Journal of Information Security and Applications 68 (2022) 103235

Contents lists available at ScienceDirect

Journal of Information Security and Applications

journal homepage: www.elsevier.com/locate/jisa

Privacy-preserving process mining: A microaggregation-based approach

Edgar Batista [a,b], Antoni Martínez-Ballesté [a], Agusti Solanas [a,*]

[a] Universitat Rovira i Virgili, Department of Computer Engineering and Mathematics, Av. Països Catalans 26, 43007 Tarragona, Catalonia, Spain
[b] SIMPPLE S.L., C. Joan Maragall 1A, 43003 Tarragona, Catalonia, Spain

ARTICLE INFO

Keywords:
Privacy-preserving process mining
Process mining
Privacy preservation
Microaggregation
k-anonymity
Confidentiality
Anonymization

ABSTRACT

The proper exploitation of vast amounts of event data by means of process mining techniques enables the discovery, monitoring and improvement of business processes, allowing organizations to develop more efficient business intelligence systems. However, event data often contain personal and/or confidential information that, unless properly managed, may jeopardize people's privacy while conducting process mining analysis. Despite its relevance, privacy aspects have barely been considered within process mining, and the field of privacy-preserving process mining is still in an embryonic stage.

With the aim to protect people's privacy, this article presents a novel privacy-preserving process mining method based on microaggregation techniques, called *k*-PPPM, that increases privacy in process mining through *k*-anonymity. Contrary to current solutions, mostly based on pseudonyms and encryption, this method averts the re-identification of targeted individuals from attacks based on the analysis of process models in combination with location-oriented attacks, such as Restricted Space Identification and Object Identification attacks. The proposed method provides adjustable parameters to tune different anonymization aspects. Six real-life event logs have been employed to evaluate the method in terms of process models quality and information loss.

1. Introduction

The data digitalization tendency has forced organizations to adapt their classical management and monitoring models towards ICT-based models. Business processes (or processes in short) are widely used in the management arena to describe the set of activities to be conducted as a way to meet business objectives [1]. Indeed, the success of an organization mostly depends on the correct design, execution and management of their processes [2]. Therefore, devoting efforts to correctly designing, monitoring and managing the execution of business processes is paramount for any organization. However, this represents a time-consuming and laborious task, especially in large organizations having lots of complex processes executed concurrently.

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 recorded by the organizational information systems and then stored in chronological order in the so-called event logs. The objective perspective of event logs on the execution of business processes makes them ideal for analytical purposes. The meaningful exploitation of event data allows discovering actual process executions, detecting misalignments between ideal and actual process executions, identifying bottlenecks and proposing corrective countermeasures, optimizing resources, minimizing costs or shortening times, among others. All in all, event logs contain valuable information for organizations that, if properly exploited, can benefit them to be more efficient, competitive and sustainable.

To meet the aforementioned challenges, the research field of *process mining* (PM) [3] aims to develop automated process-oriented analyses to extract high-level knowledge from event logs readily available in today's information systems. In general, three types of PM techniques could be observed: (i) process discovery, *i.e.,* represent and visualize process models from different perspectives (*e.g.,* control-flow, organizational, time, data...) using the information from event logs without any additional knowledge, (ii) conformance checking, *i.e.,* assess deviations between existing (and ideal) process models and those reflected in the event logs, and (iii) process enhancement, *i.e.,* extend or improve existing process models using previous processes executions information recorded in the event logs so as to better reflect reality. Many organizations from different application domains, ranging from healthcare and finance to manufacturing and logistics [4], are adopting PM techniques to conduct internal analyses and readjust themselves to the market demands.

With the advent of the Internet of Events [5], event logs have increased in both size and complexity at a dizzying pace. In general, event logs contain attributes related to the activity that has been conducted, the resource responsible for that activity (typically, an individual) and

* Corresponding author.
E-mail addresses: edgar.batista@urv.cat (E. Batista), antoni.martinez@urv.cat (A. Martínez-Ballesté), agusti.solanas@urv.cat (A. Solanas).

https://doi.org/10.1016/j.jisa.2022.103235

Available online 16 June 2022

the timestamp of its completion. Notwithstanding, in some application domains, event logs might contain personally identifiable information (*e.g.,* full name, national ID number, passport number or social security number) or confidential information (*e.g.,* health conditions, socioeconomic status, ethnicity, sexual orientation or beliefs). Unless properly managed, the release of these event logs, either premeditated or accidental, clearly jeopardizes people's privacy. To prevent these setbacks, legal privacy regulations are gaining importance to settle the proper management and processing of digital information. For instance, the GDPR in the European Union [6] forbids the release or sharing of potentially harmful information among organizations, such as unprotected event logs in this case, and encourages the application of privacy-by-design principles, such as lawful processing, data minimization, pseudonymization or encryption, among others. Privacy issues are more apparent in certain domains, such as the healthcare that deals with highly confidential data (*e.g.,* patients' health conditions, treatments, illnesses, etc.). Indeed, the continuous evolution of healthcare paradigms, such as smart and cognitive health [7,8], promotes the use of PM analyses to acquire valuable knowledge and improve care services, optimize resources and shorten treatment times [9]. Unfortunately, despite experts advice [10], privacy aspects are barely considered within this field and, when addressed, countermeasures are mainly based on pseudonymization or encryption [11].

To this end, *privacy-preserving process mining* (PPPM) [12], an emerging research direction in PM, studies means to preserve people's privacy during PM analyses so as to avoid disclosing personal or confidential data to unauthorized parties. These techniques imply the distortion (*i.e.,* transformation) of event data, which impact on the utility of the PM results, typically by lowering the quality of process models. Hence, the main challenge of PPPM is to determine the best approach to distort event data to counteract specific attacks, so that the quality of the PM results is minimally affected, while reducing or averting disclosure risks. These disclosure risks could either refer to (i) the seamless re-identification of people from either event data or process models (*i.e.,* identity disclosure), or (ii) the inference of confidential information from the event log to personally identifiable information (*i.e.,* attribute disclosure). Although this trade-off between privacy and data utility is well-reported in the literature on privacy protection, there is little literature on privacy in the PM discipline.

1.1. Contribution and plan of the article

This article presents a specific, but very serious, concern that might lead to privacy breaches if attackers properly exploit pseudonymized or encrypted event logs through location-oriented targeted attacks. To face this issue, we present a novel PPPM method, called *k*-PPPM, based on microaggregation techniques to increase privacy through *k*-anonymity during PM analyses. To the best of our knowledge, this is the first approach using microaggregation to protect people's privacy during PM. The proposed method has been tested using six real-life event logs, where different approaches to minimize information loss on the anonymized process mining results have been assessed.

The rest of the article is organized as follows. Section 2 provides some background on statistical disclosure control and microaggregation, and discusses current literature on PPPM. Next, in Section 3, we illustrate the shortcomings of applying simple (but classical) strategies to achieve confidentiality in PM, like pseudonymization or encryption, which may not be enough to protect individuals' privacy under certain situations. In particular, we describe an attacker model based on the inference of process models that enables people re-identification, when combined with location-oriented attacks against targeted individuals, such as restricted space identification (RSI) and object identification (OI) attacks. To counteract these attacks, Section 4 presents *k*-PPPM, a novel privacy-preserving technique based on microaggregation that addresses PPPM through *k*-anonymity. As a result, the proposed method modifies the event data so as to guarantee *k*-anonymity with an eye to minimize the distortion of the protected process models. Then, Section 5 evaluates the impact of the proposed *k*-PPPM method by measuring the distortion introduced to the process models discovered from the protected event logs after applying *k*-PPPM in comparison to the original event logs. Experiments have been tested with six real-life event logs, and have demonstrated the usefulness of our solution. Finally, the article concludes in Section 6 with some final remarks.

2. Background

The goal of this section is twofold. First, Section 2.1 introduces general concepts about privacy protection and privacy models, such as statistical disclosure control, microaggregation and the *k*-anonymity model. Later, Section 2.2 provides relevant literature addressing PPPM.

2.1. Privacy protection: Statistical disclosure control and microaggregation

Many privacy protection strategies have been proposed to face privacy issues at different stages, including data storage or data retrieval, such as database privacy, search engine privacy or private information retrieval, among others. In particular, we concentrate on statistical disclosure control (hereafter, SDC), which lays the foundation for the design of the presented PPPM method.

SDC comprises a group of methods aiming at reducing the risk of disclosing sensitive information about individuals [13]. These techniques are to be applied before the release of the so-called microdata sets, *i.e.,* datasets containing records with information referring to identifiable individuals. Each record contains multiple attributes: direct identifiers (*e.g.,* full name or national ID number), key attributes (or quasi-identifiers, *e.g.,* genre or birth date), and other data attributes, either confidential (*e.g.,* medical, financial or beliefs) or non-confidential. SDC techniques modify the information from microdata sets in such a way that (i) individuals cannot be re-identified or associated to any record from the released microdata set (*i.e.,* identity disclosure), or (ii) confidential data cannot be inferred to individuals based on the released microdata set (*i.e.,* attribute disclosure). In general, minimizing the disclosure risk, which increases the privacy level, results into a poor data utility. In this sense, SDC can be observed as an optimization problem due to the existing trade-off between the disclosure risk and the data utility.

There exist several categories of SDC methods, such as data suppression, swapping, noise addition or microaggregation. In particular, microaggregation [14] relies in the publication of microdata sets whose records are indistinguishable from, at least, *k*–1 other records (where *k* is the privacy threshold). This procedure implies grouping records into clusters, *i.e.,* creating a *k*-partition, according to two conditions: (i) each cluster must contain, at least, *k* records, and (ii) the within-cluster records should be as homogeneous as possible. Once all records are clustered, they are replaced by the average of the cluster's records, *i.e.,* the centroid. Nevertheless, determining the optimal *k*-partition is an NP-hard problem [15], thus heuristic approaches are used to determine an approximate optimal solution in a reasonable time. The Maximum Distance to Average Vector (MDAV) [16], *k*-member [17] and One-pass K-means Algorithm (OKA) [18] are well-known microaggregation-oriented clustering heuristics. It is worth emphasizing that these algorithms differ from classical clustering algorithms, such as *k*-means or hierarchical clustering. Whereas microaggregation-oriented clustering algorithms require a minimum number of records per cluster (the privacy threshold), classical clustering algorithms are constrained by the number of clusters, without a minimum number of records per cluster. By replacing records with centroids, microaggregation provides the microdata set with the *k*-anonymity property [19]. Despite its limitations, which are overcome in more robust models such as *l*-diversity, *t*-closeness or *p*-sensitivity, *k*-anonymity still stands as a widely accepted measure for privacy preservation at a first stage.

2.2. Privacy-preserving process mining

PPPM research is gaining momentum: the global awareness on data privacy, the recent enforcement of privacy legislations and the definition of the FACT principles (fairness, accuracy, confidentiality and transparency), aligned to conduct responsible process mining [20], have motivated researching privacy-preserving strategies for PM analyses in the recent years.

There are many application uses in which the use of PPPM is fundamental, particularly in domains dealing with confidential data. First, in situations where PM analyses need to be externalized to third parties (because the institution itself is not capable of), some pre-processing on the event logs needs to be conducted so as to limit the third-parties' ability to retrieve private information of the individuals beyond the very PM results. Also, situations where PM requires event logs from multiple organizations (called cross-organizational process mining), in which the execution of a process is not conducted in a single organization, but by a group of independent organizations, could arise privacy issues. Besides, the worldwide trend towards open data models for transparency purposes also entails the necessity to apply privacy-preserving techniques on event logs in case to be shared or released. In this case, the application of these techniques is paramount due to the observed individual uniqueness in event logs that might enable re-identification risks [21]. Last, PPPM could be seen as safeguards or preventive countermeasures against data leakages or data thefts. All in all, when privacy aspects are considered within event logs, they can be labeled as high quality [10].

The main privacy challenges associated to PM in human-centered industrial environments are classified in [22] according to their nature, either technological (*e.g.,* data minimization, data aggregation or transparent processing) or organizational (*e.g.,* processing consent, auditing or data breaches procedures). To address them, general recommendations and good practices guidelines are provided, such as storing encrypted data. Further practical guidelines about the anonymization of event data are also discussed in [23]. Recently, after synthesizing existing literature on PPPM, Elkoumy et al. [24] identified the main threats and requirements to be met in PPPM methods, and highlighted future research challenges.

Initial studies proposed the achievement of confidentiality within event logs by means of pseudonymization or encryption strategies. For instance, Burattin et al. [25] presented a complete framework for hiding confidential information from event logs using symmetric and homomorphic encryption when outsourcing PM analyses. In the same line, Tillem et al. [26] proposed a modified version of the alpha algorithm (one of the very first PM algorithms) to discover process models from encrypted event logs in a privacy-preserved fashion. Within a cross-organization context, Liu et al. [27] proposed a trusted-third-party scheme dealing with either public process models, which are shared across organizations, and private models, which limit confidential and/or additional information. More recently, Rafiei et al. [28] analyzed the weaknesses and open challenges of event data encryption, and proved that confidentiality cannot be achieved by merely encrypting all data in [29]. To this end, authors presented a confidentiality framework based on the encryption and abstraction of event logs to discover process models and social networks [29]. Last but not least, Michael et al. [30] designed a GDPR-friendly privacy-preserving user-centered system for PM using an ABAC-based authorization model, that enables tracking who does what, when, why, where and how, with personal data.

Beyond encryption, Pika et al. [31,32] recently analyzed the privacy requirements for process models in the healthcare domain, and assessed the suitability of existing data transformation techniques, namely noise addition, suppression and encryption, anonymize attributes' values within event logs. Authors observed that the proper selection of these techniques can help interpret and improve the accuracy of the PM results, although they highlight that the quality will largely depend on the very characteristics of the event logs and the goals of the PM analyses. To address these challenges, authors described a theoretical PPPM framework for supporting healthcare PM analyses as well as a privacy metadata model to capture the history of all privacy-preserving transformations applied on the event logs. Similarly, Rafiei and van der Aalst [33] also explored common transformation techniques to anonymize event log attributes, such as the suppression of activities or the generalization of temporal attributes, and presented an XML-based extension for defining the privacy metadata model used within privacy-preserved event logs.

More generally, Fahrenkrog-Petersen [34] outlined the main privacy guarantees to be achieved within PPPM, and highlighted two main strategies: event log sanitization, *i.e.,* pre-process event logs to guarantee a certain privacy level, and privatized process mining, *i.e.,* develop PM algorithms that generate PM results guaranteeing a certain privacy level. With regards to the first strategy, PRETSA [35] is an algorithm that achieves the t-closeness property within event logs, by removing personally identifiable information and discarding infrequent (and potentially identifiable) behavior. Consequently, PM analyses requiring the behavior, performance or role of particular individuals are not possible. With regards to the second strategy, Mannhardt et al. [12] presented a holistic privacy model for process discovery based on differential privacy. In this context, the privacy guarantees lie on the addition of noise to the resulting process models without distorting event data. This framework successfully correlated the utility of the protected processes models with the number of traces variants (infrequent behavior) within event logs. However, the high complexity of this method involves some constraints, such as the length of the traces or the kind of information that can be discovered from individuals. Both solutions have been integrated in the publicly available web application called ELPaaS [36].

Facing a more user-centric approach, the authors in [37] observed the feasibility to re-identify individuals by analyzing the distribution of sensitive attributes in event logs when combined with location-based attacker models. Authors proposed a novel method to distort these distributions, while minimizing the impact on the protected process models.

Although most studies concentrate on the control-flow perspective of processes, some of them also consider the privacy issues related to other perspectives. For instance, Rafiei and van der Aalst [38] highlighted the privacy issues associated to the organizational perspective, and presented a decomposition method for discovering people's roles in a private fashion against frequency-based attacks, in which attackers know the relationship between activities and individuals, the most/least frequent activities or the first/last activities. Furthermore, the privacy issues associated to the discovery of process models from the case perspective are discussed in Rafiei et al. [39]. In this investigation, authors introduced the *TLKC*-privacy model using a group-based anonymization to avert case linkage and attribute linkage attacks. Despite the remarkable results, the method was tested using only one event log, and future work implementing smarter pruning algorithms to minimize the computational time will be desirable. More comprehensively, an in-depth analysis of group-based privacy-preserving techniques considering different attacker models is provided in [40]. Similar to ELPaaS, another open-source web-based application, called PPDP-PM [41], has integrated these privacy-preserving techniques.

Last but not least, recalling the cross-organizational process mining concept, Elkoumy et al. [42] recently proposed Shareprom, a secure multi-party computation tool enabling multiple parties perform basic PM analysis over partial logs without sharing any sensitive information from other parties. To overcome potential scalability issues, a distributed divide-and-conquer scheme for parallel processing of event logs was presented in [43]. Also, the tool relies on a differential privacy mechanism to ensure that the PM results are protected against possible privacy leakages [44]. Finally, an empirical evaluation of the trade-off between disclosure risk and utility loss of the previous mechanism is detailed in [45].

Table 1
Example of an event log fragment with confidential data from a healthcare institution.

