SI on Advances in Privacy-Preserving Computing manuscript No.
(will be inserted by the editor)

A Uniformization-based Approach to Preserve Individuals' Privacy during Process Mining Analyses

Edgar Batista · Agusti Solanas

Received: date / Accepted: date

Abstract Process Mining is a set of techniques that aim at discovering, monitoring and improving real processes by using logs of events created and stored by corporate information systems. The growing use of information and communication technologies and the imminent wide deployment of the Internet of Things enable the massive collection of events, which are going to be studied so as to improve all kinds of systems efficiency. Despite its enormous benefits, analyzing event logs might endanger individuals privacy, especially when those logs contain personal and confidential information, such as healthcare data. This article contributes to an emerging research direction within the process mining field, known as Privacy-Preserving Process Mining (PPPM), which embraces the privacy-by-design principle when conducting process mining analyses. We show that current solutions based on pseudonyms and encryption are vulnerable to attacks based on the analysis of the distribution of events combined with well-known location-oriented attacks such as the restricted space identification and the object identification attacks. With the aim to counteract these attacks, we present u-PPPM, a novel privacy-preserving process mining technique based on the uniformization of events distributions. This approach protects the privacy of the individuals appearing in event logs while minimizing the information loss during process discovery analyses. Experimental results, conducted using six real-life event logs, demonstrate the feasibility of our approach in real settings.

E. Batista
Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili,
Av. Països Catalans 26, 43007 Tarragona, Catalonia, Spain
E-mail: edgar.batista@estudiants.urv.cat
SIMPPLE S.L., C. Joan Maragall 1A, 43003 Tarragona, Catalonia, Spain
E-mail: edgar.batista@simpple.com

A. Solanas
Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili,
Av. Països Catalans 26, 43007 Tarragona, Catalonia, Spain
E-mail: agusti.solanas@urv.cat

Keywords Process mining · Privacy · Privacy-preserving process mining · Distribution-based attacks · Uniformization strategies

1 Introduction

The steady growth of organizational information, gathered automatically by corporate information systems, is fostering investment in data analytics. With the proper contextualization of such data, realistic visions of the organizational workflows can be obtained, and modeled as business processes, *i.e.,* a set of logically related tasks performed to achieve a certain business goal [48]. For traceability purposes, business processes are set to leave traces in the form of events, represented as records describing well-defined steps of processes executions. All generated events are stored sequentially in event logs that are progressively growing and paving the way for the so-called Internet of Events [2].

The neutral perspective of event logs on the execution of business process instances is ideal to model the actual behavior of those processes and to conduct advanced analyses, such as discovering actual process executions, verifying the alignment between ideal and actual process executions, identifying bottlenecks, optimizing resources, or recommending countermeasures to identified problems. With the goal of meeting these challenges, the field of *process mining* (PM) [1] studies means to discover, monitor and improve real processes by extracting knowledge from event logs readily available in today's information systems, by means of intelligent a-posteriori analysis techniques, such as data analysis, process modeling and process analysis. In general, PM techniques can be grouped into three categories: (i) process discovery, aiming at visualizing process models from event logs without any additional knowledge; (ii) conformance checking, aiming at monitoring deviations by comparing models and event logs; and (iii) enhancement, by improving existing process models using previous processes executions information recorded in event logs. Thanks to the cross-cutting nature of PM, many industries noticed the opportunity to better manage their business processes and become more efficient, effective and sustainable [17]. For example, the healthcare industry exploits PM techniques to reduce medical expenditure costs by enhancing oncological and emergency care processes, among others [18,24,13]. Further, PM also helps to optimize loan application processes in the banking industry [30], improve the service quality of complaint handling service in big corporations [49], and even predict customers movements for tourism industries [7].

In general, event logs contain attributes related to the activity that has been conducted, the process instance identifier, the resource responsible for executing the activity and a timestamp, among others. Yet, in certain contexts, some attributes might contain personal data (*e.g.,* full name or national ID/social security number) and/or sensitive data (*e.g.,* beliefs or health conditions) that can jeopardize people's privacy unless properly managed. To prevent data misuse scandals [33,22], legal privacy regulations, such as GDPR

in Europe [14], are transforming the management and processing of digital data by fostering privacy-by-design principles (*e.g.,* data minimization, pseudonymization or encryption, or lawful processing), and forbidding the release of potentially harmful information (event logs in this case) among institutions, as a way to preserve people's privacy. Besides, this has also fostered the appearance of novel approaches to store and retrieve data using decentralized data structures, such as blockchains [40], as well as searching and sharing encrypted data from cloud infrastructures in a timely efficient manner [19]. Privacy issues are even more apparent in sensitive sectors, such as the medical domain, which deals with highly confidential data: patients' health conditions, treatments, illnesses, clinical trials, and so on. With the advent of wearables and IoT, the management of large volumes of healthcare data can contribute to the early detection of illnesses or the prediction of diseases, but without compromising people's privacy [20]. Indeed, within this domain, the continuous evolution of healthcare paradigms towards more effective, cost-efficient and sustainable models, such as smart and cognitive health [43, 26], promotes the use of PM techniques to improve care services, optimize resources, and reduce costs and treatments duration [42]. However, despite experts advice [4], only a small number of studies consider the privacy risks involved in their PM procedures [5] and, when addressed, countermeasures are mainly based on the pseudonymization of personally identifiable information or the encryption of event data as vehicles to achieve confidentiality.

With the aim to foster the study of the privacy implications of PM analyses and the development of privacy-preserving techniques for PM, the *Privacy-Preserving Process Mining* (PPPM) field emerged very recently [27], and it has become a hot topic in PM research. Formally, PPPM considers privacy-enhancing techniques for preventing the disclosure of people's personal or sensitive data to unauthorized parties once conducting PM analyses, so minimizing the risk of people re-identification. These techniques typically imply the transformation/distortion of event log data, which directly affect the quality of PM results (*i.e.,* mainly process models). Hence, PPPM techniques have to focus on the maximization of the quality of the process models, while reducing/averting disclosure risks. This trade-off between privacy and data utility is well-known and has been reported in the privacy protection literature for years [21]. The privacy constraints introduced by PPPM techniques on the event log data aim to prevent (i) the seamless re-identification of people from event data or from the discovered process models (*i.e.,* identity disclosure), and/or (ii) the association of confidential event log data to personally identifiable information (*i.e.,* attribute disclosure). Hence, the main challenge of PPPM is to determine the best approach to distort event log data to counteract specific attacks, so that PM results are minimally affected, while individuals' privacy is protected.

1.1 Contribution and Plan of the Article

This article presents a novel PPPM technique, called u-PPPM, based on the uniformization of sensitive distributions of event data attributes that can be exploited to re-identify people by means of location-oriented targeted attacks.

In Section 2, we justify the need for applying PPPM in practice and discuss the related work available in the literature. Next, in Section 3, we show some limitations of classical solutions, such as pseudonymization or encryption, under specific contexts, in which the privacy of individuals appearing in event logs (gathered in public areas) may be at stake. Moreover, we illustrate an attacker model based on the inference and exploitation of event data distributions that enables people re-identification, when combined with location-oriented targeted attacks, such as restricted space identification and object identification attacks. With the aim to counteract the aforementioned attacks, in Section 4, we present our PPPM technique, called u-PPPM, aiming to produce privacy-preserved event logs, in such a way that the statistical distributions of their attributes become uniform according to a privacy threshold and, hence, remain indistinguishable from the attacker's perspective.

Since the ultimate goal of PM is to obtain actionable knowledge (in the form of processes) from event data, our u-PPPM approach focuses on protecting privacy while minimizing the distortion of the protected process models. In Section 5, we evaluate our u-PPPM approach by measuring the distortion (*i.e.,* decrease of quality) introduced in the process models discovered from the protected event logs after applying u-PPPM. Experiments, conducted using six real-life event logs, show the usefulness of our solution. We conclude the article with some final comments and remarks in Section 6.

2 Privacy-Preserving Process Mining: Rationale and Related Work

Research in PPPM is on the rise. Whereas privacy aspects have barely been considered in PM for many years, they have recently attracted the attention of researchers. The global awareness on data privacy, the enforcement of privacy legislations, and the definition of FACT principles (Fairness, Accuracy, Confidentiality and Transparency) to conduct Responsible Process Mining [3] provide sufficient grounds to incorporate privacy-preserving strategies during PM analyses.

The creation of privacy-preserved event logs for conducting PPPM analyses has many valuable application uses. First of all, it is of the utmost importance within sensitive domains, which contain confidential data about personally identifiable information, such as healthcare or banking. In addition, in situations where institutions are not capable of conducting PM analyses and need to externalize PM services to third parties, it is necessary to apply PPPM techniques so as to limit the third-parties' ability to retrieve private information beyond the very process models. Globally, the worldwide trend towards open data models for transparency purposes also prompts the need for using

privacy-preserving techniques on the data to be released in order to protect people's privacy. Besides, the observed uniqueness of event data opens the door to re-identification risks [31]. Significantly enough, situations where PM involves event log data from multiple institutions (*i.e.,* the execution of a process is not conducted by a single institution, but by a group of independent institutions), known as cross-organizational process mining, could originate privacy issues. Last but not least, even if the event logs are not going to be shared or released, the application of PPPM techniques could serve as a preventive countermeasure against data breaches or data thefts. For example, [28] mentioned the privacy challenges associated with the legal requirements on data protection, recommendations and good practices guidelines for PM in human-centered industrial environments.

