IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 21 September 2023, accepted 17 November 2023, date of publication 23 November 2023,
date of current version 30 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3336413

RESEARCH ARTICLE

PLI-Assess: A Behavior Profile-Based Approach for Privacy-Preserving Log Assessment

GENG WANG[1] **AND HUAN FANG**[1,2]
[1]College of Mathematics and Big Data, Anhui University of Science and Technology, Huainan 232002, China
[2]Anhui Province Engineering Laboratory of Big Data Analysis and Early Warning Technology of Coal Mine Safety, Huainan 232001, China

Corresponding author: Huan Fang (fanghuan0307@163.com)

This work was supported in part by the National Natural Science Foundation, China, under Grant 61902002.

ABSTRACT Event logs of business processes provide a valuable starting point for process mining, and high-quality event logs can significantly enhance the quality of process mining. However, event logs often contain a substantial amount of sensitive and personal information. Therefore, the release of event logs should prioritize the model's quality while minimizing the risk of privacy exposure. Specifically, quantifying performance indicators between the original event logs and the released ones enables the operational goals. To date, privacy benefit and utility loss are two main target performance indicators, especially from the perspective of structural similarity comparison of mined process models. To the best of our knowledge, no study aims to measure the privacy-preserving performance indicators from the point of behavior differentiation between the original event logs and released ones. In this paper, we propose an approach to quantify the behavior differentiation between the original event logs and the corresponding released ones. Specifically, an approach of event log release mechanism that effectively combines behavior privacy gain and behavior utility loss is presented in this paper. Firstly, we discuss challenges in scenarios where event data is released without privacy preservation, and describe potential attacks that could occur when third-party businesses perform process mining techniques. Based on these potential attacks, we present a behavior differentiation-based event log release mechanism named PLI-Assess to combat these threats. Finally, we conduct experiments on four groups of practical event logs for comparisons with the baseline methods.The experimental results suggest feasibility of privacy-utility trade-offs.

INDEX TERMS Log release mechanism, performance indicators, privacy preserving, privacy protection, responsible process mining.

I. INTRODUCTION

In practical applications, the actual executions of a given business process system serve as the primary data object for analysis and pave the way for process mining. Event logs usually offer essential insights from multiple perspectives, such as process discovery, performance evaluation, and behavior deviation detection, etc. Additionally, log data provides decision-making perspectives for process prediction, monitoring, bottleneck identification, and resource optimization. In principle, process mining (PM) research begins with event logs and is complemented by intelligent analysis approaches, such as data analysis, process modeling, and process analysis, to extract knowledge from event logs. The primary goal of PM is to discover, monitor, and enhance real process performance. Generally, PM techniques can be categorized into three main categories: (a) process discovery, which aims to generate process models from event logs; (b) conformance checking, which identifies discrepancies between logs and models; (c) enhancement, which leverages the historical data stored in event logs to improve existing process models [1].

In general, event log contains information such as activity names, process instance identifiers, activity resources, timestamps, etc. This information can be represented as attributes of event log. However, some attributes within the log may contain private or sensitive data that, if made public without proper processing, could harm individuals' privacy. This is particularly true on the application that need to collect

The associate editor coordinating the review of this manuscript and approving it for publication was Junggab Son.

personal information. For example, in the medical and healthcare applications, there exists a large amount of healthcare data collected to aid early disease identification or prediction. However, this kind of healthcare data is inevitable contain sensitive patient information such as health status, clinical data, and attending physician information. Attackers can use some related background knowledge and chain attacks to gain access to a complete chain of private information [2]. So, when the business analyzers utilize some PM techniques to manage and enhance process model from event logs, it is necessary to pseudonymize or encrypt the privacy data in event logs to protect the privacy of the data. This is the starting point of the PPPM (Privacy Preservation Process Mining) techniques [3].

PPPM techniques aim to enhance privacy protection, prevent personal or sensitive data from being leaked to unauthorized parties, and hence to minimize the risk of reidentification of sensitive data. PPPM study is a new and active interdisciplinary of process mining and privacy protection research. PPPM techniques strengthen the integration of traditional process mining with data privacy preservation. For instance, process mining helps to build a digital twin physical model and realize the privacy protection of the underlying data [4]. Figure 1 describe the process of developing a log releasing strategy through privacy protection evaluation.

Although privacy protection technology can protect private data to some extent, it inevitably causes data distortion in released event logs, which can result in a decline in the mining quality of the process model and impact the subsequent process mining analysis. Therefore, we argue that the main challenge of PPPM techniques is how to improve the quality of the model while minimizing the privacy disclosure risk. The state-of-the-art study is focused on the measurement of performance indicators such as privacy gain, utility loss, and etc, and some structural similarity based measurements are designed to quantify the deviation between the original event logs and the released ones. These current techniques fall short in measuring the performance indicator that combining both privacy gain and utility loss indicators together. Furthermore, to the best of our knowledge, there exists no study aims to measure the deviation between the original event logs and the released ones, from the perspective of process behaviors. To this end, this paper propose an approach of privacy-utility balanced release mechanism for event logs, its greatest innovations lie in the facts that introducing a behavior differentiation [5] based approach of event log release mechanism, i.e., PLI-Assess, which is designed to combat potential third-party attacks, and combining both privacy gain and utility loss performance indicators together. The main contributions of this paper are as the follows on.

(1) Constructing a kind of privacy-utility balanced log release mechanism, comprehensively evaluate the level of log privacy protection. Through motivation example and its analysis, it is indicated that there exists no single privacy protection method that is applicable to all types of the extant event logs. Thus, this paper proposes a privacy-utility balanced log release mechanism, supporting the decision making of stakeholders in publishing companies. In other words, this proposed log release mechanism can provides suitable event log release strategies for different types of event logs.

(2) Proposing an approach of behavior differentiation based privacy preserving log assessment, named PLI-Assess method. It includes a kind of behavior privacy-utility balanced performance indicator in privacy-preserving level. The proposed performance indicator starts from the perspective of behavior differentiation between the original event logs and the released ones. To the best of our knowledge, this paper is the first work to quantify the behavior deviation between the original event logs and the released ones, and utilizes the behavior-based performance indicator considering the trade-off between privacy gain and utility loss.

Conclusively, the biggest innovation of this paper lies in the fact it is the first work to formalize a comprehensive and behavior-based performance indicator, where the performance of privacy gain and utility loss are balanced. It introduces the concepts of behavior privacy gain, behavior utility loss concepts, and containment coefficients, and prepare the ground for quantified the privacy-preserving performance indicator from the perspective of behavior differentiation. By using four groups of applicable event logs, a series of experiments are conducted for in-depth analysis. Based on the experimental results obtained from the proposed PLI-Assess method, this paper investigates the applicability characteristics of mainstream privacy preserving methods, and compares our proposed method with the existing ones, to further elucidate the interpretability and applicability of the proposed PLI-Assess method. Finally, a rational and more comprehensive evaluation technique for privacy preserving issue based on process mining is provided.

The structure of this paper is as follows. Section II explains the need for PPPM and reviews the state-of-the-art related work; Section III clarifies some basic concepts about event logs and behavioral profile; Section IV presents a motivation example, including the semi-honest attack model, and discuss the main challenges in PPPM; Section V proposes the behavior differentiation based privacy preserving log assess method, i.e. PLI-Assess method, including the specific steps of the proposed method; Section VI describes the simulation experiments carried out using four types of real logs, and the results indicate the general feasibility of the proposed method; Section VII concludes the paper and offers insights for future work.

II. RELATED WORK

The growing concern over data privacy worldwide, along with the enforcement of privacy regulations and the widespread adoption of the FACT principle [6] (i.e. Fairness, Accuracy, Confidentiality, and Transparency), highlight the need to integrate privacy protection measures into process mining analyses. In this section, we will provide a brief overview of the research conducted by PPPM in this area.

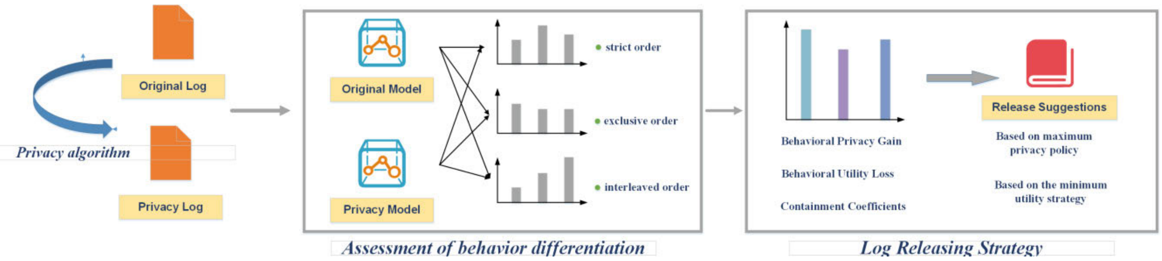


FIGURE 1. Privacy protection assessment and releasing strategy.

In recent years, there has been a growing awareness of privacy issues, particularly with regards to preventing data misuse scandals. This has led to the enactment of legal privacy regulations, such as GDPR in Europe [7], which promote privacy by design principles like data minimization, pseudonymization, encryption, and lawful processing, while prohibiting the release of potentially harmful information between institutions. The risk of privacy breaches has also spurred the development of innovative methods for storing and retrieving data, such as blockchain, as well as searching and sharing encrypted data from cloud infrastructure in various industries. As a result, there has been a gradual increase in scientific interest in privacy protection in a variety of fields.

A. MAINSTREAM PRIVACY PRESERVING METHODS

To address privacy challenges in process mining, various methods and models have been proposed in the literature. These methods can be broadly categorized into three types: encryption, data perturbation, and anonymization. The challenges in privacy protection for process mining are similar to those in pattern mining, with a focus on preserving order and maintaining anonymous trace properties. The concept of RPM (Responsible Process Mining) was introduced as a part of RDS (Responsible Data Science) to address issues related to data misuse analysis and mitigate negative impacts [6]. The privacy challenges in process mining are akin to protecting ordered pattern mining and anonymous trace attributes. In the literature, RPM (Responsible Process Mining) was first proposed as a new challenge of RDS (Responsible Data Science) to address the misuse of data analysis and avoid negative impacts.