Event ID	Case ID	Patient	Doctor	Activity	Timestamp	Disease
239561	20AME7KJN	Alice Brown	Eric Jones	X-ray test	2021-11-27 11:26:31	Lung cancer
239562	18OKY9MBE	Peter Stevenson	Fred Grimes	Admission	2021-11-27 11:27:04	Flu
239563	84FIQ6NKW	Courtney Song	Charles Smith	Blood test	2021-11-27 11:29:51	Arrhythmia
239564	44GVB2IKD	Whitney Johnson	Tom Miller	Prescribe medication	2021-11-27 11:31:28	Arthritis
239565	96UJS3AXZ	Yi Sun	Jimmy Adams	Screening	2021-11-27 11:31:46	Breast cancer
239566	18OKY9MBE	Peter Stevenson	Fred Grimes	Blood pressure test	2021-11-27 11:33:03	Flu
239567	74SXW8YTR	Alec Dorsey	Veronica Gibson	End surgery	2021-11-27 11:35:17	Femur fracture
239568	96UJS3AXZ	Yi Sun	Jimmy Adams	Mammogram	2021-11-27 11:36:03	Breast cancer
239569	77KRZ7KDU	Darcy Griffin	Frank Noah	Sputum test	2021-11-27 11:36:58	Pneumonia
239570	42ITC2UPL	Millie Hunter	Tom Miller	X-ray test	2021-11-27 11:38:12	Bone fracture
239571	84FIQ6NKW	Courtney Song	Charles Smith	Prescribe medication	2021-11-27 11:39:02	Arrhythmia
239572	20AME7KJN	Alice Brown	Eric Jones	Prescribe medication	2021-11-27 11:40:33	Lung cancer
239573	90CEO1NBO	Timmy Hunter	Ann Zigber	Blood pressure test	2021-11-27 11:41:39	Viral infection
239574	87MUW3NVX	Damian Lee	Susan Leonard	Blood test	2021-11-27 11:42:05	Diabetes
239575	54FGJ4ECZ	Kenny Johnson	Estelle Edmund	Blood pressure test	2021-11-27 11:43:40	Viral infection

3. Re-identification using event logs from public places: Towards targeted location-oriented attacker models

Pseudonymization and encryption are GDPR-friendly measures to obfuscate personal data for protecting individuals' privacy. Despite their popularity, unfortunately these techniques do not offer enough protection against certain location-oriented attacks. This section elaborates on the drawbacks of pseudonymization or encryption techniques to counteract specific attacks, and the potential privacy issues that could be raised. More specifically, we show how the individuals within an obfuscated event log could be re-identified or linked to confidential data when an attacker, who has access to the event log, has also gained physical access to a public place inside the organization. In particular, these attacks are feasible in organizations with free-public access, such as hospitals, emergency units, banks or governmental institutions, in which attackers can jeopardize people's privacy if they acquire enough background knowledge. To this end, first we formalize the notation used throughout the article in Section 3.1. Then, in Section 3.2, we demonstrate the potential privacy concerns arising from the release of event logs obfuscated with the aforementioned classical techniques. Finally, in Section 3.3, we formalize an attacker model able to disclose unauthorized information from targeted individuals within this kind of event logs, by modeling their activities and inferring their process models.

3.1. Notation

Let $\mathcal{E}$ be the universe of all events, $\mathcal{A}$ be the universe of all attribute names, and $\delta_a(e)$ be a mapping function to obtain the value of an attribute $a \in \mathcal{A}$ from an event $e \in \mathcal{E}$. Each event $e \in \mathcal{E}$ is a record with, at least, attributes $\{id, case, act, ind, time, conf\}$. More specifically, $\delta_{id}(e)$ is a unique event identifier, $\delta_{case}(e)$ is the process instance identifier related to event e,[1] $\delta_{act}(e)$ indicates the activity performed, $\delta_{ind}(e)$ indicates personally identifiable information of the individual (resource) responsible for event e, $\delta_{time}(e)$ refers to the occurring timestamp of event e, and $\delta_{conf}(e)$ indicates confidential information about the individual. Let $\mathcal{T}$ be the universe of all traces, a trace $t \in \mathcal{T}$ is a chronological sequence of unique events associated to a single process instance, *i.e.,* $t = \langle e_1, \ldots, e_n \rangle$, where n is the length of the trace t. All together, an event log L is a set of traces, $L = \{t_1, \ldots, t_m\}$, where m is the number of traces, such that each event appears only once in the entire event log L. Table 1 depicts an event log example with confidential information within the healthcare domain.

By means of process discovery algorithms, the event data within an event log L could be represented as a process model M using a modeling notation language (*e.g.,* petri nets, BPMN, graphs, transition systems, YAML...). Process models could represent a wide variety of event log information: for example, considering the event log in Table 1, one could discover a unique process model by considering all the events so as to have a global vision of the functioning of the healthcare institution, discover the process models of patients with the same disease so as to have a particular vision of the medical paths for each disease individually, or discover a process model for each patient or doctor so as to know which cases are more complex, efficient or time-consuming (*i.e.,* resource behavior).

3.2. Simple strategies for privacy protection: Common approaches and problems

When applying simple privacy protection techniques, confidentiality might not be guaranteed in some cases, and attackers could exploit particular attacker models to avert it.

A straightforward transformation would be the suppression of either personally identifiable information or confidential data from event logs, as suggested in [23,31,35]. Despite the privacy enhancements, the utility of the data decreases dramatically, and PM results might omit valuable knowledge. Also, some PM analyses, such as organizational analyses, resources-oriented analyses or performance analyses, are no longer possible.

Pseudonymization, a recommended practice in [6,46], is a popular technique to transform personally identifiable information into faux, synthetic identifiers, called pseudonyms. The strength of this technique lies in the separation of confidential data from people's information into different data sources: only authorized users are able to retrieve the identity behind each pseudonym and, therefore, associate confidential data to individuals. From a PM perspective, process models do not suffer any distortion; however, they cannot be directly associated to particular individuals, only to pseudonyms and, hence, there is no direct relationship between people's identities, their process models and people's confidential data.

Similarly, data encryption is another well-known technique to guarantee confidentiality in accordance to GDPR recommendations [22,47] and PPPM literature [25,26,28,32]. By means of cryptographic functions, the event data is encoded into unreadable data and, unless using the appropriate decryption key, people's identities and confidential data remain undisclosed. However, applying PM analyses on encrypted event data decreases the utility of the process models, since they would lack from semantics and readability.

Both pseudonymization and encryption techniques share a common fact: personally identifiable information is obfuscated into unreadable data. Unfortunately, these transformations are commonly static, *i.e.,* a certain text x in the original event log is always replaced by the same unreadable text x' in the pseudonymized/encrypted event log. So, anyone could figure out that all events with the same value x'

[1] The *case* attribute could refer to, for instance, a treatment code within the medical context, an order number within an e-commerce or a claim number within an airline company.

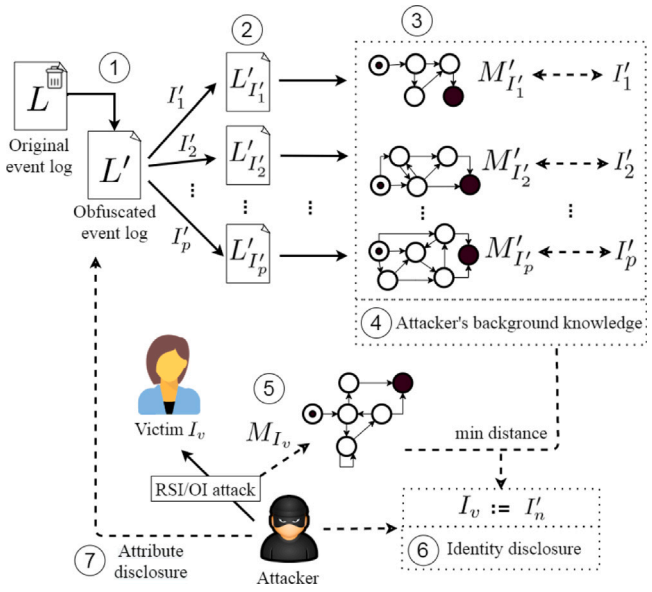

Fig. 1. Attacker model for re-identifying targeted individuals through location-oriented attacks in pseudonymized or encrypted event logs.

Fig. 2. Scenario of an RSI/OI attack in the waiting room of a hospital. An attacker is tracking the behavior/activities of potential victims in a restricted space.

would belong to the same individual, even though his/her identity is unknown. This lack of robustness leads to frequency attacks that attackers exploit to break the confidentiality of the released event logs [37]. Ideally, this could be averted by transforming an original text x into multiple and different unreadable texts $x'_1, x'_2, \ldots$. From a privacy perspective, this scheme is a clear improvement. However, it limits the PM analyses, especially for resource behavior analysis, since there is no way to group the events/traces of the same individual. Despite the weaknesses, PPPM literature dealing with encryption [25,28] focuses on static transformations.

3.3. Problem formalization: The attacker model

This section formalizes an attacker model able to disclose the identity of an individual or retrieve their confidential data from pseudonymized or encrypted event logs. More specifically, we show the potential privacy issues that could emerge when combining location-oriented attacks with the analysis of activities and the inference of process models in institutions with free-public access. This attacker model is illustrated in Fig. 1, whose steps are detailed next.

Let L be an event log describing the activities carried out by people in an institution with free-public access (*e.g.,* employees from a public hospital, public administration…) that manages confidential data. It is apparent that L cannot be released unmodified, because it associates confidential data to personally identifiable information.

Therefore, for confidentiality reasons, the institution decides to either pseudonymize or encrypt the personally identifiable information, and outputs L' (step ①). To enable the attack, we assume that the attacker has access to L', because either it has been released for transparency purposes, shared with another institution, or obtained through malicious ways (*e.g.,* data theft), and he/she aims to exploit such information for disclosing private information. Also, we assume that L cannot be recovered from L', so the attacker starts the attack with the information in L' only.

First, all the events in L' are separated according to the value of the attribute *ind*, which contains the obfuscated value of the individual's identity. If L' contains information about p individuals, namely $I'_1, I'_2, \ldots, I'_p$, then L' is decomposed into p sub-event logs, $L'_{I'_1}, L'_{I'_2}, \ldots, L'_{I'_p}$, each of them with the events of a specific individual (step ②). For each of these sub-event logs, the attacker discovers its corresponding process model $M'_{I'_1}, M'_{I'_2}, \ldots, M'_{I'_p}$, respectively, from a control-flow perspective using a specific resource behavior discovery algorithm. In this sense, each process model represents the workflow of activities (*i.e.,* behavior) performed by an unknown individual (step ③). The attacker has hence acquired a background knowledge that relates each individual's obfuscated identifier to an individual's process model (step ④). Since people's real identities cannot be retrieved from the obfuscated identifiers, the attacker takes advantage of the discovered process models to break confidentiality.

Then, the people's re-identification phase begins. The proposed attacker model is based on restricted space identification (RSI) and object identification (OI), well-known attacks against Location-Based Services that imply the approximate contact of the attacker with the targets. The attacker selects a targeted individual (victim) I_v to expose, whose identity is obfuscated in L' and, hence, unknown for the attacker. To conduct this kind of attack, the attacker stays physically close to the victim and tracks all the activities that he/she performs in the organization. Fig. 2 illustrates an attacker conducting an RSI/OI attack in the waiting room of a hospital (*i.e.,* a restricted space –RSI–) to different victims (*i.e.,* the targets of the attack –OI–). After a reasonable period of time, this observed information is cross-correlated with his/her background knowledge. As a result, the attacker would be able to describe the behavior of the victim I_v as a process model M_{I_v} (step ⑤).

Finally, the attacker compares the observed process model M_{I_v} against all the process models from his/her background knowledge $(M'_{I'_1}, M'_{I'_2}, \ldots, M'_{I'_p})$ using some distance or similarity function. The attacker can infer the obfuscated identifier of the victim I_v, namely I'_n (for $1 \leq n \leq p$), by identifying the process model $M'_{I'_n}$ most similar to M_{I_v}. Hence, the attacker has correlated the identity of a victim I_v to an obfuscated identifier I'_n from L', *i.e.,* identity disclosure (step ⑥). Additionally, the attacker can also infer the confidential data of I_v as those information associated to I'_n in L', *i.e.,* attribute disclosure (step ⑦).

4. The *k* -PPPM method

This section presents *k*-PPPM, a novel microaggregation-based method for conducting PPPM. Due to the feasibility to infer process models when conducting location-oriented attacks, in which simple obfuscation techniques fail, the proposed method distorts the event data in L to create a privacy-preserved event log version L' that renders the background knowledge acquired by the attackers useless. In addition,

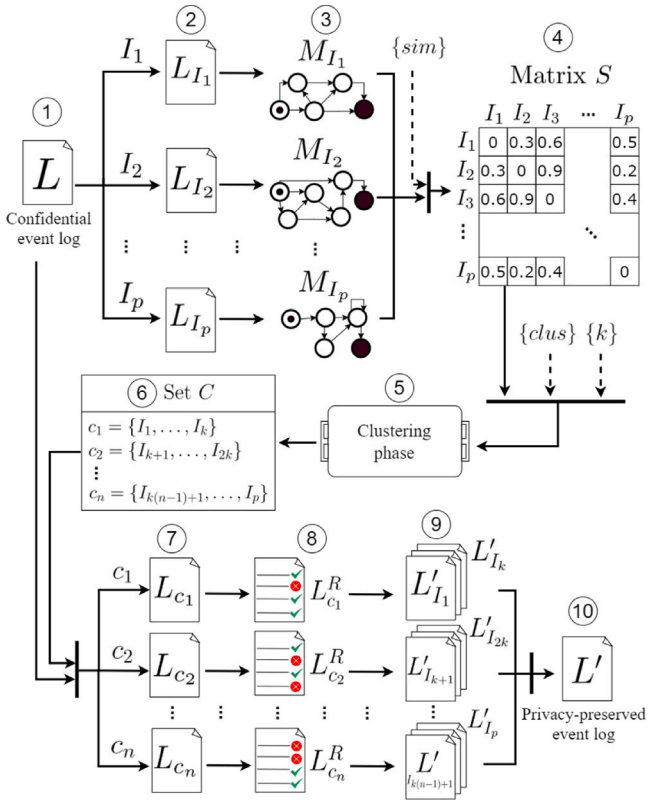

Fig. 3. Step-by-step scheme of the proposed k-PPPM method.

Algorithm 1 k-PPPM algorithm

Require:
L is a non-empty event log describing the activities of p individuals, where $p > 0$.
sim is a measure to assess the similarity between two process models.
$clus$ is a microaggregation-oriented clustering algorithm.
k is a privacy level in the range, where $2 \leq k \leq p$.
Ensure:
L' is a non-empty event log, whose events and individuals' process models are indistinguishable among k individuals.

```
function k-PPPM(EventLog L, Measure sim, Algorithm clus, Integer k)
    List⟨Individual⟩ individuals ← getAllIndividuals(L);
    HashMap⟨Individual, ProcessModel⟩ models;
    for all I_i in individuals do                                  ▷ Steps ② and ③
        EventLog L_{I_i} ← getEventsFromIndividual(L, I_i);
        ProcessModel M_{I_i} ← discoverModel(L_{I_i});
        models.put(I_i, M_{I_i});
    end for
    Matrix2D⟨Double⟩ S;
    for i ← 1 to p do                                              ▷ Step ④
        for j ← i to p do
            Individual I_i ← individuals[i];
            Individual I_j ← individuals[j];
            Double dist ← compare(models.get(I_i), models.get(I_j), sim);
            S[I_i][I_j] ← dist;
            S[I_j][I_i] ← dist;
        end for
    end for
    List⟨List⟨Individual⟩⟩ C ← clustering(S, k, clus);             ▷ Steps ⑤ and ⑥
    EventLog L';
    for i ← 1 to n do
        List⟨Individual⟩ c_i ← C[i]
        EventLog L_{c_i} ← getEventsFromCluster(L, c_i);           ▷ Step ⑦
        EventLog L^R_{c_i} ← selectRepresentative(L_{c_i});        ▷ Step ⑧
        for j ← 1 to k do                                          ▷ Step ⑨
            Individual I_j ← c_i[j];
            EventLog L'_{I_j} ← L^R_{c_i};
            for all event in L'_{I_j} do
                δ_ind(event) ←obfuscatePII(I_j);
            end for
            L'.append(L'_{I_j});
        end for
    end for
    return L';                                                     ▷ Step ⑩
end function
```

k-PPPM aims to minimize the disclosure risks while maximizing, as much as possible, utility of the event logs and the process models to be discovered.

More concretely, given an event log L, k-PPPM clusters similar individuals into groups of size k according to their process models, being k a privacy threshold ($k \geq 2$). Then, each cluster selects a representative process model, which is associated to each individual within the cluster. All the individuals within the same cluster will be represented by the same process model in L' and, therefore, are indistinguishable to an attacker conducting location-oriented attacks. Inspired by the classical microaggregation techniques, this procedure leverages the properties of k-anonymity. Although microaggregation techniques have been extensively studied in many privacy domains, to the best of our knowledge, this is the very first PPPM method founded on them. Therefore, this method enables guaranteeing k-anonymity in the protected event log.

First, Section 4.1 describes the main design decisions and implementation details of the proposed method. Then, Section 4.2 elaborates on a security and privacy analysis of the event logs protected using k-PPPM in comparison to event logs protected using pseudonymization or encryption.

4.1. Implementation details

This section describes the implementation details of k-PPPM. For the sake of clarity, the explanation is supported by Fig. 3, which depicts the step-by-step transformations applied to an event log, and the outline of the algorithm is provided in Algorithm 1.

The procedure begins with an event log L that describes the activities of p individuals, namely $I_1, I_2, \ldots, I_p$, whose identity must be preserved (step ①). First, L is decomposed into p sub-event logs, $L_{I_1}, L_{I_2}, \ldots, L_{I_p}$, each of them containing the events of each individual, respectively (step ②). Next, for each sub-event log, a process model $M_{I_1}, M_{I_2}, \ldots, M_{I_p}$ is discovered, from a control-flow perspective, using some discovery algorithm. Each process model represents the behavior of each individual as a workflow of activities, *i.e.,* resource behavior (step ③). It is noteworthy that these initial steps are identical to the ones conducted by the attackers in order to reproduce their behavior and decrease their ability to acquire the background knowledge.

With the microaggregation principles in mind, k-PPPM determines that two individuals are similar if their process models are similar. To do this, a similarity matrix S of length $p \times p$ is created, by comparing all the pairs of process models between them using a given similarity measure sim (step ④). Matrix S fulfills two properties: it is (i) symmetric, *i.e.,* $S[I_a][I_b] = S[I_b][I_a]$ because $sim(M_{I_a}, M_{I_b}) = sim(M_{I_b}, M_{I_a})$, and (ii) hollow, *i.e.,* $S[I_a][I_a] = 0$, because $sim(M_{I_a}, M_{I_a}) = 0$, in case that 0 means total similarity.