Some studies aim to achieve confidentiality in PM by means of pseudonymization or encryption of event logs. For instance, [38] presented a confidentiality framework and analyzed the weaknesses and open challenges of encryption of event logs. More comprehensively, [9] approached the outsourcing of PM analyses, where the confidentiality of event logs and the resulting processes must be guaranteed, by taking advantage of either symmetric or homomorphic encryption for hiding sensitive data. In the same line, [45] proposed a simple protocol to generate process models in a privacy-preserved fashion using the alpha algorithm (one of the very first PM algorithms) from encrypted event logs. In the context of cross-organizational settings, where PM analyses are conducted using event logs from multiple organizations, [25] presented a trusted-third-party scheme dealing with public and private business process models. In [29], a privacy-preserving system design for PM based on an ABAC-based authorization model to support privacy strategies on event logs was presented.

Recently, [34] analyzed the data privacy requirements for process models in the healthcare domain, and proposed a theoretical PPPM framework to support PM analyses of healthcare processes, based on the anonymization of healthcare data and the creation of privacy metadata. More interestingly, they evaluated the pros and cons of using different data transformation techniques, such as data swapping, suppression, generalization and noise addition, to modify the attributes' values within event logs. In a nutshell, it can be observed that the quality of PM results decreases when anonymization is required, but the magnitude of this quality loss depends on the event logs themselves and the goals of the PM analyses. More formally, these data anonymization operations are described in [37] through practical examples, such as the suppression of events according to the activity attribute, the addition of events upon certain conditions, the substitution, generalization or swap of attributes values, and the use of cryptography.

In [15], two main approaches to guarantee privacy in PM are highlighted, namely event log sanitization (*i.e.,* the pre-processing of event logs so they guarantee a certain level of privacy) and privatized process mining (*i.e.,* the development of new PM techniques that generate PM artifacts guaranteeing a level of privacy). Regarding event logs sanitization, [16] proposed PRETSA, an algorithm providing k-anonymity and t-closeness properties within event

logs. To break the link between personally identifiable information and confidential data, people's information is removed from the event log (and only event identifiers, activities and confidential data are kept). Consequently, no PM analyses involving the behavior, performance or role of particular individuals are possible. On a different approach, [27] presented a differential privacy model for a privatized process discovery, where the privacy of individuals is ensured due to the introduction of noise into each query result according to the differential privacy framework. Although very powerful, the complexity of this framework implies a number of constraints, such as in the length of the traces or the kind of information that can be discovered from individuals. Both solutions can be found in the ELPaaS web-application service [6].

Last but not least, [35] focused on the privacy concerns on resource behavior analyses, such as role mining where the roles of individuals can be represented as social networks, once an attacker has advanced knowledge on the activities that individuals are responsible for. In [39], authors presented the *TLKC*-privacy model to avert attribute linkage attacks in process discovery and performance analyses through group-based anonymization. These methods are integrated in an open-source web-based PM tool called PPDP-PM [36].

3 Re-identification using Event Logs from Public Spaces: Towards Distribution-based Attacker Models

Pseudonimization and encryption are classical techniques to obfuscate the link between personally identifiable information and confidential data, both suggested by GDPR and used by some PM techniques, as referred in the previous section. This section demonstrates the potential privacy issues that could emerge when releasing pseudonimized or encrypted event logs, and the inability of such techniques to counteract specific attacks. More specifically, we describe an attacker model that permits the re-identification of individuals by exploiting the distribution of attributes from event logs that are only protected via pseudonymization or encryption. Indeed, individuals' privacy might be at risk when this attacker model is combined with location-oriented targeted attacks. These attacks are feasible in public spaces institutions, such as hospitals, emergency rooms, banks or public administrations, which manage sensitive information (*e.g.,* medical data, socio-economic status...).

Let L be an event log with the activities carried out by people in a public space institution (*e.g.,* employees from a public hospital, public administration...). The event log L consists of a set of traces $T = \langle t_1, \ldots, t_m \rangle$, where m ($m > 0$) is the number of traces in L. Each trace $t_j \in T$, for $1 \leq j \leq m$, contains a set of events $\langle e_1, \ldots, e_n \rangle \in E$, where n ($n > 0$) is the number of events in t_j, referring to the same process execution instance (*i.e.,* the case), in chronological order. Each event $e_i \in E$, represented as a record, describes a well-defined step of a process execution with multiple attributes, namely a unique event ID, a case ID, the performed activity, a timestamp, person-

ally identifiable information of the responsible of such activity and, in some cases, confidential data associated to the person. It is apparent that institutions should not release L unmodified, because it associates confidential data to personally identifiable information that would jeopardize privacy. Therefore, in order to create a privacy-preserving event log L', institutions could decide to directly suppress the personally identifiable information and/or the confidential data from L, as suggested in [34,16]. However, if this suppression is applied, it limits the number of PM studies: for instance, resources-oriented analyses, performance analyses or organizational analyses are no longer possible. In most cases, following the recommendations of GDPR [14,28] and PM literature [38,9], institutions might take advantage of pseudonymization or encryption of personally identifiable information (among other attributes) to break the linkage between personally identifiable information and confidential data, hence releasing a protected event log L' with unreadable personally identifiable information associated to confidential data. However, we next show a potential weakness of this kind of strategies that would enable re-identification.

Despite the distortion of personally identifiable information in L' produced by means of pseudonymization or encryption strategies, the appearance distribution of attributes' values is not affected (although the actual values in L' seem unreadable). For instance, if *Alice Fisher* appears in L a total of α times, then the pseudonym or the ciphertext associated with *Alice Fisher* appears in L' a total of α times, too. Table 1 shows a toy event log file that exemplifies this situation. Although appearance distributions might not seem to be harmful, attackers could exploit them meaningfully, opening the door to the re-identifications of individuals. Trying to break this direct correlation, one could use dynamic encryption or pseudonymization, where multiple values $\{x_1, x_2, \ldots, x_b\}$ are assigned in L' to the same value x in L. This approach is a clear improvement from a privacy perspective but complicates PM analyses extremely, especially when grouping events from the same individual is required

Table 1 Example of an event log L encompassing sensitive data from a medical institution, and the resulting event log L' with the encryption of personally identifiable information.

Original event log L						
Event ID	**Case ID**	**Patient**	**Doctor**	**Activity**	**Timestamp**	**Disease**
15637	2018TGC36587	Anthony Green	Stephen Murray	Start surgery	2018-04-25 17:03:54	Fracture
15638	2018CZE65214	Alice Fisher	Tom Adams	Mammogram	2018-04-25 17:12:05	Cancer
15639	2018KPI02547	Peter Brown	Charles Thompkins	Admission	2018-04-25 17:15:26	Flu
15640	2018CZE65214	Alice Fisher	Tom Adams	Discharge	2018-04-25 17:26:20	Cancer
15641	2018KPI02547	Peter Brown	Charles Thompkins	Prescription	2018-04-25 17:32:53	Flu
15642	2018KPI02547	Peter Brown	Charles Thompkins	Discharge	2018-04-25 17:34:17	Flu
15643	2018TGC36587	Anthony Green	Stephen Murray	End surgery	2018-04-25 17:58:10	Fracture

Event log L' with the personally identifiable information encrypted						
Event ID	**Case ID**	**Patient**	**Doctor**	**Activity**	**Timestamp**	**Disease**
15637	2018TGC36587	Ok4&ff)785	38-!bC3_El68	Start surgery	2018-04-25 17:03:54	Fracture
15638	2018CZE65214	Lo569/2;M98	1Ba_:5Dfa/2	Mammogram	2018-04-25 17:12:05	Cancer
15639	2018KPI02547	B36*-f;Ms3%-fc	36:%72_f7E!l6E	Admission	2018-04-25 17:15:26	Flu
15640	2018CZE65214	Lo569/2;M98	1Ba_:5Dfa/2	Discharge	2018-04-25 17:26:20	Cancer
15641	2018KPI02547	B36*-f;Ms3%-fc	36:%72_f7E!l6E	Prescription	2018-04-25 17:32:53	Flu
15642	2018KPI02547	B36*-f;Ms3%-fc	36:%72_f7E!l6E	Discharge	2018-04-25 17:34:17	Flu
15643	2018TGC36587	Ok4&ff)785	38-!bC3_El68	End surgery	2018-04-25 17:58:10	Fracture

so as to discover specific individuals processes. This is why solutions found in the PM literature proposing encryption mechanisms use static approaches.

Within this context, we assume that the attacker has access to a pseudonymized or encrypted log file L' from an institution, because either L' has been released for transparency purposes, shared with another organization, or obtained through malicious ways (*e.g.,* data theft), and he/she aims to exploit such information for disclosing private information. By analyzing the event data in L' (*e.g.,* corresponding to the activity in a public hospital), it is likely that an attacker realizes that some people do more activities or participate in more cases than others (*e.g.,* some doctors are likely to have more patients -cases- or do more activities -events- than other doctors). Knowing that the distribution of the values remains unaltered between L and L', the attacker could model the frequency of activity of each person I' in L'. Hence, the attacker could model the distribution of activity of the people in L', so knowing which are the pseudonyms/ciphertexts with more and less activity in L'. With the knowledge of this distribution, the attacker conducts a targeted attack based on restricted space identification (RSI) and object identification (OI) (*i.e.,* well-known attacks against Location-Based Services that imply the direct or approximate contact with the targets). The attacker stays physically in the institution (*e.g.,* let's say in a waiting room of the hospital) and annotates the frequency of activity of each doctor (*e.g.,* how many patients each doctor has received). By extending this attack for a reasonable period of time, the attacker is able to know which doctors are more active within the institution. Finally, the attacker could infer from his/her observations that the doctor with more activity corresponds to the pseudonym/ciphertext from L' with more appearances, and so on. By following this strategy, the attacker is able to infer the correlation between the pseudonyms/ciphertexts associated to people in L' to real people (*i.e.,* identity disclosure). In addition, the attacker could therefore infer their confidential data: those confidential data associated with the corresponding pseudonym/ciphertext from L' (*i.e.,* attribute disclosure).

This attack (cf. Figure 1) demonstrates that static data transformations, although popular in practice, might not be enough to preserve people's privacy.