Currently, privacy preservation research in the field of process mining mainly focuses on two research areas: intra-organizational and inter-organizational. Intra-organizational privacy preservation research primarily concentrates on protecting data privacy by limiting access to sensitive information. Conversely, inter-organizational privacy preservation research necessitates more complex protection measures to address the issue of privacy leakage when sharing data across different organizations.

Intra-organizational research mainly involves three types of privacy protection methods and their derived algorithm models, including encryption, data perturbation, and anonymization. Rafiee et al. proposed a confidentiality framework and assessed the vulnerabilities, furthermore, they discussed the open difficulties of event log encryption [8], including the incompleteness of encrypting event records and the shortcomings of using a single technique. Tillem et al. proposed an Alpha method employing encryption protocols for process discovery, the proposed method may protect the privacy of users and software simultaneously [9]; Burattin et al. addressed the outsourcing of PM analytics [10], where the confidentiality of event logs and resultant processes must be ensured by concealing sensitive data with symmetric or homomorphic encryption; Liu et al. presented a trusted third party for public and private business process models [11]. To implement privacy regulations on event logs, a PM privacy-preserving system design based on an Attribute-Based Access Control (ABAC)authorization model is proposed in the literature [12]. However, under the assumption that the attacker has a limited background knowledge, these up mentioned approaches still pose potential danger of information disclosure.

In this regard, Mannhardt et al. proposed a privacy protection engine based on (ε, δ) differential privacy [13], which theoretically ensures that personal information cannot be identified regardless of whether the attacker has background knowledge or not; Fahrenkprog-Petersen et al. proposed a privacy-protected event log release Framework (PRIPEL) [14], which followed the principle of localized differential privacy and provided differential privacy guarantees at the case level rather than the entire log; Elkoumy focused on the utility loss caused by differential privacy methods, and proposed an optimized parameter setting method of ε using utility-based estimation [15].

Other studies aim to safeguard event logs using anonymization techniques. Fahrenkrog-Petersen et al. proposed an event log cleaning algorithm named *PRETSA*, which used K-anonymity and T-closure as privacy protection requirements [16]. The development of privacy protection requirements find similar traces and merge them until privacy protection requirements are satisfied; Pika et al. analyzed the data privacy requirements of process models in the healthcare domain, and proposed a theoretical PPPM framework to support PM analysis for healthcare processes, which was based on the anonymization characterizations of healthcare data and privacy metadata [17]. In addition, Rafiei and Alast [18] proposed general privacy quantification frameworks and introduced measures to evaluate the efficacy of privacy-preserving techniques. In latest studies, a kind of TLKC privacy model is formalized through group-based anonymization, in order to prevent attribute linking attacks in process discovery and performance analysis [19], and an integrated wen-based, open-source PM platform is constructed for these techniques are integrated into a web-based, open-source PM application [20].

For inter-organizational process mining, the parties involved in the process are unwilling to share execution data with each other or with third parties. To address this challenge, Tillem et al. proposed the Alpha algorithm, which uses encryption protocols for process discovery and ensures the privacy of both users and software. Liu et al. proposed a trusted third-party scheme for handling public and private business process models. Research [21] demonstrated how secure multiparty computation can be used to construct process models based on cross-organizational processes without data sharing between parties. This problem has also been addressed using specialized hardware to build a trusted execution environment between organizations.

Furthermore, several tools have been developed to support the application of privacy-preserving process mining, such as ELPaaS, Shareprom, and PC4PM.

Table 1 summarizes the privacy-preserving process mining techniques for both intra-organizational and inter-organizational scenarios based on different research perspectives, along with their representative literature, and also compares the advantages and limitations of each method. Available privacy-preserving process mining application tools are listed in Table 2.

We can find that based on different research perspectives, methods focus on different entry points, and the classification of methods also determines their wide range of limitations. It is very challenging to find a unified evaluation method to measure this difference. Similarly, these PPPM tools allow us to freely combine different methods and logs, thus establishing the convenience of the preliminary work.

TABLE 1. Privacy-preserving process mining technology from different perspectives.

Study perspective	Characteristics	Challenges
Multi-organizational	Process integration; Resource sharing	Complexity, Organizational diversity; Collaboration.
Encryption	Guaranteed log data security.	Access vulnerability; brute force susceptibility.
Data perturbation	Secure and compatible with logs.	Limited data; combinatorial attack.
Anonymization	Secure sharing.	Limited data access; Re-identification risk.

B. PERFORMANCE INDICATORS OF PRIVACY PRESERVING STUDY

Performance indicators in privacy preserving study are those quantifiable metrics that allow the evaluation of process privacy protection. There are many kinds of performance indicators for typical PPPM methods regarding to various study focuses. However, these existing performance indicators are not applicable for general analysis purposes, due to the particularity of log data structure and the individuality of analysis target. For example, regarding to privacy gain in PPPM, performance indicators such as information loss, confidence level, support level, and privacy budget are common in use; however, regarding to utility loss, other performance indicators such as accuracy loss, completeness loss, timeliness loss, and usability loss are popular, as shown in Table 3.

TABLE 2. Privacy preserving application tools based on process mining.

Application tools	Open source address	Application range
ELPaaS[22]	https://github.com/samadeusfp/elpaas	Event log sanitization and Privatized process mining
Shareprom[23]	https://github.com/Elkoumy/shareprom/releases/tag/v0.2	Inter-organizational process mining based on Secure Multi-Party Computing
PC4PM[24]	https://github.com/m4jidRafiei/PC4PM	Multi-angle privacy protection technology

Besides, in order to highlight the privacy preserving study target in process mining, which is different from other study perspectives in this field. The deviation identification between the original process models (event logs) and the privacy protected ones is another key issue in PPPM. Only a small amount of research has been conducted on this issue, such as the structural similarity based method using graph theory [25], the multi perspective group-based method [21]. Table 3 gives a summary of performance indicators in privacy preserving study. Apart from the control perspective, resource perspective, time perspective, and other event log attributes, behavior perspective is an important perspective regarding to privacy preserving method. To the best of our knowledge, there exists no work aims to this issue, and this paper is the first one regarding to behavior perspective of privacy preserving.

TABLE 3. Privacy protection evaluation index and its measurement method.

Study perspective	Performance indicators	Measurement Method	Challenges and potential limitations
Privacy Gain	Information loss	information entropy, mean square error	Due to the unique characteristics of log structures, there are limited indicators related to changes in privacy protection objects. Besides, it is necessary to quantify the multi-perspective behavior deviation between the original models(logs) and the protected ones.
	Confidence level	precision, recall	
	Support level	frequency, proportion	
	Privacy budget	privacy parameters	
Utility Loss	Accuracy loss	precision, recall	
	Completeness loss	text similarity, graph similarity	
	Timeliness loss	time measure	
	Usability loss	reproducible, operable	

This is why we hope to propose a method for evaluating the degree of privacy protection that is applicable to different scenarios and is not constrained by the log structure.

Although there exists an increasing trend in topics and studies of PPPM, however, most of the latest PPPM studies is focused on the developing new data release mechanism or methods based on classical privacy algorithms, such as K-anonymity, T-closure, and (ε, δ) differential privacy, and the main measurement targets are tuned as privacy discourse probability, utility gains, and the potential utility losses. However, in order to improve the quality of process mining, there exists little consideration given to the trade-off between privacy protection and utility gains, the potential utility losses. The reasons for this phenomenon mainly lie in that the structure of event logs in process mining is unique, which obviously different from those in classical structural dataset in privacy preservation research.

Addressing to this issue, this paper discusses the potential impact that unique structures of event logs posed on different privacy data release, and furthermore integrates several indicators to form a comprehensive one, such as utility gains, and the potential utility losses, which could be more suitable for the privacy preservation of event log data used in process mining. The proposed PLI-Assess method, which includes evaluation indicators such as behavioral privacy gain α_{P_G}, behavioral utility loss α_{U_L}, privacy containment coefficient $\eta_{\alpha_{P_G}}$ and the utility containment coefficient $\lambda_{\alpha_{U_L}}$, etc. These indicators take both privacy protection and utility gains and losses into account. The weights of the two factors are set rationally to provide the optimal solution for log release scenarios in the real world.

III. PRELIMINARIES

In this section, we introduce some basic concepts and notations about event logs, Petri nets, and behavioral profile similarity metrics briefly, which used as preliminaries for the subsequent sections.

Definition 1 (Event and Event Log [26]): An event is a tuple $e = \{c, a, r, t, attr_{\{1,2,\ldots n\}}\}$ in the business system, where c is the case id, a is the activity associated with the event, r is the resource, who is performing the activity, t is the event timestamp, and $attr_{\{1,2,\ldots n\}}$ is a list of additional attributes values, where $\forall 1 \leq i \leq n : attr_i \in attr$. The cartesian product $\xi = c \times a \times r \times t \times attr_{\{1,2,\ldots n\}}$ is called as the event universe. An event log EL satisfies $EL \subseteq \xi$, where each event can appear only once, i.e., events are uniquely identifiable by their attributes.

The event log EL shown in Table 4 contains unique case ID, event ID, and attribute values such as time stamp, activity, and resource, etc. If the resource attribute in this log segment is sensitive, it is crucial to minimize the risk of disclosure. Therefore, the log holder must responsibly handle the resource value before releasing it to the public. The released log is denoted as EL_P. To ensure the privacy of sensitive attributes in event log, appropriate privacy protection measures should be implemented to prevent unauthorized access and disclosure.