Next, the microaggregation's clustering phase begins. The aim of this phase is to group individuals into clusters, in such a way that the individuals within the same cluster are as similar as possible to maximize within-cluster similarity (and, hence, minimize information loss). Considering the information in S, a microaggregation-oriented clustering algorithm $clus$ is executed by specifying the number of

individuals per cluster k (step ⑤). In particular, number k directly refers to the privacy level of k-PPPM (for $k \geq 2$): the higher the value k, the more privacy. It is noteworthy that maximum privacy is achieved when $k = p$, since all individuals are grouped within a single cluster. Besides, note that any of the microaggregation clustering algorithms from the literature could be used. As a result, a set C of n clusters, $C = \{c_1, c_2, \ldots, c_n\}$, is obtained, in which: (i) the number of clusters n is equal to $\lfloor p/k \rfloor$, (ii) each cluster in C is a group of k to $2k-1$ individuals, (iii) all p individuals must be assigned to one and only one cluster (step ⑥).

Once clusters have been created, each cluster must select its representative (or centroid): a *virtual* individual that averagely represents all the individuals within the same cluster. In k-PPPM, this representation is observed from the process model perspective: the process model of the representative must averagely represent all the process models from all the individuals of the cluster. To create this averaged process model, the original event log L is used. This phase requires two main steps: (i) aggregation of individuals, and (ii) sampling of events, as explained below.

At first, L is decomposed into n sub-event logs, $L_{c_1}, L_{c_2}, \ldots, L_{c_n}$, each of them containing all the events associated to all the individuals within each cluster $c_1, c_2, \ldots, c_n$, respectively. Each sub-event log L_{c_i} (for $1 \leq i \leq n$) can be observed as an aggregator of events and traces of all the individuals within each cluster c_i (step ⑦). If these sub-event logs would be exploited to represent process models, they would show an aggregated behavior of the entire cluster's individuals. Next, the selection of the representative of each cluster c_i is performed using the event data in L_{c_i}. To do this, a number of traces from L_{c_i} are randomly sampled, and stored in $L^R_{c_i}$, *i.e.,* $L^R_{c_i}$ is a subset of L_{c_i} (step ⑧). The number of traces to be sampled corresponds to the average of traces per individual within the cluster. For instance, if a cluster c_i has three individuals ($k = 3$), and each of them has 19, 26 and 15 traces in L, respectively, then L_{c_i} would contain 60 traces, but only 20 of these traces would be chosen for $L^R_{c_i}$. Thus, the behavior of the representative is composed of partial behaviors of k individuals. The sampling of the traces is done at random, so k-PPPM is a non-deterministic method.

Once all clusters' representatives have been selected, all the individuals within each cluster are replaced by their representative. To this end, the event data in $L^R_{c_i}$ is replicated k times, one for each individual within the cluster c_i. This step is paramount to break the uniqueness of event data (and the individuals' process models to be discovered in them), since it duplicates the same information to k different individuals and prevent direct identity disclosure. In addition, instead of assigning each event to the personally identifiable information of the individual (as it is in L), it is assigned to an obfuscated identifier (either pseudonymized or encrypted) of the individual. The event data associated to each obfuscated individual I_i (for $1 \leq 1 \leq p$) is stored in L'_{I_i} (step ⑨). Formally, the event data in L'_{I_i} does no longer associate personally identifiable information to confidential data because its *ind* attribute is obfuscated. Finally, the method returns L', a privacy-preserved event log resulting from the union of the p sub-event logs L'_{I_i}, *i.e.,* $L' = L'_{I_1} \cup L'_{I_2} \cup \cdots \cup L'_{I_p}$ (step ⑩).

The data transformations applied by k-PPPM on L in order to obtain L' are likely to degraded the quality of the individuals' process models to be discovered. Given a certain individual I_i (for $1 \leq i \leq p$), the original process model M_{I_i} discovered from L and the protected process model M'_{I_i} discovered from L' will differ, because the latter is affected by (i) the loss of events/traces that have not been chosen to be in the cluster's representative, and (ii) the new events/traces (originally belonging to the other $k-1$ individuals within the same cluster) that have been chosen to be in the cluster's representative.

4.2. Security and privacy analysis

This section provides a throughout analysis of k-PPPM regarding data confidentiality in terms of security and privacy enhancements. More specifically, these properties are compared between the original (unprotected) event log L, a protected event log version L' obtained after executing k-PPPM, and a protected event log version L'' obtained through pseudonymization or encryption.

With regards to security aspects, k-PPPM contributes to confidentiality by obfuscating direct identifiers and breaking their direct linkage with the confidential attributes in the event logs. Other security requirements, such as integrity, availability and authentication, are beyond the scope of this method. Also, since the k-PPPM's anonymization procedure does not require external parties nor network communications, the potential security concerns are relaxed, and the security of the method lies in the distortion of the statistical properties of the event data. Recalling confidentiality, k-PPPM is supported by the obfuscation of personally identifiable information through pseudonymization or state-of-the-art encryption techniques at the last stage of the algorithm. These techniques guarantee the confidentiality of L' as long as the private cryptographic keys remain secret. As both L' and L'' rely on pseudonymization or encryption, the confidentiality level of both approaches is the same.

Significantly enough, the strength of k-PPPM lies in the privacy guarantees added to the individuals appearing in L', which contribute to minimize the impact of location-oriented attacks, where classical obfuscation techniques fail. Given an event log L', it can be noticed that the event data and the individuals' process models that can be discovered are constrained by the k-anonymity model: (i) each trace in L' is assigned to, at least, k different individuals, each of them with their own confidential data, which helps minimize attribute disclosure risk, (ii) the same process model would be discovered in, at least, k different individuals, which helps minimize identity disclosure risk, and (iii) each process model represents the behavior of a group of, at least, k individuals, instead of a single (and potentially identifiable) individual, which helps minimize the identity disclosure risk as well.

Comparing the data in L' and L'', attackers gain different knowledge when modeling people's processes during their attacks, as illustrated in Fig. 4. Whereas p different process models can be discovered in L'' (and L as well), only n different process models can be discovered in L'. Despite this information loss, k-PPPM is resilient to the attacker model described in Section 3.3, since attackers are not able to link the observed process model of the victim to their background knowledge, but to a group of k indistinguishable individuals. Hence, the re-identification risk is upper-bounded by $1/k$.

Last but not least, note that the microaggregation strategy applied in k-PPPM is slightly different in comparison to classical microaggregation techniques used in SDC. In particular, we highlight two major differences. On the one hand, there are some differences from a *data* perspective: event logs (used in k-PPPM) cannot be directly understood as microdata sets (used in SDC). Generally, SDC techniques suppress the personally identifiable information in the microdata sets, and the privacy guarantees reside on the ability to generalize or aggregate quasi-identifier attributes. However, event logs rarely contain quasi-identifiers, and the limitations of the PM analysis when completely removing the individual's information from event logs have already been discussed (see Section 3.2). On the other hand, there are also some discrepancies from an *analytical* perspective: the knowledge that can be acquired from event logs is different to that of microdata sets. In microdata sets, each record corresponds to an independent individual, so there is no relationship between records. However, events do have a relationship between them (mainly defined with the *case* and *time* attributes), and these relationships are exploited in PM, such as when discovering process models from the control-flow perspective. Consequently, in contrast to classical SDC approaches, k-PPPM must preserve individuals' privacy at the same time that preserves the relationships among events with the goal of maximizing data utility. A comparison between the two microaggregation phases conducted in classical SDC and in k-PPPM is provided in Table 2.

Table 2
Comparison between classical SDC and *k*-PPPM.

		Classical SDC	*k*-PPPM
Clustering phase	Input	Microdata sets	Event logs
	Objective	Minimize within-cluster records distance	Maximize within-cluster models similarity
	Elements	Independent records	Process models, *i.e.,* individuals
	Criteria	Distance measures: Euclidean, Manhattan, Minkowski, Chebyshev...	Process model similarity measures: VEO, VR, WD, DC...
	Output	Clusters of k to $2k-1$ records	Clusters of k to $2k-1$ individuals
Representative phase	Input	Clusters of k to $2k-1$ records	Clusters of k to $2k-1$ individuals
	Objective	Compute a virtual centroid record, a within-cluster representative record	Compute a virtual centroid individual, a within-cluster representative process model
	Selection	Average or median of all the records within the same cluster	Random sampling of the traces associated to all the individuals within the same cluster
	Replacement	Replace each record's value by the representative's value	Replicate the traces of the representative individual for all the individuals within the cluster
	Obfuscation	Remove direct identifiers and aggregate quasi-identifiers	Obfuscate personally identifiable information of individuals (attribute *ind*)
	Output	A microdata set containing, at least, k records for each combination of quasi-identifiers	An event log containing, at least, k individuals with the same events and traces, *i.e.,* process models

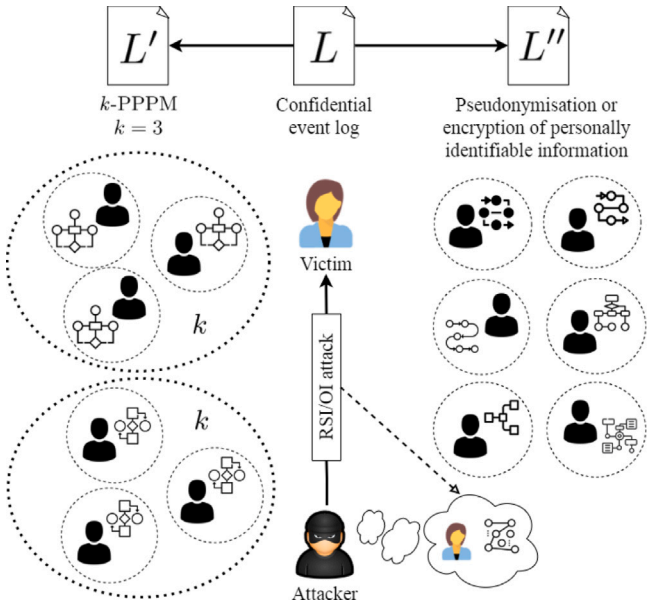

Fig. 4. Privacy enhancements between the proposed *k*-PPPM method in comparison to pseudonymization or encryption methods.

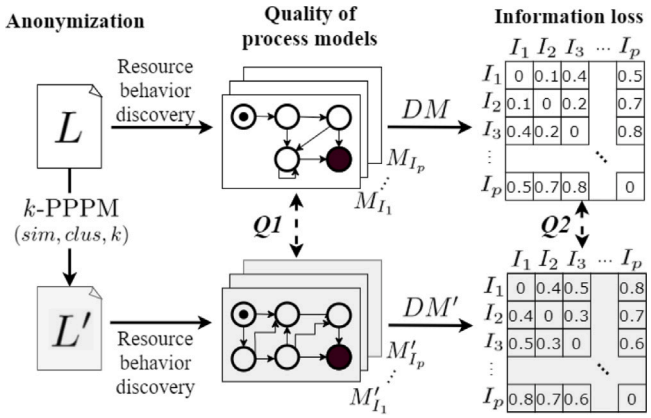

Fig. 5. Evaluation methodology.

5. Evaluation and discussion

This section presents the experimental evaluation of the *k*-PPPM method. This evaluation aims to assess the impact of our privacy protection model by focusing on the quality of the process models discovered from the protected event logs L' in comparison to the corresponding process models discovered from unprotected event logs L. First, Section 5.1 elaborates on the experimental setup designed to evaluate the proposed *k*-PPPM method, and Section 5.2 discusses the results obtained.

5.1. Experimental setup

The user-centric nature of *k*-PPPM leads to evaluate the process models associated to each individual from a control-flow perspective. More specifically, the quality of the results obtained using *k*-PPPM is measured according to the following two research questions (see Fig. 5):

- *Q1 — Individual distortion*: How similar is the process model of an individual I when discovered from the original event log (M_I), and when discovered from the protected event log (M'_I)?
- *Q2 — Inter-individual distortion*: Are the differences among the individuals' process models discovered from the original event log ($M_{I_1}, M_{I_2}, \ldots, M_{I_p}$) also maintained among the individuals' process models discovered from the protected event log ($M'_{I_1}, M'_{I_2}, \ldots, M'_{I_p}$)?

5.1.1. Process modeling notation and resource behavior discovery

In this article, process models are modeled as D/F-graphs, a generic notation widely used in the PM field, intended to represent the relationships between event data from dependency/frequency tables created from event logs [48–50]. The use of graphs, despite their limitations (*e.g.,* concurrency), enables the discovery of process models from a generic and high-level perspective, and avoids the restrictions and constraints introduced by advanced modeling notations, such as Petri nets or BPMN. This less-restrictive notation allows observing the very impact of the proposed method on the quality of the process models. Hence, a process model $M = (V, E)$ is characterized as a weighted directed graph, where V is the set of vertices representing the activities of the event log L, and E is the set of edges representing the transitions between two vertices from V with a certain dependency $w \in [0, 1]$.

Table 3
Properties of the event logs used to evaluate the *k*-PPPM method.

Name	Availability	Number of events	Number of traces	Number of activities	Number of resources	Events per trace (Avg.)	Traces per resource (Avg.)
BPI12 [54]	Public	262 200	13 087	23	68	20.03	192.46
BPI13 [55]	Public	6660	1487	7	584	4.48	2.55
BPI14 [56]	Public	466 155	46 507	39	242	10.02	192.18
BPI15 [57]	Public	262 628	5649	356	72	46.49	78.46
CoSeLoG [58]	Public	8577	1434	27	48	5.98	29.88
TGN	Proprietary[a]	122 179	58 836	36	280	2.08	210.13

[a]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.

Table 4
Summary of the QS results grouped by parameter.

			Event logs						
			BPI12	BPI13	BPI14	BPI15	CoSeLoG	TGN	Avg.
Grouping criteria	*sim*	**VEO**	0.4 ± 0.177	0.248 ± 0.198	0.491 ± 0.113	0.626 ± 0.149	0.469 ± 0.164	0.431 ± 0.123	0.444 ± 0.153
		VR	0.393 ± 0.183	0.26 ± 0.202	0.495 ± 0.117	0.624 ± 0.154	0.467 ± 0.166	0.438 ± 0.12	0.446 ± 0.157
		WD	0.392 ± 0.189	0.257 ± 0.202	0.496 ± 0.118	0.639 ± 0.157	0.48 ± 0.162	0.439 ± 0.129	0.452 ± 0.155
		DC	0.387 ± 0.182	0.245 ± 0.198	0.5 ± 0.119	0.649 ± 0.153	0.485 ± 0.169	0.435 ± 0.128	0.45 ± 0.158
	clus	**MDAV**	0.363 ± 0.182	0.195 ± 0.188	0.479 ± 0.114	0.631 ± 0.151	0.466 ± 0.15	0.399 ± 0.121	0.422 ± 0.151
		KM	0.368 ± 0.176	0.195 ± 0.186	0.482 ± 0.111	0.632 ± 0.149	0.465 ± 0.154	0.404 ± 0.113	0.424 ± 0.148
		OKA	0.373 ± 0.204	0.235 ± 0.208	0.496 ± 0.127	0.634 ± 0.153	0.474 ± 0.176	0.412 ± 0.142	0.437 ± 0.168
		BL	0.474 ± 0.168	0.386 ± 0.219	0.525 ± 0.116	0.642 ± 0.161	0.498 ± 0.181	0.529 ± 0.124	0.509 ± 0.162
	k	**2**	0.268 ± 0.161	0.169 ± 0.183	0.407 ± 0.129	0.496 ± 0.193	0.337 ± 0.188	0.324 ± 0.128	0.334 ± 0.164
		3	0.335 ± 0.181	0.212 ± 0.195	0.459 ± 0.122	0.587 ± 0.17	0.424 ± 0.175	0.389 ± 0.126	0.401 ± 0.162
		4	0.365 ± 0.187	0.243 ± 0.199	0.485 ± 0.12	0.628 ± 0.162	0.46 ± 0.18	0.426 ± 0.125	0.435 ± 0.162
		5	0.389 ± 0.186	0.257 ± 0.205	0.502 ± 0.117	0.654 ± 0.149	0.5 ± 0.167	0.448 ± 0.126	0.459 ± 0.158
		10	0.476 ± 0.192	0.298 ± 0.208	0.544 ± 0.112	0.704 ± 0.127	0.543 ± 0.149	0.494 ± 0.124	0.51 ± 0.152
		20	0.534 ± 0.187	0.337 ± 0.211	0.576 ± 0.102	0.736 ± 0.12	0.588 ± 0.132	0.533 ± 0.121	0.551 ± 0.146

5.1.2. Event log anonymization

Three parameters are required to execute *k*-PPPM. The selection of these parameters' values directly affects the quality of the event log anonymization and, therefore, the quality of the resulting process models. To assess this impact, our method is tested with different combinations of its three parameters. First, concerning the privacy level k, we evaluate the method for $k = 2, 3, 4, 5, 10$ and 20, common values within the *k*-anonymity literature. Second, with regards to the clustering algorithm $clus$, we use four strategies, in which three of them are popular heuristics from the literature, namely MDAV [16], *k*-member (KM) [17] and OKA [18]. Besides, we also use a naive algorithm as a baseline (BL) that groups individuals according to their number of traces, *i.e.,* two individuals are similar if they have a similar number of traces in L. The objective of BL is to evaluate whether there is a significant difference between using process-oriented clustering algorithms and event log-oriented clustering algorithms. And third, regarding the similarity measure sim, we use four graph similarity measures from the literature, namely Vertex Edge Overlap (VEO) [51], Vertex Ranking (VR) [51], Weight Distance (WD) [52] and DeltaCon (DC) [53]. Since these measures are bounded between 0 and 1, for the sake of consistency, we have standardized total similarity (*i.e.,* no distortion) as 0. Thus, the lower the sim's output, the more similar two process models are.

To observe the impact of these parameters, we have executed *k*-PPPM for all the combinations of these parameters: (s_i, c_i, k_i), $\forall s_i \in sim, \forall c_i \in clus, \forall k_i \in k$. Therefore, each event log L is anonymized 96 times ($4 \times 4 \times 6$), resulting into 96 privacy-preserved event log versions L': (VEO, MDAV, 2), (VEO, MDAV, 3),..., (DC, BL, 20).

5.1.3. Event logs

For the sake of completeness, experiments have been conducted using six real-life event logs from multiple domains, as described in Table 3.

