# **Providing Privacy Guarantees in Process Mining**

Stephan A. Fahrenkrog-Petersen

Humboldt-Universitt zu Berlin, Berlin, Germany stephan.fahrenkrog-petersen@hu-berlin.de

**Abstract.** Information systems record data while executing business processes. This data can be analyzed, by process mining, to gain knowledge about the business processes underlying the information systems. Data recorded by the information systems is often personal data belonging to individuals such as customers or process workers. Such data has become a strong focus of recent regulations like the GDPR. These new legal developments force organizations that process personal data to ensure a certain level of privacy. Unlike in other fields of data science, in the field of process mining there are no existing solutions to guarantee such privacy.

This research aims to provide such solutions that enable organizations to do process mining while giving privacy guarantees to individuals, such as employees, that contribute their data. In this work, we present privacy challenges in the area of process mining and outline privacy guarantees we aim to provide for process mining. We want to follow the design science paradigm to achieve our goals. We describe our preliminary results, an algorithm, called PRETSA, to sanitize event logs for privacy-aware process discovery and show the next steps we want to take in our research.

Keywords: Process Mining · Privacy · Privacy-aware Data Mining

# 1 Introduction

Information systems, like ERP or CRM systems, are used to execute business processes. While doing so, these system record data. Process mining [22] allows an organization to utilize this data to produce insights into the processes of the organization. Data records of information systems are called event logs and contain personal information of individuals involved in the underlying business process, e.g. about process worker, and customers. Personal data is protected by privacy legislation, such as the GDPR in the European Union, the Health insurance Portability and Accountability Act [3] in the United States, or the Personal Information Protection and Electronic Documents Act [17] in Canada.

Besides those legal requirements, there is also a motivation from a business point of view to ensure privacy. Violations against privacy regulations can result in expensive fines, up to 4% of the annual revenue of a company [10], and have negative impact on the market value of a company [2]. Therefore, it makes sense from a business point of view to invest in privacy-enhancing technologies to minimize the risk for the business.

If an organization wants to ensure the protection of personal data, it may strive for compliance, to certain privacy guarantees [26], e.g. k-anonymity [20] or differential privacy [7]. Much research has aimed to providing such guarantees in areas like machine learning or sequence mining. However, techniques to ensure such guarantees do not yet exists for process mining. Since privacy-enhancing technologies come with a utility loss [6], it is beneficiary to customize these technologies for each application, to preserve as much utility as possible. For this reason, we plan to develop techniques suited for process mining. We outline our research plans in more detail later in this paper.

The remainder of the paper is structured as follows: In Section 2 we give an overview about related lines of research and describe the existing privacy notions we plan to build on. In Section 3 we explain in detail what research problems we want to answer. Our research methodology is explained in Section 4. We outline the results we already achieved so far and our short term plan in Section 5.

# 2 Related Work and Background

In the area of privacy, notions [26] such as k-anonymity [20] or differential privacy [7] are used to guarantee a certain level of privacy. These guarantees can either limit the information an adversary can gain from an attack or bound the chance of success that an attack will succeed. To achieve these goals, kanonymity, for example, gives a lower bound of entries in a data set that need to have the same values for their identifying attributes. In this way, k-anonymity limits the chance of relating one entry in a data set to a specific individual. The approach of k-anonymity and its extending concepts, e.g. l-diversity [14] or t-closeness [13], are used in the context of data publishing. The mentioned extensions of k-anonymity provide additional protection against information release about so called sensitive attributes of an individual. Differential privacy, on the other hand, guarantees an upper bound of privacy impact of a query, that is evaluated on a data set. This goal is achieved by adding noise to the data. The goal of differential privacy is to make it impossible to determine if the data of a certain individual is part of the data set. Differential privacy was widely adapted by industry, e.g., by Apple [21]. While k-anonymity and its enhancements are constraints on the data set, differential privacy is usually a constraint applied to a query that is evaluated over the data set [9]. A stronger privacy guarantee, for either k-anonymity or differential privacy, usually comes with higher utility loss [6], i.e., it lowers the ability to answer some analysis question based on the data.

