Preprint Data & Knowledge Engineering 134 (2021) 101908

# Group-based privacy preservation techniques for process mining

Majid Rafiei [*], Wil M.P. van der Aalst

*Chair of Process and Data Science, RWTH Aachen University, Aachen, Germany*

ARTICLE INFO

*Keywords:*
Responsible process mining
Privacy preservation
Result utility
Data utility
Event data

ABSTRACT

Process mining techniques help to improve processes using event data. Such data are widely available in information systems. However, they often contain highly sensitive information. For example, healthcare information systems record event data that can be utilized by process mining techniques to improve the treatment process, reduce patient's waiting times, improve resource productivity, etc. However, the recorded event data include highly sensitive information related to treatment activities. Responsible process mining should provide insights about the underlying processes, yet, at the same time, it should not reveal sensitive information. In this paper, we discuss the challenges regarding directly applying existing well-known group-based privacy preservation techniques, e.g., $k$-anonymity, $l$-diversity, etc, to event data. We provide formal definitions of attack models and introduce an effective *group-based privacy preservation technique* for process mining. Our technique covers the main perspectives of process mining including *control-flow*, *time*, *case*, and *organizational* perspectives. The proposed technique provides interpretable and adjustable parameters to handle different privacy aspects. We employ real-life event data and evaluate both data utility and result utility to show the effectiveness of the privacy preservation technique. We also compare this approach with other group-based approaches for privacy-preserving event data publishing.

## 1. Introduction

Process mining employs event data to discover, analyze, and improve the real processes [1]. Indeed, it provides fact-based insights into the actual processes using event logs. There are many algorithms and techniques in the field of process mining. However, the three basic types of process mining are (1) *process discovery*, where the goal is to learn real process models from event logs, (2) *conformance checking*, where the aim is to find commonalities and discordances between a process model and an event log, and (3) *process re-engineering* (*enhancement*), where the aim is to extend or improve a process model using different aspects of the available data.

An event log is a collection of events where each event is described by its attributes [1]. The typical attributes required for the main process mining algorithms are *case identifier*, *activity*, *timestamp*, and *resource*. The *case identifier* refers to the entity that the event belongs to, the *activity* refers to the activity associated with the event, the *timestamp* is the time that the event occurred, and the *resource* is the activity performer. In the human-centered processes, case identifiers refer to persons. For example, in a patient treatment process, the case identifiers refer to the patients whose data are recorded. Moreover, the *resource* attribute often refers to the persons performing activities, e.g., in the healthcare context, the resources refer to the doctors or nurses performing activities for the patients. The event attributes are not limited to the above-mentioned ones, and an event may also carry other case-related attributes, so-called case attributes, e.g., *age*, *salary*, *disease*, etc, which could be considered as sensitive person-specific information. Table 1 shows a sample event log.

[*] Corresponding author.
*E-mail address:* majid.rafiei@pads.rwth-aachen.de (M. Rafiei).

https://doi.org/10.1016/j.datak.2021.101908
Received 30 November 2020; Received in revised form 1 April 2021; Accepted 25 May 2021
Available online 7 June 2021

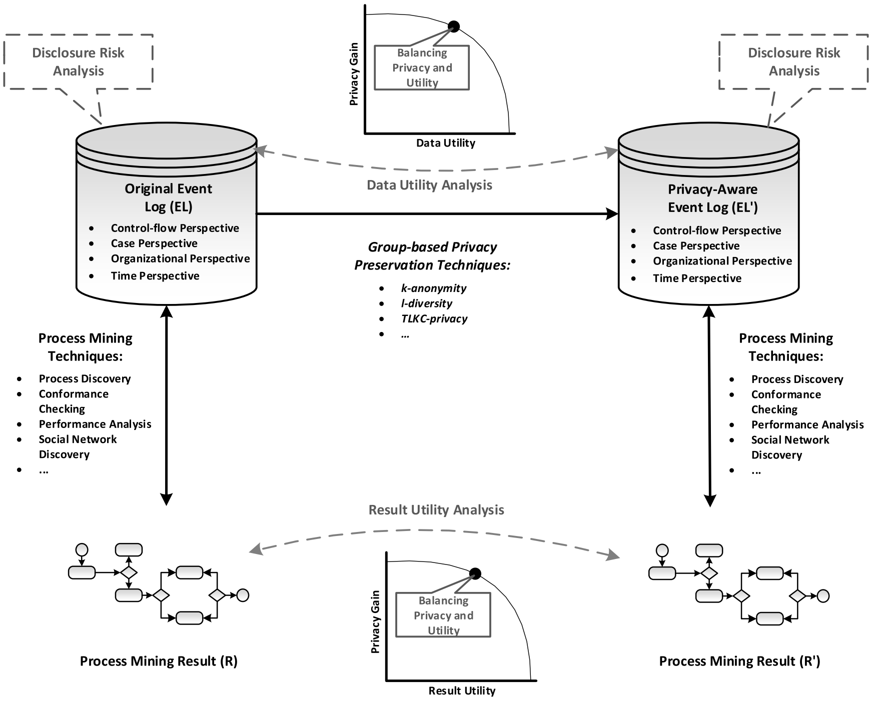


**Fig. 1.** The general overview of privacy-related activities in process mining. Privacy preservation techniques are applied to event logs to mitigate disclosure risks. The data and result utility analyses are used to evaluate the effectiveness of the techniques where the goal is to balance utility loss and privacy gain.