It is noteworthy that these event logs have very different properties, such as the number of traces or the number of resources, so as to evaluate the global behavior of the proposed method regardless of these characteristics. Indeed, the number of resources in each event log corresponds to p, the number of individuals and process models that would be discovered from both L and L'. According to these properties, for example, it is also expected that the process models from BPI13 are simpler than those from BPI15, due to the significant differences of traces per resource and events per trace.

5.1.4. Quality of the process models

First, the individual quality of the process models ($Q1$) is evaluated by means of a model-by-model comparison. This evaluation, which determines the distortion that *k*-PPPM introduces to the very process model of each individual, is helpful to estimate how different a given process model is with regards to its original version.

The procedure works as follows. For each *k*-PPPM execution with a certain combination of parameters, we 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.,* $M'_{I_1}, M'_{I_2}, \ldots, M'_{I_p}$). Then, each protected process model is compared to its corresponding original version (*i.e.,* $\{M_{I_1}, M'_{I_1}\}, \{M_{I_2}, M'_{I_2}\}, \ldots, \{M_{I_p}, M'_{I_p}\}$) using a similarity measure. To prevent bias, process models are compared using the four aforementioned similarity measures. As a result, the comparison of each pair of process models leads to four independent similarity measures, thus obtaining a total of $4 \times p$ similarity values for the entire set of process models. It is worth mentioning that the similarity measure used in the evaluation is not necessarily the same as the one used for the anonymization. For example, for *k*-PPPM with parameters (VEO, MDAV, 2), p similarity values are obtaining using VEO for both anonymization and evaluation, but $3 \times p$ similarity values are obtained from evaluating the process models with another measure. Taking the average of all these values, a *quality score* (QS) can be associated to the *k*-PPPM execution. This QS value, bounded between 0 and 1 (the higher, the more distortion), indicates the averaged distortion that a given *k*-PPPM execution introduces to the protected process models individually.

Table 5
Summary of the ILS results grouped by parameter.

			Event logs						
			BPI12	BPI13	BPI14	BPI15	CoSeLoG	TGN	Avg.
Grouping criteria	*sim*	**VEO**	0.154 ± 0.03	0.094 ± 0.022	0.074 ± 0.027	0.13 ± 0.026	0.172 ± 0.032	0.089 ± 0.031	0.119 ± 0.051
		VR	0.145 ± 0.026	0.097 ± 0.024	0.074 ± 0.026	0.13 ± 0.025	0.166 ± 0.031	0.091 ± 0.031	0.118 ± 0.047
		WD	0.156 ± 0.033	0.1 ± 0.025	0.075 ± 0.029	0.136 ± 0.037	0.173 ± 0.034	0.094 ± 0.035	0.123 ± 0.055
		DC	0.143 ± 0.025	0.093 ± 0.021	0.078 ± 0.029	0.141 ± 0.039	0.172 ± 0.034	0.092 ± 0.033	0.12 ± 0.052
	clus	**MDAV**	0.136 ± 0.025	0.072 ± 0.02	0.071 ± 0.025	0.125 ± 0.033	0.158 ± 0.031	0.076 ± 0.025	0.107 ± 0.049
		KM	0.136 ± 0.024	0.072 ± 0.019	0.072 ± 0.026	0.127 ± 0.036	0.162 ± 0.032	0.077 ± 0.024	0.108 ± 0.05
		OKA	0.136 ± 0.028	0.086 ± 0.021	0.075 ± 0.03	0.126 ± 0.031	0.164 ± 0.033	0.081 ± 0.028	0.112 ± 0.048
		BL	0.19 ± 0.037	0.154 ± 0.031	0.083 ± 0.031	0.159 ± 0.028	0.197 ± 0.035	0.133 ± 0.051	0.154 ± 0.058
	k	**2**	0.071 ± 0.016	0.069 ± 0.017	0.047 ± 0.022	0.044 ± 0.023	0.075 ± 0.025	0.058 ± 0.023	0.061 ± 0.026
		3	0.103 ± 0.025	0.081 ± 0.019	0.058 ± 0.026	0.064 ± 0.028	0.096 ± 0.029	0.071 ± 0.028	0.08 ± 0.034
		4	0.116 ± 0.027	0.093 ± 0.021	0.066 ± 0.028	0.077 ± 0.032	0.116 ± 0.03	0.082 ± 0.032	0.093 ± 0.037
		5	0.131 ± 0.029	0.095 ± 0.025	0.071 ± 0.029	0.118 ± 0.033	0.131 ± 0.028	0.089 ± 0.034	0.108 ± 0.04
		10	0.201 ± 0.037	0.11 ± 0.025	0.087 ± 0.032	0.173 ± 0.041	0.236 ± 0.031	0.112 ± 0.038	0.155 ± 0.063
		20	0.275 ± 0.037	0.128 ± 0.029	0.123 ± 0.03	0.328 ± 0.035	0.371 ± 0.054	0.137 ± 0.039	0.229 ± 0.107

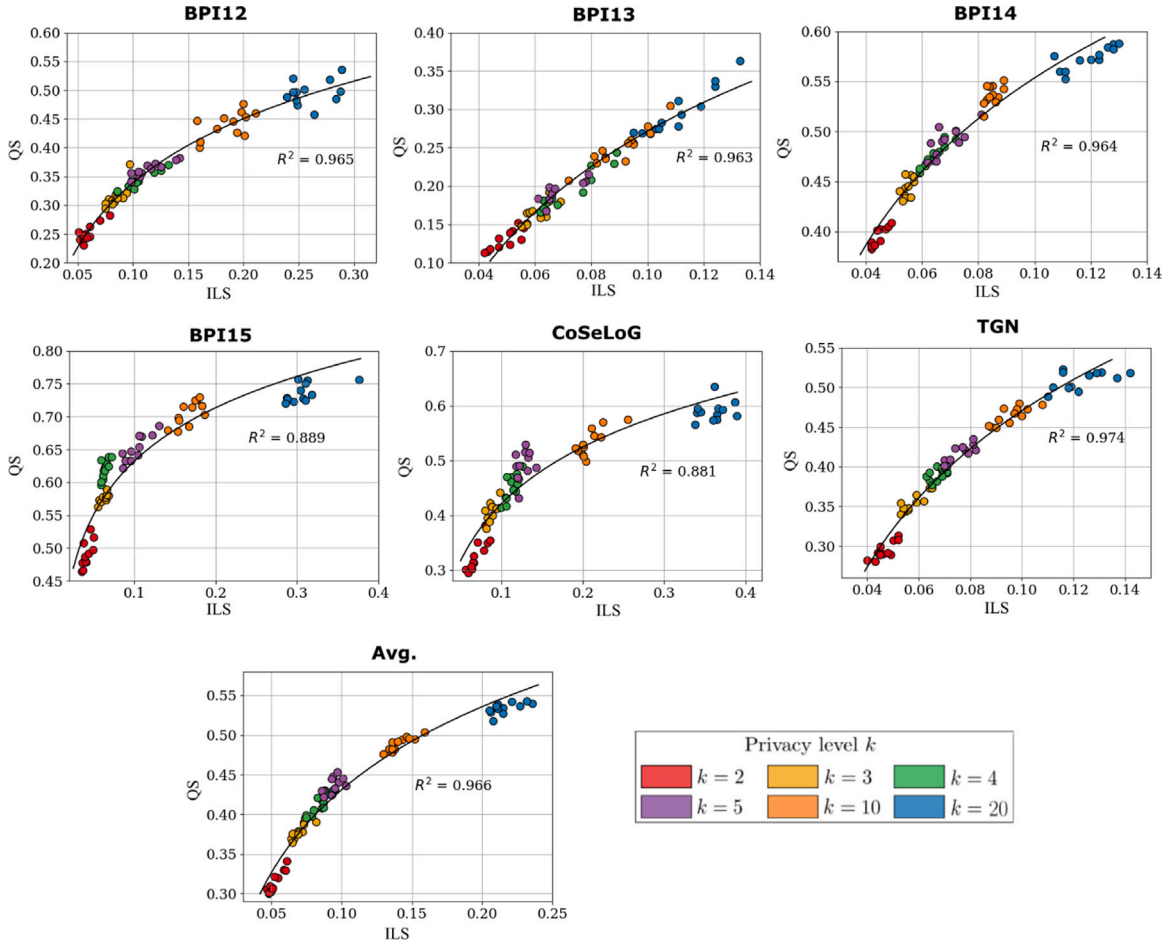

Fig. 6. Correlation between the QS and ILS results: each point represents a *k*-PPPM execution with a certain configuration of parameters. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.1.5. Information loss

Assuming the unavoidable distortion introduced by *k*-PPPM in the process models individually, it is also important to assess how this distortion affects the entire event log and the process models as a whole, this is, the inter-individual distortion (*Q2*). For instance, if the process models of two individuals, M_{I_1} and M_{I_2}, are very similar in L, it would be desirable to preserve, as much as possible, this similarity in the protected process models M'_{I_1} and M'_{I_2} in L'. Small inter-individual distortions would enable acquiring similar knowledge when comparing protected process models, such as for performance analysis, as if using the original process models.

The procedure works as follows. For each *k*-PPPM execution with a certain combination of parameters, we compute the distance matrices DM and DM' (with size $p \times p$). These matrices contain the similarity between all pairs of process models in L and L', respectively. Specifically, as before, the similarity values in DM and DM' are computed independently using the four aforementioned similarity measures. Thus, considering an event log L and a protected event

log L', four distance matrices are obtained, *i.e.,* $\{DM_{\text{VEO}}, \ldots, DM_{\text{DC}}\}$ and $\{DM'_{\text{VEO}}, \ldots, DM'_{\text{DC}}\}$, respectively. Note that the similarity measure used to calculate these matrices for the evaluation is not necessarily the same as the one used during the anonymization. Then, 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 MAE function (Mean Absolute Error, in Eq. (1)). The MAE, bounded between 0 and 1 (the higher, the lower data utility), serves as an indicator of the information loss. Taking the average of the four MAE results, an *information loss score* (ILS) can be associated to the k-PPPM execution. This ILS value, bounded between 0 and 1, indicates the averaged information loss incurred when releasing the protected event log.

$$\text{MAE} = \frac{1}{p^2} \sum_{i=1}^{p} \sum_{j=1}^{p} |DM[i][j] - DM'[i][j]| \tag{1}$$

5.2. Results and discussion

This section presents the experimental results of k-PPPM as well as the discussion of the impact of the proposed method on the resulting process models. Due to the non-deterministic nature of the method, all the results reported correspond to the average of five executions for each combination of parameters. Unfortunately, due to the absence of further microaggregation-based methods for PPPM, comparative experiments cannot be conducted.

Appendix describes the complete experimental results. More specifically, Tables A.1 and A.2 contain the averaged QS and ILS results, respectively, for each k-PPPM execution with a combination of its parameters for all the event logs. Besides the six event logs evaluated, an averaged result from all of them is also provided. For the sake of comprehensiveness, results are grouped by k and, within each group, the best and the worst QS/ILS values are highlighted in green and red, respectively. To better evaluate the experimental results and the contribution of each parameter, QS and ILS results have been grouped and averaged, according to each parameter value, in Tables 4 and 5, respectively. Also, the best and the worst values are highlighted in green and red, respectively, within each parameter group.

Given the data-dependence nature of k-PPPM, the quality of the protected process models depends on the properties of the original event logs. More specifically, since k-PPPM is a group-based anonymization, the quality highly depends on the profile of the individuals within the event logs. If the behavior of the individuals is very heterogeneous and distant, the quality of the anonymization will worsen significantly. In contrast, if event logs contain individuals who behave similarly (*i.e.,* there are individuals from the same department or with the same role), the quality loss will not be that severe because clusters will maintain certain similarity. For this reason, although executing k-PPPM with the same combination of parameters, the QS/ILS results can differ among event logs. For instance, executing k-PPPM with sim = VEO, $clus$ = MDAV and k = 2 in event logs BPI13, BPI15 and CoSeLoG, the QS results are 0.12, 0.467 and 0.302, respectively, and the ILS results are 0.047, 0.037 and 0.064, respectively. For this reason, computing the average quality of the k-PPPM executions (last column of the tables) can be insightful to evaluate the high-level behavior of k-PPPM. Despite the above, both Tables 4 and 5 reflect similar tendencies as the k-PPPM's parameters vary. Next, we discuss the contribution of each of these parameters.

According to the results, increasing the privacy level k leads to a decrease of the quality of the protected process models. The larger k, the more individuals share the same process model, so the more difficult for the attacker to re-identify them. However, achieving high privacy implies a worsening on the quality of the protected process models associated to each individual. This is aligned with classical observations from the privacy protection literature. In average, both QS/ILS results steadily increase together with value k. However, note that this fact is not always true, and the other two parameters can slightly help maximize the quality at the same privacy level.

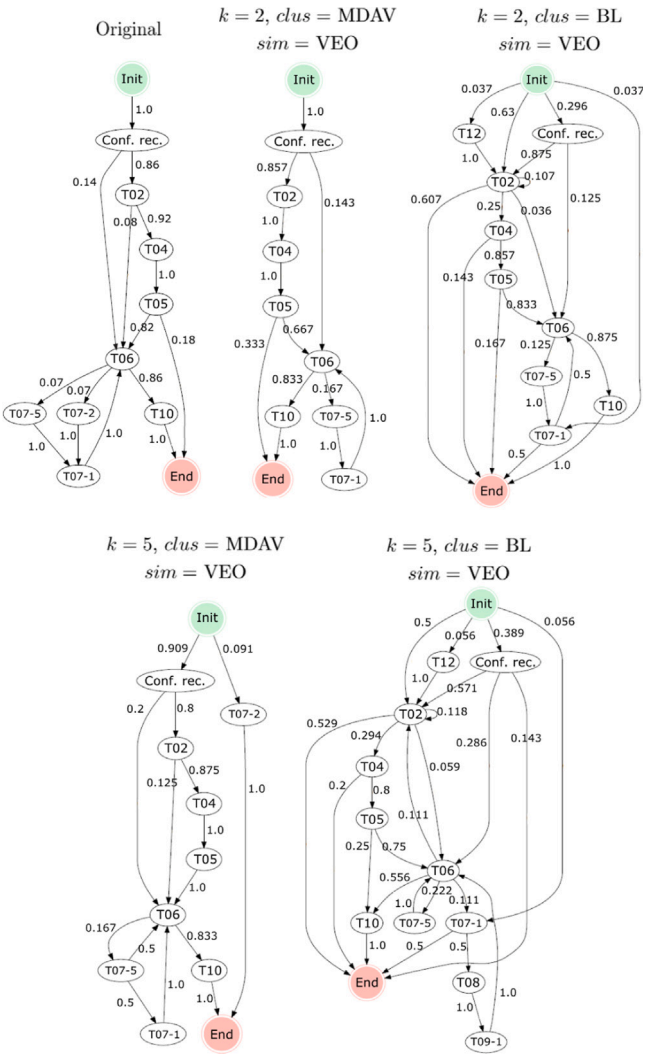

Fig. 7. Versions of the process model discovered from a certain individual from the CoSeLoG event logs for different k-PPPM executions.

Our experiments show that the proper selection of the clustering algorithm can contribute to reduce the process models distortion for a certain privacy level. Better results are feasible at higher privacy levels if heuristic clustering algorithms (*i.e.,* MDAV, KM or OKA) are used, rather than non-optimal clustering algorithms (*i.e.,* BL) at lower privacy levels. For instance, both QS and ILS results obtained with sim = VEO, $clus$ = MDAV and k = 3 in BPI12 are better in comparison to the ones obtained using sim = VEO, $clus$ = BL and k = 2. The BL algorithm demonstrates that a clustering criteria based on an event log property (*i.e.,* the number of traces per individual) is inefficient. Although process models are discovered from the event logs, it seems that forming the clusters according to an event log criteria is not optimal. On the other hand, the three heuristic algorithms, which use a process models similarity criteria, are more efficient to this end. Despite the small differences between them, note that MDAV and KM behave slightly better than the OKA algorithm. We state that this aspect derives from the very design of the algorithm, as both MDAV and KM start clustering the most distant records, which are more likely to be part of less cohesive clusters. The aim of starting clustering these records is precisely to find the most appropriate clusters for them in order to maximize the within-cluster similarity, already low by default. Hence, clustering the most abnormal records first leads to better anonymization results.

Table A.1
QS experimental results.

