Practical Aspect of Privacy-Preserving Data Publishing in Process Mining*

Majid Rafiei[0000−0001−7161−6927] and Wil M.P. van der Aalst[0000−0002−0955−6940]

Chair of Process and Data Science, RWTH Aachen University, Aachen, Germany

Abstract. Process mining techniques such as process discovery and conformance checking provide insights into actual processes by analyzing event data that are widely available in information systems. These data are very valuable, but often contain sensitive information, and process analysts need to balance confidentiality and utility. Privacy issues in process mining are recently receiving more attention from researchers which should be complemented by a tool to integrate the solutions and make them available in the real world. In this paper, we introduce a Python-based infrastructure implementing state-of-the-art privacy preservation techniques in process mining. The infrastructure provides a hierarchy of usages from single techniques to the collection of techniques, integrated as web-based tools. Our infrastructure manages both standard and non-standard event data resulting from privacy preservation techniques. It also stores explicit privacy metadata to track the modifications applied to protect sensitive data.

Keywords: Responsible process mining · Privacy preservation · Process mining · Event data

1 Introduction

Process mining provides fact-based insights into actual business processes using event data, which are often stored in the form of event logs. The three basic types of process mining are *process discovery*, *conformance checking*, and *process enhancement* [1]. An event log is a collection of events, and each event is described by its attributes. The main attributes required for process mining are *case id*, *activity*, *timestamp*, and *resource*. Some of the event attributes may refer to individuals, e.g., in the health-care context, the *case id* attribute may refer to the patients whose data are recorded, and the *resource* attribute may refer to the employees performing activities for the patients, e.g., nurses or surgeons.

Privacy issues in process mining are highlighted when the individuals' data are included in the event logs. According to the regulations such as the European General Data Protection Regulation (GDPR) [12], organizations are compelled to take the privacy of individuals into account while analyzing their data. The

necessity of responsibly analyzing private data has recently resulted in more attention for privacy issues in process mining [10,6,4,5]. In [2], the authors introduce a web-based tool, ELPaaS, implementing the privacy preservation techniques introduced in [4] and [5]. ELPaaS gets the required parameters from users and provides results, as CSV files, in email addresses of the users.

Figure 1 shows the general approach of privacy in process mining including two main activities: *Privacy-Preserving Data Publishing* (PPDP) and *Privacy-Preserving Process Mining* (PPPM). PPDP aims to hide the identity and the sensitive data of record owners in event data to protect their privacy. PPPM aims to extend traditional process mining algorithms to work with the non-standard data resulting from some PPDP techniques. Note that PPPM algorithms are tightly coupled with the corresponding PPDP techniques.

In this paper, we introduce a tool which mainly focuses on PPDP and offers state-of-the-art privacy preservation techniques including the *connector*

Fig. 1: The general approach of privacy in process mining.

method for securely discovering processes [9,10], the *decomposition method* for privacy-aware role mining [6], and *TLKC-privacy model* for process mining [8]. The *privacy metadata* proposed in [7] are also embedded in the offered privacy preservation techniques. Moreover, privacy in the context of process mining is presented through PM4Py-WS (PMTK) [3] with a web-based interface which is a particular example to show that the provided privacy preservation techniques can be added to the existing process mining tools for supporting PPPM.

The remainder of the paper is organized as follows. In Section 2, we demonstrate the functionality and characteristics of the tool. Section 3 outlines the maturity and availability of the tool, and Section 4 concludes the paper.

