Quantifying the Re-identification Risk of Event Logs for Process Mining

Empiricial Evaluation Paper

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], and Matthias Weidlich[2]

[1] Technische Universität Berlin, Berlin, Germany
{saskia.nunezvonvoigt,florian.tschorsch}@tu-berlin.de
[2] Humboldt-Universität zu Berlin, Berlin, Germany
{stephan.fahrenkrog-petersen,matthias.weidlich}@hu-berlin.de
[3] Kiel University, Kiel, 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

Abstract. Event logs recorded during the execution of business processes constitute a valuable source of information. Applying process mining techniques to them, event logs may reveal the actual process execution and enable reasoning on quantitative or qualitative process properties. However, event logs often contain sensitive information that could be related to individual process stakeholders through background information and cross-correlation. We therefore argue that, when publishing event logs, the risk of such re-identification attacks must be considered. In this paper, we show how to quantify the re-identification risk with measures for the individual uniqueness in event logs. We also report on a large-scale study that explored the individual uniqueness in a collection of publicly available event logs. Our results suggest that potentially up to all of the cases in an event log may be re-identified, which highlights the importance of privacy-preserving techniques in process mining.

1 Introduction

Process mining uses data recorded in the form of event logs by information systems to, for example, reveal the actual execution of business processes [1]. Since most activities in modern organization are supported by technology, each process execution produces a digital footprint indicating the occurrence and timing of activities. Consequentially, event logs may contain sensitive information and are vulnerable to adversarial attacks. Unfortunately, there is no general method

S. Dustdar et al. (Eds.): CAiSE 2020, LNCS 12127, pp. 252–267, 2020.
https://doi.org/10.1007/978-3-030-49435-3_16

on how to safely remove personal and sensitive references. Since the existence of privacy threats are generally known, the willingness to publish event logs is low. Publicly available event logs, however, are necessary to evaluate process mining models [2–4] and therefore discussions are needed on how to safely publish event logs. Against this background, we argue that it is crucial to understand the risk of data re-identification in event logs and process mining. With this insight, we can balance how much information of an event log can be shared and how much should be anonymized to preserve privacy. While many examples confirm the general risk of data re-identification [5–7], the re-identification risk of event logs has not received much attention yet.

The intention of this paper is to raise awareness to the re-identification risk of event logs and therefore provide measures to quantify this risk. To this end, we provide an approach to express the *uniqueness* of data, which is derived from models that are commonly adopted by process mining techniques. Each event recorded in an event log consists of specific data types, such as the activity name of the respective process step, the timestamp of its execution, and event attributes that capture the context and the parameters of the activity. Additionally, sequences of events that relate to the same case of a process, also known as traces, come with data attributes, so-called case attributes that contain general information about the case. To extract sensitive information, an adversary uses background knowledge to link a target's attributes with the case/event attributes in the event log, e.g., by cross-correlating publicly-available sources. The higher the uniqueness of an event log, the higher an adversary's chances to identify the target. 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.

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 for these logs 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.

The contributions of this paper can be summarized as follows:

- We present an approach to quantify the privacy risk associated to event logs. In this way, we support the identification of information that should be suppressed when publishing an event log, thereby fostering the responsible use of logs and paving the way for novel use cases based on event log analysis.
- By reporting the results of a large-scale evaluation study, we highlight the need to develop privacy-preserving techniques for event logs with high utility for process analytics. Our notions of individual uniqueness may serve as a catalyst for such efforts, since they make the inherent privacy risks explicit.

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

This paper is structured as follows. Section 2 illustrates privacy threats in process mining. Section 3 presents our approach for quantifying the re-identification risk. We analyze publicly available event logs and discuss the results in Sect. 4. We review related work in Sect. 5, before Sect. 6 concludes this paper.

Table 1. Event log example

case id	activity	timestamp	case attributes	event attributes
1000	registration	03/03/19 23:40:32	{age: 26, sex: m}	{arrival: check-in}
1000	triage	03/04/19 00:27:12	{age: 26, sex: m}	{status: uncritical}
1000	liquid	03/04/19 00:47:44	{age: 26, sex: m}	{liquid: NaCl}
...	...	...	...	...
1001	registration	03/04/19 00:01:24	{age: 78, sex: f}	{arrival: ambulance}
1001	antibiotics	03/04/19 00:09:06	{age: 78, sex: f}	{drug: penicillin}
...	...	...	...	...

2 Privacy Threats in Process Mining

Process mining uses event logs to discover and analyze business processes. Event logs capture the execution of activities as events. A finite sequence of such events forms a trace, representing a single process instance (aka case). For example, the treatment of patients in an emergency room includes a number of events, such as blood sampling and analysis, which together follow a certain structure as determined by the process. Accordingly, the events related to an individual patient form a case. In addition, case attributes provide general information about a case, e.g., place of birth of a patient. Each event consists of various data types, such as the name of the respective *activity*, the *timestamp* of the execution, and *event attributes*. Event attributes are event-specific and may be changing over time, e.g., a temperature or the department performing a treatment. The key difference between case attributes and event attributes is that case attributes do *not* change their value for a case during the observed period of time. We show a synthetic event log example capturing an emergency room process in Table 1.