Parameters			Event logs						Avg.
sim	*clus*	*k*	BPI12	BPI13	BPI14	BPI15	CoSeLoG	TGN	
VEO	**MDAV**	**2**	0.242 ± 0.145	0.12 ± 0.155	0.39 ± 0.113	0.467 ± 0.174	0.302 ± 0.17	0.289 ± 0.112	0.302 ± 0.145
VR	**MDAV**	**2**	0.248 ± 0.152	0.142 ± 0.173	0.385 ± 0.12	0.478 ± 0.177	0.326 ± 0.157	0.282 ± 0.11	0.31 ± 0.148
WD	**MDAV**	**2**	0.244 ± 0.145	0.131 ± 0.166	0.386 ± 0.118	0.48 ± 0.188	0.306 ± 0.173	0.291 ± 0.114	0.306 ± 0.151
DC	**MDAV**	**2**	0.24 ± 0.151	0.113 ± 0.146	0.391 ± 0.122	0.488 ± 0.212	0.315 ± 0.163	0.281 ± 0.123	0.303 ± 0.153
VEO	**KM**	**2**	0.231 ± 0.14	0.118 ± 0.151	0.386 ± 0.124	0.465 ± 0.19	0.313 ± 0.157	0.29 ± 0.109	0.301 ± 0.145
VR	**KM**	**2**	0.25 ± 0.164	0.132 ± 0.167	0.386 ± 0.125	0.478 ± 0.176	0.3 ± 0.177	0.291 ± 0.104	0.306 ± 0.152
WD	**KM**	**2**	0.246 ± 0.139	0.124 ± 0.157	0.382 ± 0.122	0.479 ± 0.179	0.295 ± 0.162	0.289 ± 0.114	0.303 ± 0.145
DC	**KM**	**2**	0.249 ± 0.153	0.115 ± 0.151	0.403 ± 0.123	0.495 ± 0.216	0.336 ± 0.182	0.289 ± 0.116	0.316 ± 0.157
VEO	**OKA**	**2**	0.254 ± 0.163	0.139 ± 0.171	0.401 ± 0.131	0.487 ± 0.19	0.35 ± 0.209	0.3 ± 0.118	0.322 ± 0.164
VR	**OKA**	**2**	0.274 ± 0.182	0.152 ± 0.175	0.405 ± 0.142	0.492 ± 0.196	0.351 ± 0.215	0.307 ± 0.141	0.33 ± 0.175
WD	**OKA**	**2**	0.264 ± 0.173	0.146 ± 0.18	0.408 ± 0.147	0.498 ± 0.207	0.354 ± 0.213	0.308 ± 0.15	0.33 ± 0.178
DC	**OKA**	**2**	0.283 ± 0.181	0.152 ± 0.191	0.402 ± 0.138	0.516 ± 0.199	0.381 ± 0.225	0.313 ± 0.142	0.341 ± 0.179
VEO	**BL**	**2**	0.316 ± 0.169	0.279 ± 0.237	0.445 ± 0.129	0.514 ± 0.196	0.366 ± 0.194	0.417 ± 0.147	0.389 ± 0.179
VR	**BL**	**2**	0.321 ± 0.177	0.284 ± 0.232	0.447 ± 0.135	0.51 ± 0.197	0.37 ± 0.204	0.414 ± 0.154	0.391 ± 0.183
WD	**BL**	**2**	0.314 ± 0.171	0.283 ± 0.237	0.447 ± 0.137	0.518 ± 0.188	0.367 ± 0.2	0.414 ± 0.15	0.39 ± 0.18
DC	**BL**	**2**	0.319 ± 0.169	0.28 ± 0.237	0.446 ± 0.135	0.526 ± 0.198	0.366 ± 0.21	0.413 ± 0.147	0.392 ± 0.183
VEO	**MDAV**	**3**	0.302 ± 0.186	0.161 ± 0.174	0.431 ± 0.109	0.571 ± 0.138	0.376 ± 0.168	0.348 ± 0.113	0.365 ± 0.148
VR	**MDAV**	**3**	0.303 ± 0.178	0.166 ± 0.18	0.445 ± 0.116	0.578 ± 0.169	0.415 ± 0.175	0.348 ± 0.112	0.376 ± 0.155
WD	**MDAV**	**3**	0.313 ± 0.183	0.158 ± 0.175	0.44 ± 0.124	0.574 ± 0.166	0.409 ± 0.124	0.357 ± 0.125	0.375 ± 0.149
DC	**MDAV**	**3**	0.305 ± 0.166	0.147 ± 0.168	0.434 ± 0.131	0.596 ± 0.163	0.415 ± 0.154	0.34 ± 0.121	0.374 ± 0.15
VEO	**KM**	**3**	0.308 ± 0.163	0.15 ± 0.167	0.436 ± 0.118	0.574 ± 0.136	0.399 ± 0.164	0.346 ± 0.114	0.369 ± 0.144
VR	**KM**	**3**	0.308 ± 0.164	0.165 ± 0.179	0.443 ± 0.108	0.563 ± 0.146	0.388 ± 0.168	0.347 ± 0.114	0.369 ± 0.147
WD	**KM**	**3**	0.321 ± 0.173	0.16 ± 0.18	0.433 ± 0.117	0.574 ± 0.164	0.422 ± 0.138	0.355 ± 0.114	0.378 ± 0.148
DC	**KM**	**3**	0.311 ± 0.17	0.149 ± 0.175	0.45 ± 0.12	0.573 ± 0.159	0.395 ± 0.183	0.344 ± 0.117	0.37 ± 0.154
VEO	**OKA**	**3**	0.371 ± 0.225	0.167 ± 0.184	0.457 ± 0.125	0.58 ± 0.183	0.413 ± 0.205	0.354 ± 0.128	0.39 ± 0.175
VR	**OKA**	**3**	0.319 ± 0.189	0.192 ± 0.197	0.455 ± 0.142	0.581 ± 0.191	0.462 ± 0.203	0.364 ± 0.138	0.395 ± 0.176
WD	**OKA**	**3**	0.313 ± 0.205	0.181 ± 0.2	0.457 ± 0.126	0.589 ± 0.189	0.412 ± 0.189	0.373 ± 0.144	0.388 ± 0.175
DC	**OKA**	**3**	0.294 ± 0.187	0.18 ± 0.199	0.473 ± 0.131	0.577 ± 0.172	0.441 ± 0.19	0.373 ± 0.145	0.39 ± 0.171
VEO	**BL**	**3**	0.401 ± 0.177	0.357 ± 0.232	0.499 ± 0.119	0.591 ± 0.183	0.43 ± 0.187	0.491 ± 0.136	0.461 ± 0.173
VR	**BL**	**3**	0.396 ± 0.181	0.354 ± 0.233	0.499 ± 0.124	0.605 ± 0.191	0.462 ± 0.192	0.493 ± 0.132	0.468 ± 0.175
WD	**BL**	**3**	0.396 ± 0.184	0.35 ± 0.236	0.499 ± 0.119	0.605 ± 0.183	0.472 ± 0.168	0.493 ± 0.134	0.469 ± 0.171
DC	**BL**	**3**	0.398 ± 0.174	0.351 ± 0.237	0.495 ± 0.127	0.589 ± 0.187	0.469 ± 0.198	0.491 ± 0.131	0.465 ± 0.176
VEO	**MDAV**	**4**	0.342 ± 0.187	0.171 ± 0.182	0.465 ± 0.104	0.61 ± 0.131	0.461 ± 0.156	0.382 ± 0.115	0.405 ± 0.146
VR	**MDAV**	**4**	0.325 ± 0.171	0.184 ± 0.184	0.459 ± 0.119	0.601 ± 0.168	0.414 ± 0.18	0.396 ± 0.119	0.396 ± 0.157
WD	**MDAV**	**4**	0.358 ± 0.2	0.192 ± 0.192	0.463 ± 0.126	0.634 ± 0.171	0.417 ± 0.164	0.388 ± 0.124	0.409 ± 0.163
DC	**MDAV**	**4**	0.327 ± 0.178	0.168 ± 0.186	0.471 ± 0.125	0.62 ± 0.156	0.47 ± 0.143	0.393 ± 0.12	0.401 ± 0.151
VEO	**KM**	**4**	0.333 ± 0.172	0.178 ± 0.178	0.458 ± 0.111	0.596 ± 0.154	0.443 ± 0.158	0.377 ± 0.118	0.398 ± 0.149
VR	**KM**	**4**	0.323 ± 0.168	0.182 ± 0.181	0.467 ± 0.113	0.604 ± 0.157	0.432 ± 0.174	0.388 ± 0.114	0.399 ± 0.151
WD	**KM**	**4**	0.36 ± 0.2	0.176 ± 0.183	0.472 ± 0.107	0.618 ± 0.161	0.431 ± 0.166	0.392 ± 0.117	0.408 ± 0.156
DC	**KM**	**4**	0.34 ± 0.192	0.166 ± 0.181	0.48 ± 0.104	0.628 ± 0.153	0.47 ± 0.193	0.382 ± 0.118	0.403 ± 0.157
VEO	**OKA**	**4**	0.37 ± 0.208	0.244 ± 0.21	0.478 ± 0.133	0.618 ± 0.173	0.475 ± 0.2	0.397 ± 0.144	0.43 ± 0.178
VR	**OKA**	**4**	0.345 ± 0.196	0.227 ± 0.214	0.485 ± 0.129	0.624 ± 0.165	0.446 ± 0.196	0.401 ± 0.126	0.421 ± 0.171
WD	**OKA**	**4**	0.352 ± 0.205	0.208 ± 0.206	0.492 ± 0.135	0.639 ± 0.16	0.458 ± 0.185	0.399 ± 0.142	0.425 ± 0.172
DC	**OKA**	**4**	0.34 ± 0.201	0.229 ± 0.213	0.495 ± 0.126	0.639 ± 0.132	0.489 ± 0.2	0.393 ± 0.133	0.431 ± 0.168
VEO	**BL**	**4**	0.439 ± 0.171	0.389 ± 0.218	0.523 ± 0.113	0.635 ± 0.175	0.478 ± 0.194	0.531 ± 0.126	0.499 ± 0.166
VR	**BL**	**4**	0.433 ± 0.181	0.395 ± 0.22	0.521 ± 0.123	0.631 ± 0.185	0.495 ± 0.187	0.53 ± 0.123	0.501 ± 0.17
WD	**BL**	**4**	0.432 ± 0.18	0.39 ± 0.218	0.516 ± 0.129	0.633 ± 0.178	0.49 ± 0.193	0.531 ± 0.128	0.499 ± 0.171
DC	**BL**	**4**	0.427 ± 0.179	0.387 ± 0.222	0.519 ± 0.126	0.646 ± 0.169	0.493 ± 0.192	0.534 ± 0.126	0.501 ± 0.169
VEO	**MDAV**	**5**	0.364 ± 0.193	0.19 ± 0.188	0.47 ± 0.107	0.644 ± 0.143	0.511 ± 0.138	0.4 ± 0.114	0.43 ± 0.147
VR	**MDAV**	**5**	0.356 ± 0.191	0.197 ± 0.193	0.477 ± 0.12	0.633 ± 0.149	0.467 ± 0.16	0.402 ± 0.117	0.422 ± 0.155
WD	**MDAV**	**5**	0.379 ± 0.197	0.204 ± 0.194	0.492 ± 0.107	0.642 ± 0.158	0.482 ± 0.155	0.417 ± 0.134	0.436 ± 0.157
DC	**MDAV**	**5**	0.349 ± 0.177	0.168 ± 0.182	0.489 ± 0.126	0.642 ± 0.145	0.515 ± 0.136	0.401 ± 0.129	0.436 ± 0.149
VEO	**KM**	**5**	0.367 ± 0.178	0.183 ± 0.183	0.469 ± 0.102	0.633 ± 0.132	0.487 ± 0.15	0.408 ± 0.109	0.425 ± 0.142
VR	**KM**	**5**	0.369 ± 0.18	0.199 ± 0.198	0.49 ± 0.113	0.622 ± 0.158	0.489 ± 0.179	0.409 ± 0.106	0.43 ± 0.156
WD	**KM**	**5**	0.383 ± 0.198	0.206 ± 0.201	0.501 ± 0.104	0.652 ± 0.135	0.491 ± 0.119	0.421 ± 0.118	0.435 ± 0.146
DC	**KM**	**5**	0.354 ± 0.18	0.184 ± 0.189	0.495 ± 0.125	0.646 ± 0.151	0.505 ± 0.167	0.421 ± 0.125	0.431 ± 0.156
VEO	**OKA**	**5**	0.372 ± 0.205	0.237 ± 0.21	0.5 ± 0.124	0.633 ± 0.14	0.431 ± 0.176	0.423 ± 0.157	0.433 ± 0.169
VR	**OKA**	**5**	0.356 ± 0.201	0.268 ± 0.23	0.505 ± 0.119	0.647 ± 0.135	0.469 ± 0.183	0.425 ± 0.132	0.445 ± 0.167
WD	**OKA**	**5**	0.359 ± 0.183	0.256 ± 0.217	0.488 ± 0.115	0.654 ± 0.164	0.529 ± 0.152	0.435 ± 0.159	0.454 ± 0.165
DC	**OKA**	**5**	0.34 ± 0.203	0.215 ± 0.216	0.517 ± 0.128	0.651 ± 0.136	0.514 ± 0.175	0.428 ± 0.139	0.448 ± 0.166
VEO	**BL**	**5**	0.467 ± 0.17	0.404 ± 0.224	0.545 ± 0.108	0.667 ± 0.161	0.536 ± 0.204	0.546 ± 0.122	0.528 ± 0.165
VR	**BL**	**5**	0.468 ± 0.17	0.399 ± 0.224	0.537 ± 0.112	0.671 ± 0.154	0.534 ± 0.189	0.548 ± 0.123	0.526 ± 0.162
WD	**BL**	**5**	0.465 ± 0.175	0.406 ± 0.217	0.525 ± 0.132	0.666 ± 0.162	0.537 ± 0.19	0.548 ± 0.116	0.524 ± 0.165
DC	**BL**	**5**	0.471 ± 0.171	0.403 ± 0.221	0.53 ± 0.123	0.661 ± 0.159	0.506 ± 0.195	0.545 ± 0.123	0.519 ± 0.165

(*continued on next page*)

Besides, experiments demonstrate that the quality of the process models is not strongly affected by the similarity measure used during the anonymization. Although some measures could particularly behave better in certain executions or in certain event logs, the general differences between the quality results are relatively insignificant. Therefore, the similarity measure does not significantly influence the anonymization results, and users are free to use the measure that best fit with their interests, such as according to the process models modeling notation or based on computational criteria.

Table A.1 (*continued*).

Parameters			Event logs						Avg.
sim	*clus*	*k*	BPI12	BPI13	BPI14	BPI15	CoSeLoG	TGN	
VEO	**MDAV**	**10**	0.445 ± 0.192	0.239 ± 0.203	0.515 ± 0.119	0.698 ± 0.126	0.542 ± 0.129	0.456 ± 0.129	0.483 ± 0.15
VR	**MDAV**	**10**	0.463 ± 0.199	0.246 ± 0.211	0.528 ± 0.101	0.677 ± 0.125	0.51 ± 0.11	0.473 ± 0.113	0.483 ± 0.143
WD	**MDAV**	**10**	0.422 ± 0.201	0.26 ± 0.208	0.545 ± 0.107	0.685 ± 0.132	0.546 ± 0.177	0.468 ± 0.128	0.488 ± 0.159
DC	**MDAV**	**10**	0.41 ± 0.199	0.207 ± 0.189	0.542 ± 0.118	0.696 ± 0.133	0.546 ± 0.173	0.45 ± 0.126	0.478 ± 0.156
VEO	**KM**	**10**	0.433 ± 0.18	0.23 ± 0.195	0.53 ± 0.097	0.679 ± 0.117	0.525 ± 0.122	0.459 ± 0.115	0.476 ± 0.138
VR	**KM**	**10**	0.452 ± 0.196	0.235 ± 0.2	0.532 ± 0.1	0.685 ± 0.121	0.518 ± 0.105	0.474 ± 0.107	0.483 ± 0.138
WD	**KM**	**10**	0.46 ± 0.212	0.255 ± 0.207	0.537 ± 0.106	0.714 ± 0.136	0.554 ± 0.148	0.48 ± 0.117	0.503 ± 0.154
DC	**KM**	**10**	0.4 ± 0.201	0.233 ± 0.202	0.535 ± 0.121	0.706 ± 0.128	0.566 ± 0.133	0.464 ± 0.129	0.492 ± 0.152
VEO	**OKA**	**10**	0.477 ± 0.208	0.269 ± 0.216	0.535 ± 0.131	0.709 ± 0.137	0.508 ± 0.145	0.451 ± 0.144	0.493 ± 0.163
VR	**OKA**	**10**	0.426 ± 0.229	0.304 ± 0.224	0.545 ± 0.132	0.703 ± 0.135	0.522 ± 0.171	0.472 ± 0.154	0.493 ± 0.174
WD	**OKA**	**10**	0.454 ± 0.237	0.256 ± 0.228	0.546 ± 0.129	0.696 ± 0.125	0.529 ± 0.159	0.449 ± 0.143	0.492 ± 0.17
DC	**OKA**	**10**	0.447 ± 0.209	0.278 ± 0.223	0.551 ± 0.111	0.695 ± 0.112	0.499 ± 0.15	0.478 ± 0.146	0.491 ± 0.158
VEO	**BL**	**10**	0.584 ± 0.159	0.431 ± 0.206	0.572 ± 0.098	0.728 ± 0.124	0.557 ± 0.162	0.589 ± 0.108	0.574 ± 0.143
VR	**BL**	**10**	0.586 ± 0.153	0.451 ± 0.206	0.557 ± 0.116	0.718 ± 0.145	0.578 ± 0.162	0.579 ± 0.106	0.564 ± 0.148
WD	**BL**	**10**	0.578 ± 0.15	0.432 ± 0.201	0.571 ± 0.104	0.72 ± 0.117	0.583 ± 0.168	0.585 ± 0.109	0.585 ± 0.141
DC	**BL**	**10**	0.585 ± 0.155	0.435 ± 0.207	0.563 ± 0.098	0.723 ± 0.122	0.566 ± 0.167	0.582 ± 0.111	0.582 ± 0.143
VEO	**MDAV**	**20**	0.518 ± 0.15	0.275 ± 0.215	0.552 ± 0.1	0.728 ± 0.128	0.595 ± 0.14	0.495 ± 0.138	0.527 ± 0.145
VR	**MDAV**	**20**	0.497 ± 0.179	0.283 ± 0.206	0.576 ± 0.091	0.725 ± 0.118	0.588 ± 0.109	0.519 ± 0.096	0.531 ± 0.133
WD	**MDAV**	**20**	0.498 ± 0.245	0.293 ± 0.216	0.577 ± 0.105	0.755 ± 0.13	0.585 ± 0.131	0.501 ± 0.136	0.538 ± 0.161
DC	**MDAV**	**20**	0.457 ± 0.199	0.269 ± 0.218	0.572 ± 0.11	0.733 ± 0.123	0.587 ± 0.127	0.488 ± 0.147	0.518 ± 0.154
VEO	**KM**	**20**	0.535 ± 0.144	0.275 ± 0.216	0.571 ± 0.093	0.74 ± 0.129	0.584 ± 0.137	0.5 ± 0.099	0.534 ± 0.136
VR	**KM**	**20**	0.474 ± 0.214	0.311 ± 0.208	0.56 ± 0.093	0.728 ± 0.114	0.574 ± 0.133	0.523 ± 0.101	0.528 ± 0.144
WD	**KM**	**20**	0.521 ± 0.159	0.27 ± 0.194	0.572 ± 0.102	0.726 ± 0.123	0.584 ± 0.136	0.515 ± 0.12	0.534 ± 0.144
DC	**KM**	**20**	0.501 ± 0.189	0.278 ± 0.228	0.567 ± 0.106	0.727 ± 0.133	0.573 ± 0.137	0.519 ± 0.106	0.539 ± 0.145
VEO	**OKA**	**20**	0.488 ± 0.233	0.304 ± 0.239	0.588 ± 0.121	0.727 ± 0.104	0.593 ± 0.122	0.5 ± 0.124	0.533 ± 0.157
VR	**OKA**	**20**	0.496 ± 0.223	0.337 ± 0.209	0.582 ± 0.116	0.72 ± 0.107	0.565 ± 0.111	0.519 ± 0.146	0.537 ± 0.152
WD	**OKA**	**20**	0.485 ± 0.23	0.363 ± 0.21	0.56 ± 0.109	0.751 ± 0.115	0.581 ± 0.12	0.512 ± 0.168	0.542 ± 0.159
DC	**OKA**	**20**	0.483 ± 0.22	0.33 ± 0.225	0.584 ± 0.102	0.723 ± 0.107	0.595 ± 0.122	0.518 ± 0.137	0.539 ± 0.152
VEO	**BL**	**20**	0.647 ± 0.139	0.447 ± 0.194	0.578 ± 0.093	0.761 ± 0.123	0.603 ± 0.153	0.6 ± 0.108	0.613 ± 0.135
VR	**BL**	**20**	0.645 ± 0.158	0.433 ± 0.213	0.589 ± 0.094	0.764 ± 0.125	0.602 ± 0.138	0.6 ± 0.102	0.608 ± 0.139
WD	**BL**	**20**	0.646 ± 0.155	0.471 ± 0.199	0.585 ± 0.108	0.773 ± 0.122	0.607 ± 0.158	0.613 ± 0.1	0.621 ± 0.141
DC	**BL**	**20**	0.646 ± 0.155	0.458 ± 0.187	0.585 ± 0.093	0.759 ± 0.118	0.617 ± 0.141	0.606 ± 0.104	0.609 ± 0.133

