Anonymization Techniques for Privacy-preserving Process Mining (Extended Abstract)

Stephan A. Fahrenkrog-Petersen[1,2]

[1]Humboldt-Universität zu Berlin, Berlin, Germany
[2]Weizenbaum Institute, Berlin, Germany

Abstract

Privacy concerns arise when analyzing event logs in process mining, as they may contain sensitive information, such as clinical treatment details. Recently, through the work covered in this thesis, the risk of re-identification in event logs was quantified, and privacy threats within process mining were modeled. Anonymization emerged as a key strategy to address these privacy issues. The aim is to strike a balance between preserving utility and providing a predetermined level of privacy. This thesis introduced several novel anonymization techniques that provide superior utility compared to the state-of-the-art.

Keywords

Privacy, Anonymization, Process Mining, Privacy-aware Process Mining

1. Introduction

Event logs are commonly used for process mining, enabling the analysis of business processes [1]. These logs contain fine-granular information about the execution of a business process. Sometimes these logs cover extremely private information, such as the clinical treatment of a patient [2]. Consequently, there are privacy concerns over the sensitive information within an event log. These issue sometimes block the adoption of process mining in sensitive domains such as healthcare or human resource management.

While the issue of privacy in process mining was already mentioned in the process mining manifesto [1], it was not well-studied. This changed recently, partly due to the work covered in the thesis [3], and the risk of re-identification in event logs [4] and also process models [5] was quantified. Furthermore, the threats to privacy within process mining have been modeled [6]. Through this approach, anonymization was established as a key strategy to handle the privacy issues in process mining.

Within the thesis discussed in this paper, several novel anonymization techniques have been introduced. These techniques are suited to preserving as much utility as possible while providing a pre-determined privacy guarantee. This privacy-utility trade-off is a major challenge addressed within the thesis. Overall, the covered thesis contributes to the field of process mining in three ways: (i) The thesis studies the re-identification risk of event logs and explores the

Proceedings of the Best BPM Dissertation Award, Doctoral Consortium, and Demonstrations & Resources Forum co-located with 22nd International Conference on Business Process Management (BPM 2024), Krakow, Poland, September 1st to 6th, 2024.

stephan.fahrenkrog-petersen@hu-berlin.de (S. A. Fahrenkrog-Petersen)
0000-0002-1863-8390 (S. A. Fahrenkrog-Petersen)

challenges that exist for privacy-preserving process mining; (ii) The thesis introduces novel anonymization strategies that provide better utility than state-of-the-art techniques; (iii) The algorithms of this thesis were made available to the community through integration into the famous framework PM4PY [7].

The remaining paper follows the same structure as the underlying thesis: In Section 2 we highlight how the thesis contributed towards understanding privacy threats in process mining. Next, we outline the anonymization techniques introduced in the thesis in Section 3. Finally, we discuss future impact of the thesis for the field in Section 4.

2. The Case for Anonymization

Within section, we show studies that support the argument why anonymization techniques for privacy-preserving process mining is needed. We do this by two studies, one qualitative study and one quantitative study:

Qualitative Discussion on the Threats and Requirements of Privacy-preserving Process Mining. While there exists a general understanding of the need to consider privacy within process mining, the concrete requirements and threats have not been analysed. We fill this gap in the thesis with this chapter, that is based on a collaboration of several privacy-preserving process mining experts [6] (including the author of the thesis). This group collected the privacy threats to process mining and derived requirements for privacy-preserving process mining based on these threats. Furthermore, we argue that anonymization can help to address a majority of the requirements spelled out and therefore, motivate the need for anonymization techniques for event logs.

Quantitative Analysis of Re-identification Risk in Event Logs. Until now the discussion of threats to the privacy in event logs have been completely qualitative, within this part of the thesis we provided a qualitative study that estimates the re-identification risk in event logs [4]. In this study we used well-established techniques for re-identification risk estimation and adjusted them to the specifics of event logs. We applied the new method to event logs and considered also only subsets of the data within these event logs. All publicly available event logs have been considered and were analyzed in a pseudonymized manner. Through our experiments we could show that event logs posses significant re-identification threats and therefore can reveal sensitive information about the individuals involved within the event logs.