Recently, the problem of privacy in the context of process mining was discussed in [16] by Mannhardt et al. The paper introduced a framework that explains challenges arising from the GDPR for the design of process mining systems. A major contribution of the paper was the discussion of the *primary use* and the *secondary use* in the context of process mining. In the case of business processes, customers usually agree to the usage of their data e.g. for executing the business process, but they usually never agreed to the secondary use of their data, e.g., for process mining. However, the framework by Mannhardt et al. [16] did not provide any techniques to ensure privacy for process mining. We plan to fill that gap with our work.

A related concept to privacy is the concept of confidentiality. In [1], this concept was described in the context of process mining. In [18], an approach based on encryption to achieve confidentiality was introduced. However, no guarantee for confidentiality was given. The concept of confidentiality also differs from privacy, since some information might be confidential but not relevant in terms of privacy. An example are aggregated performance information about a business process. Such performance information might be useful to a competitor and an organization therefore might want to protect them. However, aggregated information itself is usually not critical from a privacy point of view, as long as it not possible to link such data to an individual. While confidentiality is a nice-to-have, privacy is often a must due to legal regulations, like the GDPR [24]. However, even if the goals of confidentiality and privacy differ, some future techniques in both areas might also be applicable to the other area.

# 3 Research Aim



Fig. 1. Approaches to ensure privacy in Process Mining

Our overall goal is to provide privacy guarantees for process mining by preserving as much utility as possible. We think it is possible to provide the guarantees in two approaches, as visualized in Figure 1. These two different approaches can be described as follows:

**Event Log Sanitization** One way of ensuring privacy is by preprocessing an event log in such a way, that the event log itself guarantees certain levels of

privacy e.g. k-anonymity. As such, ideas from privacy-aware data publishing [20] are adapted for event logs. We call the event log that is generated by this preprocessing step a *sanitized event log*. Such a *sanitized event log* can be used as input data for existing process mining techniques. Event log sanitization would also allow an organization to share its event log with another organization, e.g., a consulting firm, while at the same time being able to provide privacy guarantees to the individuals involved in the process.

**Privatized Process Mining** Alternatively, it is possible to develop new process mining techniques, that guarantee a certain level of privacy for the generated process mining artifacts. This can for example be done, by using queries on the event data that fulfill differential privacy. If an algorithm is changed in a way that it only uses such queries, the resulting artifact of the algorithm also guarantees differential privacy.

The two possible approaches come with specific advantages and disadvantages that we list in Table 1.

	Event Log Sanitization	Privatized Process Min- ing
Is a privacy guarantees provided for the process mining artifact?		Yes
Is a privacy guarantees provided for raw event data?		No
Is it compatible with known Process Mining techniques?		Depends on the implemen- tation/technique, could be a completely new technique or could enhance known tech- niques
Can the event data be outsourced to another party and privacy is still guaranteed?		No
Will the utility be pre- served?	Depends on technique	Depends on technique, but we assume more utility as can be preserved with <i>event</i> <i>log sanitization</i>
	processing is necessary, this	- Depends on technique, but s an increase in runtime for v each technique is quite likely

Table 1. Comparison of approaches to ensure privacy in Process Mining

We plan to develop unique techniques for each of the process mining sub fields listed below, because we think it is necessary to tackle each sub field with their own approaches to maintain as much utility as possible for each sub field. Since each sub field has it own definition of utility we need different approaches to achieve that goal.

Usually process mining is structured in the sub fields *process discovery, con*formance checking, and enhancement. In process discovery [5], models that describe the business process are automatically generated from the event data. Sensitive information, like performance information, is also often part of the input data and displayed in the resulting models. The information about each instance of a process itself is also related to one specific individual and is therefore sensitive information. These examples clearly show that it is necessary to ensure privacy in the context of process discovery, to protect such information.

In conformance checking the goal is to check if an event log complies to a known process model. We think it is not useful to develop privacy techniques for conformance checking, because the whole point of conformance checking is to identify process instances that differ from the standard. Therefore it would not be meaningful to hide unusual information.

For enhancement, the extraction of additional information from an event log, we want to highlight the sub fields predictive process monitoring [15], i.e. the construction of models to predict properties of running process instances, and queue mining [19], the analysis of queueing effects in resource driven business processes. For both fields, it is necessary to process events individually to give an online prediction. Therefore, it would not be enough to just sanitize an event log. Instead, it must also be sanitize to privatize individual events. Hence, the requirements to meet by any sanitization technique are different from those imposed by process discovery.