4 Our Uniformization Approach: *u*-PPPM

The distribution of attribute values in a sensitive event log L is a factor to consider at the time of protecting individuals' privacy in a privacy-preserved event log L', since distributions could be exploited with enough background knowledge by means of RSI/OI attacks, as previously explained in Section 3.

As a countermeasure against this distribution-based attack, we propose u-PPPM, a Privacy-Preserving Process Mining technique based on the uniformization of potentially identifiable distributions from a sensitive event log L. Assuming that attackers could infer the frequency of individuals' activities (*e.g.,* the number of cases an individual participates on, the number of activities an individual performs...), the proposed uniformization procedure averts

Fig. 1 Scenario of an RSI/OI attack in the waiting room of a public hospital.

the attack by distorting the event data in L in such a way that prevents the direct re-identification of individuals because it renders the distribution knowledge acquired by the attacker through RSI and OI attacks useless. The basis of the proposed technique is to group similar individuals from L in groups of size k (*i.e.,* a privacy threshold), and exchange events among the individuals in the same group, until all individuals within the same group $\{I_1, \ldots, I_k\}$ are uniform from a distribution perspective and, hence, are indistinguishable to an attacker. Finding a good balance between information loss (caused by distortion/exchanges of event data) and privacy (achieved by uniformization) is of utmost importance to preserve the quality of the process models. The details of the method are explained next:

The *privacy level* of the method is directly related to the individuals group size (k): the higher k, the more privacy is achieved. Thus, from a distribution perspective, the maximum privacy level is achieved when the size of the groups (k) equals the number of individuals in L, meaning that all the individuals share the same distribution and will be indistinguishable from the attacker's perspective. However, higher privacy levels negatively affect the quality of the discovered process models since they require more exchanges and distortion of the original distributions. Regarding the *creation of groups of k individuals*, it is well-known that grouping similar individuals allows the creation of homogeneous groups and, hence, it helps to reduce information loss. In our method, the similarity between individuals is measured according to the individual's frequency of activity in L (*i.e.,* the knowledge attackers try to exploit). Hence, from this perspective, we assume that k individuals are similar if their activity frequencies are similar. Since privacy is guaranteed by making this distribution uniform, this criteria allows minimizing the distortion of the event data, since fewer events exchanges are required to reach a uniform state within the group of k individuals.

Once the k individuals belonging to each group are determined, events from those individuals in the same group are interchanged so as to make their distributions uniform. To this end, first *the selection of individuals* within the same group is necessary: in particular, an individual is selected as a "provider" of events and another individual is selected as a "receiver" of events. Efficient selections of individuals allow minimizing the number of event exchanges, which reduces distortion. Although u-PPPM allows for the use of any selection strategy, we suggest and test four different strategies (S1...S4):

1. *Roulette-wheel* (S1): Each individual has a probability of being selected, according to their need for providing or receiving events from the other members of the group. Individuals with a higher-than-average activity have a higher probability of being selected as "providers", while individuals with lower-than-average activity have a higher probability of being selected as "receivers", during the exchange.
2. *Max-Min* (S2): This strategy chooses the individual with the highest frequency in the group as the "provider", and the individual with the lowest frequency in the group as the "receiver".
3. *Random* (S3): The individuals are randomly selected, without considering their needs to be "receivers" or "providers". This is a blind strategy that does not consider the actual distribution/frequency of activity.
4. *Lateral* (S4): Individuals are sorted in descending order according to their frequency of activity. The first individual (*i.e.,* the one with the highest activity) provides events to the second individual, until their distribution is uniform. Next, the third individual acts as receiver from the previous two, until the three individuals are uniform. This procedure is repeated for all individuals within the group until all individuals activity distributions are uniform.

Once two individuals I_a and I_b are selected by following any of the above strategies, a number of events are exchanged from the "provider" to the "receiver" so that their frequency of activity distribution becomes uniform. For instance, if individuals I_a and I_b are responsible for ten cases and four cases, respectively, in L (attacker's distribution knowledge), the events associated to three cases from I_a (the "provider") are assigned to I_b (the "receiver"), so both have seven cases each in L'. Thus, I_a and I_b are indistinguishable from the attacker's distribution point of view in L'. The very *selection of the events* to be exchanged from an individual to another is done at random. This results in u-PPPM being a non-deterministic method. In a nutshell, the exchange of events among individuals results in a modification of their individuals' information.

Once all groups are made uniform, u-PPPM obfuscates the personally identifiable information associated to the event data by means of pseudonymization or encryption (like other methods in the literature). Finally, the method returns the created event log L'.

The distortion added to the protected event logs L' implies a decrease on the quality of the process models. More specifically, this directly affects the

discovery of process models associated to each individual, which is important in some PM analyses used to evaluate the performance or behavior of specific people. Therefore, the process model M_a of an individual I_a discovered from L (*i.e.,* the unmodified event log) will be different to the process model M'_a discovered from L' (after applying u-PPPM) because the process model M'_a is affected by (i) the loss of events that I_a has provided to the rest of the individuals within the same group, and (ii) the events that I_a has received from the rest of the $k-1$ individuals within the same group. To the best of our knowledge, in the area of PM, this is the very first article that focuses on the quality of the discovered individuals' process models once applying a privacy-preserving technique on event log data. The source code of u-PPPM, outlined in Algorithm 1, is open and available from the Smart Technologies Research Group at `http://www.smarttechresearch.com/research/upppm/code.zip`.

For the sake of completeness and clarity, in Figure 2, we provide an illustrative example of the effect of u-PPPM on a privacy-preserved event log L' depending on the privacy threshold (k). Let L be the activities performed by four individuals, namely Bob, Pete, Marie and Sam, in five independent cases $t_1 \dots t_5$ (*e.g.,* medical treatments). It can be observed that some individuals are responsible for more cases than others, *e.g.,* Bob participates in five cases, while Sam participates in one, only. If their names were only replaced by pseudonyms or encrypted in the event log L', the attacker could exploit this distribution knowledge to re-identify them by using RSI/OI attacks. By applying u-PPPM with a privacy level $k = 2$, the method creates groups of two individuals (*e.g.,* Bob with Pete, and Marie with Sam), and exchanges the events corresponding to a (random) case from Bob to Pete (*e.g.,* events $\langle X, Y, Z \rangle$) and a case from Marie to Sam (*e.g.,* events $\langle E, F \rangle$). Therefore, Bob and Pete are now responsible for four cases each, and Marie and Sam for two

Algorithm 1 u-PPPM: Uniformization of potentially identifiable distributions from a sensitive event log L

Require:
The event log L has p individuals, $I_1, I_2, \dots, I_p$, whose personally identifiable information must be pseudonymized or encrypted.
The function u-PPPM receives two parameters: (1) a value k for the group size ($1 \leq k \leq p$), and (2) the individuals selection strategy (*e.g.,* S1–S4).
Ensure:
A privacy-preserved event log L', whose distribution is indistinguishable among k individuals.

```
1: function u-PPPM(k, selection strategy)
2:     num_groups ← ⌊p/k⌋
3:     groups ← create num_groups groups of similar k individuals from L
4:     for each group in groups do
5:         while group is not uniform do
6:             provider_ind ← select individual from group according to the selection strategy
7:             receiver_ind ← select individual from group according to the selection strategy
8:             events ← select random event/group of events from provider_ind
9:             transfer resource information in events from provider_ind to receiver_ind
10:        end while
11:    end for
12:    L' ← obfuscate personally identifiable information of all groups
13:    return L'
14: end function
```

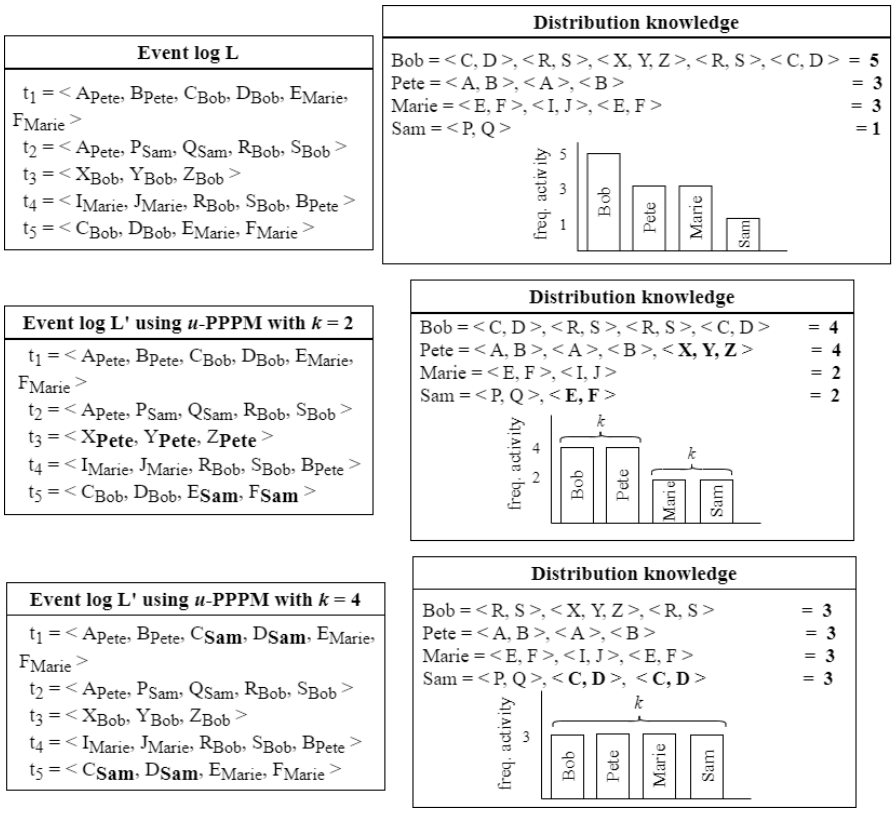

Fig. 2 A sensitive event log L and the resulting privacy-preserved event logs L' once applying u-PPPM with $k = 2$ and $k = 4$.

cases each, thus being indistinguishable in L' from the activity distribution perspective. If we applied a privacy level $k = 4$ then a single group with all four members would be created. It is likely that Bob provides the events corresponding to two (random) cases to Sam (*e.g.,* two cases with events $\langle C, D \rangle$), thus all four individuals appear to be responsible for three cases each. As a result, each individual has the same number of cases in L', and we prevent their distribution-based re-identification.