3. Anonymization Techniques

In this section, we give an overview of the anonymization techniques included in the underlying thesis. All of the algorithms are made available on Github. Additionally, the differential privacy algorithms have been integrated into PM4Py [7]. The anonymization techniques covered in the thesis can be described as follows:

Semantic-aware Control-flow Anonymization. Many process mining tasks, such as process discovery, mainly focus on the control-flow. This data might encode sensitive information of service consumers, such as medical treatments they underwent. It is therefore important that

such control-flows analyzed in a privacy-preserving manner, but at the same time it is important that the semantics of the control-flow is preserved. With *SaCoFa* [8] and *SaPa* [9] we introduce two mechanisms that achieve the privacy notion *differential privacy* [10] through noise insertion. However, compared to state-of-the-art techniques they are able to add noise that is semantically sensible and therefore preserve a higher utility for process discovery. *SaCoFa* mainly focuses on semantics in terms of a logical ordering of the activities, i.e. by preventing the adding of obviously false behavior such a traces where a patient is released from a hospital before she is admitted. *SaPa* focussed on preserving the original log size, since alternative anonymization techniques sometimes distort the log size by order of magnitudes. Within the thesis we show that *SaCoFa* and *SaPa* outperform the state of the art technique [11].

PRIPEL. Certain process mining tasks require more information than the control-flow. PRIPEL [12] is an anonymization technique that provides a differential privacy guarantee for this contextual information. The algorithm takes the output of one control-flow anonymization technique, such as *SaCoFa* or *SaPa*, and enriches it with contextual information, that is anonymized through local differential privacy. Resulting in an event log that is completely protected by differential privacy. We showed that the event logs created by PRIPEL are able to retain general properties that can be used to analyze certain aspects of of the process such as bottlenecks.

PRETSA-Algorithm Family. The above mentioned techniques focused on the privacy protection of individuals represented as one case. However, in process mining the service providers, also called resources, play an important role. The *PRETSA-Algorithm family* [13, 14] offers protection for these individuals, through the privacy guarantees k-anonymity [15] and t-closeness [16]. Both guarantees are based on the idea of hiding individuals within groups so that no individual can be singled-out. The anonymization techniques represent the control-flows of the event log as prefix trees. The algorithms travel the prefix tree and merge behavior that violates privacy guarantees with similar behavior. The algorithm family con sits of three algorithms: *PRETSA* is the basic version ans travels the prefix in an ad-hoc manner; *PRETSA** formalizes removal of behavior as a search problem and uses A*-search to find an optimal solution; *BF-PRETSA* uses the same search heuristic as *PRETSA** for a best-first-approach, but does not exhaustively explores the complete search space. Overall, we could show that *PRETSA* outperforms naive baselines. While *BF-PRETSA* outperforms *PRETSA* and other state of the art techniques. However, PRETSA* nearly never terminates for real world event logs.

4. Conclusion

The thesis described in this paper contributed through the field of process mining by giving a greater understanding of privacy issues and through anonymization techniques that preserve utility for common process mining tasks. Furthermore, it introduced leading anonymization algorithms for event logs. Therefore, the thesis enables the analysis of event logs that so far have been considered to sensitive. The rising numbers of papers [17, 18, 19, 20, 21] concerned with privacy-preserving process mining shows how this thesis opened up an important topic of research to the community.

References

[1] W. Van Der Aalst, A. Adriansyah, A. K. A. De Medeiros, F. Arcieri, T. Baier, T. Blickle, J. C. Bose, P. Van Den Brand, R. Brandtjen, J. Buijs, et al., Process mining manifesto, in: Business Process Management Workshops: BPM 2011 International Workshops, Clermont-Ferrand, France, August 29, 2011, Revised Selected Papers, Part I 9, Springer, 2012, pp. 169–194.