Definition 2 (Petri Net and Net System): [27]A Petri net is a 3-tuple $N = (S, T; F)$,where S is a nonempty finite set of places, T is a nonempty finite set of transitions, $F \subseteq (S \times T) \cup (T \times S)$ is a set of arcs, and $S \cap T = \emptyset$. In a net $N, \forall x \in S \cup T, \dot{x} = y|y \in S \cup T \wedge (y, x) \in F$ and $\dot{x} = y|y \in S \cup T \wedge (x, y) \in F$ are the preset and postset of x, respectively.

TABLE 4. An example of event log *EL*.

Case id	Event id	Attributes				
		Timestamp	Activity	Resource	Cost	...
1	35654423	30-12-2010: 11.02	A	Pete	50	...
	35654424	31-12-2010: 10.06	B	Pete	400	...
	35654425	05-01-2011: 15.12	C	Sue	100	...
	35654426	06-01-2011: 11.18	D	Sue	200	...
	35654427	07-01-2011: 14.24	E	Mike	200	...
2	35654483	30-12-2010: 11.32	D	Sue	50	...
	35654485	30-12-2010: 12.12	C	Sue	100	...
	35654487	30-12-2010: 14.16	F	Pete	400	...
	35654488	05-01-2011: 11.22	E	Mike	200	...
	35654489	08-01-2011: 12.05	G	Mike	200	...
3	35654521	06-01-2011: 15.02	A	Pete	50	...
	35654522	07-01-2011: 12.06	I	Pete	100	...
	35654523	08-01-2011: 14.43	P	Sue	400	...
	35654525	09-01-2011: 12.02	Q	Sue	200	...
	35654526	12-01-2011: 15.44	J	Pete	200	..
...	...	...	...	...	...	...

This notation can be extended to a set of nodes, that is $\forall x \subset S \cup T, \dot{X} = \cup_{x \in X} \dot{x}$ and $\dot{X} = \cup_{x \in X} \dot{x}$.

Definition 3 (Behavior Profiles [24]): Let $\sum = (N, M_0)$ be a net system, $N = (S, T; F)$, and $T' \subseteq T$ be a transition set, the weak order relationship $x \succ y$ between transitions x and $y((x, y) \subseteq \left(T' \times T'\right))$ is defined as follows:

(1) strict order: $x \succ y$ and $y \nsucc x$,denoted as $x \rightarrow y$;

(2) exclusive order: $x \nsucc y$ and $y \nsucc x$,denoted as $x + y$;

(3) interleaved order: $x \succ y$ and $y \succ x$,denoted as $x||y$.

The above relationships form the behavior profile of net N, denoted as $BP = (\rightarrow, +, ||)$,and the set of all transition pairs form the behavior profile set BP_s.

Definition 4 (Similarity Measurement Based On Behavior Profile): Let EL and EL_P be the original event log and the privacy preserved event log, respectively, with their corresponding Petri models denoted as S and S_P. The strict-order-relationship-based similarity represented as $Sim_{\rightarrow}(S, S_P)$, exclusive-order-relationship-based similarity represented as $Sim_{+}(S, S_P)$,and interleaved-order-relationship-based similarity represented as $Sim_{||}(S, S_P)$, respectively [28], where:

$$Sim_{\rightarrow}(S, S_P) = \frac{|S_{\rightarrow} \cap S_{P\rightarrow}|}{|S_{\rightarrow} \cup S_{P\rightarrow}|}$$

$$Sim_{+}(S, S_P) = \frac{|S_{+} \cap S_{P+}|}{|S_{+} \cup S_{P+}|}$$

$$Sim_{||}(S, S_P) = \frac{\left|S_{||} \cap S_{P||}\right|}{\left|S_{||} \cup S_{P||}\right|}$$

The weight coefficients corresponding to the three types of relationships are assigned according the following equations:

$$\omega_{\rightarrow} = \frac{|S_{\rightarrow} \cup S_{P\rightarrow}|}{|S_{\rightarrow} \cup S_{P\rightarrow}| + |S_{+} \cup S_{P+}| + \left|S_{||} \cup S_{P||}\right|}$$

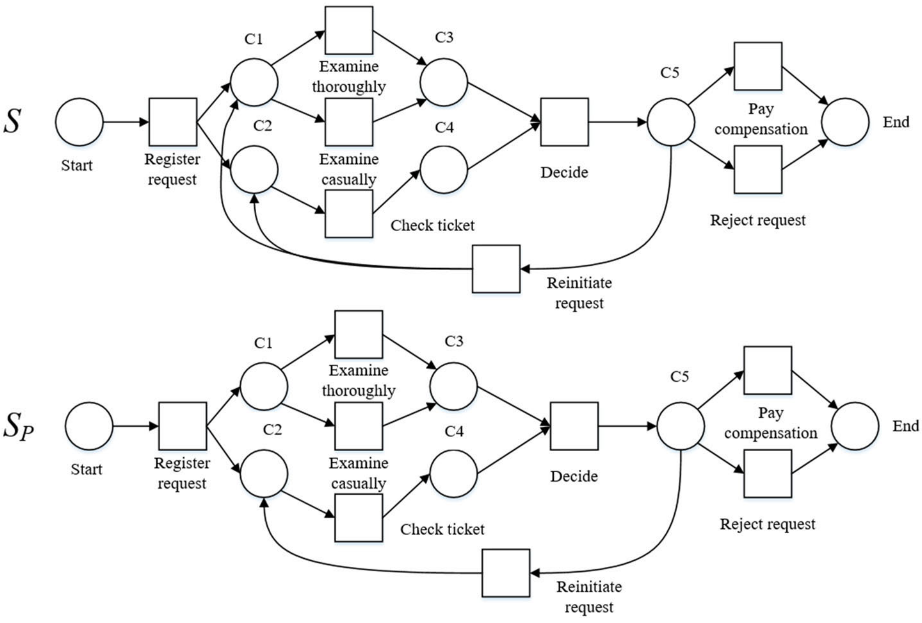


FIGURE 2. Corresponding process models mined from event logs.

$$\omega_{+} = \frac{|S_{+} \cup S_{P+}|}{|S_{\rightarrow} \cup S_{P\rightarrow}| + |S_{+} \cup S_{P+}| + \left|S_{||} \cup S_{P||}\right|}$$

$$\omega_{||} = \frac{\left|S_{||} \cup S_{P||}\right|}{|S_{\rightarrow} \cup S_{P\rightarrow}| + |S_{+} \cup S_{P+}| + \left|S_{||} \cup S_{P||}\right|}$$

The behavior profiled based similarity of models S and S_P is denoted as $Sim\,(S, S_P)$, where $Sim\,(S, S_P) = \omega_{\rightarrow} * Sim_{\rightarrow}\,(S, S_P) + \omega_{+} * Sim_{+}\,(S, S_P) + \omega_{||} * Sim_{||}\,(S, S_P)$.

In order to clarify the calculation procedures, an example of the event log *Example.xes* is taken as an illustration. The event logs and other related dataset have been made available online.

Let EL be the original log, EL_P be the protected log generated using TLKC [21], S and S_P are the process models of the corresponding logs excavated by PROcess MiningFramework(http://www.promtools.org/doku.php?id=prom69), as shown in Figure 2. Table 5 lists all the pairs of strict order relationships between the two models.

In the generated process Petri net models S and S_P, the numbers of pairs of transitions in strict order relationship are not equal, for the reason that compared to the original log EL, the log EL_P is processed by the privacy algorithm. Table 5 gives the evidence that the released protected event log by privacy preservation method probably cannot maintain the same log behavior characteristics as the original one. Generally speaking, in order to resist frequency attacks, the privacy protection algorithm increases the level of privacy protection at the expense of model distortion. we consider the retained strict order relationship pairs as privacy gain and the lost strict order relationship pairs as utility loss. Obviously, utility loss could diminish the utility value for process analysis in some aspects.

TABLE 5. Strict order relationship pairs of models S and S_P.

Event Type	Model Type	Strictly Ordered Pair
EL	***S***	{(examine casually,decide),(decide,pay compensation),(register request,examine thoroughly),(reinitiate request,examine casually),(check ticket,decide),(register request,examine casually),(reinitiate request,examine thoroughly),(decide,reject request),(examine thoroughly,decide),(reinitiate request,check ticket),(register request,check ticket),(decide,reinitiate request)}
EL_P	***S_P***	{(examine casually,decide),(decide,pay compensation),(register request,examine thoroughly),(check ticket,decide),(register request,examine casually),(decide,reject request),(examine thoroughly,decide),(reinitiate request,check ticket),(register request,check ticket),(decide,reinitiate request)}

According to the calculation method outlined in Definition 4, as displayed in Table 2, $|S_{\rightarrow}| = 12, |S_{P\rightarrow}| = 10, Sim_{\rightarrow}\,(S, S_P) = 0.833$, and the similarity of the strict order relationship between S and S_P is 0.833. Similarly, it is possible to calculate the similarities of the exclusive order relationship and interleaved order relationship, as well as the behavioral-profile-based similarity between the models S and S_P.

IV. SEMI-HONEST ATTACKER MODEL AND MOTIVATION EXAMPLE

In order to protect personal information, encryption is a commonly used technique to prevent the linkage between identifiable information and confidential data. recommends this technique due to its high utility in PPPM analysis while

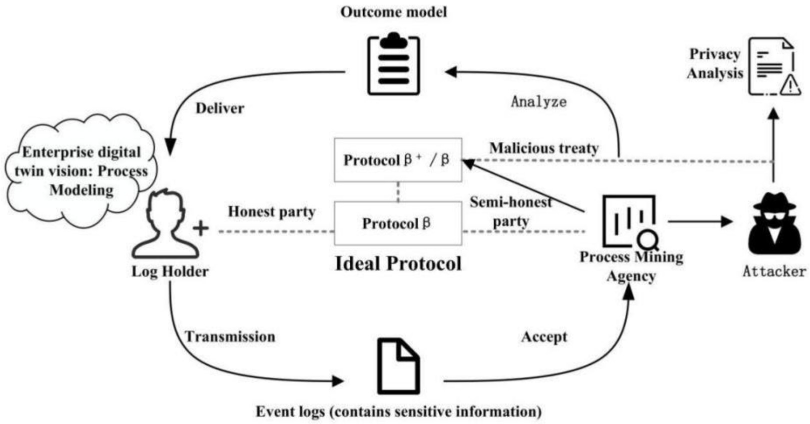


FIGURE 3. Semi-honest attacker model.

maintaining confidentiality. This section focuses on a type of event log attack that can occur in an enter- prise scenario, known as the semi-honest attacker model. We discuss that how commonly used encryption techniques may not be sufficient to completely resist such attacks through simple cases. Additionally, we demonstrate through comparative experiments that the utility index should not be the only factor used to measure the indicators, and the use of impact of privacy gain should also be taken into considered.