Considering the structure of an event log, several privacy threats are identified. Linking a case to an individual can reveal sensitive 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 [8] or the work schedule of hospital staff. Such kind of staff surveillance is a critical privacy threat. Clearly, it is essential to include privacy considerations in process mining projects. We assume that an adversary's goal is to identify an individual in an event log linking external information. Depending on the type of background information, different adversary models are possible. We assume a targeted re-identification, i.e., an adversary has information about specific individuals, which includes a subset of the attribute values. Based thereon, the adversary aims to reveal sensitive information, e.g., a diagnosis. Here, we assume that an adversary knows that an individual is present in the event log. In this paper, we consider the uniqueness measure to quantify the re-identification risk of sensitive information, thereby providing a basis for managing privacy considerations.

Table 2. Preparation of event log

case id	sex	age	activity	timestamp	arrival channel
10	male	26	[reg., liquid, . . .]	[3/3/19, 3/4/19, . . .]	[check-in, none, . . .]
11	female	78	[reg., antibiotics]	[3/4/19, 3/4/19]	[ambulance, none]
12	female	26	[reg., liquid, . . .]	[3/5/19, 3/7/19, . . .]	[check-in, none, . . .]
. . .	. . .	. . .	. . .	. . .	. . .

3 Re-identifications of Event Logs

To apply our uniqueness measure to cases, we summarize all occurring event data to its corresponding case. This assumption eases handling multiple events belonging to the same case. Since case attributes are invariant over time, they only need to be taken into consideration once, whereas event attributes may be different for every event and therefore their temporal change needs to be considered. Table 2 provides a respective example. Each row in this table belongs to one case. The case attributes “sex” and “age” are listed in separate columns. The columns “activity”, “timestamp”, and “arrival channel” contain an ordered list of the respective attributes. For example, the case id 11 has only two events and therefore two activities. The second activity “antibiotics” on March 04, 2019 has no “arrival channel” (i.e., it is “none”).

The uniqueness of an event log serves as a basis for estimating how likely a case can be re-identified. We investigate a number of so-called projections that can be considered as a data minimization technique, effectively reducing the potential risks of re-identification in an event log. Projections refer to a subset of attributes in the event log. They can easily be adopted to assess the risk in different scenarios. Table 3 summarizes the projections for event logs and their potential usage in process mining. Projection A contains the sequence of all executed activities with their timestamps, while projection F only contains the case attributes. It has been shown that even sparse projections of event logs hold privacy risks [4]. Therefore, in our evaluation, we will consider the re-identification risk for various projections.

3.1 Uniqueness Based on Case Attributes

In addition to unique identifiers (UID), so-called quasi-identifiers are information that can be linked to individuals as well. A combination of quasi-identifiers may be sufficient to create a UID. In event logs, the case attributes can be seen as quasi-identifier. For example, in the event log of the BPI Challenge 2018 [9], the area of all parcels and the ID of the local department can be considered as case attributes. Measuring the uniqueness based on case attributes is a common way to quantify the re-identification risk [10]. Case uniqueness and thus an individual uniqueness highly increases the risk of re-identification. A single value of a case attribute does not lead to identification. The combination with other attributes, however, may lead to a unique case. In particular, when linking attributes to other sources of information, it may result in successful re-identification.

Table 3. Projections of event logs

projection	data included	exemplary usage in process mining
A	activities, timestamps	analysis of bottlenecks
B	activities, event and case attributes	predictive process monitoring
C	activities, event attributes	decision mining
D	activities, case attributes	trace clustering
E	activities	process discovery
F	case attributes	traditional data mining

We define uniqueness as the fraction of unique cases in a event log. Let f_k be the frequency of the kth combination of case attributes values in a sample. One case is unique if $f_k = 1$, i.e., there is no other case with the same values of case attributes. Accordingly, uniqueness for case attributes is defined as

$$U_{\text{case}} = \frac{\sum I(f_k = 1)}{N}, \tag{1}$$

where the indicator function $I(f_k = 1)$ is 1, if the kth combination is unique, and N is the total number of cases in the event log. Referring to our data in Table 2, the attribute value "sex: female" leads to two possible case candidates (id:10 and id:11), i.e., $f_k = 2$, which implies that the combination is not unique. Taking "age" as an additional quasi-identifier into account, makes all three listed cases unique, i.e., $U_{\text{case}} = 1$. Since often a sample of the event log is published, we distinguish between sample uniqueness and population uniqueness. The number of unique cases in the sample is called sample uniqueness. With population uniqueness, we refer to the amount of unique cases in the complete event log (i.e., population). Based on the disclosed event log we can measure the sample uniqueness. The population uniqueness is the number of cases that are unique within the sample and are also unique in the underlying population from which the data has been sampled. Usually the event log is a sample from a population and the original event log is not available. Therefore, the population uniqueness cannot be measured and must be estimated.