## 4 Research Approach

Our approach to develop solutions for the problem space mentioned above will be based on the methodology of *design science* [25]. This means that we will build prototypes of our proposed solutions and then evaluate these prototypes in an experimental setup. In these experiments, we will evaluate our approaches on real-world event logs to show applicability and usefulness, and on synthetic data to study scalability and sensitivity of our techniques.

To complement our experimental evaluations, we plan to conduct case studies with organizations to test our developed solutions in practice. Case studies would allow us to assess, if our solutions are feasible in a real-world setting. We would also be able to examine, if we solved the most pressing privacy issues of our industry partners. Therefore, these case studies might lead to directions for further research.

Finally we want our techniques to be available for others to use. We plan to make our implementations available as open source code. Additionally, we plan to integrate our approaches into existing process mining solutions, like ProM [23] or Apromore [11] and provide event log sanitization as a service on a website.

# 5 Preliminary Results

In this section, we introduce our results achieved so far and the tasks we are currently working on.

#### 5.1 Providing k-Anonymity and t-Closeness for Process Discovery

Our first contribution is an algorithm that provides a privacy guarantee for process discovery, called PRETSA [8]. It provides k-anonymity and t-closeness guarantees, by a sanitization of an event log. Our algorithm works on a prefix-tree based representation of the event log and modifies the prefix-tree until the privacy constrain is fulfilled. The resulting event log can be used to generate an annotated process model, e.g., a model with performance information. We aimed to preserve as much utility as possible for process discovery with PRETSA. Based on experiments with the *inductive miner* [12] on three real-world event logs we showed that PRETSA preserves more utility than a baseline. We even can provide an event log with reasonable utility for settings in which a baseline fails to provide any sanitized event log.

As mentioned earlier it is desirable to conducts a case study to test our approaches. We already reached out to two organizations to check if they would be interested to test *PRETSA* in a case study.

We made our PRETSA implementation available as a stand-alone python program on Github<sup>1</sup> under the MIT licence. Next, we plan to integrate it in a process mining solution.

#### 5.2 Providing Differential Privacy for Process Discovery

We started working on a mechanism to provide  $(\epsilon, \delta)$ -differential privacy for process discovery. We plan to provide these guarantees for a query that returns the directly follows relationships of an event log, since directly follows relations are widely used by process discovery techniques. Our research aims to build a specialized noise function for such queries that provides noise with low utility loss for process discovery.

## 5.3 Predictive Process Monitoring

*Predictive process monitoring* is a sub field of process mining that aims to predict future outcomes of ongoing cases, like the remaining time of the case or the next event. This field is an applies machine learning to achieve it goals. In the field of machine learning various work to achieve privacy guarantees exists [4] provides

<sup>&</sup>lt;sup>1</sup> https://github.com/samadeusfp/PRETSA

an broad overview about privacy issues and their solution in the area of machine learning.

Moreover, our current work considers privacy and confidentiality issues in the area of *predictive process monitoring*. We plan to provide an overview about these issues and set them in context with issues and solutions for privacy and confidentiality in the context of machine learning in general.

# 6 Conclusion

In this work, we explained the importance of privacy guarantees for the field of process mining. We outlined related lines of research and our own research plans to achieve such privacy guarantees for process discovery, predictive process monitoring, and queue mining. We explained that we will conduct this research based on the design science methodology. Our preliminary results include an algorithm, PRETSA, ensure k-anonymity and t-closeness for process discovery by preprocessing an event log. We also gave an overview about our current work and future plans.