A. SEMI-HONEST ATTACKER MODEL

The semi-honest attacker model refers to a scenario where the participants of a multi-party agreement act honestly, but one of the parties may have hidden potential attackers, who seek to obtain additional information based on the input of other parties or the intermediate results of calculations for the purpose of stealing information [29]. In this paper, we define a model consisting of two parties: the log holder, who we refer to as Alice, and a third-party process mining agency, who we refer to as Bob.

Definition 5 (Ideal Protocol [29]): Let f : $(\{0, 1\}^*)^n \rightarrow (\{0, 1\}^*)^n$ be an n-ary function,, and $f(\overline{x}) = (f_1(\overline{x}), f_2(\overline{x}), \ldots f_n(\overline{x}))$, x_i be the private data($x_i \in Alice$) and f be the process mining technology($f \in Bob$). For any $i \in (1, 2 \ldots n)$, β is executed, which is an ideal protocol.

To ensure secure process mining analysis, an ideal protocol assumes that Bob performs analysis on log data $\bar{x}$ provided by Alice, and returns the final result M to Alice, without either party receiving additional input or output information. Both parties are required to be truthful to maintain the security of the protocol. However, if one party becomes dishonest, the protocol β becomes insecure and may be vulnerable to attack.

Definition 6 (Semi-Honest Attacker Model [30]): Let β^+ be a malicious protocol($\beta \in \beta^+$), y be an extra operation function. If $\exists x_i \in\rightarrow y(x_i) = \delta_i$,then $\delta = (\delta_1, \delta_2 \ldots \delta_n)$ is called as the sensitive information disclosure under malicious protocol β^+.

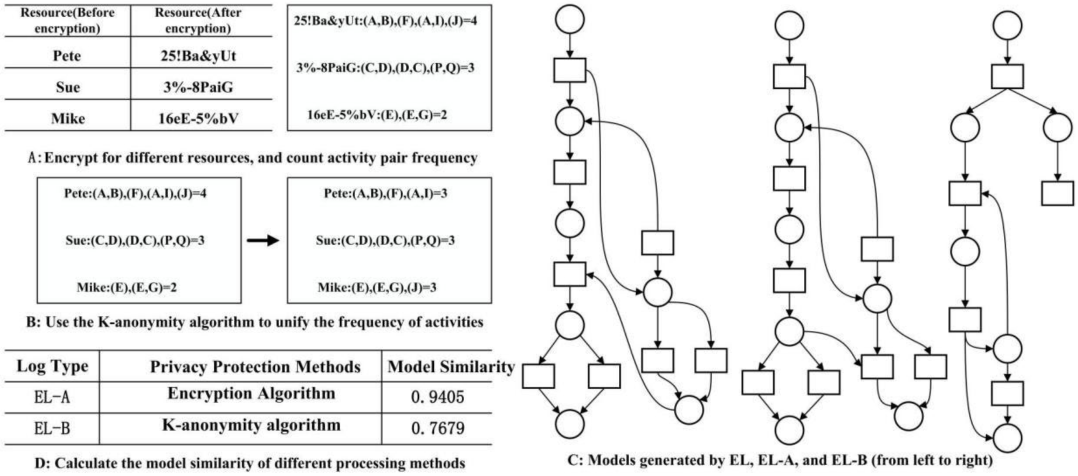
In the semi-honest attacker model, Bob still delivers the result M to Alice honestly, based on the $\bar{x}$ provided by Alice. However, Bob now executes a malicious contract as part of protocol β^+, instead of the original protocol β. While still fulfilling the terms of the original agreement, the semi-honest party adds malicious contracts and other means to the original agreement in order to obtain sensitive information sources. The semi-honest attacker model is depicted as Figure 2.

When an enterprise creates a digital twin and utilizes process mining techniques for physical model construction, the semi-honest attacker model poses a potential threat. As shown in Figure 3, under normal circumstances, the log holder is deemed trustworthy and possesses the legal right to retain the log, but lacks process mining capabilities. Conversely, the third-party process mining organization is considered untrustworthy. If the original log is not protected during transmission for privacy, there is a risk of privacy breach, even if the log holder obtains the desired result model. Hence, effective privacy protection measures are crucial. Generally, processing event logs for privacy preservation target is definitely to reduce the quality of process mining outcomes, with the degree of degradation influenced by the event log and PM analysis objectives. Therefore, the log holder is accountable for assessing the loss and its significance.

Organizations need an assessment technique to quantify their privacy breaches for log releases, depending on the method they use.

B. MOTIVATION EXAMPLE

In some applicable scenario, we assume that event log EL contains certain event attributes that may include sensitive personal information and personally identifiable information. It is evident that the enterprise cannot provide the original unprocessed log EL to a third party that is not trusted. Hence, it becomes crucial to implement appropriate privacy protection measures to break the linkage between personally identifiable information and sensitive personal data, which allows for the release of a protected event log EL_P.



Resource(Before encryption)	Resource(After encryption)
Pete	25!Ba&yUt
Sue	3%-8PaiG
Mike	16eE-5%bV

25!Ba&yUt:(A,B),(F),(A,I),(J)=4

3%-8PaiG:(C,D),(D,C),(P,Q)=3

16eE-5%bV:(E),(E,G)=2

A: Encrypt for different resources, and count activity pair frequency

Pete:(A,B),(F),(A,I),(J)=4

Sue:(C,D),(D,C),(P,Q)=3

Mike:(E),(E,G)=2

→

Pete:(A,B),(F),(A,I)=3

Sue:(C,D),(D,C),(P,Q)=3

Mike:(E),(E,G),(J)=3

B: Use the K-anonymity algorithm to unify the frequency of activities

Log Type	Privacy Protection Methods	Model Similarity
EL-A	Encryption Algorithm	0. 9405
EL-B	K-anonymity algorithm	0. 7679

D: Calculate the model similarity of different processing methods

C: Models generated by EL, EL-A, and EL-B (from left to right)

FIGURE 4. Comparisons of logs and models before and after encryption.

Currently, the commonly used techniques for privacy protection include encryption, anonymization, and data perturbation [31] (represented by differential privacy), as well as some other related algorithm variants. Here, we employed encryption techniques and the k-anonymity algorithm (with privacy level $k = 3$) to process the example log in Table 4 for privacy protection. Two sets of encrypted logs named EL_{P1} and EL_{P2} respectively, as well as their corresponding Petri net models named S_{P1} and S_{P2} respectively can be obtained, as shown in Figure 4. It is can be drawn that encrypting resource attributes induces information distortion, which is aimed to obfuscate the connection between identifiable individuals and confidential information. However, the overall distribution of resource attribute remained unchanged, as shown in Figure 4A. Before encryption, the activity pairs for the three types of resources were *{(A,B),(F),(A,J),(J)}*, *{(C,D),(D,C),(P,Q)}*, and *{(E),(E,J)}*, and the frequency characteristic values after encryption remained unchanged, with values of 4, 3, and 2, respectively. In contrast, the k-anonymity algorithm in EL_{P2} did not pseudonymize the resource attribute values, but instead normalized the frequency of activity pairs, as shown in Figure 4B. The frequency of occurrence of the three types of resources, *{(A,B),(F),(A,I)}*,*{(C,D),(D,C),(P,Q)}*,and *{(E),(E,G),(J)}*, were all adjusted to 3.

Through encryption and the k-anonymity algorithms, event logs EL_{P1} and EL_{P2} can be obtained. EL_{P1} and EL_{P2}, as well as original event log EL are used as input for process mining on platform PROM. The mined Petri net models S_P, S_{P1} and S_{P2} are depicted in Figure 3.C(from left to right). Through Def.4, the similarities between S_{P1}, S_{P2} and S_P can be calculated as $SIM_{(S_P,S_{P1})} = 0.9405$, $SIM_{(S_P,S_{P2})} = 0.7679$. Hence, we conclude that the encryption algorithm was more effective than the k-anonymity algorithm in terms of utility loss measurement. However, the logs EL_{P1} processed by the encryption algorithm can still be re-identified by attackers. For instance, in the semi-honest attacker model mentioned earlier, a third-party mining organization can maliciously modify the protocol, allowing the use of frequency attacks or chain attacks to obtain related feature values of resource attributes. This is because the encryption algorithm only pseudonymizes the target value, and if the activity of the target value appears m times in the original log EL, it will also appear m times in EL_{P1}. Consequently, attackers can exploit the unaltered distribution feature to mine specific information through frequency attacks. After obtaining the plaintext of one resource attribute, chain attacks can be used to deduce the entire original sequence. As a result, EL_{P1} can only resist a limited range of attacks. Conversely, EL_{P2}, published based on the k-anonymity algorithm, can effectively resist the aforementioned attacks, as each resource attribute has the same activity-to-frequency ratio in the log, which avoids re-identification.

Therefore, this motivation example gives the evidence that evaluating privacy protection methods solely based on the model similarity index is inadequate, as it only considers the degree of utility loss. Our investigation reveals that in the state-of-the-art PPPM studies, utility loss is a crucial factor for experimental evaluation, while privacy gain is often neglected. Given the unique characteristics of process mining technology, maintaining high utility requires retaining more analysis opportunities, but this goal should not be achieved at the expense of increased privacy disclosure risks.