[2] E. Rojas, J. Munoz-Gama, M. Sepúlveda, D. Capurro, Process mining in healthcare: A literature review, Journal of biomedical informatics 61 (2016) 224–236.

[3] S. A. Fahrenkrog-Petersen, Anonymization techniques for privacy-preserving process mining (2023).

[4] S. N. von Voigt, S. A. Fahrenkrog-Petersen, D. Janssen, A. Koschmider, F. Tschorsch, F. Mannhardt, O. Landsiedel, M. Weidlich, Quantifying the re-identification risk of event logs for process mining - empiricial evaluation paper, in: S. Dustdar, E. Yu, C. Salinesi, D. Rieu, V. Pant (Eds.), Advanced Information Systems Engineering - 32nd International Conference, CAiSE 2020, Grenoble, France, June 8-12, 2020, Proceedings, volume 12127 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 252–267. URL: https://doi.org/10.1007/978-3-030-49435-3_16. doi:10.1007/978-3-030-49435-3_16.

[5] K. Maatouk, F. Mannhardt, Quantifying the re-identification risk in published process models, in: J. Munoz-Gama, X. Lu (Eds.), Process Mining Workshops - ICPM 2021 International Workshops, Eindhoven, The Netherlands, October 31 - November 4, 2021, Revised Selected Papers, volume 433 of *Lecture Notes in Business Information Processing*, Springer, 2021, pp. 382–394. URL: https://doi.org/10.1007/978-3-030-98581-3_28. doi:10.1007/978-3-030-98581-3_28.

[6] G. Elkoumy, S. A. Fahrenkrog-Petersen, M. F. Sani, A. Koschmider, F. Mannhardt, S. N. von Voigt, M. Rafiei, L. von Waldthausen, Privacy and confidentiality in process mining: Threats and research challenges, ACM Trans. Manag. Inf. Syst. 13 (2022) 11:1–11:17. URL: https://doi.org/10.1145/3468877. doi:10.1145/3468877.

[7] H. Kirchmann, S. A. Fahrenkrog-Petersen, M. Kabierski, H. van der Aa, M. Weidlich, Privacy-preserving process mining with pm4py (extended abstract), in: M. Hassani, A. Koschmider, M. Comuzzi, F. M. Maggi, L. Pufahl (Eds.), Proceedings of the ICPM Doctoral Consortium and Demo Track 2022 co-located with 4th International Conference on Process Mining (ICPM 2022), Bolzano, Italy, October, 2022, volume 3299 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 85–89. URL: https://ceur-ws.org/Vol-3299/Paper18.pdf.

[8] S. A. Fahrenkrog-Petersen, M. Kabierski, F. Rösel, H. van der Aa, M. Weidlich, Sacofa: Semantics-aware control-flow anonymization for process mining, in: C. D. Ciccio, C. D. Francescomarino, P. Soffer (Eds.), 3rd International Conference on Process Mining, ICPM 2021, Eindhoven, The Netherlands, October 31 - Nov. 4, 2021, IEEE, 2021, pp. 72–79. URL: https://doi.org/10.1109/ICPM53251.2021.9576857. doi:10.1109/ICPM53251.2021.9576857.

[9] S. A. Fahrenkrog-Petersen, M. Kabierski, H. van der Aa, M. Weidlich, Semantics-aware mechanisms for control-flow anonymization in process mining, Inf. Syst. 114 (2023) 102169. URL: https://doi.org/10.1016/j.is.2023.102169. doi:10.1016/j.is.2023.102169.

[10] C. Dwork, Differential privacy, in: International colloquium on automata, languages, and

programming, Springer, 2006, pp. 1–12.