There are several models to estimate the population uniqueness from a sample. These methods model the population uniqueness based on extrapolations of the contingency table to fit specific distributions to frequency counts [10]. We adopt the method of Rocher and Hendrickx [7] to estimate the population uniqueness.[2] The authors use Gaussian copulas to model population uniqueness, approximate the marginals from the sample, and estimate the likelihood for a sample unique being a population unique. For this analysis, we assume that the event log is a published sample. By applying the method, we estimate the population uniqueness of cases in terms of their case attributes.

3.2 Uniqueness Based on Traces

Most of the published event logs for process mining do not have many case attributes, only event attributes. For example, the Sepsis event log [11] has only one case attribute ("age"). However, a case can also be unique based on the events. We measure the uniqueness using the traces. A trace consists of an ordered set of activities $a_1, a_2, \dots a_n$, their timestamps $t_1, t_2, \dots t_n$ and l event attributes $e_{11}, \dots e_{ln}$. A tuple $p_j = (a_j, t_j, e_{1j}, \dots, e_{lj})$ represents a point from the trace $[(a_1, t_1, e_{11}, ..., e_{l1}), (a_2, t_2, e_{12}, ..., e_{l2}), \dots, (a_n, t_n, e_{1n}, ..., e_{ln})]$. We assume that an adversary's main goal is to re-identify an individual given a number of points and to reveal other sensitive points. We argue that an adversary has a certain knowledge and knows some points, which she is able to link with the event log. In particular, we assume that an adversary knows that a certain person is contained in the event log. In other words, we consider the published event log as population. As our example in Table 2 shows, even without considering the case attributes, all cases are unique: Case 11 is uniquely identifiable by its second activity "antibiotics". The Cases 10 and 12 are uniquely identified by combining the activity with the respective timestamp. An adversary for example might have information about a patient's arrival (e.g., "check-in: 3/5/19"). Given this information as a point from the trace it is sufficient for an adversary to identify the patient and reveal additional information from the event log.

Accordingly, we express the re-identification risk as the ratio of unique cases. The uniqueness of a trace can be measured similarly to location trajectories [12,13]. In location trajectories, points consist only of a location and a timestamp. In contrast, we have not only two-dimensional but multi-dimensional points with i.a. an activity, a resource, and a timestamp. Let $\{c_i\}_{i=1,\dots,N}$ be the event consisting of a set of N traces. Given a set of m random points, called M_p we compute the number of traces that include the set of points. A trace is unique if the set of points M_p is only contained in a single trace. The uniqueness of traces given M_p is defined as

$$U_{\text{trace}} = \frac{\sum \delta_i}{N}, \tag{2}$$

where $\delta_i = 1$, if a trace is unique $|\{c_i | M_p \subseteq c_i\}| = 1$, otherwise $\delta_i = 0$.

[2] Code available at https://github.com/computationalprivacy.

4 Results

For our evaluation we used the publicly available event logs from the 4TU.Centre for Research Data. We classified the event logs into real-life-individuals (R) and software (S) event logs. The case identifier of real-life-individuals refers to a natural person, e.g., the ADL event log [14] includes activities of daily living activities of individuals. In event logs referring to software activities, events do not directly refer to a natural person, but to technical components. For instance, the BPI Challenge 2013 event log [15] consists of events from an incident management system. Some of the software related event logs even consist of a single case, which makes measuring uniqueness of cases more difficult. However, if a suitable identifier can be linked to the cases, it will also be possible to measure the uniqueness for software related event logs. For example, the incidents in the BPI Challenge 2013 event log are processed by a natural person. By using an appropriate transformation, this natural person could serve as a case identifier.

In the following, we apply our methods to estimate the uniqueness of the real-life-individuals event logs (R) only. We measure the uniqueness of case attributes for event logs with more than one case attribute only. Table 4 summarizes the results of our classification, provides some basic metrics on the number of cases and activities, and indicates the applied uniqueness measures.

For improved readability and for ethical considerations (see Sect. 4.3 for details), we will apply our methods and discuss intermediate results in detail only for the BPI Challenge 2018 [9] and the Sepsis [11] event logs. For all other event logs, we provide condensed and pseudonymized results. Note that the pseudonymized event logs in the following sections have not the same order as in Table 4, but the pseudonymization is consistent across the evaluation.

4.1 Uniqueness Results Based on Case Attributes

The BPI Challenge 2018 event log is provided by the German company "data experts". It contains events related to application of payments process of EU's Agricultural Guarantee Fund. The event log consists of 43,809 cases, each representing a farmer's direct payments application over a period of three years. We identified "payment_actual" (PYMT), "area" (ARA), "department" (DPT), "number_parcels" (#PCL), "smallfarmer" (SF), "youngfarmer" (YF), "year" (Y) and "amount_applied" (AMT) as case attributes. The data contributor generalized the attributes PYMT, #PCL, and AMT by grouping the values in 100 bins, where the bins are identified by the minimum value [9].