4.1 Security analysis

This section provides a security analysis of the proposed u-PPPM method in terms of confidentiality and privacy guarantees. Other security requirements, such as integrity, availability and authentication, are beyond the scope of the proposed anonymization method. Due to its offline nature, no online communication is required with external entities to conduct the anonymization procedure. In this sense, the avoidance of communications with further com-

puter systems relaxes the potential security threats, and the security of the method resides in the distortion of the statistical properties of the event data introduced by the very method. For the sake of completeness, it is compared with the security achieved by means of pseudonymization or encryption mechanisms. Let L be the original event log containing sensitive information in clear text, L' be a privacy-preserved event log version of L using u-PPPM, and L'' be an event log version of L with the personally identifiable information pseudonymized or encrypted. L is supposed to be secret and only accessible to the authorized parties, and both L' and L'' are shared with third parties or publicly released.

The confidentiality of the proposed method is supported by the obfuscation of personally identifiable information, by means of pseudonymization or encryption techniques. This property is achieved at the later stage of u-PPPM once all groups are uniformed. Using state-of-the-art encryption algorithms or pseudonymization strategies ensures the confidentiality of the event log L' as long as the private cryptographic keys remain secret. The management of these keys is beyond the scope of this method. Hence, users exploiting L' are not able to decrypt the personally identifiable information and, therefore, it cannot be directly associated to sensitive information in L'. As both L' and L'' rely on pseudonymization or encryption, the confidentiality level of both approaches is the same.

The main benefit of the proposed method is the added privacy guarantees that are not supported when using pseudonymization or encryption only. Comparing the data in L' and L'', the knowledge that an attacker's gains from the

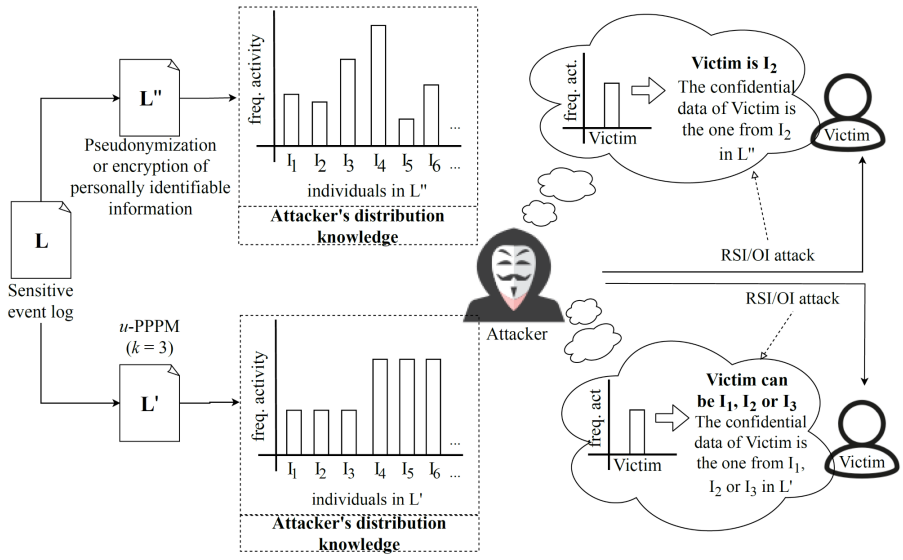

Fig. 3 Illustration of the differences, from the attacker's perspective, between privacy preservation with u-PPPM or with pseudonimization/encryption only.

distribution is significantly different. Whereas the attributes distribution has not been changed in L'' (enabling the re-identification of individuals), in L' the attacker's distribution knowledge has been limited, and re-identification is constrained to groups of k individuals. Formally, whereas L'' describes its attributes with an unknown distribution (*e.g.,* binomial, Poisson...) that might be susceptible to location-oriented attacks, the proposed method reshapes this distribution in L' towards a uniform distribution. The main advantage of using this kind of distribution is that it renders the distribution knowledge that attackers could gain useless, because groups of k individuals are indistinguishable among them. In case of an attacker would be able to infer the probability distribution of a given individual in L', the re-identification risk is upper-bounded by $1/k$. This privacy enhancement is graphically illustrated in Figure 3, in which attackers are unable to re-identify individuals from potentially identifiable distributions. Thus, it can be seen that applying u-PPPM prevents the attack described in Section 3.

Although there is some resemblance between our approach and those studied in related privacy fields such as Statistical Disclosure Control (SDC), we must emphasize that u-PPPM can hardly be compared to the existing SDC privacy-preserving models (*e.g.,* k-anonymity, l-diversity,...) [21]. First, SDC-related protection models aim to protect micro-data sets by modifying their statistical properties with an eye on the utility of the data sets themselves. On the contrary, u-PPPM focuses on minimizing the risks of identity/attribute disclosure against RSI/OI attacks, and evaluates the utility of the discovered process models. Second, data sources in SDC protection models and u-PPPM are different: records in a micro-data set belong to different individuals, while multiple records in an event log could eventually belong to the same individual. Certainly, similarities could be found between u-PPPM and the k-anonymity model: although both approaches create clusters of k individuals, k-anonymity uses distance functions (*e.g.,* Euclidean, Manhattan...) to group homogeneous individuals/records, while u-PPPM creates groups based on the activity distribution knowledge that an attacker could infer by means of RSI/OI attacks, and it does not consider the distance among events in L. Moreover, k-anonymity results in k indistinguishable records in the protected micro-data set, while u-PPPM results in k indistinguishable individuals from a distribution perspective.

5 Evaluation and Discussion

This section evaluates our proposed technique, u-PPPM, and discusses the experimental results obtained using real-life event logs. This evaluation aims to test the impact of u-PPPM on the quality of the process models discovered from protected event logs, by comparing them with the corresponding process models discovered from the original (unprotected) event logs. In particular, since u-PPPM distorts the individuals' information in the event data, the process models to be evaluated are those associated to each individual from

a control-flow perspective, which can be valuable for people's performance analyses. We have measured the impact of u-PPPM on the quality of the process models by answering three questions, as follows:

- *Q1 - Individual distortion:* How similar are the process models (M_I) of an individual I when discovered from the original event log, and those (M'_I) obtained from the protected event log.
- *Q2 - Inter-individual distortion:* Are the differences among the individuals' process models from the original event log ($M_{I_1}, M_{I_2}, \ldots, M_{I_p}$) maintained among the individuals' process models obtained from the protected event log ($M'_{I_1}, M'_{I_2}, \ldots, M'_{I_p}$)?
- *Q3 - Conformance:* How well the event data of an individual in the original event log L conforms with the process model of that individual when discovered from the protected event log (M'_I).

Whereas many evaluation methodologies only focus on the individual distortion of the process models or the conformance, this evaluation methodology aims to provide a holistic view on the impact of u-PPPM with regards to the quality of the process models from three independent perspectives. The rationality of each perspective is discussed in Section 5.1. Figure 4 summarizes and illustrates this evaluation methodology in accordance with the previous questions.

5.1 Experimental Setup

Since the evaluation of u-PPPM is conducted at the level of the very process models, it is necessary to *discover* the process models and represent them using

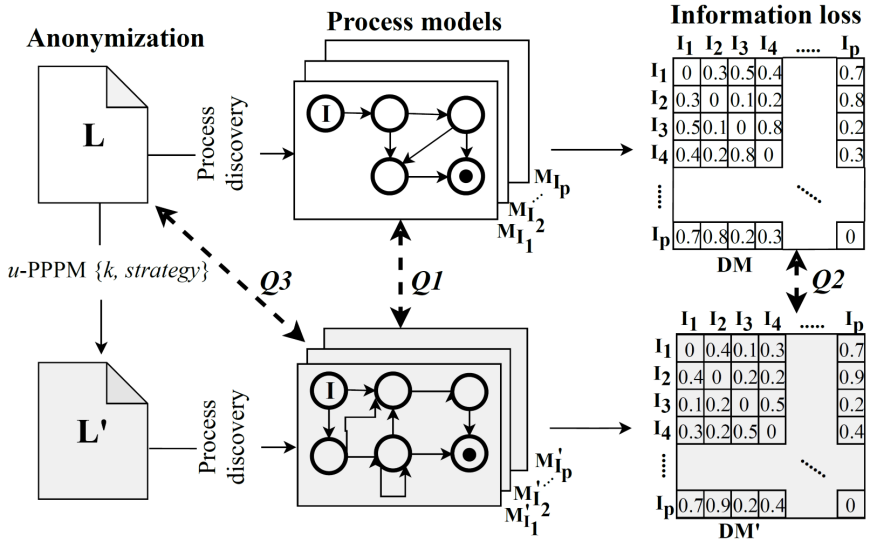

Fig. 4 Evaluation methodology to evaluate the impact of u-PPPM.

Table 2 Properties of the event logs used for the evaluation of u-PPPM.