In the upcoming section, we will propose a measurement method that is based on behavior profile, which integrating privacy gain and utility loss together as evaluation indicators, and defining the similarity between the original log EL and the protected log EL_P through weight coefficients assignment. Our focus will be on maximizing model quality while minimizing the risk of privacy disclosure.

V. PROPOSED PLI-ASSESSS METHOD

In this section, we introduce the Privacy Level Index (*PLI* for short), a novel evaluation index for the level of log protection measurements. Through *PLI* analysis, we redefine the level of log protection in terms of both privacy gain and loss of utility, resulting in a more comprehensive evaluation approach than previous methods. We first define the relevant parameters and algorithmic steps, then illustrate the application of *PLI* through a practical event log instance.

Definition 7 (Behavioral Privacy Gain and Behavioral Utility Loss): Let EL and EL_P be the original event log and the protected event log, with their corresponding Petri models named S and S_P, respectively. $\alpha_{P_G} = |S_{BP_S} \cap S_{P-BP_S}|/|S_{BP_S} \cup S_{P-BP_S}|$ is represented as the behavioral privacy gain($0 \leq \alpha_{P_G} \leq 1$), and $\alpha_{U_L} = 1 - \alpha_{P_G}$ is represented as the behavioral utility loss($0 \leq \alpha_{U_L} \leq 1$), where S_{BP_S} and S_{P-BP_S} represent the number of behavior profile sets contained in S and S_P respectively.

Behavioral privacy gain pertains to the quantity of retained behavioral profile relationship sets post-oversensitive data processing. Conversely, behavioral utility loss quantifies the number of forfeited behavioral profile relationship sets attributed to privacy protection measures. In the field of PPPM, the magnitude relationship between privacy gain and utility loss requires special attention. For instance, protection methods with high levels of privacy typically lead to greater privacy gains but also greater utility losses. Thus, comparisons based on a single criterion are inadequate. It is accurate to conclude that as privacy gains increase, utility losses also increase. In Definition 8, we introduce the containment coefficient and assign weight coefficients to the two variables for enable a comprehensive evaluation.

Definition 8 (Containment Coefficients): Let α_{P_G} be the behavioral privacy gain ($0 \leq \alpha_{P_G} \leq 1$), and $\alpha_{U_L} = 1 - \alpha_{P_G}$ be the behavioral utility loss ($0 \leq \alpha_{U_L} \leq 1$). The privacy containment coefficient is denoted as $\eta_{\alpha_{P_G}}$, and the utility containment coefficient is denoted as $\lambda_{\alpha_{U_L}}$, where $\eta_{\alpha_{P_G}} = \alpha_{U_L}/(\alpha_{P_G} + \alpha_{U_L})$, $\lambda_{\alpha_{U_L}} = \alpha_{P_G}/(\alpha_{P_G} + \alpha_{U_L})$, with $\eta_{\alpha_{P_G}} + \lambda_{\alpha_{U_L}} = 1$ and $\eta_{\alpha_{P_G}}, \lambda_{\alpha_{U_L}} \in [0, 1]$.

The relationship between utility loss and privacy gain is analyzed, and the game relationship is quantitatively studied, which can be regarded as a trade-off between the two concepts of utility loss and privacy gain. Obviously, it is impossible to achieve a common optimization trend due to this trade-off relationship. For example, in the traditional k-anonymity privacy-preserving algorithm, as the value of k increases, the privacy level also increases, in the meanwhile, however, the degree of data distortion also increases.

Definition 9 (Privacy Level Index): Let α_{P_G} be the behavioral privacy gain($0 \leq \alpha_{P_G} \leq 1$), $\alpha_{U_L} = 1 - \alpha_{P_G}$ be the behavioral utility loss($0 \leq \alpha_{U_L} \leq 1$), the privacy containment coefficient be $\eta_{\alpha_{P_G}}$,and the utility containment coefficient be $\lambda_{\alpha_{U_L}}$. The privacy level indicators *PLI* is calculated as $PLI = 2 * (\eta_{\alpha_{P_G}} * \alpha_{P_G} + \lambda_{\alpha_{U_L}} * \alpha_{U_L})$ ($PLI \in [0, 1]$).

The closer the *PLI* is to 1, the higher the level of privacy protection and the lower the utility loss, while a score of *PLI* closer to 0 indicates lower privacy protection and greater utility loss. Unlike other methods that focus on a single goal of privacy gain or utility loss, the proposed indicator *PLI* takes a balanced approach to evaluate the overall performance of privacy-preserving methods.

Algorithm 1 PLI-Assess Algorithm for Releasing Event Log

```
    Input: Original log $EL$;$M_{set}$ (set of privacy preserving
      methods)
    Output: Privacy protected log $EL_P$.
 1. Adopting privacy protection algorithm $M_i$ to process $EL$,
      and generate $EL_P$;
 2. Mining models $S$ and $S_P$ corresponds to logs $EL$ and $EL_P$
      respectively;
 3. Establishing $M_1$ as the initial indicator method criterion;
 4. for $M_i \in M_{set}$ $(i \geq 2)$:
 5.   Generating $S_{BP_S}$ and $S_{P\_BP_S}$;
 6.   Calculating $Sim(S, S_P)$ by $S_{BP_S}$ and $S_{P\_BP_S}$;
 7.   Calculating privacy gain $\alpha_{P_G}$ and utility loss $\alpha_{U_L}$;
 8.   Assigning containment scores $\eta_{\alpha_{P_G}}$ and $\lambda_{\alpha_{U_L}}$;
 9.   Calculating the comprehensive index score $PLI_{M_{set}}$;
10.   if $PLI_{M_i} \leq PLI_{M_1}$:
11.       $i = i + 1$;
12.       Go to line 3;
13. else
14.   $PLI_{M_1} = PLI_{M_i}$;
15.   $i = i + 1$;
16.   Go to line 3;
17. end
18. Selecting the highest $PLI_{M_i}$ and use the privacy protection
      algorithm $M_i$ as benchmark method.
19. Releasing the protected event log $EL_P$;
20. The end.
```

Algorithm 1 demonstrates the event log privacy level evaluation procedures and event log publishing steps based on the preceding definitions.

As shown in Algorithm 1, based on the event log EL that needs to be made public, we first adopt a privacy protection method to generate EL_P, then discover the Petri net models corresponding to EL and EL_P, named S and S_P for simplicity. We arbitrarily select a kind of privacy protection method as the benchmark method, and calculate the privacy gain and utility loss for all privacy protection methods, as well as simultaneously assign weight coefficients, as outlined in steps 1-8 in the algorithm 1. Referring to values of calculated *PLI*, we further compare the privacy gain and utility loss. If the *PLI* value is greater than the value of selected benchmark method, then we set the method with higher *PLI* value as the new benchmark method, and repeat the preceding steps 3-16 until all methods are validated; If it is lower, we discard the method and move on to the subsequent screening phase. Finally, we select the method with the highest *PLI* score from a given methods set as a baseline method for event log releasing, and the released log EL_P is served as the foundation for open source use.

In order to demonstrate the procedures of the proposed PLI-Assess algorithm, a simple example is taken to case study. Example.xes introduced in section IV is used as a selected sample event log. From Figure 3.D, we can observe that the similarities between S_{P1}, S_{P2} and S_P can be calculated as $SIM_{(S_P,S_{P1})} = 0.9405$, $SIM_{(S_P,S_{P2})} = 0.7679$.The effect of the encryption algorithm is improved, but it cannot withstand the specific analysis in the aforementioned attacker model.

The experimental results show that the event log through k-anonymity processing has a higher comprehensive score, indicating a better balance between privacy gain and utility loss. This characteristic enhances its resistance against diverse attack types. Additionally, its final score is comparable to that of the general model algorithm, further illustrating its effectiveness., making it a more reasonable choice for log protection.

VI. EXPERIMENTAL EVALUATION

In this section, we present the evaluation of privacy level indicators using four sets of practical event logs. The selected primary methods for protecting privacy are pseudonymization, anonymization, and data distortion. To investigate the varying effects on the original log from the perspective of behavioral profiles, we select the state of the art publishing algorithms in PPPM studies, named Hybrid Encryption [10], PRIPEL [14], and TLKC [19] respectively. Our target is to address the following two issues and provide practical recommendations to businesses log holders, who have event log release requirements:

Q1: What is the balance of privacy utility between original event log EL and released log EL_P?

Q2: How are the selection criteria for current privacy protection methods defined?

In section VI-A, we provide an in-depth introduction to the experimental preparations, which includes an overview of the algorithm and the dataset preparation; In section VI-B, we present and visualize the results of 12 groups of different experiments, providing a comprehensive analysis of the effectiveness of various privacy protection methods. In section VI-C, we give an industrial application of underground locomotive dispatching system. Finally, in section VI-D, we thoroughly analyze and discuss the findings of our experiments, and provide reasoned recommendations for businesses log holders to event log releasing.

The data sets used in this question are all open source(https://github.com/lengyilan/PLI-data-set.git).

A. EXPERIMENT PREPARATION

Hybrid Encryption [10] is the state-of-the-art encryption scheme that combines AES and Paillier cryptosystems to enable process mining outsourcing while ensuring dataset and process privacy. PRIPEL [14], on the other hand, adheres to the principle of localized differential privacy, which guarantees privacy at the case level rather than the entire log; TLKC [19] adopts a group-based anonymization method and a greedy algorithm to prevent attribute linking attacks during process discovery and performance analysis. In our experiments, we set T = hours, $L = 2$, $K = 10$, $C = 0.5$, $\theta = 0.7$ for better results based on previous work [10].