To determine the impact of case attributes, we evaluate their uniqueness using various combinations. Specifically, we investigate which combinations of attribute values make cases more distinct and thus unique. The more extensive an adversary's background knowledge is, the more likely it is that this individual

Table 4. Classification of event logs

event log	category	#cases	#activities	uniqueness	
				case attr	traces
ADL [14]	R	75	34	no	yes
BPIC 2012 [16]	R	13,087	24	yes	yes
BPIC 2015 [17]	R	1,199	398	yes	yes
BPIC 2017 [18]	R	31,509	26	yes	yes
BPIC 2018 [9]	R	43,809	14	yes	yes
CCC 2019 [19]	R	10,035	8	no	yes
Credit [20]	R	20	29	no	yes
HB [21]	R	100,000	18	no	yes
RlH [22]	R	1,143	624	no	yes
WABO [23]	R	1,434	27	yes	yes
RTFM [24]	R	150,370	11	no	yes
Sepsis [11]	R	1,049	16	no	yes
Apache [25]	S	3	74	–	–
BPIC 2013 [15]	S	1,487	4	–	–
BPIC 2014 [26]	S	46,616	39	–	–
BPIC 2016 [27]	S	25,647	600	–	–
BPIC 2019 [28]	S	251,734	42	–	–
JUnit [29]	S	1	182	–	–
NASA [30]	S	2,566	47	–	–
SWA [31]	S	1	106	–	–

becomes unique and thus identifiable. For each combination, we count the number of unique cases. As expected, the more case attributes are known, the more unique the cases become. Table 5 (left) shows that when considering PYMT only, there are 40.9% unique cases. In combination with #PCL, uniqueness increases to 69.8%. With all case attributes, 84.5% of the cases are unique in the sample.

However, the sample uniqueness alone does not lead to a high re-identification risk. Therefore we also have to consider the population uniqueness We used the method described in Sect. 3.1 to estimate the population uniqueness and approximate the marginals from the published event log. In Table 5 (left), we present the average estimated population uniqueness of five runs. Interestingly, the population uniqueness with a single case attribute (PYMT) is already 16.1%. Considering all case attributes, a population uniqueness of around 97% is observed. We measure the sample uniqueness and estimate the population uniqueness for all event logs with more than one case attribute resulting in four event logs for the analysis. We do not consider case attributes that contain activities of the event log (i.e., the first executed activity), since we assume that an adversary does not know the exact order of executed activities. Table 5 (right) lists the average

Table 5. Sample uniqueness and population uniqueness (estimated) based on case attributes (left for BPI Challenge 2018; right for all event logs)

combination	sample	population
PYMT	0.409	0.161
PYMT, ARA	0.476	0.164
PYMT, DPT	0.528	0.419
PYMT, #PCL	0.698	0.594
PYMT, ARA, #PCL	0.747	0.649
PYMT, DPT, #PCL	0.788	0.718
PYMT, DPT, #PCL, ARA, SF	0.845	0.971

event log	sample	population
3.	0.011	0.005
6.	0.035	0.071
7.	0.152	0.146
8.	1.000	0.952

sample uniqueness and the average estimated population uniqueness after five runs. We notice that not all event logs show a high uniqueness based on the case attributes. In case of the BPI Challenge 2018 event log, it can be observed that even a small number of case attributes produces a high uniqueness and thus a high re-identification risk.

4.2 Uniqueness Results Based on Traces

The Sepsis event log is obtained from the information system of a Dutch hospital. It contains events related to logistics and treatment of patients that enter the emergency room and are suspected to suffer from sepsis, which is a life-threatening condition that warrants immediate treatment. Originally, the event log was analyzed regarding the adherence to guidelines on timely administration of antibiotics and, more generally, related to the overall trajectory of patients [32]. The data was made publicly available for research purposes [11]. Several measures were taken to prevent identification, including:

- randomization of timestamps by perturbing the start of cases and adjusting timestamps of respective subsequent events accordingly
- pseudonymization of discharge related activities, e.g., “Release A”
- generalization of employee information by stating the department only
- pseudonymization of the working diagnosis
- generalization of age to groups of 5 years and at least 10 people.

The event log consists of 1,049 cases with 16 different activities. Each case represents the pathway through the hospital of a natural person. The traces have an average length of 14 points (min = 3, max = 185). In contrast to the BPI Challenge 2018 event log, the Sepsis event log has only one attribute that can be used as a case attribute.

To estimate the uniqueness of traces, we use the method described in Sect. 3.2. The points in the Sepsis event log consist of activities, timestamps, and departments that are currently responsible for a patient’s treatment. The “age” serves as case attribute. Since patients are treated in different departments, the “department” does not satisfy the time-invariant criteria of a case attribute (cf. Sect. 2).