Name	Availability	#events	#cases	#activities	#resources	Events/case	Cases/resource
BPI12 [10]	Public	262.200	13.087	23	68	20,03	192,46
BPI13 [44]	Public	6.660	1.487	7	584	4,48	2,55
BPI14 [11]	Public	466.155	46.507	39	242	10,02	192,18
BPI15 [12]	Public	262.628	5.649	356	72	46,49	78,46
CoSeLoG [8]	Public	8.577	1.434	27	48	5,98	29,88
Hosp. TGN*	Propietary	122.179	58.836	36	280	2,08	210,13

* This event log was collected by the authors in a real hospital institution in the area of Tarragona (Catalonia, Spain) for the purposes of this research.

a *modeling notation*. This article represents process models as D/F-graphs, a generic notation used to represent the relationships between event data in accordance to dependency/frequency tables created from event logs, as explained in [46,47]. Despite graphs limitations (*i.e.,* concurrency), this notation enables discovering the process models from a generic and high-level perspective, and avoids the constraints and restrictions introduced in subsequent modeling notations, such as Petri nets or BPMN. As a first step towards our novel PPPM technique, this less-restrictive notation enables observing the very impact of u-PPPM on the quality of process models, without considering the restrictive particularities introduced by other modeling notations.

The *parameters of u-PPPM* affect the quality of the obtained process models. Therefore, we have tested u-PPPM with different combinations of its two parameters: the size of the groups k, where $k = \{2, 3, 4, 5, 8, 10\}$, and the selection strategy, *e.g.,* the aforementioned S1, S2, S3 and S4 strategies. With the aim to observe the impact on discovered process models, u-PPPM is executed for all combinations of these parameters. Therefore, each event log L is protected 24 (6×4) times, resulting in 24 different privacy-preserved event logs $L'_{(2,S1)}, L'_{(3,S1)}, \ldots, L'_{(10,S4)}$. In addition, we have evaluated the behavior of u-PPPM in very different event log settings, specifically, u-PPPM has been tested on six real-life *event logs* from multiple domains, as described in Table 2. Note the different nature and properties of those real-life events logs (*i.e.,* number of cases, events, resources/individuals...).

First, we evaluate how u-PPPM affects the *quality of the process models individually* (*Q1*). To do it quantitatively, we proceed in a model-by-model comparison. In this context, this evaluation determines the distortion that u-PPPM introduces in the very process model of each individual. In short, this evaluation is useful to estimate how different a given process model is with regards to its original version. Big differences could jeopardize personalized analyses. The evaluation procedure works as follows: For each u-PPPM execution, with a certain combination of parameters, we first discover the p original process models from L, *i.e.,* $\{M_{I_1}, M_{I_2}, \ldots, M_{I_p}\}$, and the p protected process models from L', *i.e.,* $\{M'_{I_1}, M'_{I_2}, \ldots, M'_{I_p}\}$. Next, the pairs of process models belonging to the same individual, *i.e.,* $\{M_{I_1}, M'_{I_1}\}, \{M_{I_2}, M'_{I_2}\}, \ldots, \{M_{I_p}, M'_{I_p}\}$, are compared using a similarity measure. To do so, we use four well-known similarity measures from the graph-theory literature and compare structural differences according to graph properties: VEO [32], VR [32], Weight Distance (WD) [41] and DeltaCon (DC) [23]. These measures are defined in $[0-1]$,

where 0 means total similarity (*i.e.,* no distortion in the process models). Therefore, each pair of process models results in four similarity measures, thus obtaining a total of $4 \times p$ similarity values for the complete set of process models to evaluate. Taking the average of all these values, a *quality score* (QS) can be associated to the u-PPPM execution. This QS value, between 0 and 1 (the lower, the more similarity), indicates the averaged distortion that a given u-PPPM execution introduces in the protected process models.

Assuming the unavoidable distortion introduced by u-PPPM in the process model of each individual, it is also important to evaluate how this distortion affects the entire event data and the process models as a whole, this is the *inter-individual distortion*. For instance, if the process models of two individuals, M_{I_1} and M_{I_2}, are similar in the original event log L, it is desirable to preserve this similarity in the protected process models M'_{I_1} and M'_{I_2} too (*Q2*). In this case, the insights gathered in case of comparing process models (*e.g.,* for performance analysis) would be similar either comparing the protected process models or the original process models. This evaluation works as follows: For each u-PPPM execution, with a certain combination of parameters, we compute the distance matrices DM and DM' (both with size $p \times p$), which contain the similarity between all pairs of process models in L and L', respectively. These similarity values within DM and DM' are computed using the aforementioned four similarity measures. Thus, given an event log L and a protected event log L', a total of four distance matrices are obtained, $\{DM_{\text{VEO}}, \ldots, DM_{\text{DC}}\}$ and $\{DM'_{\text{VEO}}, \ldots, DM'_{\text{DC}}\}$. Pairs of matrices calculated with the same similarity measure, *i.e.,* $\{DM_{\text{VEO}}, DM'_{\text{VEO}}\}, \ldots \{DM_{\text{DC}}, DM'_{\text{DC}}\}$, can be compared using the well-known Mean Absolute Error (MAE), which is defined in $[0-1]$, where lower values indicate a lower information loss. By taking the average of the four MAE results, an *information loss score* (ILS) can be associated to the u-PPPM execution. The value of the ILS shows the averaged information loss in the protected event log in comparison to the original event log.

Last but not least, the quality of the protected process models can also be measured against the original event data (*Q3*), which resembles *conformance checking* techniques. This evaluation shows whether the original behavior (the event data of an individual I_1 in L) can be replayed in the distorted process model M'_{I_1} of that individual. From a quality perspective, it is ideal that all (or most) of the original behaviors could be preserved, despite the distortion introduced by u-PPPM, and reproduced in the protected process models. This evaluation procedure works as follows: For each u-PPPM execution with a given combination of parameters, we discover the p protected process models from L', $\{M'_{I_1}, M'_{I_2}, \ldots, M'_{I_p}\}$. Also, we extract the original event data for each individual in L, which are then grouped by case. Then, the different cases of each individual $I_1, I_2, \ldots, I_p$ are replayed in their corresponding process models $\{M'_{I_1}, M'_{I_2}, \ldots, M'_{I_p}\}$. Since processes are modeled as D/F-graphs, a case is replayed only if there exists a graph connection between all the consecutive events in the case. By replaying all the cases of an individual, a replay score can be associated to an individual by dividing the number of cases that have been successfully replayed by the total number of possible cases. This value,

defined in $[0-1]$, shows the degree of conformance of a given individual: high values indicate that the original behavior is preserved and can be replayed in the protected process models. By taking the average of the replay scores of all the individuals in L, a *conformance score* (CS) can be associated to the u-PPPM execution. The CS shows the averaged conformance level of the original event data with regard to the protected process models.

5.2 Experimental Results and Discussion

The results obtained by following the evaluation methodology described above are shown in Figures 5 to 10, which illustrates the QS, ILS and CS results after applying the proposed u-PPPM technique using different combinations of parameters on the six event logs described in Table 2. Each figure consists of four charts: the first indicates the correlation between ILS-CS results, the second represents the correlation between QS-CS results, the third represents the correlation between QS-ILS results, and the fourth combines the aforementioned planes to build a 3D space and shows the correlation between QS, ILS, and CS results. The background color of each projection plane helps to compose the planes and visualize the 3D perspective. In each chart, every point corresponds to a u-PPPM execution with a certain combination of parameters: a group size k (represented with different colors) and a selection strategy (represented with different shapes). To globally evaluate the generic impact of u-PPPM regardless of the particularities of each event log, Figure 11 shows the averaged results of all event logs. It is worth mentioning that, since the execution of u-PPPM is non-deterministic, these results correspond to the average of five executions for each combination of parameters.

Experimental results show that u-PPPM behaves consistently regardless the event log to which it is applied, and similar trends can be observed when u-PPPM parameters vary. Notwithstanding, the data-dependence nature of u-PPPM makes that the quality of the protected process models (either evaluated using QS, ILS or CS) depends on the event data. For instance, although executing the method with the same parameters, the results from QS, ILS or CS vary with each event log (*e.g.,* by using $k = 5$ and S1 in BPI12 and BPI15, the QS results are 0.232 and 0.31, respectively). Therefore, to evaluate the high-level impact of the u-PPPM parameters on the quality of the process models, the averaged results (Figure 11) are specially useful.

With regards to the u-PPPM parameters, it is clearly observed that their selection directly affects the quality of the process models. Indeed, increasing the privacy level k negatively affects the process models quality. The larger k, the more individuals share the same distribution, so the more difficult for the attacker is to re-identify them. However, this introduces more distortion in the protected process models, as observed. This result is aligned with those of privacy-preserving techniques applied in other fields (*e.g.,* SDC, PPDM...): data utility decreases at higher privacy levels. However, it is worth noting that this is not always true in this method, since the selection strategy can

slightly contribute to minimize the distortion at a given privacy level. In this case, it can be observed that S2 is the best strategy in all cases since it is designed to select the two most appropriate individuals (from a distribution perspective), avoiding the use of probabilistic or random guesses, which results into a minimization of the event data distortion. In opposition to S2, the worst strategy is S4, hence, demonstrating that its iterative design propagates the distortion among all individuals in the group, resulting in a quality decrease. For instance, better results can be achieved at higher privacy levels when using the right strategy, *e.g.,* QS, ILS and CS results using $k = 5$ and S2 in BPI15 are better than those using $k = 4$ and S3.

