also highlighting the need to develop privacy-preserving techniques for event logs. We believe that this work will foster the trust and increases the willingness for sharing event logs while providing privacy guarantees.

Literaturverzeichnis

[FaAW19] Fahrenkrog-Petersen, S.A., van der Aa, H., Weidlich, M.: PRETSA: Event Log Sanitization for Privacy-Aware Process Discovery. In: ICPM '19: Proceedings of the International Conference on Process Mining. (2019) 1-8

[MKBW19] Mannhardt, F., Koschmider, A., Baracaldo, N., Weidlich, M., Michael, J.: Privacy-Preserving Process Mining. Business & Information Systems Engineering 61(5) (2019) 595-614

[PLWF17] Pika, A.; Leyer, M.; Wynn, M.T.; Fidge, C.J.; Ter Hofstede, A.H.; van der Aalst, W.M.: Mining Resource Profiles from Event Logs. TMIS '17: Proceedings of ACM Transactions on Management Information Systems 8(1) (2017)

[ZBBC17] Zook, M.; Barocas, S.; Boyd, D.; Crawford, K.; Keller, E.; Gangadharan, S.P.; Pasquale, F.: Ten Simple Rules for Responsible Big Data Research (2017)

[1] https://data.4tu.nl/repository/collection:event_logs_real