For each case, we randomly select m points of the trace and count the number of traces with identical points. In other words, we look for other traces that for example include the same activities by the same department. We opt for a random point selection to avoid making assumptions on the adversary's knowledge. We are aware that this may underestimate the re-identification risk. As a consequence, a high uniqueness in our results emphasizes the re-identification risk as a more sophisticated and optimized point selection would likely lead to an even higher uniqueness.

In Fig. 1, we show the uniqueness of traces for different values of m points and different projections.[3] As expected, we generally observe that more points lead to a higher uniqueness. Assuming that timestamps are correct (which they are not), projection A shows that four points including the activity and the timestamp are sufficient to identify all traces. By generalizing timestamps, i.e., reducing the resolution to days, only 31% of traces are unique when considering four points and 70% when considering all points of a trace. Hence, the results clearly show the impact of generalization on the re-identification risk.

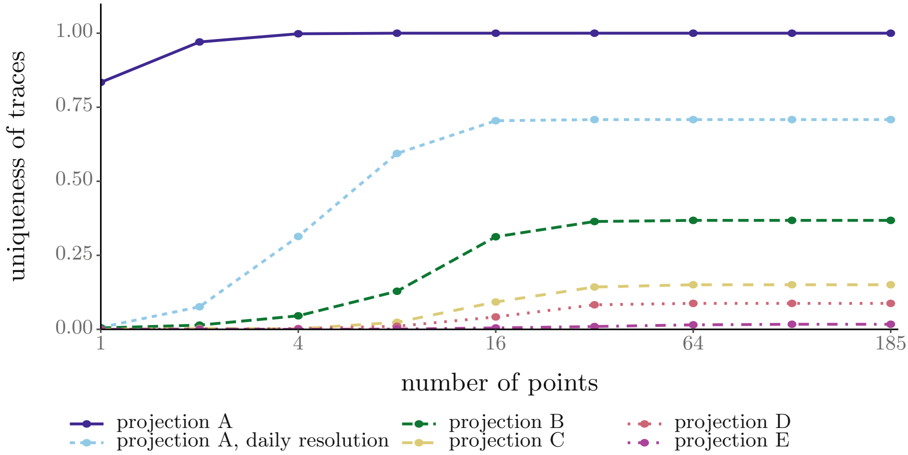


Fig. 1. Uniqueness based on traces for Sepsis event log.

The privacy-enhancing effect of removing values from the event log becomes apparent, when considering the other projections. Projection B, for example, omits timestamps but otherwise assumes that an adversary has background knowledge on all activities, case and event attributes. Yet, it is able to significantly limit the uniqueness to approximately 37%. Projection D, where case attributes and activities are still included, is even able to limit the uniqueness of traces to a maximum of 9%. The uniqueness of traces remains stable for more than 64 points since only 2% of the traces have more than 64 points.

[3] Code available at https://github.com/d-o-m-i-n-i-k/re-identification-risk.

Our method of estimating the uniqueness based on traces can be applied to all event logs categorized as real-life-individuals (R). Figure 2 presents the uniqueness for all event logs for different projections. We evaluate the uniqueness given 10%, 50%, and 90% of possible points per trace, i.e., an adversary knows this number of points per case. Grey fields without numbers imply that this projection could not be evaluated due to missing attributes.

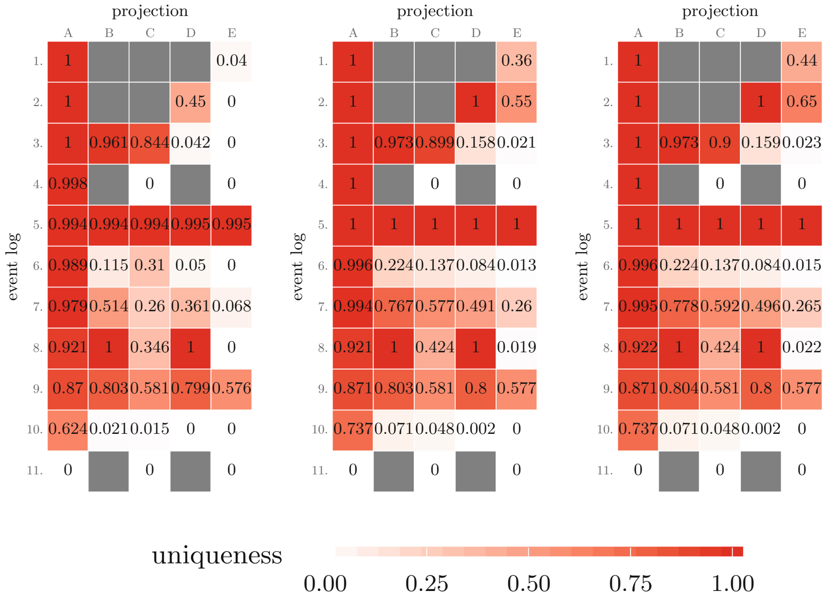
In Fig. 2 we observe a similar trend as before for the Sepsis event log: Projection A generally leads to a high uniqueness. By omitting information, expressed by the various projections, the uniqueness decreases. This becomes apparent when comparing projection B to C, where the case attributes are removed. Projection E, i.e., considering the activities only, leads to a small uniqueness, with the exception of event log 5 and 9. We explain this by the fact that these event logs have many different activities and have a varying trace length per case. For event log 10, we can already see a clear reduction of the uniqueness for projection B. This can be explained by the small number of case attributes and small number of unique activities.