Interestingly enough, the evaluation of the process models from different perspectives (*e.g.,* QS, ILS and CS) has an apparent direct correlation. This suggests that the protection of event logs using u-PPPM introduces a uniform/homogeneous distortion in the process models: the more distortion in the process models individually (QS), the more distant the relationships among them (ILS), and less conformed with the original event data (CS). In particular, despite the individual distortion of the protected process models (*i.e.,* QS results could be relatively high at large privacy levels), the relationships between the different process models are notably preserved (*i.e.,* ILS results are much lower) and the original event data mostly conforms with the protected process models (*i.e.,* CS results are slightly worse). In this sense, the protected event logs preserve most of the patterns from the original event logs at the same time that the individual results have been protected/anonymized,

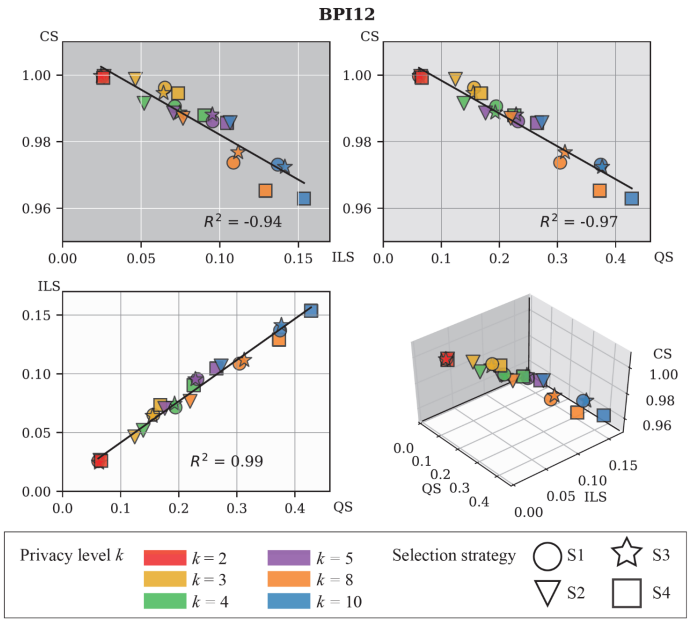

Fig. 5 Correlation between the QS, ILS and CS results using BPI12.

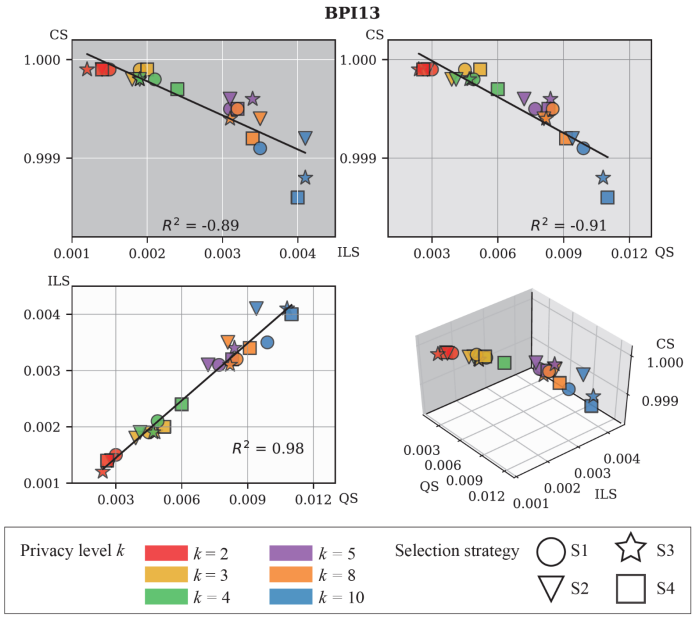

Fig. 6 Correlation between the QS, ILS and CS results using BPI13.

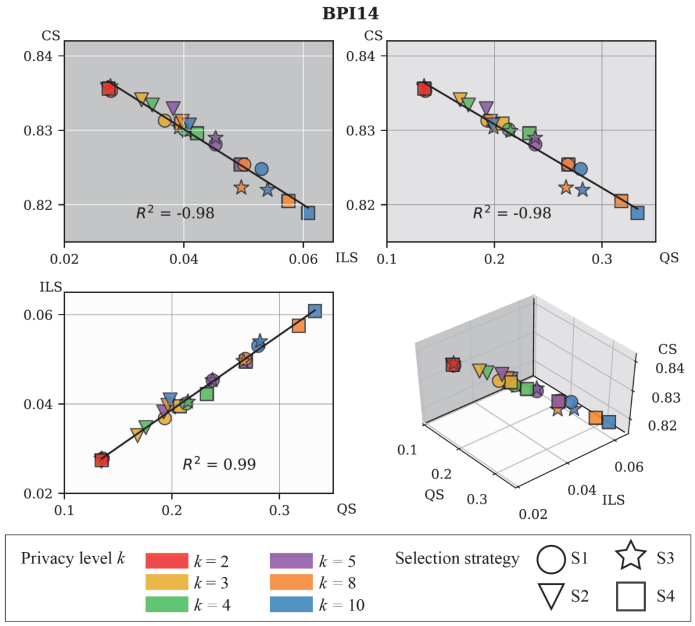

Fig. 7 Correlation between the QS, ILS and CS results using BPI14.

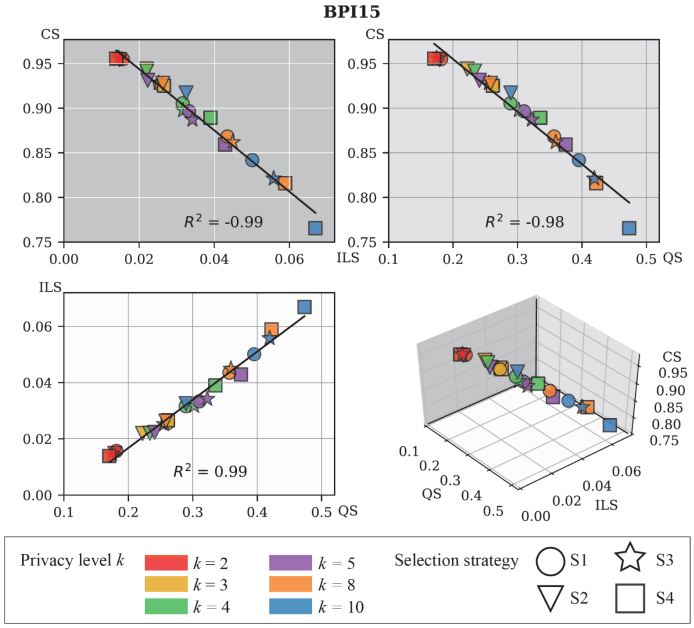

Fig. 8 Correlation between the QS, ILS and CS results using BPI15.

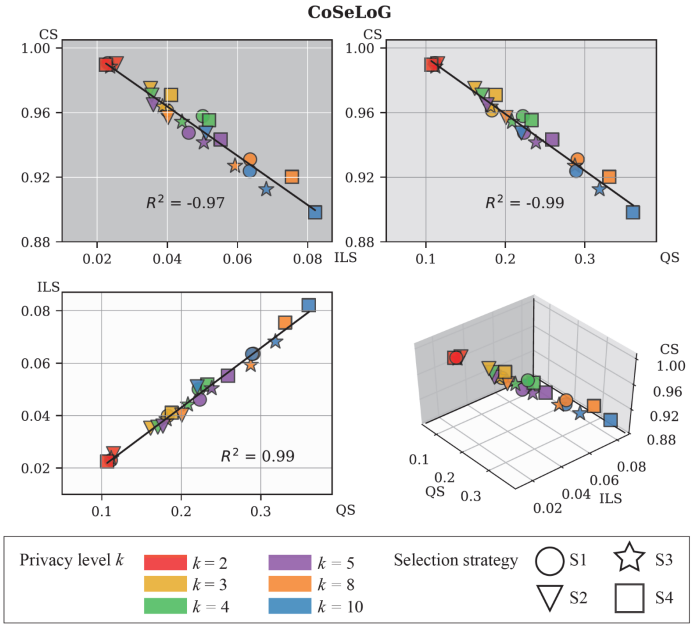

Fig. 9 Correlation between the QS, ILS and CS results using CoSeLoG.

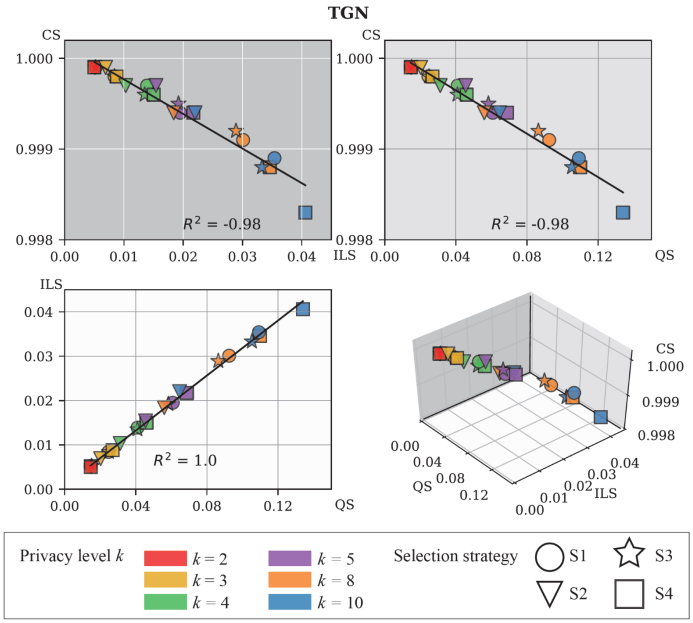

Fig. 10 Correlation between the QS, ILS and CS results using TGN.

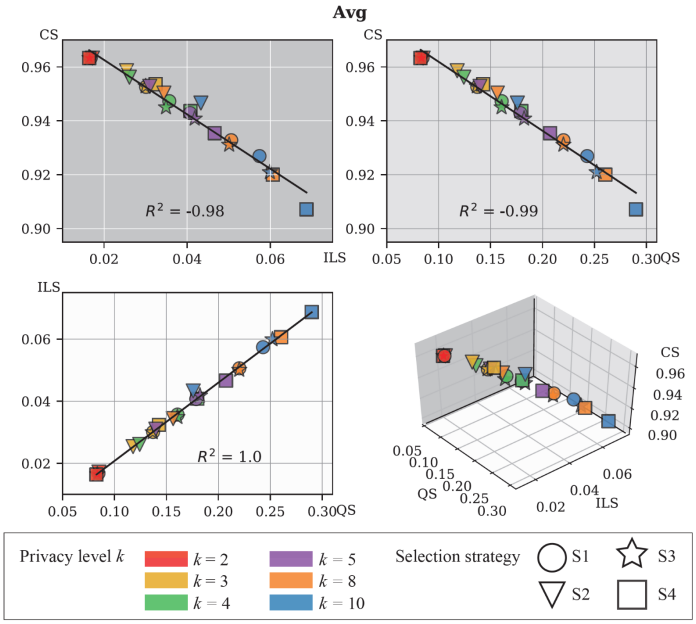

Fig. 11 Average correlation between all the QS, ILS and CS results.

thus preserving individuals' privacy. From a privacy perspective, this property enables the gathering of similar information or conclusions from the protected event logs (and the discovered protected process models) as if they were obtained from the original event logs (and the original process models too), but without releasing sensitive information.