Four kinds of real-world datasets are used for our experimental evaluation, which are BPIC_2014 [22], BPIC_2015 [23], Sepsis Cases [24], and Road Traffic Fines [33]. BPIC_2014 contains detailed record information of Rabobank Group ICT extracted from the HP Service Manager ITIL service management tool. Some sensitive attributes have been anonymized in the initial dataset; BPIC_2015 is provided by five cities in the Netherlands and contains information regarding the primary application and opposition procedure at various stages. In this paper, we focus on Municipality 1 for our research; The sepsis log (SC) is a hospital event log documenting suspected cases of life-threatening sepsis, with 846 unique trace variants out of a total of 1050 traces; The road traffic fine log (RTF) is the event log obtained by Italian local police while enforcing road traffic laws. In a sample of approximately 150,000 traces, only 231 different trace variants are observed. Some detailed datasets information is listed in Table 7.

TABLE 6. Similarity scores calculated using two distinct algorithms.

Log Type	Model Type	Model-Similarity Method (Graph Similarity Metric VR and VEO[32])	PLI-Assess Method
EL-A	S_{p1}	$SIM_{(S_P,S_{P1})} = 0.9405$	$SIM_{(S_P,S_{P1})} = 0.2238$
EL-B	S_{p2}	$SIM_{(S_P,S_{P2})} = 0.7679$	$SIM_{(S_P,S_{P2})} = 0.7129$

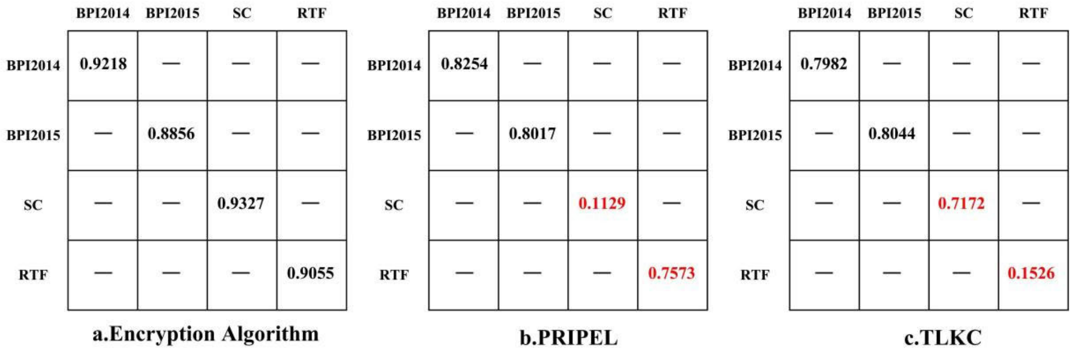
TABLE 7. Event log properties utilized in the calculation of privacy level index.

Event log	Availability	Number of activities	Number of events	Number of traces	Resource quantity
BPIC_2014[22]	Public	39	466,155	46507	242
BPIC_2015[23]	Public	356	262,628	5649	72
SC[24]	Public	16	15,012	1050	39
RTF[33]	Public	11	561,470	150,370	62

B. EXPERIMENTAL RESULTS

As described in Section IV, the first step in completing PLI-Assess is to calculate the values of α_{P_G} and α_{U_L} between the model corresponding to EL and the model corresponding to EL_P, as determined by PLI-Assess algorithm. From the behavior-based perspective, the behavioral privacy gain α_{P_G} and the behavioral utility loss α_{U_L} of their respective models can be obtained. Figures 5 and 6 depict the α_{P_G} and α_{U_L} values obtained utilizing three privacy algorithms based on the selected datasets. In order to control the effect of the experimental procedure on each set of data and to minimize the variance of the subsequent calculation results, we employed the following measures: the experimental algorithm parameters of data sets are set to the same value.

Table 8 presents the Privacy Level Index (PLI) obtained from various event logs using three different groups of privacy protection algorithms. The highest PLI scores are highlighted in red font, while the lowest scores are highlighted in green font. Based on the Hybrid Encryption algorithm, BPIC_2015 obtains the highest score of 0.4054, while the lowest score of 0.2511 is observed in the SC dataset; based on the PRIPEL algorithm, the highest and lowest scores are observed in RTF and SC datasets, with scores of 0.7352 and 0.4006, respectively. In contrast, the TLKC algorithm results in the highest and lowest scores for



a.Encryption Algorithm

	BPI2014	BPI2015	SC	RTF
BPI2014	0.9218	—	—	—
BPI2015	—	0.8856	—	—
SC	—	—	0.9327	—
RTF	—	—	—	0.9055

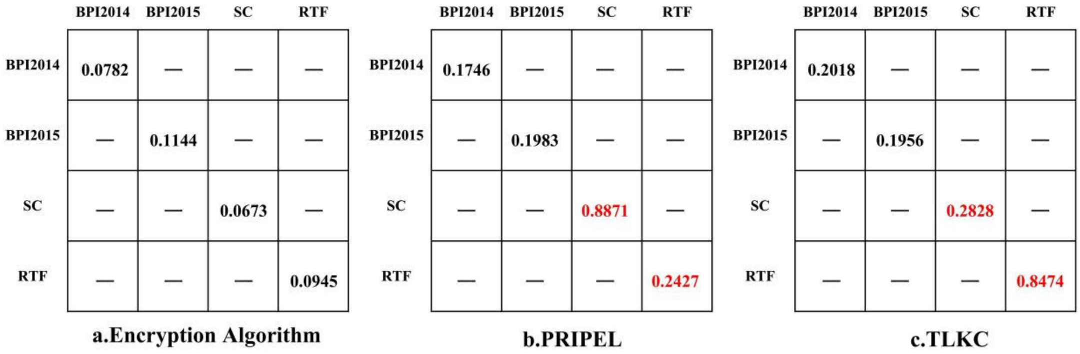
b.PRIPEL

	BPI2014	BPI2015	SC	RTF
BPI2014	0.8254	—	—	—
BPI2015	—	0.8017	—	—
SC	—	—	0.1129	—
RTF	—	—	—	0.7573

c.TLKC

	BPI2014	BPI2015	SC	RTF
BPI2014	0.7982	—	—	—
BPI2015	—	0.8044	—	—
SC	—	—	0.7172	—
RTF	—	—	—	0.1526

FIGURE 5. α_{P_G} under different privacy algorithms.



a.Encryption Algorithm

	BPI2014	BPI2015	SC	RTF
BPI2014	0.0782	—	—	—
BPI2015	—	0.1144	—	—
SC	—	—	0.0673	—
RTF	—	—	—	0.0945

b.PRIPEL

	BPI2014	BPI2015	SC	RTF
BPI2014	0.1746	—	—	—
BPI2015	—	0.1983	—	—
SC	—	—	0.8871	—
RTF	—	—	—	0.2427

c.TLKC

	BPI2014	BPI2015	SC	RTF
BPI2014	0.2018	—	—	—
BPI2015	—	0.1956	—	—
SC	—	—	0.2828	—
RTF	—	—	—	0.8474

FIGURE 6. α_{U_L} under different privacy algorithms.

TABLE 8. Calculate privacy level index value.

Algorithm Classification	Event log	α_{P_G}	α_{U_L}	C_{score}
Hybrid Encryption[12]	BPIC_2014	0.9218	0.0782	0.2883
	BPIC_2015	0.8856	0.1144	0.4054
	SC	0.9327	0.0673	0.2511
	RTF	0.9055	0.0945	0.3423
PRIPEL Framework[16]	BPIC_2014	0.8254	0.1746	0.5765
	BPIC_2015	0.8017	0.1983	0.6359
	SC	0.1129	0.8871	0.4006
	RTF	0.7573	0.2427	0.7352
TLKC Model[21]	BPIC_2014	0.7982	0.2018	0.6443
	BPIC_2015	0.8044	0.1956	0.6294
	SC	0.7172	0.2828	0.8113
	RTF	0.1562	0.8474	0.5295

TABLE 9. Statistical validation results.

Comparison method	Data set	p-value
[Hybrid Encryption,PRIPEL]	[BPIC_2014,BPIC-2015,SC,RTF]	0.031(<0.05)
[Hybrid Encryption,TLKC]		0.037(<0.05)
[PRIPEL,PRIPEL]		0.078(>0.05)

the SC and RTF datasets, respectively, which is opposite to the findings of the PRIPEL algorithm.

Furthermore, Table 9 indicates that there is marked significant difference by statistical pairwise comparisons for selected methods, as the significance probability expressed as p-values don't arrived the significance level, i.e. 0.05.

In the field of PPPM, current research tends to prioritize privacy gain maximization over analyzing utility loss, leading to difficulties in utility analysis during process mining and restricting its application. Compared to the mainstream performance analysis (such as precision and accuracy) [25] and privacy-gain indicators evaluation [21], the pro- posed PLI_Assess method has three significant advantages listed as follows on:

(1) Convenience: Current research in the field of PPPM tends to prioritize privacy gain maximization, neglecting the analysis of utility loss. These approaches limit the application of process mining due to the difficulty in analyzing utility loss. The proposed PLI_Assess method provides an advantageous evaluation system that operates directly on the process model obtained from the log and can calculate the final index score with minimal effort. Unlike other performance metrics, PLI_Assess focuses on the objective of the process utility analysis model, reflecting the final utility loss rate.

(2) Versatility: The proposed PLI_Assess method is universal and can evaluate process models generated from different types of event logs. For instance, the privacy-utility evaluation system in the TLKC method can be applied to a large number of infrequent trace logs, but it may be less efficient on the remaining structured logs.

(3) Interpretability: The *PLI* value depends on the type of event log and the privacy protection algorithm used. Different process models are derived from distinct event logs using various privacy protection algorithms. The proposed PLI_Assess method provides comprehensive reflection of the applicable algorithm type and the modification of privacy utility.