The most surprising event log is 11. It has no unique cases. The prime reason for this difference is the result of a timestamp in daily resolution and the small number of unique activities. It is worth adding that increasing the number of points from 10% to 50% is significant with respect to uniqueness compared to the number of points from 50% to 90%. For example, the uniqueness of projection A for event log 10 increases from 62.4% in Fig. 2a to 73.7% in Fig. 2b. Given 90% of points of the trace, we cannot observe an increase of the uniqueness for event log 10. This can also be observed for other event logs and other projections. The prime cause of this is the high variance of the trace length.

Overall in our study, we find that the uniqueness based on traces is higher than on case attributes (cf. results in Table 5). For example, event log 3 has a sample uniqueness based on case attributes of 1.1%. Based on traces, however, it reaches for projection C a case uniqueness of 84.4%. We conclude that traces are particularly vulnerable to data re-identification attacks.

4.3 Discussion

Our results demonstrate that 11 of 12 evaluated event logs have a uniqueness greater than 62%, even for a random selection of trace points. More specific information, e.g., the order of individual activities, can lead to a greater uniqueness with fewer points. Additional knowledge about the process in general could be used by an adversary to predict certain activities, which was also confirmed in [33]. The random selection, however, clearly shows that little background knowledge is sufficient and already induces a considerable re-identification risk for event logs. In contrast, generalization of attributes helps to reduce the risk [34]. The results, however, show that combining several attributes, such as case attributes and activities, still yields unique cases. In combination with lowering the resolution of values, e.g., publishing only the year of birth instead of the full birthday, reduces the re-identification risk. Such generalization techniques can also be applied to timestamps, activities, or case attributes.



event log	A	B	C	D	E
1.	1				0.04
2.	1			0.45	0
3.	1	0.961	0.844	0.042	0
4.	0.998		0		0
5.	0.994	0.994	0.994	0.995	0.995
6.	0.989	0.115	0.31	0.05	0
7.	0.979	0.514	0.26	0.361	0.068
8.	0.921	1	0.346	1	0
9.	0.87	0.803	0.581	0.799	0.576
10.	0.624	0.021	0.015	0	0
11.	0		0		0

(a) For 10% of points.

event log	A	B	C	D	E
1.	1				0.36
2.	1			1	0.55
3.	1	0.973	0.899	0.158	0.021
4.	1		0		0
5.	1	1	1	1	1
6.	0.996	0.224	0.137	0.084	0.013
7.	0.994	0.767	0.577	0.491	0.26
8.	0.921	1	0.424	1	0.019
9.	0.871	0.803	0.581	0.8	0.577
10.	0.737	0.071	0.048	0.002	0
11.	0		0		0

(b) For 50% of points.

event log	A	B	C	D	E
1.	1				0.44
2.	1			1	0.65
3.	1	0.973	0.9	0.159	0.023
4.	1		0		0
5.	1	1	1	1	1
6.	0.996	0.224	0.137	0.084	0.015
7.	0.995	0.778	0.592	0.496	0.265
8.	0.922	1	0.424	1	0.022
9.	0.871	0.804	0.581	0.8	0.577
10.	0.737	0.071	0.048	0.002	0
11.	0		0		0

(c) For 90% of points.

Fig. 2. Uniqueness based on traces for all event logs.

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.

We apply our methods to already published event logs to point out the risk of re-identification in the domain of process mining. To this end, we only quantify the risk and refrain from cross-correlating other event logs, which might re-identify individuals. In addition, we take measures such as pseudonymizing event logs in our evaluation to neither expose nor blame specific event logs.

5 Related Work

Re-identification Attacks. Re-identification attacks were addressed and successfully carried out in the past by a large number of researchers [6,7,12,13,35,36]. Narayanan and Shmatikov [35] de-anonymize a data set from Netflix containing movie ratings by cross-correlating multiple data sets. In [36], they modified their approach to apply it to social networks. In contrast, our adversary's goal is to re-identify an individual (also known as singling out) and not reconstruct all attribute values of an individual. We therefore measure the uniqueness. We base our uniqueness measures on two well-known approaches [7,12,13] and adapt them for the domain of event logs and process mining. Rocher et al. [7] estimate the population uniqueness based on given attribute values. We employ their method to estimate the uniqueness based on case attributes. Our method to estimate the uniqueness based on traces relies on the approach presented in [12,13], where uniqueness in mobility traces with location data is estimated. Due to the structure of an event log, both methods alone are not sufficient to determine the uniqueness in event logs and require data preparation. For example, event logs have a specific format that requires transformation in order to apply uniqueness measures on traces.