Finally, as an illustrative example, with the aim to visually analyze the distortion from a qualitative perspective, Figure 12 depicts the resulting process models of a given individual within the CoSeLoG event log after applying the u-PPPM method with different parameters.

6 Conclusion

Process mining is an emerging discipline that studies and creates solutions to analyze vast amounts of event log data, which enables the proper management of business processes within organizations. Due to its important results, PM is attracting increasing attention, and many efforts are being devoted to developing process discovery algorithms, approaching processes from multiples perspectives, and providing strategic process visualizations. However, there are still important challenges to be met. Among those, considering privacy throughout the entire PM analysis is one of utmost interest, since there are serious privacy risks associated with the modeling of business processes that may allow an attacker to disclose confidential data (*e.g.,* especially when dealing with confidential information such as that in the healthcare sector).

We have observed that in order to lessen risks for people's privacy, privacy-preserving methodologies must be applied during PM analyses. Hence, in this article, we have shown the importance of *Privacy-Preserving Process Mining* (PPPM) techniques, which is a very young and promising research direction within the PM field. This article represents a step forward towards the understanding of PPPM and aims at motivating researchers to continue with further in-depth studies on novel privacy-preserving techniques to guarantee individuals privacy during PM analyses under a wide variety of attacker models.

We have shown that common approaches to protect privacy in event logs, such as pseudonimization and attributes encryption, are not robust against attacker models built upon distribution-based attacks because they cannot break the link between personally identifiable information and confidential data. This is specially relevant for event logs that are collected in public places (*e.g.,* hospitals) in which attackers can easily perform location-oriented targeted attacks, such as RSI and OI attacks, and acquire background knowledge based on the distribution of events' attributes. With the aim to protect users' privacy against the aforementioned attacks, in this article, we have presented u-PPPM, a uniformization-based PPPM technique that distorts attributes distributions in event logs, and averts distribution-based re-identification. By defining a privacy level and an individuals selection strategy (four strategies have been defined in this article), the proposed method conducts a group-based

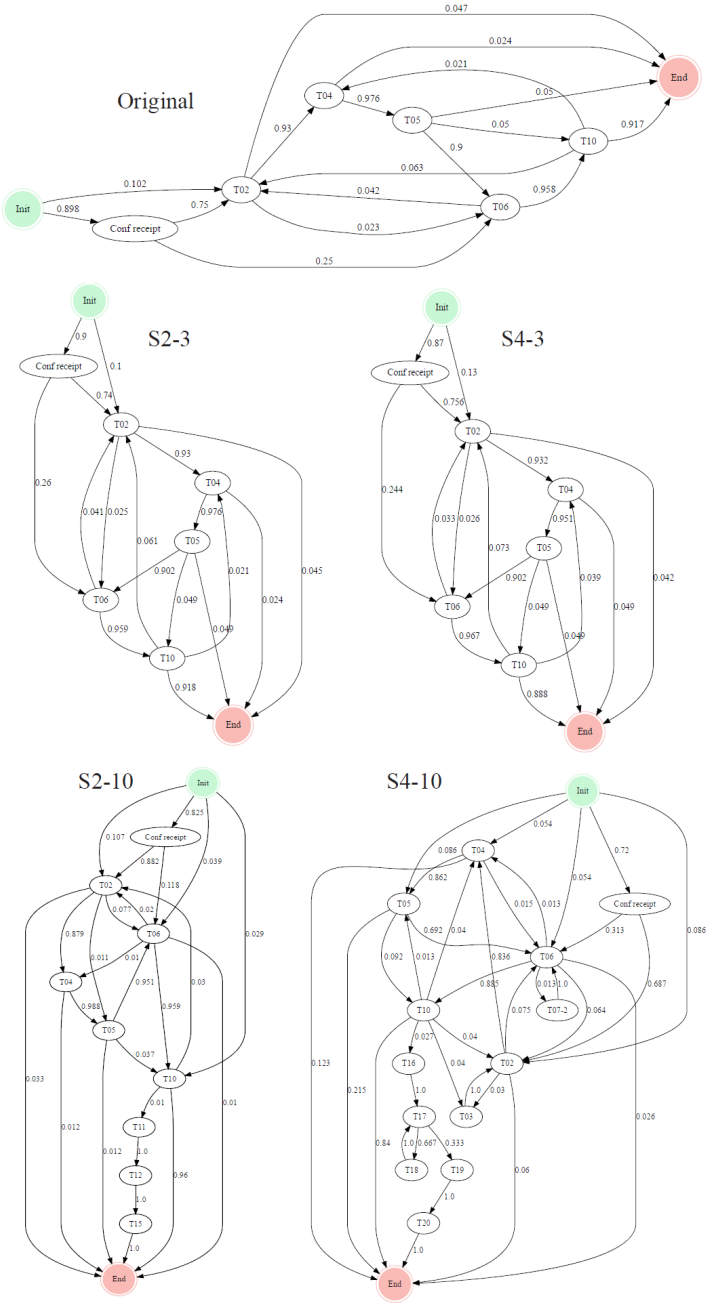

Fig. 12 Process models of a certain individual discovered from the different event logs protected with different combinations of u-PPPM parameters ($k = 3$ and 10, and selection strategy S2 and S4).

anonymization that exchanges events among individuals with the aim to distort these potentially identifiable distributions, thus rendering the attackers background knowledge useless.

As any privacy protection method, the distortion of (event) data worsens the quality of the PM results. Consequently, we have assessed the impact of u-PPPM over the obtained people's process models by using six real-life event logs from multiple domains. To do so, we have designed a quality assessment methodology based on the measurement of dissimilarities among process models discovered in the original event log and their corresponding version discovered in privacy-preserved event logs (obtained after applying u-PPPM), which allows us to quantify the distortion introduced by our solution. In particular, these measurements have been analyzed from three perspectives: the individual distortion suffered by process models individually, the inter-individual distortion, this is, the differences among all process models, and the conformance level with the original event data. Experimental results are aligned with those in related disciplines and have shown that increasing the privacy level increases the distortion (hence, reducing the utility of the process models) and the proper selection strategy helps to keep the distortion under control. In particular, it is shown that probabilistic and iterative strategies should be avoided to mitigate propagating the distortion introduced in all the resulting process models. Correlations between the results suggest that the proposed method introduces an homogeneous distortion on the process models.

Although the contributions in this article are a step forward in the field of PPPM, the research community is only scratching the surface of the many opportunities within this field. Future work will focus on the development of novel PPPM techniques able to cope with even more complex attacker models. We foresee the creation and application of robust privacy-preserving models, which incorporate properties such as k-anonymity or l-diversity into PM to prevent targeted attacks. Moreover, extending the proposed technique beyond graphs models, such as Petri nets or BPMN, will be a primary research line in the field. Finally, the differences between the original and the protected models could be assessed qualitatively, by asking experts and practitioners whether those differences could change their understanding and the decisions they would make. This, too, will be an interesting research line to pursue in the near future.