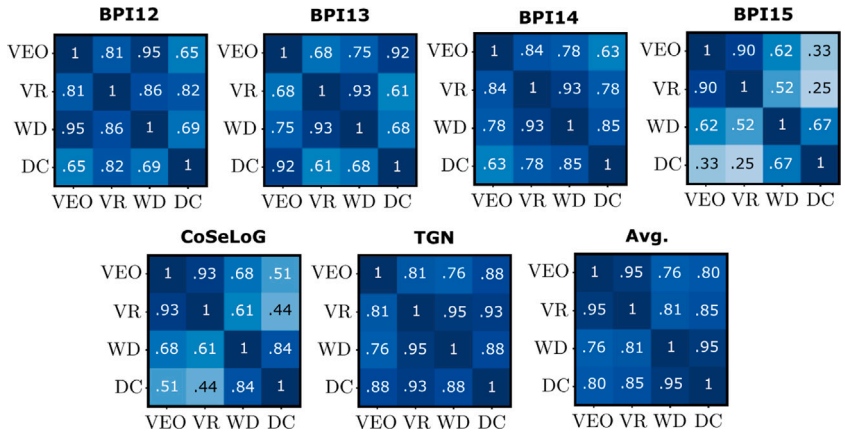

Fig. A.1. Results of the *p*-values from the t-Tests according to the *sim* parameter in the QS results.

5.2.1. *Relationship between QS and ILS results*

To contextualize the experimental results reported in Tables 4 and 5, Fig. 6 depicts the relationship between the QS and ILS results for all event logs. Also, note that the last chart illustrates the averaged results from all the event logs in order to evaluate the general impact of *k*-PPPM from a high-level perspective, regardless the particularities of each event log. Each dot in the chart, colored according to its *k* value, represents the execution of *k*-PPPM with a certain combination of parameters. Also, results obtained using the BL clustering algorithm have been discarded to prevent the appearance of non-optimal solutions.

Interestingly enough, although evaluating the process models from different perspectives, there exists an apparent direct correlation between the results: the more individual distortion in the process models (QS), the more distant the relationships among them (ILS). Due to the microaggregation nature of *k*-PPPM, the process models might suffer a notable individual distortion at the very beginning. But, as the privacy level increases, these distortions are not that severe. Contrary, inter-individual distortions are greatly preserved, *i.e.,* ILS results are generally low. However, increasing the privacy level has a larger impact on these inter-individual distortions rather than the individual distortions. This is because little, but continuous, distortions of all the process models individually have a higher effect in all the relationships among all process models. Notwithstanding, the low ILS results suggests that the method preserves most of the patterns from the original event logs, while protecting individuals' privacy (*i.e.,* QS) at the same time. From a privacy perspective, this property enables acquiring similar insights and knowledge from the protected event logs (and protected process models) as if they were obtained from the original event logs (and original process models), but without disclosing confidential information.

In addition to the quantitative results, Fig. 7 depicts different versions of the process model associated to a certain individual from the

Table A.2
ILS experimental results.

Parameters			Event logs						Avg.
sim	*clus*	*k*	BPI12	BPI13	BPI14	BPI15	CoSeLoG	TGN	
VEO	**MDAV**	**2**	0.055 ± 0.011	0.047 ± 0.013	0.042 ± 0.018	0.037 ± 0.014	0.064 ± 0.019	0.045 ± 0.018	0.048 ± 0.018
VR	**MDAV**	**2**	0.054 ± 0.012	0.052 ± 0.017	0.042 ± 0.016	0.036 ± 0.013	0.066 ± 0.019	0.04 ± 0.013	0.049 ± 0.018
WD	**MDAV**	**2**	0.058 ± 0.014	0.055 ± 0.017	0.043 ± 0.019	0.04 ± 0.018	0.063 ± 0.02	0.048 ± 0.021	0.051 ± 0.02
DC	**MDAV**	**2**	0.052 ± 0.009	0.042 ± 0.01	0.045 ± 0.019	0.038 ± 0.034	0.065 ± 0.022	0.043 ± 0.016	0.051 ± 0.022
VEO	**KM**	**2**	0.055 ± 0.011	0.044 ± 0.012	0.042 ± 0.018	0.035 ± 0.012	0.066 ± 0.019	0.046 ± 0.017	0.048 ± 0.018
VR	**KM**	**2**	0.057 ± 0.014	0.047 ± 0.015	0.042 ± 0.017	0.037 ± 0.012	0.056 ± 0.012	0.044 ± 0.015	0.047 ± 0.016
WD	**KM**	**2**	0.061 ± 0.016	0.051 ± 0.016	0.042 ± 0.019	0.041 ± 0.017	0.059 ± 0.015	0.049 ± 0.02	0.05 ± 0.019
DC	**KM**	**2**	0.054 ± 0.011	0.043 ± 0.011	0.045 ± 0.02	0.046 ± 0.041	0.078 ± 0.031	0.045 ± 0.016	0.053 ± 0.027
VEO	**OKA**	**2**	0.051 ± 0.004	0.051 ± 0.012	0.044 ± 0.022	0.039 ± 0.018	0.083 ± 0.03	0.045 ± 0.016	0.052 ± 0.024
VR	**OKA**	**2**	0.07 ± 0.024	0.054 ± 0.014	0.048 ± 0.022	0.044 ± 0.021	0.071 ± 0.021	0.05 ± 0.017	0.055 ± 0.026
WD	**OKA**	**2**	0.061 ± 0.013	0.056 ± 0.015	0.049 ± 0.027	0.049 ± 0.03	0.086 ± 0.031	0.052 ± 0.023	0.06 ± 0.027
DC	**OKA**	**2**	0.079 ± 0.018	0.057 ± 0.015	0.047 ± 0.022	0.05 ± 0.025	0.08 ± 0.031	0.052 ± 0.021	0.061 ± 0.026
VEO	**BL**	**2**	0.103 ± 0.024	0.124 ± 0.025	0.056 ± 0.028	0.053 ± 0.028	0.087 ± 0.031	0.094 ± 0.04	0.086 ± 0.039
VR	**BL**	**2**	0.105 ± 0.026	0.125 ± 0.026	0.058 ± 0.03	0.053 ± 0.028	0.094 ± 0.037	0.095 ± 0.041	0.088 ± 0.041
WD	**BL**	**2**	0.102 ± 0.023	0.126 ± 0.026	0.056 ± 0.027	0.052 ± 0.028	0.086 ± 0.029	0.091 ± 0.039	0.085 ± 0.039
DC	**BL**	**2**	0.103 ± 0.025	0.124 ± 0.025	0.056 ± 0.028	0.054 ± 0.029	0.089 ± 0.031	0.093 ± 0.04	0.086 ± 0.039
VEO	**MDAV**	**3**	0.081 ± 0.016	0.062 ± 0.017	0.053 ± 0.022	0.059 ± 0.019	0.081 ± 0.02	0.056 ± 0.021	0.065 ± 0.022
VR	**MDAV**	**3**	0.075 ± 0.01	0.058 ± 0.015	0.055 ± 0.022	0.061 ± 0.018	0.088 ± 0.023	0.054 ± 0.019	0.065 ± 0.022
WD	**MDAV**	**3**	0.091 ± 0.024	0.062 ± 0.016	0.052 ± 0.023	0.065 ± 0.026	0.08 ± 0.019	0.062 ± 0.027	0.069 ± 0.026
DC	**MDAV**	**3**	0.083 ± 0.014	0.055 ± 0.012	0.056 ± 0.024	0.066 ± 0.052	0.089 ± 0.026	0.053 ± 0.019	0.07 ± 0.033
VEO	**KM**	**3**	0.083 ± 0.018	0.057 ± 0.015	0.054 ± 0.022	0.057 ± 0.019	0.089 ± 0.027	0.056 ± 0.02	0.066 ± 0.025
VR	**KM**	**3**	0.081 ± 0.016	0.057 ± 0.016	0.054 ± 0.021	0.055 ± 0.015	0.085 ± 0.024	0.054 ± 0.018	0.064 ± 0.023
WD	**KM**	**3**	0.094 ± 0.028	0.064 ± 0.019	0.054 ± 0.023	0.062 ± 0.024	0.086 ± 0.026	0.059 ± 0.023	0.07 ± 0.028
DC	**KM**	**3**	0.078 ± 0.015	0.055 ± 0.013	0.057 ± 0.024	0.064 ± 0.045	0.083 ± 0.026	0.055 ± 0.019	0.069 ± 0.029
VEO	**OKA**	**3**	0.097 ± 0.067	0.059 ± 0.013	0.056 ± 0.025	0.068 ± 0.03	0.097 ± 0.03	0.053 ± 0.017	0.076 ± 0.051
VR	**OKA**	**3**	0.084 ± 0.023	0.065 ± 0.017	0.057 ± 0.023	0.066 ± 0.028	0.088 ± 0.043	0.059 ± 0.02	0.074 ± 0.034
WD	**OKA**	**3**	0.086 ± 0.026	0.065 ± 0.017	0.054 ± 0.027	0.065 ± 0.036	0.094 ± 0.03	0.065 ± 0.028	0.073 ± 0.031
DC	**OKA**	**3**	0.075 ± 0.017	0.069 ± 0.018	0.064 ± 0.03	0.065 ± 0.027	0.098 ± 0.033	0.064 ± 0.026	0.073 ± 0.028
VEO	**BL**	**3**	0.147 ± 0.034	0.143 ± 0.028	0.065 ± 0.031	0.07 ± 0.029	0.108 ± 0.03	0.111 ± 0.047	0.107 ± 0.047
VR	**BL**	**3**	0.145 ± 0.032	0.144 ± 0.029	0.067 ± 0.032	0.069 ± 0.028	0.112 ± 0.034	0.112 ± 0.046	0.108 ± 0.047
WD	**BL**	**3**	0.144 ± 0.031	0.142 ± 0.028	0.066 ± 0.032	0.069 ± 0.03	0.109 ± 0.031	0.111 ± 0.047	0.107 ± 0.046
DC	**BL**	**3**	0.146 ± 0.033	0.144 ± 0.03	0.066 ± 0.032	0.068 ± 0.029	0.12 ± 0.036	0.111 ± 0.046	0.109 ± 0.048
VEO	**MDAV**	**4**	0.105 ± 0.025	0.063 ± 0.015	0.061 ± 0.024	0.072 ± 0.02	0.112 ± 0.028	0.067 ± 0.023	0.08 ± 0.031
VR	**MDAV**	**4**	0.086 ± 0.013	0.066 ± 0.018	0.059 ± 0.022	0.072 ± 0.019	0.1 ± 0.021	0.069 ± 0.023	0.075 ± 0.024
WD	**MDAV**	**4**	0.119 ± 0.038	0.077 ± 0.023	0.059 ± 0.025	0.073 ± 0.037	0.106 ± 0.025	0.069 ± 0.026	0.085 ± 0.036
DC	**MDAV**	**4**	0.101 ± 0.021	0.064 ± 0.015	0.064 ± 0.026	0.074 ± 0.057	0.107 ± 0.027	0.071 ± 0.026	0.082 ± 0.038
VEO	**KM**	**4**	0.096 ± 0.017	0.064 ± 0.017	0.059 ± 0.023	0.072 ± 0.019	0.117 ± 0.03	0.065 ± 0.023	0.079 ± 0.03
VR	**KM**	**4**	0.086 ± 0.012	0.063 ± 0.017	0.061 ± 0.023	0.072 ± 0.018	0.105 ± 0.023	0.063 ± 0.02	0.075 ± 0.025
WD	**KM**	**4**	0.125 ± 0.039	0.068 ± 0.019	0.062 ± 0.027	0.081 ± 0.033	0.106 ± 0.025	0.071 ± 0.028	0.086 ± 0.037
DC	**KM**	**4**	0.102 ± 0.022	0.062 ± 0.015	0.066 ± 0.028	0.078 ± 0.057	0.113 ± 0.031	0.064 ± 0.023	0.086 ± 0.039
VEO	**OKA**	**4**	0.132 ± 0.042	0.089 ± 0.022	0.064 ± 0.028	0.082 ± 0.032	0.119 ± 0.037	0.069 ± 0.028	0.094 ± 0.042
VR	**OKA**	**4**	0.099 ± 0.022	0.08 ± 0.018	0.068 ± 0.028	0.07 ± 0.021	0.114 ± 0.024	0.067 ± 0.024	0.083 ± 0.029
WD	**OKA**	**4**	0.103 ± 0.025	0.08 ± 0.017	0.072 ± 0.034	0.077 ± 0.039	0.12 ± 0.032	0.069 ± 0.028	0.088 ± 0.035
DC	**OKA**	**4**	0.103 ± 0.025	0.088 ± 0.021	0.068 ± 0.03	0.077 ± 0.026	0.126 ± 0.034	0.064 ± 0.024	0.088 ± 0.035
VEO	**BL**	**4**	0.149 ± 0.03	0.154 ± 0.031	0.073 ± 0.035	0.084 ± 0.032	0.13 ± 0.029	0.126 ± 0.053	0.117 ± 0.047
VR	**BL**	**4**	0.15 ± 0.032	0.158 ± 0.033	0.072 ± 0.033	0.085 ± 0.032	0.131 ± 0.041	0.126 ± 0.054	0.12 ± 0.05
WD	**BL**	**4**	0.153 ± 0.035	0.153 ± 0.03	0.073 ± 0.033	0.086 ± 0.033	0.126 ± 0.037	0.127 ± 0.053	0.12 ± 0.049
DC	**BL**	**4**	0.151 ± 0.033	0.154 ± 0.031	0.072 ± 0.032	0.084 ± 0.032	0.126 ± 0.037	0.124 ± 0.052	0.118 ± 0.048
VEO	**MDAV**	**5**	0.116 ± 0.025	0.066 ± 0.019	0.065 ± 0.024	0.085 ± 0.021	0.117 ± 0.021	0.07 ± 0.024	0.086 ± 0.032
VR	**MDAV**	**5**	0.108 ± 0.019	0.067 ± 0.021	0.065 ± 0.023	0.089 ± 0.022	0.121 ± 0.022	0.07 ± 0.023	0.087 ± 0.03
WD	**MDAV**	**5**	0.139 ± 0.039	0.077 ± 0.026	0.068 ± 0.028	0.104 ± 0.035	0.132 ± 0.042	0.079 ± 0.032	0.103 ± 0.047
DC	**MDAV**	**5**	0.106 ± 0.018	0.064 ± 0.015	0.073 ± 0.029	0.122 ± 0.051	0.129 ± 0.027	0.071 ± 0.025	0.095 ± 0.04
VEO	**KM**	**5**	0.125 ± 0.027	0.065 ± 0.019	0.062 ± 0.022	0.094 ± 0.028	0.143 ± 0.035	0.07 ± 0.022	0.093 ± 0.041
VR	**KM**	**5**	0.113 ± 0.019	0.065 ± 0.02	0.067 ± 0.025	0.086 ± 0.019	0.118 ± 0.019	0.072 ± 0.022	0.087 ± 0.03
WD	**KM**	**5**	0.142 ± 0.038	0.078 ± 0.027	0.072 ± 0.031	0.109 ± 0.04	0.122 ± 0.023	0.082 ± 0.033	0.101 ± 0.041
DC	**KM**	**5**	0.102 ± 0.013	0.071 ± 0.017	0.075 ± 0.03	0.13 ± 0.058	0.133 ± 0.031	0.078 ± 0.029	0.098 ± 0.043
VEO	**OKA**	**5**	0.12 ± 0.034	0.085 ± 0.024	0.072 ± 0.032	0.096 ± 0.03	0.121 ± 0.02	0.074 ± 0.026	0.095 ± 0.034
VR	**OKA**	**5**	0.098 ± 0.02	0.101 ± 0.033	0.066 ± 0.024	0.096 ± 0.028	0.12 ± 0.02	0.077 ± 0.025	0.093 ± 0.031
WD	**OKA**	**5**	0.105 ± 0.025	0.095 ± 0.029	0.063 ± 0.026	0.106 ± 0.039	0.13 ± 0.029	0.081 ± 0.033	0.097 ± 0.037
DC	**OKA**	**5**	0.098 ± 0.021	0.079 ± 0.02	0.071 ± 0.035	0.106 ± 0.037	0.134 ± 0.03	0.081 ± 0.031	0.094 ± 0.035
VEO	**BL**	**5**	0.184 ± 0.043	0.157 ± 0.034	0.079 ± 0.036	0.167 ± 0.03	0.143 ± 0.034	0.132 ± 0.055	0.166 ± 0.053
VR	**BL**	**5**	0.181 ± 0.042	0.16 ± 0.036	0.078 ± 0.035	0.168 ± 0.031	0.135 ± 0.028	0.129 ± 0.05	0.16 ± 0.052
WD	**BL**	**5**	0.178 ± 0.038	0.154 ± 0.031	0.077 ± 0.033	0.167 ± 0.031	0.137 ± 0.029	0.129 ± 0.055	0.159 ± 0.05
DC	**BL**	**5**	0.177 ± 0.039	0.154 ± 0.032	0.076 ± 0.033	0.166 ± 0.03	0.139 ± 0.031	0.129 ± 0.052	0.159 ± 0.05

(*continued on next page*)

CoSeLoG event log obtained for different executions of k-PPPM. More specifically, the original process model can be qualitatively compared to protected process models obtained using heuristic or non-heuristic algorithms (MDAV or BL, respectively) and for different privacy levels ($k = 2$ and $k = 5$). By comparing process models with the same privacy level k, it can be noticed that the model obtained with MDAV is more similar to the original model (in terms of number of nodes, edges and weights), rather than the one obtained with BL. Moreover, this quality decrease is also noticeable when comparing models of different privacy levels. Asking experts and practitioners whether those differences could

Table A.2 (*continued*).