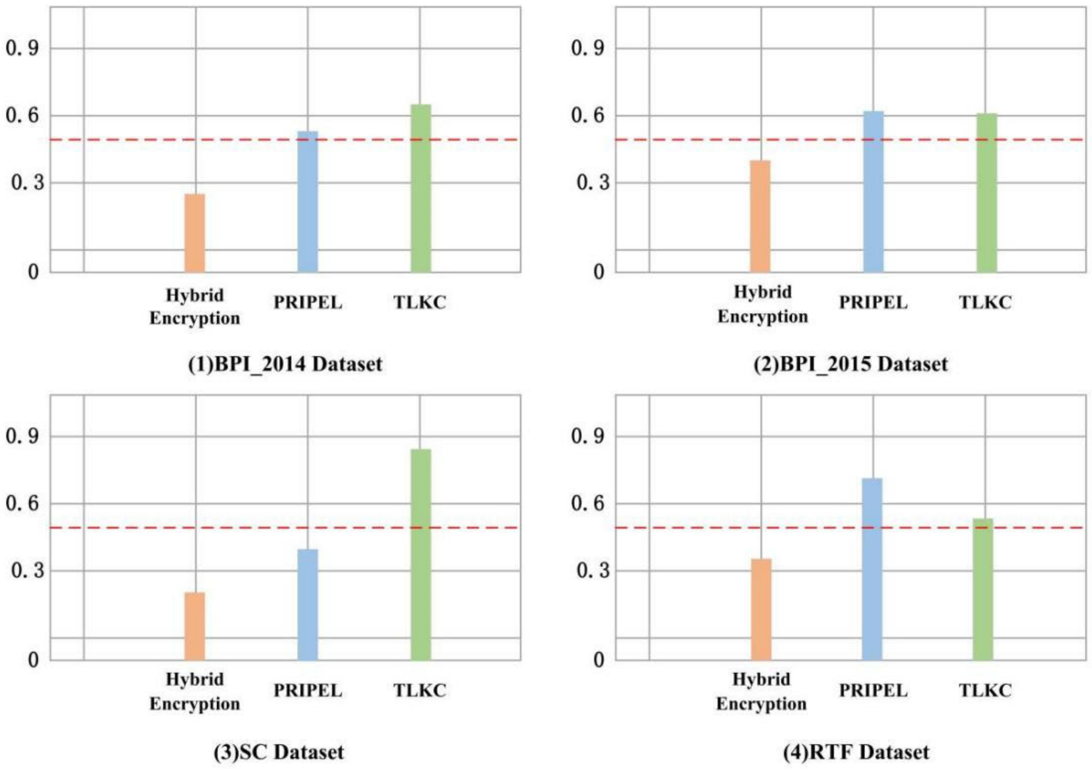


FIGURE 7. Comparison of PLI under various privacy algorithms.

C. INDUSTRIAL APPLICATION

In this section, the proposed PLI_Assess method is adopted to an industrial application of underground locomotive dispatching system in intelligent coal mine, providing recommendation for safe-critical data publication.

As an important part of the intelligent coal mine, locomotive dispatching plays an important role, such as the functions of grasping the real-time status, tracing locomotive operations. However, underground locomotive operation data and road section information are sensitive information, and unprotected release may lead to the problem of exploiting safe-critical data. Therefore, it is necessary to protect the privacy of mine locomotive running data and section log information.

The experiment in this section takes the real event log of the underground locomotive dispatching of a coal mine in Huainan as the research object. The underground locomotive route contains 15 road sections and III mining points. The sensitive information parameters are set as the road section number, mine point number and locomotive operating parameters, using the above three methods to process the original event log *Locomotive dispatch.xes*, and the obtained process model diagram is shown in Figure 8.

For the above process model, PLI_Assess is used to calculate it. From Table 10, it can be found that the event log of locomotive dispatching is most suitable for the privacy protection algorithm of data disturbance to publish it privately.

TABLE 10. Locomotive dispatch event log score data.

Algorithm Classification	α_{P_G}	α_{U_L}	C_{score}
Hybrid Encryption [10]	0.9752	0.0248	0.0967
PRIPEL Framework [14]	0.7836	0.2164	0.6783
TLKC Modell [19]	0.1817	0.8183	0.1487

D. RECOMMENDATION AND EVALUATION OF EVENT LOG RELEASE MECHANISM

In order to answer Q1, we can evaluate the changes in the privacy level index value of *PLI*. After the privacy protection algorithm processing, the original log *EL* differs from EL_P in varying degrees. This distinction is reflected in the Petri net models' behavioral privacy enhancement. In terms of privacy gain and utility loss, the direct consequence of this distinction is that some process mining and analysis techniques are lost their quality, such as fitness, accuracy in process discovery and process enhancement. Consequently, it is crucial to minimize this disparity by considering the privacy-utility balanced relationship. Instead of applying a privacy algorithm to all categories of event logs, we should choose the most proper method with a high *PLI* value, aimed to complete our log protection release under identical technical conditions.

In order to answer Q2, first and foremost, we need to categorize the event logs. In general, event logs can be classified into structured and unstructured categories. Among the four groups of practical event logs selected for this experiment, for example, the SC dataset contains a large number of infrequent traces, indicating the presence of unstructured event logs; on

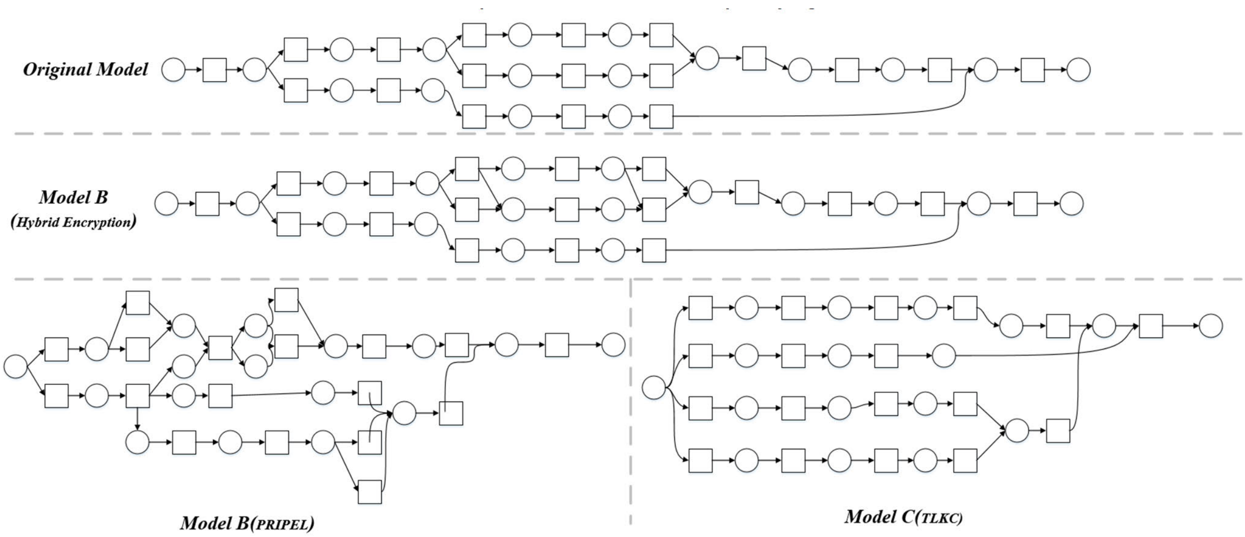


FIGURE 8. Comparison of locomotive scheduling process models.

the other hand, the RTF dataset has a large number of traces, with only 69 unique trace variants, which is a typical kind of structured event logs. The BPIC_2014 and BPIC_2015 datasets are not as obvious in this categorization, so we consider them to be intermediate.

Based on the nature of the three types of privacy protection algorithms, it is indicated that the TLKC model is designed to address the high-dimensional sparsity of event logs and uses a greedy algorithm to suppress events to obtain anonymous logs, which is more effective at processing unstructured event logs, such as those generated by social media platforms. Specifically, on one hand, as for the SC dataset, there are various medical methods and medical means for different patients, making it unsuitable for typical structured event logs like the RTF dataset. However, PRIPEL privacy protection requires the introduction of a large amount of noise, which leads to the accumulation of errors and a higher error rate. As a result, the differential privacy algorithm is not suitable for privacy protection of unstructured event logs, but performs well on structured event logs, such as those generated by a database. On the other hand, as for the RTF dataset, there is usually a uniform and standard processing method for vehicles and pedestrians committing traffic violations, or perform equally well on the locomotive scheduling dataset. Most pseudonymization algorithms offer privacy protection through natural language, maintaining the event log's internal structure unchanged. This processing method retains greater utility at the expense of an increased risk of privacy leakage. In practice, it can be compared with other privacy algorithms before making decision.

Therefore, for example, in an enterprises with a digital twin vision, different types of privacy protection algorithms can be selected for log protection and release based on their internal process structure and event log type. In addition, when evaluating the approximate results of the privacy magnitude index, multiple factors such as calculation cost and release method should be taken into account for a comprehensive analysis.

VII. CONCLUSION

In this paper, we present a semi-honest attacker model and illustrate the potential risk of privacy disclosure if an enterprise releases its original event log without protection during the public data stage. To address this issue, we formalize a privacy level indicator named *PLI*, which combines privacy gain and utility loss together. More specifically, we propose the PLI-Assess method along with its detailed calculation steps. Our approach investigates the balance between privacy gain and utility loss, from the perspective of event log releasing based on behavior profiles of models. The biggest innovation lies in that the proposed PLI-Assess method plays a game balance between privacy gain and utility loss, and provides an in-depth event log release mechanism recommendation through comprehensive experimental analysis. compare different privacy.

The log holders can make a choice of log privacy publishing techniques based on the log type under specific scenarios. Specifically, the conducted experimental results demonstrated that the anonymization algorithm is more suitable for unstructured event logs, whereas the data perturbation algorithm is more suitable for structured event logs, and the pseudonymization algorithm requires exhaustive consideration of the remaining costs.