Acknowledgements The authors are supported by the Government of Catalonia (GC) with grant 2017-DI-002. A. Solanas is supported by the GC with project 2017-SGR-896, and by Fundació PuntCAT with the Vinton Cerf Distinction, and by the Spanish Ministry of Science & Technology with project IoTrain - RTI2018-095499-B-C32, and by the EU with project LOCARD (Grant Agreement no. 832735). Pictures designed by Freepik.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. van der Aalst, W.M.P.: Process Mining: Discovery, Conformance and Enhancement of Business Processes. Springer (2011)
2. van der Aalst, W.M.P.: Process Mining: Data Science in Action. Springer (2016)
3. van der Aalst, W.M.P.: Responsible Data Science: Using Event Data in a "People Friendly" Manner. In: Proceedings of the 18th International Conference on Enterprise Information Systems, pp. 3–28. Rome, Italy (2016)
4. van der Aalst, W.M.P., Adriansyah, A., Alves de Medeiros, A.K., Arcieri, F., Baier, T., Blickle, T., Bose, J.C., van den Brand, P., Brandtjen, R., and Buijs, J., et al.: Process Mining Manifesto. In: Proceedings of the 9th International Conference on Business Process Management, pp. 169–194. Clermont-Ferrand, France (2011)
5. Batista, E., Solanas, A.: Process Mining in Healthcare: A Systematic Review. In: Proceedings of the 9th International Conference on Information, Intelligence, Systems Applications, pp. 1–6. Zakynthos, Greece (2018)
6. Bauer, M., Fahrenkrog-Petersen, S., 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 the 17th International Conference on Business Process Management, pp. 1–5. Vienna, Austria (2019)
7. Brunk, J., Riehle, D.M. and Delfmann, P.: Prediction of Customer Movements in Large Tourism Industries by the Means of Process Mining. Research Papers **40**, pp. 1–16 (2018)
8. Buijs, J.C.A.M.: Receipt phase of an environmental permit application process ('WABO'), CoSeLoG project (2014). Eindhoven University of Technology. Dataset. `https://doi.org/10.4121/uuid:a07386a5-7be3-4367-9535-70bc9e77dbe6`
9. Burattin, A., Conti, M., Turato, D.: Toward an Anonymous Process Mining. In: Proceedings of the 3rd International Conference on Future Internet of Things & Cloud, pp. 58–63. Rome, Italy (2015)
10. van Dongen, B.F.: BPI Challenge 2012 (2012). 4TU. Centre for Research Data. Dataset. `https://doi.org/10.4121/uuid:3926db30-f712-4394-aebc-75976070e91f`
11. van Dongen, B.F.: BPI Challenge 2014: Activity log for incidents (2014). 4TU. Centre for Research Data. Dataset. `https://doi.org/10.4121/uuid:86977bac-f874-49cf-8337-80f26bf5d2ef`
12. van Dongen, B.F.: BPI Challenge 2015 (2015). 4TU. Centre for Research Data. Dataset. `https://doi.org/10.4121/uuid:31a308ef-c844-48da-948c-305d167a0ec1`
13. Duma, D., Aringhieri, R.: An ad hoc process mining approach to discover patient paths of an Emergency Department. Flexible Services and Manufacturing Journal **32**(1), pp. 6–34 (2020)
14. European Union: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union **L119**, pp. 1–88 (2016)
15. Fahrenkrog-Petersen, S.A.: Providing Privacy Guarantees in Process Mining. In: Proceedings of the 31st International Conference on Advanced Information Systems Engineering – Doctoral Consortium, pp. 23–32. Rome, Italy (2019)
16. Fahrenkrog-Petersen, S.A., van der Aa, H., Weidlich, M.: PRETSA: Event Log Sanitization for Privacy-aware Process Discovery. In: Proceedings of the 1st International Conference on Process Mining, pp. 1–8. Aachen, Germany (2019)
17. Garcia, C.d.S., Meincheim, A., Faria Junior, E.R., Dallagassa, M.R., Sato, D.M.V., Carvalho, D.R., Santos, E.A.P., Scalabrin, E.E.: Process mining techniques and applications – A systematic mapping study. Expert Systems with Applications **133**, pp. 260–295 (2019)
18. Gatta, R., Vallati, M., Fernandez-Llatas, C., Martinez-Millana, A., Orini, S., Sacchi, L., Lenkowicz, J., Marcos, M., Munoz-Gama, J., Cuendet, M., de Bari, B., Marco-Ruiz, L., Stefanini, A., Castellano, M.: Clinical Guidelines: A Crossroad of Many Research Areas. Challenges and Opportunities in Process Mining for Healthcare. In: Proceedings of the 17th International Conference on Business Process Management, pp. 545–556. Vienna, Austria (2019)

19. Ge, C., Susilo, W., Liu, Z., Xia, J., Szalachowski, P., Liming, F.: Secure Keyword Search and Data Sharing Mechanism for Cloud Computing. IEEE Transactions on Dependable and Secure Computing, pp. 1–14 (2020)
20. Ge, C., Yin, C., Liu, Z., Fang, L., Zhu, J., Ling, H.: A Privacy Preserve Big Data Analysis System for Wearable Wireless Sensor Network. Computers & Security, pp. 101887 (2020)
21. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E.S., Spicer, K., de Wolf, P.P.: Statistical Disclosure Control. John Wiley & Sons (2012)
22. Isaak, J., Hanna, M.J.: User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection. Computer **51**(8), pp. 56–59 (2018)
23. Koutra, D., Vogelstein, J.T., Faloutsos, C.: DeltaCon: A Principled Massive-Graph Similarity Function. In: Proceedings of the SIAM International Conference on Data Mining, pp. 162–170. Austin, USA (2013)
24. Kurniati, A.P., Hall, G., Hogg, D., Johnson, O.: Process mining in oncology using the MIMIC-III dataset. Journal of Physics: Conference Series **971**, pp. 012008 (2018)
25. Liu, C., Duan, H., Zeng, Q., Zhou, M., Lu, F., Cheng, J.: Towards Comprehensive Support for Privacy Preservation Cross-organization Business Process Mining. IEEE Transactions on Services Computing **12**(4), pp. 1–15 (2016)
26. Machin, J., Solanas, A.: Conceptual Description of Nature-Inspired Cognitive Cities: Properties and Challenges. In: Proceedings of the International Work-Conference on the Interplay Between Natural and Artificial Computation, pp. 212–222. Almeria, Spain (2019)
27. Mannhardt, F., Koschmider, A., Baracaldo, N., Weidlich, M., Michael, J.: Privacy-Preserving Process Mining. Business & Information Systems Engineering **61**(5), pp. 595–614 (2019)
28. Mannhardt, F., Petersen, S.A., de Oliveira, M.F.D.: Privacy Challenges for Process Mining in Human-centered Industrial Environments. In: Proceedings of the 14th International Conference on Intelligent Environments, pp. 1–8. Rome, Italy (2018)
29. Michael, J., Koschmider, A., Mannhardt, F., Baracaldo, N., Rumpe, B.: User-Centered and Privacy-Driven Process Mining System Design for IoT. In: Proceedings of the 31st International Conference on Advanced Information Systems Engineering, pp. 194–206. Rome, Italy (2019)
30. Moreira, C., Haven, E., Sozzo, S., Wichert, A.: Process mining with real world financial loan applications: Improving inference on incomplete event logs. PloS ONE **13**(12), pp. e0207806 (2018)
31. Nuñez von Voigt, S., Fahrenkrog-Petersen, S.A., Janssen, D., Koschmider, A., Tschorsch, F., Mannhardt, F., Landsiedel, O., Weidlich, M.: Quantifying the Re-identification Risk of Event Logs for Process Mining. In: Proceedings of the 32nd International Conference on Advanced Information Systems Engineering, pp. 252–267. Grenoble, France (2020)
32. Papadimitriou, P., Dasdan, A., Garcia-Molina, H.: Web graph similarity for anomaly detection. Journal of Internet Services and Applications **1**(1), pp. 19–30 (2010)
33. Papageorgiou, A., Strigkos, M., Politou, E., Alepis, E., Solanas, A., Patsakis, C.: Security and Privacy Analysis of Mobile Health Applications: The Alarming State of Practice. IEEE Access **6**, pp. 9390–9403 (2018)
34. Pika, A., Wynn M.T., Budiono, S., ter Hofstede, A.H.M., van der Aalst, W.M.P., Reijers, H.A.: Towards Privacy-Preserving Process Mining in Healthcare. In: Proceedings of the 2nd International Workshop on Process-Oriented Data Science for Healthcare, pp. 1–12. Vienna, Austria (2019)
35. Rafiei, M., van der Aalst, W.M.P.: Mining Roles From Event Logs While Preserving Privacy. In: Proceedings of the 17th International Conference on Business Process Management – Workshop Security and Privacy-enhanced Business Process Management, pp. 1–12. Vienna, Austria (2019)
36. Rafiei, M., van der Aalst, W.M.P.: Practical Aspect of Privacy-Preserving Data Publishing in Process Mining. arXiv preprint arXiv:2009.11542, pp. 1–5 (2020)
37. Rafiei, M., van der Aalst, W.M.P.: Privacy-Preserving Data Publishing in Process Mining. In: Proceedings of the 18th International Conference on Business Process Management, pp. 122–138. Seville, Spain (2020)

38. 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 & Analysis, pp. 3–17. Seville, Spain (2018)
39. Rafiei, M., Wagner, M., van der Aalst, W.M.P.: *TLKC*-Privacy Model for Process Mining. In: Proceedings of the 14th International Conference on Research Challenges in Information Science, pp. 398–416. Limassol, Cyprus (2020)
40. Ren, Y., Zhu, F., Sharma, P.K., Wang, T., Wang, J., Alfarraj, O., Tolba, A.: Data Query Mechanism Based on Hash Computing Power of Blockchain in Internet of Things. Sensors **20**(1), pp. 207 (2020)
41. Shoubridge, P., Kraetzl, M., Wallis, W.A.L., Bunke, H.: Detection of abnormal change in a time series of graphs. Journal of Interconnection Networks **3**(01n02), pp. 85–101 (2002)
42. Solanas, A., Casino, F., Batista, E., Rallo, R.: Trends and challenges in smart healthcare research: A journey from data to wisdom. In: Proceedings of the 3rd International Forum on Research and Technologies for Society and Industry, pp. 1–6. Trento, Italy (2017)
43. Solanas, A., Patsakis, C., Conti, M., Vlachos, I.S., Ramos, V., Falcone, F., Postolache, O., Pérez-Martínez, P.A., Di Pietro, R., Perrea, D.N., Martínez-Ballesté, A.: Smart health: A context-aware health paradigm within smart cities. IEEE Communications Magazine **52**(8), pp. 74–81 (2014)
44. Steeman, W.: BPI Challenge 2013, closed problems (2013). Ghent University. Dataset. `https://doi.org/10.4121/uuid:c2c3b154-ab26-4b31-a0e8-8f2350ddac11`
45. Tillem, G., Erkin, Z., Lagendijk, R.L.: Privacy-Preserving Alpha Algorithm for Software Analysis. In: Proceedings of the International Symposium on Information Theory and Signal Processing in the Benelux, pp. 136–143. Louvain-la-Neuve, Belgium (2016)
46. Weijters, A.J.M.M., van der Aalst, W.M.P.: Rediscovering Workflow Models from Event-Based Data using Little Thumb. Integrated Computer-Aided Engineering **10**(2), pp. 151–162 (2003)
47. Weijters, A.J.M.M., van der Aalst, W.M.P., Alves de Medeiros, A.K.: Process Mining with the Heuristics Miner Algorithm. Technische Universiteit Eindhoven, Tech. Rep. WP **166**, pp. 1–34 (2006)
48. Weske, M.: Business Process Management – Concepts, Languages, Architectures. Springer (2007)
49. Wu, Q., He, Z., Wang, H., Wen, L., Yu, T.: A Business Process Analysis Methodology Based on Process Mining for Complaint Handling Service Processes. Applied Sciences **9**(16), pp. 3313 (2019)