Parameters			Event logs						Avg.
sim	*clus*	*k*	BPI12	BPI13	BPI14	BPI15	CoSeLoG	TGN	
VEO	**MDAV**	**10**	0.191 ± 0.034	0.081 ± 0.025	0.082 ± 0.026	0.154 ± 0.027	0.222 ± 0.032	0.095 ± 0.03	0.138 ± 0.063
VR	**MDAV**	**10**	0.198 ± 0.038	0.084 ± 0.027	0.082 ± 0.024	0.153 ± 0.024	0.2 ± 0.029	0.098 ± 0.032	0.136 ± 0.059
WD	**MDAV**	**10**	0.201 ± 0.039	0.094 ± 0.032	0.085 ± 0.036	0.175 ± 0.074	0.225 ± 0.055	0.097 ± 0.035	0.146 ± 0.084
DC	**MDAV**	**10**	0.161 ± 0.02	0.072 ± 0.018	0.089 ± 0.036	0.183 ± 0.049	0.214 ± 0.031	0.088 ± 0.028	0.136 ± 0.062
VEO	**KM**	**10**	0.176 ± 0.024	0.082 ± 0.014	0.086 ± 0.028	0.141 ± 0.018	0.201 ± 0.028	0.091 ± 0.026	0.13 ± 0.052
VR	**KM**	**10**	0.182 ± 0.026	0.085 ± 0.018	0.083 ± 0.025	0.167 ± 0.035	0.194 ± 0.028	0.093 ± 0.026	0.134 ± 0.055
WD	**KM**	**10**	0.211 ± 0.044	0.095 ± 0.022	0.085 ± 0.033	0.171 ± 0.063	0.255 ± 0.051	0.099 ± 0.033	0.159 ± 0.078
DC	**KM**	**10**	0.16 ± 0.022	0.092 ± 0.021	0.087 ± 0.034	0.177 ± 0.058	0.211 ± 0.029	0.1 ± 0.032	0.143 ± 0.06
VEO	**OKA**	**10**	0.2 ± 0.044	0.101 ± 0.025	0.084 ± 0.034	0.18 ± 0.087	0.201 ± 0.026	0.087 ± 0.023	0.152 ± 0.075
VR	**OKA**	**10**	0.194 ± 0.043	0.108 ± 0.03	0.084 ± 0.038	0.186 ± 0.049	0.191 ± 0.029	0.102 ± 0.032	0.148 ± 0.057
WD	**OKA**	**10**	0.202 ± 0.046	0.093 ± 0.021	0.083 ± 0.037	0.16 ± 0.032	0.2 ± 0.024	0.09 ± 0.027	0.14 ± 0.059
DC	**OKA**	**10**	0.158 ± 0.019	0.1 ± 0.025	0.089 ± 0.031	0.155 ± 0.03	0.204 ± 0.027	0.108 ± 0.038	0.136 ± 0.049
VEO	**BL**	**10**	0.237 ± 0.04	0.167 ± 0.031	0.091 ± 0.033	0.196 ± 0.029	0.324 ± 0.031	0.165 ± 0.063	0.196 ± 0.062
VR	**BL**	**10**	0.239 ± 0.044	0.172 ± 0.033	0.092 ± 0.033	0.193 ± 0.027	0.311 ± 0.026	0.153 ± 0.056	0.193 ± 0.06
WD	**BL**	**10**	0.242 ± 0.045	0.169 ± 0.032	0.091 ± 0.032	0.191 ± 0.025	0.316 ± 0.029	0.162 ± 0.06	0.195 ± 0.062
DC	**BL**	**10**	0.264 ± 0.062	0.168 ± 0.032	0.092 ± 0.031	0.19 ± 0.025	0.311 ± 0.025	0.159 ± 0.061	0.197 ± 0.068
VEO	**MDAV**	**20**	0.278 ± 0.041	0.104 ± 0.026	0.111 ± 0.025	0.307 ± 0.025	0.366 ± 0.058	0.122 ± 0.034	0.215 ± 0.112
VR	**MDAV**	**20**	0.248 ± 0.035	0.105 ± 0.029	0.107 ± 0.023	0.31 ± 0.029	0.345 ± 0.053	0.116 ± 0.03	0.205 ± 0.106
WD	**MDAV**	**20**	0.284 ± 0.039	0.112 ± 0.03	0.123 ± 0.034	0.313 ± 0.075	0.362 ± 0.056	0.119 ± 0.034	0.232 ± 0.128
DC	**MDAV**	**20**	0.264 ± 0.042	0.098 ± 0.022	0.12 ± 0.031	0.318 ± 0.03	0.34 ± 0.053	0.11 ± 0.028	0.208 ± 0.108
VEO	**KM**	**20**	0.289 ± 0.044	0.103 ± 0.027	0.116 ± 0.029	0.304 ± 0.029	0.365 ± 0.053	0.112 ± 0.024	0.215 ± 0.113
VR	**KM**	**20**	0.249 ± 0.034	0.111 ± 0.032	0.109 ± 0.025	0.287 ± 0.028	0.365 ± 0.057	0.116 ± 0.026	0.206 ± 0.106
WD	**KM**	**20**	0.288 ± 0.044	0.111 ± 0.031	0.128 ± 0.038	0.31 ± 0.092	0.36 ± 0.051	0.131 ± 0.037	0.236 ± 0.129
DC	**KM**	**20**	0.255 ± 0.03	0.095 ± 0.022	0.123 ± 0.032	0.376 ± 0.075	0.387 ± 0.068	0.126 ± 0.032	0.227 ± 0.13
VEO	**OKA**	**20**	0.239 ± 0.025	0.119 ± 0.022	0.13 ± 0.038	0.288 ± 0.018	0.372 ± 0.053	0.118 ± 0.03	0.211 ± 0.102
VR	**OKA**	**20**	0.245 ± 0.029	0.124 ± 0.026	0.128 ± 0.036	0.286 ± 0.024	0.338 ± 0.053	0.142 ± 0.044	0.21 ± 0.092
WD	**OKA**	**20**	0.245 ± 0.027	0.133 ± 0.03	0.111 ± 0.027	0.31 ± 0.022	0.389 ± 0.052	0.137 ± 0.043	0.221 ± 0.109
DC	**OKA**	**20**	0.248 ± 0.028	0.124 ± 0.025	0.126 ± 0.035	0.296 ± 0.024	0.341 ± 0.05	0.129 ± 0.036	0.211 ± 0.095
VEO	**BL**	**20**	0.323 ± 0.046	0.176 ± 0.038	0.133 ± 0.028	0.387 ± 0.021	0.397 ± 0.051	0.178 ± 0.053	0.266 ± 0.097
VR	**BL**	**20**	0.313 ± 0.04	0.176 ± 0.037	0.132 ± 0.026	0.385 ± 0.023	0.409 ± 0.059	0.177 ± 0.054	0.265 ± 0.099
WD	**BL**	**20**	0.32 ± 0.044	0.179 ± 0.038	0.135 ± 0.029	0.392 ± 0.019	0.397 ± 0.051	0.175 ± 0.051	0.268 ± 0.097
DC	**BL**	**20**	0.311 ± 0.039	0.172 ± 0.035	0.138 ± 0.027	0.39 ± 0.021	0.4 ± 0.05	0.19 ± 0.06	0.266 ± 0.096

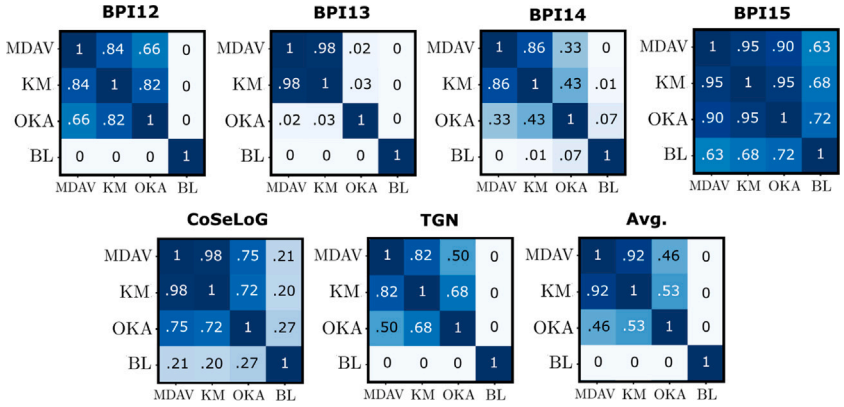

Fig. A.2. Results of the *p*-values from the t-Tests according to the *clus* parameter in the QS results.

change their understanding or the decisions they would make might of great interest too. However, this is beyond the scope of this research.

5.2.2. Significance of the anonymization parameters

Experimental results are clearly affected by the choice of the parameters. To verify the significance of the results differences, statistical analyses of two sampled t-Tests are performed. This test, widely used in statistics, compares the means of two independent groups (*i.e.,* populations) as a way to determine whether there is statistical evidence that the two means are significantly different. If statistical differences are found, it is assumed that choosing a certain parameter value instead of another affects significantly the quality of the process models.

These tests are conducted for the three parameters individually. For each of them, the population of QS results obtained when using a certain value is compared to the population of QS results obtained when using another value. As a result, the t-Test evaluates whether the means of the two populations are significantly different. This procedure is repeated for the ILS results. The settings of our t-Tests use the standard form: the *null hypothesis* states that there is no statistically significant difference in the mean of two populations (*i.e.,* means are equal); the *alternative hypothesis* states that there is statistically significant difference in the mean of the two populations (*i.e.,* means are different); and the *significance level* α is set to $\alpha = 0.05$ indicating the probability of rejecting the null hypothesis when it is true. The t-Tests, which return a *p*-value, indicate that the null hypothesis is rejected (and the alternative hypothesis is accepted) if the *p*-value is lower than α; otherwise, the null hypothesis cannot be rejected. Figs. A.1 to A.6 from Appendix illustrate the *p*-value results obtained from the t-Tests between each pair of parameter's values for the QS and ILS results, respectively.

According to the results, t-Tests confirm that the mean differences, for both the QS and ILS results, between using a similarity measure or another during the anonymization are not statistically significant. Hence, this parameter does not significantly contribute to the quality of the process models. Although no significant difference are found, it is

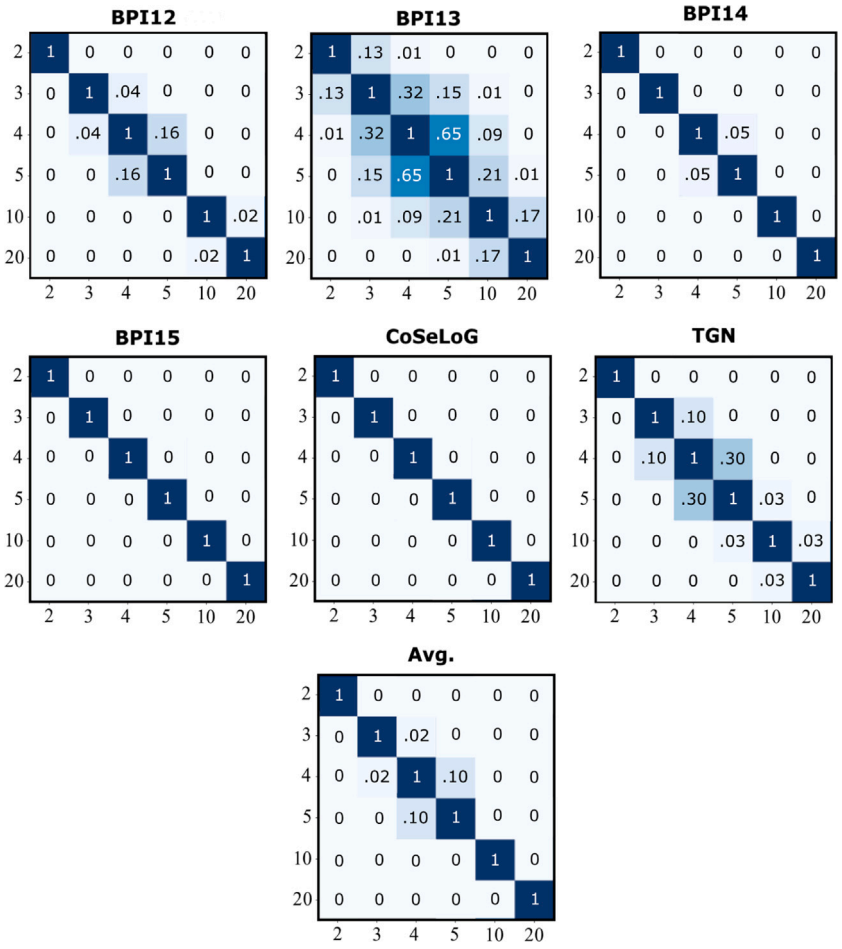

Fig. A.3. Results of the p-values from the t-Tests according to the k parameter in the QS results.

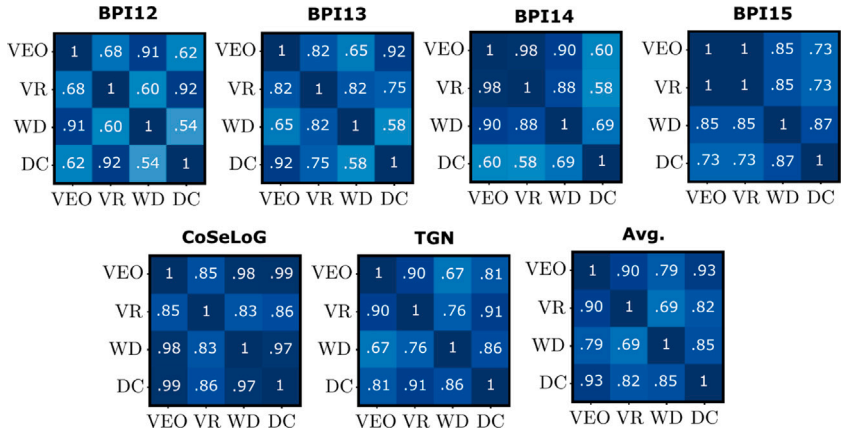

Fig. A.4. Results of the p-values from the t-Tests according to the sim parameter in the ILS results.

noteworthy that experimental results have shown that some measures lead to (slightly) better outcomes than others in particular event logs.

Regarding the clustering algorithm, t-Tests generally agree that the differences between heuristic and non-heuristic algorithms are statistically significant. Therefore, we can state that heuristic algorithms contribute to achieve better anonymization results. However, it is worth noting that the particularities of the event logs might affect this significance. In BPI15 and CoSeLoG event logs, t-Tests have not been able to detect significant differences between these two kinds of clustering algorithms. Also, it can be observed that, in BPI13 event log, significant difference were detected within the heuristic algorithms themselves, in which results obtained using MDAV and KM are significantly better than the ones obtained using OKA. Hence, despite the use of heuristic algorithms, they are not a silver bullet to guarantee the best possible results.

Finally, it can easily be observed that the privacy level highly affects to the quality of the process models, as expected. However, note that t-Tests do not sometimes detect significant difference between consecutive privacy levels (*e.g.*, between $k = 4$ and $k = 5$ in BPI12 and TGN event logs, or for multiple combinations in BPI13 event log, to name a few). These scenarios can be beneficial to increase the privacy constraints with a negligible information loss. Despite the

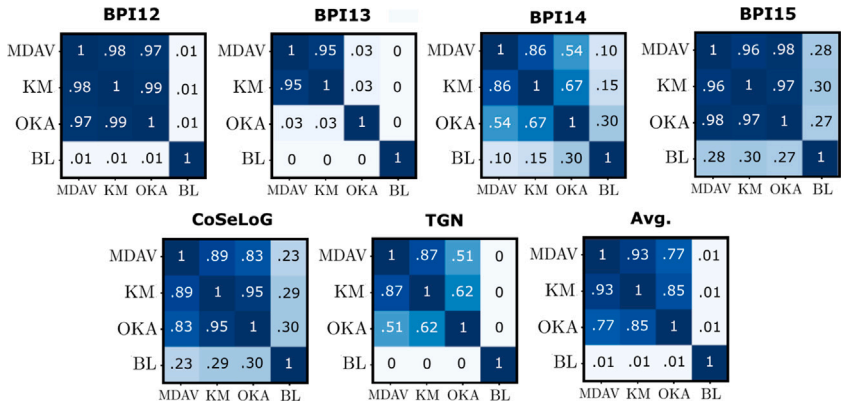

Fig. A.5. Results of the *p*-values from the t-Tests according to the *clus* parameter in the ILS results.