2   Functionality and Characteristics

In this section, we demonstrate the main functionalities and characteristics of our stand-alone web-based tool, PPDP-PM, which is written in Python using *Django* framework[1]. Our tool has four main modules: *event data management*, *privacy-aware role mining*, *connector method*, and *TLKC-privacy*. The *event data management* module has two tabs to upload and manage the event data that could be standard XES event logs[2] or non-standard event data, called *Event Log Abstraction* (ELA) [7]. In this module, an event log can be set as the input
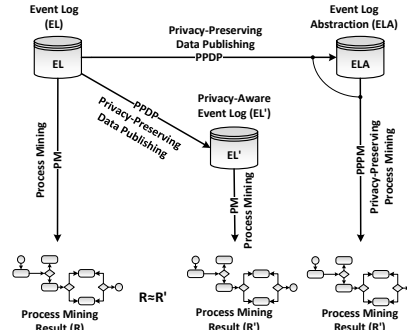
---

[1] https://www.djangoproject.com/
[2] http://www.xes-standard.org/

Fig. 2: The privacy-aware role mining page in PPDP-PM.

for the privacy preservation techniques. The *privacy-aware role mining* module (Figure 2) implements the decomposition method supporting three different techniques: *fixed-value*, *selective*, and *frequency-based* [6]. After applying a technique, the privacy-aware event log in the XES format is provided in the corresponding "Outputs" section. The generated event log preserves the data utility for mining roles from *resources* without exposing who performs what.

The *connector method* implements an encryption-based method for discovering directly follows graphs [9,10]. It breaks the traces down into the collection of directly-follows relations which are securely stored in a data structure. After applying the method, the privacy-aware event data are provided in the corresponding "Outputs" section as an XML file with the ELA format [7]. The $TLKC$-*privacy* module implements the $TLKC$-privacy model for process mining [8] that provides group-based privacy guarantees assuming four types of background knowledge: *set*, *multiset*, *sequence*, and *relative*. T refers to the accuracy of timestamps in the privacy-aware event log, L refers to the power of background knowledge, K refers to the k in the k-anonymity definition [11], and C refers to the bound of confidence regarding the sensitive attribute values in an equivalence class. Applying this method results in a privacy-aware event log in the XES format that preserves data utility for process discovery and performance analysis. We also provide the same privacy preservation techniques in the context of an open-source process mining tool. Figure 3 shows a snippet of the home page of the privacy integration in PMTK where process mining algorithms can directly be applied to the privacy-aware event data.

Each privacy preservation technique in the tool is implemented as a *Django application* that enables the simultaneous running of different techniques on an event log. This architecture makes the whole project easy to maintain, and new techniques can simply be integrated as independent applications. The outputs for the privacy preservation techniques are provided independently for each technique and can be downloaded or stored in the event data repository. PPDP-PM is designed in a way that provides a cycle of privacy preservation techniques, i.e., the privacy-aware event data, added to the event data repository, can be set as the input for the techniques again as long as they are in the form of standard XES event logs. To keep the process analysts aware of the modifications applied
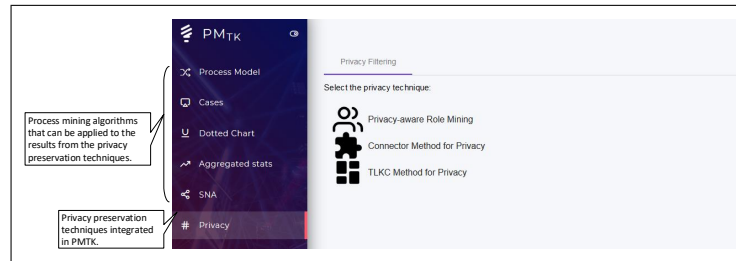
Fig. 3: The home page of the privacy integration in PM4Py-WS (PMTK).

to the privacy-aware event logs, the *privacy metadata* [7] specify the order of the applied privacy preservation techniques. Moreover, the tool follows a naming approach to uniquely identify the privacy-aware event data based on name of the technique, the creation time, and name of the event log.