[11] F. Mannhardt, A. Koschmider, N. Baracaldo, M. Weidlich, J. Michael, Privacy-preserving process mining: Differential privacy for event logs, Business & Information Systems Engineering 61 (2019) 595–614.

[12] S. A. Fahrenkrog-Petersen, H. van der Aa, M. Weidlich, PRIPEL: privacy-preserving event log publishing including contextual information, in: D. Fahland, C. Ghidini, J. Becker, M. Dumas (Eds.), Business Process Management - 18th International Conference, BPM 2020, Seville, Spain, September 13-18, 2020, Proceedings, volume 12168 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 111–128. URL: https://doi.org/10.1007/978-3-030-58666-9_7. doi:10.1007/978-3-030-58666-9_7.

[13] S. A. Fahrenkrog-Petersen, H. van der Aa, M. Weidlich, PRETSA: event log sanitization for privacy-aware process discovery, in: International Conference on Process Mining, ICPM 2019, Aachen, Germany, June 24-26, 2019, IEEE, 2019, pp. 1–8. URL: https://doi.org/10.1109/ICPM.2019.00012. doi:10.1109/ICPM.2019.00012.

[14] S. A. Fahrenkrog-Petersen, H. van der Aa, M. Weidlich, Optimal event log sanitization for privacy-preserving process mining, Data Knowl. Eng. 145 (2023) 102175. URL: https://doi.org/10.1016/j.datak.2023.102175. doi:10.1016/j.datak.2023.102175.

[15] L. Sweeney, k-anonymity: A model for protecting privacy, International journal of uncertainty, fuzziness and knowledge-based systems 10 (2002) 557–570.

[16] N. Li, T. Li, S. Venkatasubramanian, t-closeness: Privacy beyond k-anonymity and l-diversity, in: 2007 IEEE 23rd international conference on data engineering, IEEE, 2006, pp. 106–115.

[17] R. Hildebrant, S. A. Fahrenkrog-Petersen, M. Weidlich, S. Ren, PMDG: privacy for multi-perspective process mining through data generalization, in: M. Indulska, I. Reinhartz-Berger, C. Cetina, O. Pastor (Eds.), Advanced Information Systems Engineering - 35th International Conference, CAiSE 2023, Zaragoza, Spain, June 12-16, 2023, Proceedings, volume 13901 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 506–521. URL: https://doi.org/10.1007/978-3-031-34560-9_30. doi:10.1007/978-3-031-34560-9_30.

[18] M. Kabierski, S. A. Fahrenkrog-Petersen, M. Weidlich, Hiding in the forest: Privacy-preserving process performance indicators, Inf. Syst. 112 (2023) 102127. doi:10.1016/J.IS.2022.102127.

[19] M. Schulze, Y. Zisgen, M. Kirschte, E. Mohammadi, A. Koschmider, Differentially private inductive miner, CoRR abs/2407.04595 (2024). URL: https://doi.org/10.48550/arXiv.2407.04595. doi:10.48550/ARXIV.2407.04595. arXiv:2407.04595.

[20] J. Cao, C. Wang, W. Guan, S. Qian, H. Zhao, Remaining time prediction for collaborative business processes with privacy preservation, in: F. Monti, S. Rinderle-Ma, A. R. Cortés, Z. Zheng, M. Mecella (Eds.), Service-Oriented Computing - 21st International Conference, ICSOC 2023, Rome, Italy, November 28 - December 1, 2023, Proceedings, Part II, volume 14420 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 38–53.

[21] M. Rafiei, F. Wangelik, M. Pourbafrani, W. M. P. van der Aalst, Travag: Differentially private trace variant generation using gans, in: S. Nurcan, A. L. Opdahl, H. Mouratidis, A. Tsohou (Eds.), RCIS 2023, Corfu, Greece, May 23-26, 2023, Proceedings, volume 476 of *Lecture Notes in Business Information Processing*, Springer, 2023, pp. 415–431. doi:10.1007/978-3-031-33080-3_25.