Privacy in Process Mining. Awareness of privacy issues in process mining has increased [37], particularly since the General Data Protection Regulation (GDPR) was put into effect. Although the Process Mining Manifesto [38] demands to balance utility and privacy in process mining applications, the number of related contributions is still rather small. To preserve privacy in event logs while still discovering the correct main process behavior has been addressed by Fahrenkrog-Petersen et al. [4]. Their algorithm guarantees k-anonymity and t-closeness while maximizing the utility of the sanitized event log. In general, k-anonymity aggregates the data in such a way that each individual cannot be distinguished based on its values from at least $k-1$ other individuals of the data set [39,40]. Yet, it has been shown in the past that neither k-anonymity, nor t-closeness are sufficient to provide strong privacy guarantees [41].

The strongest privacy model available to date, which provides provable privacy guarantees, is differential privacy. It was recently incorporated in a first privacy-preserving technique for process mining [2]. The approach presents a privacy engine capable of keeping personal data private by adding noise to queries. The privacy techniques of [2,4] have been combined in a web-based tool [3]. Pseudonymization of data sets related to process mining has been discussed in [42,43]. Values of the original data set is replaced with pseudonyms. However, the encryption still allows for a potential re-identification by an adversary with knowledge about the domain and the statistical distribution of the encrypted data. Beside technological privacy challenges for process mining, the approach of [44] also discuss organizational privacy challenges by means of a framework. Although, the approach points to several privacy concerns in process mining, no technical solution is presented. Pika et al. [33] assess the suitability of existing privacy-preserving approaches for process mining data. They propose a framework to support privacy-preserving process mining analysis. While Pika et al. analyze the suitability of existing data transformation approaches to anonymize process data, they do not provide an approach to support the identification of information, e.g., atypical process behavior, that should be suppressed to reduce the re-identification risk of subjects. Our metric fills this gap and helps data owners to identify the unique cases with atypical process behavior.

In comparison to existing related works on privacy-aware approaches for process mining, this paper makes an attempt to quantify the re-identification risk. Data publishers can determine which information should be suppressed before releasing an event log for process mining. If a high re-identification risk is detected, the approaches mentioned above might be able to lower the risk of re-identification and therefore to provide higher privacy guarantees.

6 Conclusion

This paper identifies and evaluates the risk of re-identification in event logs for process mining. We reveal that there is a serious privacy leakage in the vast majority of the event logs used widely in the community. To address this issue, we argue for the use of methods to estimate the uniqueness that allow event log publishers to carefully evaluate their event logs before release and if need to suppress certain information. Overall, real-world data traces are an essential means to evaluate and compare algorithms. 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.

Acknowledgement. We thank Yves-Alexandre de Montjoye for providing his code to estimate data uniqueness, which we used in our evaluation.

References

1. van der Aalst, W.M.P.: Process Mining - Data Science in Action, 2nd edn. Springer, Heidelberg (2016). https://doi.org/10.1007/978-3-662-49851-4
2. Mannhardt, F., Koschmider, A., Baracaldo, N., Weidlich, M., Michael, J.: Privacy-preserving process mining. Bus. Inf. Syst. Eng. **61**(5), 595–614 (2019). https://doi.org/10.1007/s12599-019-00613-3
3. Bauer, M., Fahrenkrog-Petersen, S.A., Koschmider, A., Mannhardt, F.: ELPaaS: event Log Privacy as a service. In: Proceedings of the Dissertation Award, 17th International Conference on Business Process Management, BPM 2019, p. 5 (2019)
4. Fahrenkrog-Petersen, S.A., van der Aa, H., Weidlich, M.: PRETSA: event log sanitization for privacy-aware process discovery. In: Proceedings of the International Conference on Process Mining, ICPM 2019, pp. 1–8 (2019)
5. Lavrenovs, A., Podins, K.: Privacy violations in Riga open data public transport system. In: Proceedings of the IEEE 4th Workshop on Advances in Information, Electronic and Electrical Engineering, AIEEE 2016, pp. 1–6 (2016)
6. Douriez, M., Doraiswamy, H., Freire, J., Silva, C.T.: Anonymizing NYC taxi data: does it matter? In: Proceedings of the IEEE International Conference on Data Science and Advanced Analytics, DSAA 2016, pp. 140–148 (2016)
7. Rocher, L., Hendrickx, J., Montjoye, Y.A.: Estimating the success of re-identifications in incomplete datasets using generative models. Nat. Commun. **10**, 1–9 (2019)
8. 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. In: Proceedings of ACM Transactions on Management Information Systems, TMIS 2017, vol. 8, no. 1, p. 1 (2017)
9. van Dongen, B., Borchert, F.: BPI Challenge 2018. TU Eindhoven, Dataset (2018)