3   Availability and Maturity

As mentioned, PPDP-PM is a web-based application written in Python. The source code, a screencast, and other information are available in a GitHub repository: `https://github.com/m4jidRafiei/PPDP-PM`. The privacy preservation techniques, explained in Section 2, and the integration into PMTK are also available as separate GitHub repositories.[3] To facilitate the usage and integration of the privacy preservation techniques, they are also published as standard Python packages (https://pypi.org/): *pp-role-mining*, *p-connector-dfg*, *p-tlkc-privacy*, and *p-privacy-metadata*. Our infrastructure provides a hierarchy of usages such that users can use each technique independently, they can use PPDP-PM which integrates a set of privacy preservation techniques as a standalone web-based application, and they can also use the provided techniques in a process mining tool where the privacy preservation techniques are integrated. The scalability of the tool varies w.r.t. the privacy preservation technique and the size of the input event log. Based on our experiments, our tool can handle real-world event logs, e.g., the BPI challenge datasets[4]. However, it can still be improved for industry-scale usage. PPDP-PM and its integration in PMTK are also provided as Docker containers which can simply be hosted by the users: `https://hub.docker.com/u/m4jid`.

4   Conclusion

Event data often include highly sensitive information that needs to be considered by process analysts w.r.t. the regulations. In this paper, we introduced a Python-based infrastructure for dealing with privacy issues in process mining. A web-based application was introduced implementing privacy-preserving data

---

[3] https://github.com/m4jidRafiei/

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

publishing techniques in process mining. We also showed the privacy integration in PMTK as an open-source web-based process mining tool. The infrastructure was designed in such a way that other privacy preservation techniques can be integrated. We plan to cover different perspectives of privacy and confidentiality issues in process mining, and novel techniques are supposed to be integrated into the introduced framework. We also invite other researchers to integrate their solutions as independent applications in the provided framework.

References

1. van der Aalst, W.M.P.: Process Mining - Data Science in Action, Second Edition. Springer (2016). https://doi.org/10.1007/978-3-662-49851-4
2. Bauer, M., Fahrenkrog-Petersen, S.A., Koschmider, A., Mannhardt, F., van der Aa, H., Weidlich, M.: Elpaas: Event log privacy as a service. In: Proceedings of the Dissertation Award, Doctoral Consortium, and Demonstration Track at BPM 2019 (2019)
3. Berti, A., van Zelst, S.J., van der Aalst, W.M.P.: Pm4py web services: Easy development, integration and deployment of process mining features in any application stack. In: Proceedings of the Dissertation Award, Doctoral Consortium, and Demonstration Track at BPM 2019 (2019)
4. Fahrenkrog-Petersen, S.A., van der Aa, H., Weidlich, M.: PRETSA: event log sanitization for privacy-aware process discovery. In: International Conference on Process Mining, ICPM 2019, Aachen, Germany (2019)
5. Mannhardt, F., Koschmider, A., Baracaldo, N., Weidlich, M., Michael, J.: Privacy-preserving process mining - differential privacy for event logs. Business & Information Systems Engineering **61**(5), 595–614 (2019)
6. Rafiei, M., van der Aalst, W.M.P.: Mining roles from event logs while preserving privacy. In: Business Process Management Workshops - BPM 2019 International Workshops, Vienna, Austria. pp. 676–689 (2019)
7. Rafiei, M., van der Aalst, W.M.P.: Privacy-preserving data publishing in process mining. In: Business Process Management Forum - BPM Forum 2020, Sevilla, Spain, September 13-18, 2020, Proceedings (2020)
8. Rafiei, M., Wagner, M., van der Aalst, W.M.P.: TLKC-privacy model for process mining. In: 14th International Conference on Research Challenges in Information Science, RCIS 2020 (2020)
9. Rafiei, M., von Waldthausen, L., van der Aalst, W.M.P.: Ensuring confidentiality in process mining. In: Proceedings of the 8th International Symposium on Data-driven Process Discovery and Analysis (SIMPDA 2018), Seville, Spain (2018)
10. Rafiei, M., von Waldthausen, L., van der Aalst, W.M.P.: Supporting condentiality in process mining using abstraction and encryption. In: Data-Driven Process Discovery and Analysis - 8th IFIP WG 2.6 International Symposium, SIMPDA 2018, and 9th International Symposium, SIMPDA 2019, Revised Selected Papers (2019)
11. Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **10**(05), 557–570 (2002)
12. Voss, W.G.: European union data privacy law reform: General data protection regulation, privacy shield, and the right to delisting. Business Lawyer **72**(1) (2016)