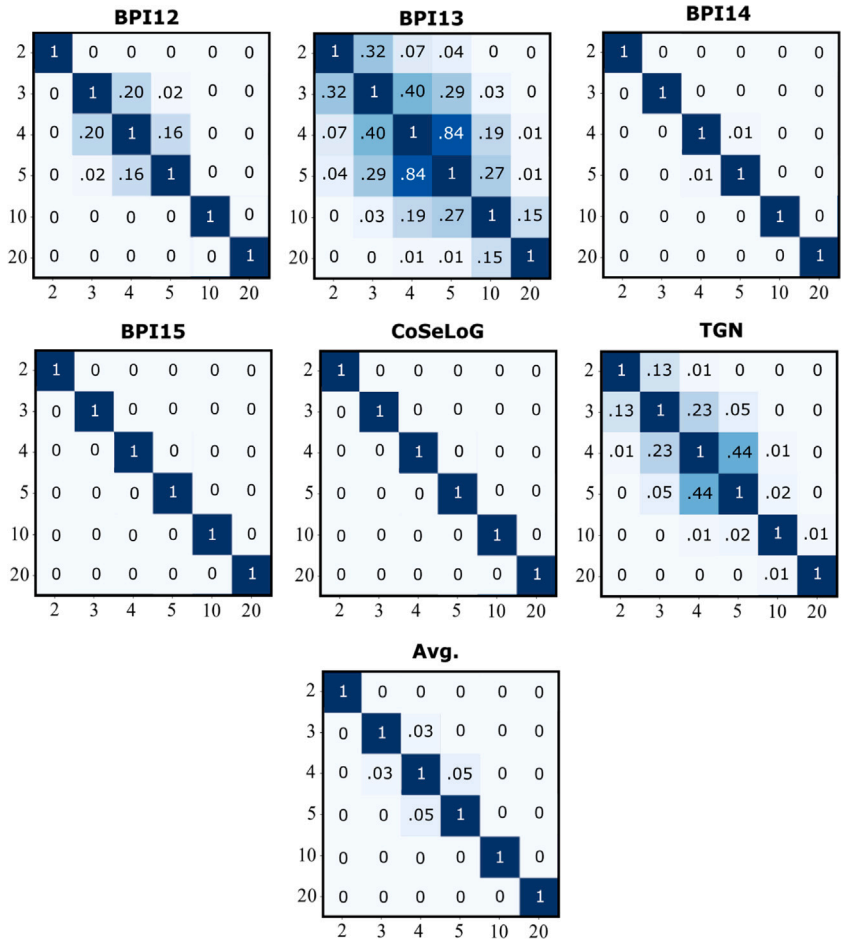

Fig. A.6. Results of the *p*-values from the t-Tests according to the *k* parameter in the ILS results.

above, averaged results show that the quality of the process models significantly improves by lowering the privacy level.

6. Conclusion

Process mining (PM) is a growing research discipline aiming at exploiting vast amounts of event data to obtain knowledge about the execution of business processes within organizations. Many advancements have been achieved in the latest years concerning the development of novel algorithms, the use of different modeling notation languages, the study of processes from diverse perspectives or the design of powerful and strategic visualizations. However, there are still several challenges to be addressed. Among those, this article has addressed some privacy risks during PM analysis, that might enable the re-identification of individuals and/or the inference of confidential data from process models. This kind of attacks, feasible in institutions with public access, may allow attackers to conduct location-oriented attacks, such as restricted space identification and object identification attacks, against targeted individuals. Unfortunately, current privacy-related solutions within PM, mostly focused on pseudonymization or encryption techniques, are not robust against these attacks, because they are unable to break the link between personally identifiable information and confidential data.

To this end, this article has proposed a novel privacy-preserving process mining method based on microaggregation, called *k*-PPPM, that distorts people's process models individually, and avoids the direct re-identification of individuals according to their process models. To the best of our knowledge, this method is the first technical contribution to PPPM using microaggregation techniques. This method is, indeed, inspired by the well-known *k*-anonymity model. The *k*-PPPM method creates a privacy-preserved event log version so that any process model could be discovered from k or more individuals, where k is the privacy level, and moreover, each process model represents the behavior of, at least, k individuals, instead of a single (and potentially identifiable) individual. Notwithstanding, the privacy constraints introduced to the protected event logs results in a loss of data utility.

To evaluate data utility, the quality of the protected process models has been evaluated from two perspectives. On the one hand, regarding the individual distortion suffered by each process model individually (QS) and, on the other hand, regarding the preservation of the relationships among all pairs of process models (ILS). Results, tested with six real-life event logs and different *k*-PPPM parameters, demonstrate the introduction of a homogeneous distortion in the quality of the process models according to the privacy level. Although the individual distortion suffered by all process models, *k*-PPPM preserves the relationships between the different process models, which allows the extraction of similar insights from the protected event logs, instead of using the original (and confidential) event logs.

Although the contributions in this article are a step forward in this novel research direction, there is still room for improvements. Indeed, gathering ideas from the classical privacy protection techniques and applying them into the PM field may bring great research opportunities. Future work will focus on the development of novel PPPM techniques facing more complex attacker models, where attackers can gain advanced knowledge. Also, we foresee the creation and application of more robust privacy-preserving models, which incorporate properties, such as *l*-diversity or *p*-sensitivity. With regards to the proposed method, it could be valuable to eliminate the non-deterministic nature of the methods and prevent the results randomness that may affect the quality results. Last but not least, evaluating the suitability of the proposed methods using advanced modeling notations, such as petri nets or BPMN, might be of interest too.

CRediT authorship contribution statement

Edgar Batista: Conceptualization, Methodology, Formal analysis, Software, Visualization, Writing – original draft, Writing – review & editing. **Antoni Martínez-Ballesté:** Conceptualization, Methodology, Writing – review & editing, Funding acquisition. **Agusti Solanas:** Conceptualization, Methodology, Formal analysis, Writing – review & editing, Funding acquisition, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by grant IoTrain RTI2018-095499-B-C32 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe", by Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) with grants 2017-DI-002, 2017-SGR-896 and ACTUA 2020PANDE00103, by Universitat Rovira i Virgili with project 2017PFR-URV-B2-41, and by the European Commission (EU) with LOCARD project under grant number 832735 and by GoodBrother COST action 19121. Pictures designed with Freepik.

Appendix. Details of the experimental results

- Table A.1 presents the averaged QS results for each execution of *k*-PPPM with a combination of its parameters.
- Table A.2 presents the averaged ILS results for each execution of *k*-PPPM with a combination of its parameters.
- Figs. A.1, A.2 and A.3 illustrate the *p*-value results from the statistical t-Tests between each pair of parameter's values (*sim*, *clus* and *k*, respectively) in the QS results.
- Figs. A.4, A.5 and A.6 illustrate the *p*-value results from the statistical t-Tests between each pair of parameter's values (*sim*, *clus* and *k*, respectively) in the ILS results.

References

[1] Weske M. Business process management – concepts, languages, architectures. 1st ed.. Berlin Heidelberg: Springer; 2007.
[2] Becker J, Kugeler M, Rosemann M. Process management: A guide for the design of business processes. 2nd ed.. Berlin Heidelberg: Springer; 2011.
[3] van der Aalst WMP. Process mining: Discovery, conformance and enhancement of business processes. 1st ed.. Berlin Heidelberg: Springer; 2011.
[4] dos Santos Garcia C, Meincheim A, Faria Junior ER, Rosano Dallagassa M, Vecino Sato DM, Ribeiro Carvalho D, Portela Santos EA, Scalabrin EE. Process mining techniques and applications – A systematic mapping study. Expert Syst Appl 2019;133:260–95.
[5] van der Aalst WMP. Process mining: Data science in action. 2nd ed.. Berlin Heidelberg: Springer; 2016.
[6] 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). Off J Eur Union 2016;L119:1–88.
[7] Solanas A, Patsakis C, Conti M, Vlachos IS, Ramos V, Falcone F, Postolache O, Pérez-Martínez PA, Di Pietro R, Perrea DN, Martínez-Ballesté A. Smart health: A context-aware health paradigm within smart cities. IEEE Commun Mag 2014;52(8):74–81.
[8] 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. Almeria, Spain: Springer; 2019, p. 212–22.
[9] 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 IEEE international forum on research and technologies for society and industry, Modena, Italy; 2017, p. 1–6.
[10] van der Aalst WMP, Adriansyah A, Alves de Medeiros AK, Arcieri F, Baier T, Blickle T, Bose JC, van den Brand P, Brandtjen R, Buijs JCAM, et al. Process mining manifesto. In: Proceedings of the 9th international conference on business process management, Clermont-Ferrand, France; 2011, p. 169–94.
[11] Batista E, Solanas A. Process mining in healthcare: A systematic review. In: Proceedings of the 9th international conference on information, intelligence, systems applications, Zakynthos, Greece; 2018, p. 1–6.
[12] Mannhardt F, Koschmider A, Baracaldo N, Weidlich M, Michael J. Privacy-preserving process mining. Bus Inform Syst Eng 2019;61(5):595–614.
[13] Hundepool A, Domingo-Ferrer J, Franconi L, Giessing S, Nordholt ES, Spicer K, de Wolf P-P. Statistical disclosure control. 1st ed.. Chichester: John Wiley & Sons; 2012.
[14] Domingo-Ferrer J, Mateo-Sanz JM. Practical data-oriented microaggregation for statistical disclosure control. IEEE Trans Knowl Data Eng 2002;14(1):189–201.
[15] Oganian A, Domingo-Ferrer J. On the complexity of optimal microaggregation for statistical disclosure control. J UN Econ Comm Eur 2001;18(4):345–53.
[16] Domingo-Ferrer J, Torra V. Ordinal, continuous and heterogeneous k-anonymity through microaggregation. Data Min Knowl Discov 2005;11(2):195–212.
[17] Byun J-W, Kamra A, Bertino E, Li N. Efficient k-anonymization using clustering techniques. In: Proceedings of the 12th international conference on database systems for advanced applications, Bangkok, Thailand; 2007, p. 188–200.
[18] Lin J-L, Wei M-C. An efficient clustering method for k-anonymization. In: Proceedings of the international workshop privacy and anonymity in information society, Nantes, France; 2008, p. 46–50.
[19] Sweeney L. k-anonymity: A model for protecting privacy. Int J Uncertain Fuzziness Knowl-Based Syst 2002;10(5):557–70.
[20] van der Aalst WMP. Responsible data science: Using event data in a "People Friendly" manner. In: Proceedings of the 18th international conference on enterprise information systems, Rome, Italy; 2016, p. 3–28.

[21] Nuñez von Voigt S, Fahrenkrog-Petersen SA, 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, Grenoble, France; 2020, p. 252–67.
[22] Mannhardt F, Petersen SA, de Oliveira MFD. Privacy challenges for process mining in human-centered industrial environments. In: Proceedings of the 14th international conference on intelligence environments, Rome, Italy; 2018, p. 1–8.
[23] Rozinat A, Günther CW. Privacy, security and ethics in process mining. Fluxicon, tech. rep., 2017, p. 1–10.
[24] Elkoumy G, Fahrenkrog-Petersen SA, Sani MF, Koschmider A, Mannhardt F, Nuñez von Voigt S, Rafiei M, von Waldthausen L. Privacy and confidentiality in process mining – threats and research challenges. 2021, p. 1–17, arXiv preprint arXiv:2106.00388.
[25] 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, Rome, Italy; 2015, p. 58–63.
[26] Tillem G, Erkin Z, Lagendijk RL. Privacy-preserving alpha algorithm for software analysis. In: Proceedings of the international symposium on information theory and signal processing in the benelux, Louvain-la-Neuve, Belgium; 2016, p. 136–43.
[27] Liu C, Duan H, Zeng Q, Zhou M, Lu F, Cheng J. Towards comprehensive support for privacy preservation cross-organization business process mining. IEEE Trans Serv Comput 2016;12(4):1–15.
[28] Rafiei M, Von Waldthausen L, van der Aalst WMP. Ensuring confidentiality in process mining. In: Proceedings of the 8th international symposium on data-driven process discovery and analysis, Seville, Spain; 2018, p. 3–17.
[29] Rafiei M, von Waldthausen L, van der Aalst WMP. Supporting confidentiality in process mining using abstraction and encryption. In: Proceedings of the 8th international symposium on data-driven process discovery and analysis. Seville, Spain: Springer; 2018, p. 101–23.
[30] 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, Rome, Italy; 2019, p. 194–206.
[31] Pika A, T. WM, Budiono S, ter Hofstede AHM, van der Aalst WMP, Reijers HA. Towards privacy-preserving process mining in healthcare. In: Proceedings of the 2nd international workshop on process-oriented data science for healthcare, Vienna, Austria; 2019, p. 1–12.
[32] Pika A, Wynn MT, Budiono S, ter Hofstede AH, van der Aalst WM, Reijers HA. Privacy-preserving process mining in healthcare. Int J Environ Res Public Health 2020;17(5):1612.
[33] Rafiei M, van der Aalst WMP. Privacy-preserving data publishing in process mining. In: Proceedings of the 18th international conference on business process management. Seville, Spain: Springer; 2020, p. 122–38.
[34] Fahrenkrog-Petersen SA. Providing privacy guarantees in process mining. In: Proceedings of the 31st international conference on advanced information systems engineering – doctoral consortium, Rome, Italy; 2019, p. 23–30.
[35] Fahrenkrog-Petersen SA, van der Aa H, Weidlich M. PRETSA: Event log sanitization for privacy-aware process discovery. In: Proceedings of the 1st IEEE international conference of process mining, Aachen, Germany; 2019, p. 1–8.
[36] 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 17th international conference on business process management, Vienna, Austria; 2019, p. 1–5.
[37] Batista E, Solanas A. A uniformization-based approach to preserve individuals' privacy during process mining analyses. Peer-To-Peer Netw Appl 2021;1–20.
[38] Rafiei M, van der Aalst WMP. 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, Vienna, Austria; 2019, p. 1–12.
[39] Rafiei M, Wagner M, van der Aalst WMP. TLKC-privacy model for process mining. In: Proceedings of the 14th international conference on research challenges in information science. Limassol, Cyprus: Springer; 2020, p. 398–416.
[40] Rafiei M, van der Aalst WMP. Group-based privacy preservation techniques for process mining. Data Knowl Eng 2021;134:101908.
[41] Rafiei M, van der Aalst WMP. Practical aspect of privacy-preserving data publishing in process mining. 2020, p. 1–5, arXiv preprint arXiv:2009.11542.
[42] Elkoumy G, Fahrenkrog-Petersen SA, Dumas M, Laud P, Pankova A, Weidlich M. Shareprom: A tool for privacy-preserving inter-organizational process mining. In: Proceedings of the 18th international conference on business process management – phd/demos; 2020, p. 72–6.
[43] Elkoumy G, Fahrenkrog-Petersen SA, Dumas M, Laud P, Pankova A, Weidlich M. Secure multi-party computation for inter-organizational process mining. In: Enterprise, business-process and information systems modeling. Springer; 2020, p. 166–81.
[44] Elkoumy G, Pankova A, Dumas M. Mine me but don't single me out: Differentially private event logs for process mining. 2021, p. 1–16, arXiv preprint arXiv:2103.11739.
[45] Elkoumy G, Pankova A, Dumas M. Privacy-preserving directly-follows graphs: Balancing risk and utility in process mining. 2020, p. 1–35, arXiv preprint arXiv:2012.01119.
[46] Data Protection Focus Group for the Safety, Protection, and Trust Platform for Society and Businesses. White paper on pseudonymization: Guidelines for the legally secure deployment of pseudonymization solutions in compliance with the general data protection regulation. Tech. rep., Ludwigshafen: Digital Summit; 2017.
[47] Spindler G, Schmechel P. Personal data and encryption in the European general data protection regulation. J Intell Prop Inform Technol E-Commer Law 2016;7:163–77.
[48] Weijters AJMM, van der Aalst WMP. Rediscovering workflow models from event-based data using little thumb. Integr Comput-Aided Eng 2003;10(2):151–62.
[49] Weijters AJMM, van der Aalst WMP, Alves de Medeiros AK. Process mining with the heuristics miner algorithm. Tech. rep. WP 166, Technische Universiteit Eindhoven; 2006, p. 1–34.
[50] Agrawal R, Gunopulos D, Leymann F. Mining process models from workflow logs. In: Proceedings of the international conference on extending database technology; 1998, p. 467–83.
[51] Papadimitriou P, Dasdan A, Garcia-Molina H. Web graph similarity for anomaly detection. J Internet Serv Appl 2010;1(1):19–30.
[52] Shoubridge P, Kraetzl M, Wallis W, Bunke H. Detection of abnormal change in a time series of graphs. J Interconnect Netw 2002;3(01n02):85–101.
[53] Koutra D, Vogelstein JT, Faloutsos C. DeltaCon: A principled massive-graph similarity function. In: Proceedings of the SIAM international conference on data mining, Austin, USA; 2013, p. 162–70.
[54] van Dongen BF. BPI challenge 2012, 4TU. Centre for Research Data. Dataset; 2012, http://dx.doi.org/10.4121/uuid:3926db30-f712-4394-aebc-75976070e91f.
[55] Steeman W. BPI challenge 2013, closed problems. Ghent University. Dataset; 2013, http://dx.doi.org/10.4121/uuid:c2c3b154-ab26-4b31-a0e8-8f2350ddac11.
[56] van Dongen BF. BPI challenge 2014: Activity log for incidents, 4TU. Centre for Research Data. Dataset; 2014, http://dx.doi.org/10.4121/uuid:86977bac-f874-49cf-8337-80f26bf5d2ef.
[57] van Dongen BF. BPI challenge 2015, 4TU. Centre for Research Data. Dataset; 2015, http://dx.doi.org/10.4121/uuid:31a308ef-c844-48da-948c-305d167a0ec1.
[58] Buijs JCAM. Receipt phase of an environmental permit application process ('WABO'), CoSeLoG project. Eindhoven University of Technology. Dataset; 2014, http://dx.doi.org/10.4121/uuid:a07386a5-7be3-4367-9535-70bc9e77dbe6.