Unlike previous similarity evaluation methods, our method considers the relative importance of privacy gain and utility loss, making it more reasonable and practical.

In the future work, we aim to introduce a trusted federated learning model and leverage concepts from block-chain algorithms, and target to construct a multi-party secure platform model, with the objective of investigating the delicate adaptive balance between model performance, privacy protection, and utility loss, etc. Furthermore, different security levels from various perspectives should be considered in order to design a more comprehensive privacy preservation framework.

REFERENCES

[1] W. M. van der Aalst, *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Heidelberg, Germany: Springer, 2011.

[2] S. S. Al-Nawafah, H. M. Al-Shorman, F. L. Y. Aityassine, F. A. Khrisat, M. F. A. Hunitie, A. Mohammad, and S. I. S. Al-Hawary, ‘‘The effect of supply chain management through social media on competitiveness of the private hospitals in Jordan,’’ *Uncertain Supply Chain Manag.*, vol. 10, no. 3, pp. 737–746, 2022.

[3] Y. Liu, L. Yang, A. Ghasemkhani, H. Livani, V. A. Centeno, P.-Y. Chen, and J. Zhang, ‘‘Robust event classification using imperfect real-world PMU data,’’ *IEEE Internet Things J.*, vol. 10, no. 9, pp. 7429–7438, May 2023.

[4] V. Kapoor, P. Poncelet, F. Trousset, and M. Teisseire, ‘‘Privacy preserving sequential pattern mining in distributed databases,’’ in *Proc. 15th ACM Int. Conf. Inf. Knowl. Manag.*, 2006, pp. 758–767.

[5] J. Guo and H. Fang, ‘‘Behavior differentiation of process variants with invisible tasks,’’ *IEEE Access*, vol. 11, pp. 64815–64830, 2023.

[6] A. Olteanu, J. Garcia-Gathright, M. de Rijke, M. D. Ekstrand, A. Roegiest, A. Lipani, A. Beutel, A. Olteanu, A. Lucic, A. A. Stoica, and A. Das, ‘‘FACTS-IR: Fairness, accountability, confidentiality, transparency, and safety in information retrieval,’’ *ACM SIGIR Forum*, vol. 53, no. 2, pp. 20–43, 2021.

[7] General Data Protection Regulation. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council*. [Online]. Available: https://dvbi.ru/Portals/0/DOCUMENTS_SHARE/RISK_MANAGEMENT/EBA/GDPR_eng_rus.pdf

[8] M. Rafiei, L. V. Waldthausen, and W. M. van der Aalst, ‘‘Ensuring confidentiality in process mining,’’ in *Proc. SIMPDA*, 2018.

[9] G. Tillem, Z. Erkin, and R. L. Lagendijk, ‘‘Privacy-preserving alpha algorithm for software analysis,’’ in *Proc. 37th WIC Symp. Inf. Theory Benelux/6th WIC/IEEE SP Symp. Inf. Theory Signal Process.*, 2016, pp. 136–143.

[10] A. Burattin, M. Conti, and D. Turato, ‘‘Toward an anonymous process mining,’’ in *Proc. 3rd Int. Conf. Future Internet Things Cloud*, Aug. 2015, pp. 58–63.

[11] C. Liu, H. Duan, Q. Zeng, M. Zhou, F. Lu, and J. Cheng, ‘‘Towards comprehensive support for privacy preservation cross-organization business process mining,’’ *IEEE Trans. Services Comput.*, vol. 12, no. 4, pp. 639–653, Jul. 2019.

[12] J. Michael, A. Koschmider, F. Mannhardt, N. Baracaldo, and B. Rumpe, ‘‘User-centered and privacy-driven process mining system design for IoT,’’ in *Information Systems Engineering in Responsible Information Systems*. Rome, Italy: Springer, 2019, pp. 194–206.

[13] F. Mannhardt, A. Koschmider, N. Baracaldo, M. Weidlich, and J. Michael, ‘‘Privacy-preserving process mining: Differential privacy for event logs,’’ *Bus. Inf. Syst. Eng.*, vol. 61, pp. 595–614, Oct. 2019.

[14] S. A. Fahrenkrog-Petersen, H. van der Aa, and M. Weidlich, ‘‘PRIPEL: Privacy-preserving event log publishing including contextual information,’’ in *Business Process Management*. Berlin, Germany: Springer, 2020, pp. 111–128.

[15] G. Elkoumy, A. Pankova, and M. Dumas, ‘‘Privacy-preserving directly-follows graphs: Balancing risk and utility in process mining,’’ 2020, *arXiv:2012.01119*.

[16] S. A. Fahrenkrog-Petersen, H. van der Aa, and M. Weidlich, ‘‘PRETSA: Event log sanitization for privacy-aware process discovery,’’ in *Proc. Int. Conf. Process Mining (ICPM)*, Jun. 2019, pp. 1–8.

[17] A. Pika, M. T. Wynn, S. Budiono, A. H. Ter Hofstede, W. M. van der Aalst, and H. A. Reijers, ‘‘Towards privacy-preserving process mining in healthcare,’’ in *Business Process Management Workshops*. Berlin, Germany: Springer, 2019, pp. 483–495.

[18] M. Rafiei and W. M. van der Aalst, ‘‘Towards quantifying privacy in process mining,’’ in *Process Mining Workshops*. Padua, Italy: Springer, 2021, pp. 385–397.

[19] M. Rafiei, M. Wagner, and W. M. van der Aalst, ‘‘TLKC-privacy model for process mining,’’ in *Proc. Int. Conf. Res. Challenges Inf. Sci.* Cham, Switzerland: Springer, 2020, pp. 398–416.

[20] M. Rafiei and W. M. P. van der Aalst, ‘‘Practical aspect of privacy-preserving data publishing in process mining,’’ 2020, *arXiv:2009.11542*.

[21] M. Rafiei and W. M. P. van der Aalst, ‘‘Group-based privacy preservation techniques for process mining,’’ *Data Knowl. Eng.*, vol. 134, Jul. 2021, Art. no. 101908.

[22] B. C. P. Brandao, G. N. Lopes, and P. H. P. Richetti. (2011). *BPIC 2014: Insights From the Analysis of Rabobank Service Desk Processes*. [Online]. Available: https://www.win.tue.nl/bpi/2014/bpic2014_submission_10.pdf

[23] I. Teinemaa, A. Leontjeva, and K.-O. Masing. (2015). *BPIC 2015: Diagnostics of Building Permit Application Process in Dutch Municipalities*. [Online]. Available: https://www.researchgate.net/profile/Anna-Leontjeva/publication/315685781_BPIC_2015_Diagnostics_of_Building_Permit_Application_Process_in_Dutch_Municipalities/links/58db738b45851578dff817e5/BPIC-2015-Diagnostics-of-Building-Permit-Application-Process-in-Dutch-Municipalities.pdf

[24] F. Mannhardt and D. Blinde, ‘‘Analyzing the trajectories of patients with sepsis using process mining,’’ in *Proc. RADAR+EMISA*, Essen, Germany, 2017, pp. 72–80.

[25] E. Batista and A. Solanas, ‘‘A uniformization-based approach to preserve individuals' privacy during process mining analyses,’’ *Peer-Peer Netw. Appl.*, vol. 14, no. 3, pp. 1500–1519, May 2021.

[26] L. Reinkemeyer, *Process Mining in Action: Principles, Use Cases and Outloook*. Cham, Switzerland: Springer, 2020.

[27] H. Fang, W. Liu, W. Wang, and S. Zhang, ‘‘Discovery of process variants based on trace context tree,’’ *Connection Sci.*, vol. 35, no. 1, Dec. 2023, Art. no. 2190499.

[28] M. Ayub, M. A. Ghazanfar, T. Khan, and A. Saleem, ‘‘An effective model for Jaccard coefficient to increase the performance of collaborative filtering,’’ *Arabian J. Sci. Eng.*, vol. 45, no. 12, pp. 9997–10017, Dec. 2020.

[29] O. Goldreich, ‘‘On the foundations of cryptography,’’ in *Providing Sound Foundations for Cryptography: on the Work of Shafi Goldwasser and Silvio Micali*. New York, NY, USA: Association for Computing Machinery, 2019, pp. 411–496.

[30] H. Liu, N. Ruan, and J. K. Liu, ‘‘Catfish effect between internal and external attackers: Being semi-honest is helpful,’’ 2019, *arXiv:1907.03720*.

[31] M. Rafiei and W. M. van der Aalst, ‘‘Privacy-preserving data publishing in process mining,’’ in *Business Process Management Forum*. Seville, Spain: Springer, 2020, pp. 122–138.

[32] P. Papadimitriou, A. Dasdan, and H. Garcia-Molina, ‘‘Web graph similarity for anomaly detection,’’ *J. Internet Services Appl.*, vol. 1, no. 1, pp. 19–30, May 2010.

[33] A. Bolt, M. De Leoni, and W. M. van der Aalst, ‘‘A visual approach to spot statistically-significant differences in event logs based on process metrics,’’ in *Advanced Information Systems Engineering*. Ljubljana, Slovenia: Springer, 2016, pp. 151–166.

GENG WANG received the B.S. degree in information security from Anhui University, in 2016. He is currently pursuing the M.S. degree in information security with the Anhui University of Science and Technology. His current research interests include process mining, Petri nets, and privacy protection.

HUAN FANG received the B.S. degree in information and computing from the Anhui University of Science and Technology, Huainan, China, in 2003, the M.S. degree in computer science and engineering from the Shandong University of Science and Technology, China, in 2006, and the Ph.D. degree in computer science and engineering from the Hefei University of Technology, China, in 2013. From 2018 to 2021, she was a Postdoctoral Researcher with the Anhui University of Science and Technology. She is currently a Professor with the School of Mathematics and Big Data, Anhui University of Science and Technology. Her current research interests include Petri nets, process mining, change mining, and intelligent control. Her awards and honors include the Excellent Teachers of Anhui Province, China.