10. Dankar, F.K., El Emam, K., Neisa, A., Roffey, T.: Estimating the re-identification risk of clinical data sets. BMC Med. Inform. Decis. Mak. **12**(1), 66 (2012)
11. Mannhardt, F.: Sepsis Cases - Event Log. TU Eindhoven, Dataset (2016)
12. Song, Y., Dahlmeier, D., Bressan, S.: Not so unique in the crowd: a simple and effective algorithm for anonymizing location data. In: Proceeding of the 1st International Workshop on Privacy-Preserving IR, PIR@SIGIR 2014, vol. 2014, pp. 19–24 (2014)
13. de Montjoye, Y.A., Hidalgo, C.A., Verleysen, M., Blondel, V.D.: Unique in the crowd: the privacy bounds of human mobility. Sci. Rep. **3**, 1376 (2013)
14. Sztyler, T., Carmona, J.: Activities of Daily Living of Several Individuals. University of Mannheim, Germany. Dataset (2015)
15. Steeman, W.: BPI Challenge 2013. Ghent University, Dataset (2013)
16. van Dongen, B.: BPI Challenge 2012. 4TU.Centre for Research Data. Dataset (2012)
17. van Dongen, B.: BPI Challenge 2015. 4TU.Centre for Research Data. Dataset (2015)
18. van Dongen, B.: BPI Challenge 2017. TU Eindhoven, Dataset (2017)
19. Munoz-Gama, J., de la Fuente, R., Sepúlveda, M., Fuentes, R.: Conformance Checking Challenge 2019 (CCC19). 4TU.Centre for Research Data. Dataset (2019)
20. Djedović, A.: Credit Requirement Event Logs. 4TU.Centre for Research Data. Dataset (2017)
21. Mannhardt, F.: Hospital Billing - Event Log. TU Eindhoven, Dataset (2017)
22. van Dongen, B.: Real-life Event Logs - Hospital log. TU Eindhoven, Dataset (2011)
23. Buijs, J.: Receipt Phase of an Environmental Permit Application Process ('WABO'). Eindhoven University of Technology, Dataset (2014)
24. de Leoni, M., Mannhardt, F.: Road Traffic Fine Management Process. TU Eindhoven, Dataset (2015)
25. Leemans, M.: Apache Commons Crypto 1.0.0 - Stream CbcNopad Unit Test Software Event Log. TU Eindhoven. Dataset (2017)
26. van Dongen, B.: BPI Challenge 2014. 4TU.Centre for Research Data. Dataset (2014)
27. Dees, M., van Dongen, B.: BPI Challenge 2016. 4TU.Centre for Research Data. Dataset (2016)
28. van Dongen, B.: BPI Challenge 2019. 4TU.Centre for Research Data. Dataset (2019)
29. Leemans, M.: JUnit 4.12 Software Event Log. TU Eindhoven. Dataset (2016)
30. Leemans, M.: NASA Crew Exploration Vehicle (CEV) Software Event Log. TU Eindhoven, Dataset (2017)
31. Leemans, M.: Statechart Workbench and Alignments Software Event Log. TU Eindhoven, Dataset (2018)
32. Mannhardt, F., Blinde, D.: Analyzing the trajectories of patients with sepsis using process mining. In: Joint Proceedings. Volume 1859 of CEUR Workshop Proceedings, RADAR+EMISA 2017, pp. 72–80. CEUR-WS.org (2017)
33. Pika, A., Wynn, M.T., Budiono, S.: Towards privacy-preserving process mining in healthcare. Proceedings of International Workshop on Process-Oriented Data Science for Healthcare, PODS4H 2019, p. 12 (2019)
34. Zook, M., et al.: Ten simple rules for responsible big data research (2017)
35. Narayanan, A., Shmatikov, V.: Robust de-anonymization of large sparse datasets. In: Proceedings of the 29th IEEE Symposium on Security and Privacy, S&P 2008, pp. 111–125 (2008)

36. Narayanan, A., Shmatikov, V.: De-anonymizing Social Networks. In: Proceedings of the 30th IEEE Symposium on Security and Privacy, S&P 2009, pp. 173–187 (2009)
37. Spiekermann, S., Cranor, L.: Engineering privacy. IEEE Trans. Softw. Eng. **35**(1), 67–82 (2009)
38. van der Aalst, W., et al.: Process mining manifesto. In: Daniel, F., Barkaoui, K., Dustdar, S. (eds.) BPM 2011. LNBIP, vol. 99, pp. 169–194. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-28108-2_19
39. Sweeney, L.: k-anonymity: a model for protecting privacy. Int. J. Uncertainty Fuzziness Knowl. Based Syst. **10**(05), 557–570 (2002)
40. Samarati, P., Sweeney, L.: Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. SRI International (1998)
41. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, pp. 265–284 (2006)
42. 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, pp. 3–17 (2018)
43. Burattin, A., Conti, M., Turato, D.: Toward an anonymous process mining. In: Proceedings of the 3rd International Conference on Future Internet of Things and Cloud, FiCloud 2015, Rome, Italy, pp. 58–63 (2015)
44. Mannhardt, F., Petersen, S.A., Oliveira, M.F.: Privacy challenges for process mining in human-centered industrial environments. In: Proceedings of the 14th International Conference on Intelligent Environments, IE 2018, pp. 64–71 (2018)