# References

- van der Aalst, W.M.P.: Responsible data science: Using event data in a "people friendly" manner. In: Enterprise Information Systems - 18th International Conference, ICEIS 2016, Rome, Italy, April 25-28, 2016, Revised Selected Papers. pp. 3–28 (2016). https://doi.org/10.1007/978-3-319-62386-3\_1, https://doi.org/10.1007/978-3-319-62386-3\_1
- 2. Acquisti, A., Friedman, A., Telang, R.: Is there a cost to privacy breaches? an event study. ICIS 2006 Proceedings p. 94 (2006)
- Act, A.: Health insurance portability and accountability act of 1996. Public law 104, 191 (1996)
- Al-Rubaie, M., Chang, J.M.: Privacy preserving machine learning: Threats and solutions. arXiv preprint arXiv:1804.11238 (2018)
- Augusto, A., Conforti, R., Dumas, M., La Rosa, M., Maggi, F.M., Marrella, A., Mecella, M., Soo, A.: Automated discovery of process models from event logs: Review and benchmark. IEEE Transactions on Knowledge and Data Engineering (2018)
- Brickell, J., Shmatikov, V.: The cost of privacy: destruction of data-mining utility in anonymized data publishing. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 70–78. ACM (2008)
- 7. Dwork, C.: Differential privacy: A survey of results. In: International Conference on Theory and Applications of Models of Computation. pp. 1–19. Springer (2008)
- Fahrenkrog-Petersen, S.A., van der Aa, H., Weidlich, M.: Pretsa: Event log sanitization for privacy-aware process discovery. IEEE International Conference of Process Mining (2019)
- 9. Holohan, N., Antonatos, S., Braghin, S., Mac Aonghusa, P.:  $(k, \epsilon)$ -anonymity: k-anonymity with  $\epsilon$ -differential privacy. arXiv preprint arXiv:1710.01615 (2017)

- 10. Houser, K., Voss, W.G.: Gdpr: The end of google and facebook or a new paradigm in data privacy? (2018)
- La Rosa, M., Reijers, H.A., Van Der Aalst, W.M., Dijkman, R.M., Mendling, J., Dumas, M., García-Bañuelos, L.: Apromore: An advanced process model repository. Expert Systems with Applications 38(6), 7029–7040 (2011)
- Leemans, S.J., Fahland, D., van der Aalst, W.M.: Discovering block-structured process models from event logs-a constructive approach. In: International conference on applications and theory of Petri nets and concurrency. pp. 311–329. Springer (2013)
- Li, N., Li, T., Venkatasubramanian, S.: t-closeness: Privacy beyond k-anonymity and l-diversity. In: 2007 IEEE 23rd International Conference on Data Engineering. pp. 106–115. IEEE (2007)
- Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M.: l-diversity: Privacy beyond k-anonymity. In: 22nd International Conference on Data Engineering (ICDE'06). pp. 24–24. IEEE (2006)
- Maggi, F.M., Di Francescomarino, C., Dumas, M., Ghidini, C.: Predictive monitoring of business processes. In: International conference on advanced information systems engineering. pp. 457–472. Springer (2014)
- Mannhardt, F., Petersen, S.A., Oliveira, M.F.: Privacy challenges for process mining in human-centered industrial environments. In: 2018 14th International Conference on Intelligent Environments (IE). pp. 64–71. IEEE (2018)
- 17. Parliament, C.: Personal information protection and electronic documents act. Consolidated Acts, SC 2000, c 5, 13 (2000)
- Rafiei, M., von Waldthausen, L., van der Aalst, W.M.P.: Ensuring confidentiality in process mining. In: Proceedings of the 8th International Symposium on Datadriven Process Discovery and Analysis (SIMPDA 2018), Seville, Spain, December 13-14, 2018. pp. 3–17 (2018), http://ceur-ws.org/Vol-2270/paper1.pdf
- Senderovich, A., Weidlich, M., Gal, A., Mandelbaum, A.: Queue mining for delay prediction in multi-class service processes. Information Systems 53, 278–295 (2015)
- Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(05), 557–570 (2002)
- Tang, J., Korolova, A., Bai, X., Wang, X., Wang, X.: Privacy loss in apple's implementation of differential privacy on macos 10.12. arXiv preprint arXiv:1709.02753 (2017)
- 22. Van Der Aalst, W.: Process mining: discovery, conformance and enhancement of business processes, vol. 2. Springer (2011)
- Van Dongen, B.F., de Medeiros, A.K.A., Verbeek, H., Weijters, A., van Der Aalst, W.M.: The prom framework: A new era in process mining tool support. In: International conference on application and theory of petri nets. pp. 444–454. Springer (2005)
- 24. Voigt, P., Von dem Bussche, A.: The eu general data protection regulation (gdpr). A Practical Guide, 1st Ed., Cham: Springer International Publishing (2017)
- Von Alan, R.H., March, S.T., Park, J., Ram, S.: Design science in information systems research. MIS quarterly 28(1), 75–105 (2004)
- Wagner, I., Eckhoff, D.: Technical privacy metrics: a systematic survey. ACM Computing Surveys (CSUR) 51(3), 57 (2018)