Orthogonal to the three mentioned types of process mining, different perspectives are also defined including *control-flow*, *organizational*, *case*, and *time* perspective [1]. The *control-flow perspective* focuses on activities and their order, which are often utilized by *process discovery* and *conformance checking* techniques. The *organizational perspective* focuses on resources and their relations, which are exploited by *social network discovery* techniques. The *case perspective* is focused on case-related attributes, and the *time perspective* is concerned with the time-related information, which can be used for *performance and bottleneck analyses*.

With respect to the main attributes of events, two different perspectives for privacy in process mining can be considered in the human-centered processes; *resource perspective* and *case perspective*. The *resource perspective* focuses on the privacy rights of the individuals performing activities, and the *case perspective* concerns the privacy rights of the individuals whose data are recorded and analyzed. Depending on the context, the relative importance of these perspectives may differ. However, often the *case perspective* is more critical for privacy than the *resource perspective*. For example, in the healthcare context, activity performers could be publicly available. However, what happens for a specific patient and her/his personal information should be kept private. In this paper, we are focused on the *case perspective*. In principle, when event logs explicitly or implicitly include personal data, *privacy concerns* appear which should be taken into account according to regulations such as the European General Data Protection Regulation (GDPR) [2].

In this paper, we describe *disclosure risks* and *linkage attacks* against event logs. The attack models are formally defined based on the available event attributes. We discuss the challenges regarding directly applying group-based privacy preservation techniques, e.g., $k$-anonymity [3], $l$-diversity [4], etc, to event logs. We extend the work described in [5], where the $TLKC$-privacy is introduced as an effective group-based privacy preservation technique for process mining. The $TLKC$-privacy exploits some restrictions regarding the availability of background knowledge in the real world to deal with process mining-specific challenges. This technique is focused on *control-flow*, *time*, and *case* perspectives. $TLKC$-privacy generalizes several traditional privacy preservation techniques, such as $k$-anonymity, confidence bounding [6], ($\alpha$, $k$)-anonymity [7], and $l$-diversity.

The extended privacy preservation technique covers all the main perspectives of process mining including *control-flow*, *time*, *case*, and *organizational* perspectives. It empowers the adjustability of the proposed technique by adding new parameters to adjust privacy guarantees and the loss of accuracy. Moreover, a new utility measure is defined to tackle the drawbacks of the current approach. To evaluate the extended technique, we employ real-life event logs and evaluate both *data utility* and *result utility*. We also compare the extended $TLKC$-privacy with the main algorithm and other group-based approaches for privacy-preserving event data publishing. Our experiments show that the proposed approach maintains high data and result utility, assuming realistic types of background knowledge. Fig. 1 shows a general overview of privacy-related activities in process mining which are discussed in this paper.

The rest of the paper is organized as follows. In Section 2, we explain the motivation and challenges. Section 3 provides preliminaries on event logs and different types of background knowledge. In Section 4, we provide formal models of the attacks. Privacy preservation techniques are discussed in Section 5. In Section 6, the experiments are presented. Section 7 outlines related work, and Section 8 concludes the paper.

**Table 1**
Sample event log (each row represents an event).

| Case Id | Activity | Timestamp | Resource | Age | Disease |
|---|---|---|---|---|---|
| 1 | Registration (RE) | 01.01.2019-08:30:00 | Employee 4 (E4) | 22 | Flu |
| 1 | Visit (VI) | 01.01.2019-08:45:00 | Doctor 3 (D3) | 22 | Flu |
| 2 | Registration (RE) | 01.01.2019-08:46:00 | Employee 1 (E1) | 30 | Infection |
| 3 | Registration (RE) | 01.01.2019-08:50:00 | Employee 1 (E1) | 32 | Infection |
| 4 | Registration (RE) | 01.01.2019-08:55:00 | Employee 4 (E4) | 29 | Poisoning |
| 1 | Release (RL) | 01.01.2019-08:58:00 | Employee 6 (E6) | 22 | Flu |
| 5 | Registration (RE) | 01.01.2019-09:00:00 | Employee 1 (E1) | 35 | Cancer |
| 2 | Hospitalization (HO) | 01.01.2019-09:01:00 | Employee 3 (E3) | 30 | Infection |
| 6 | Registration (RE) | 01.01.2019-09:05:00 | Employee 4 (E4) | 35 | Corona |
| 4 | Visit (VI) | 01.01.2019-09:10:00 | Doctor 2 (D2) | 29 | Poisoning |
| 5 | Visit (VI) | 01.01.2019-09:20:00 | Doctor 2 (D2) | 35 | Cancer |
| 4 | Infusion (IN) | 01.01.2019-09:30:00 | Nurse 2 (N2) | 29 | Poisoning |
| 5 | Hospitalization (HO) | 01.01.2019-09:55:00 | Employee 6 (E6) | 35 | Cancer |
| 3 | Hospitalization (HO) | 01.01.2019-10:00:00 | Employee 3 (E3) | 32 | Infection |
| 2 | Blood Test (BT) | 01.01.2019-10:02:00 | Nurse 1 (N1) | 30 | Infection |
| 5 | Blood Test (BT) | 01.01.2019-10:10:00 | Nurse 2 (N2) | 35 | Cancer |
| 3 | Blood Test (BT) | 01.01.2019-10:15:00 | Nurse 1 (N1) | 32 | Infection |
| 6 | Visit (VI) | 01.01.2019-10:20:00 | Doctor 3 (D3) | 35 | Corona |
| 4 | Release (RL) | 01.01.2019-10:30:00 | Employee 6 (E6) | 29 | Poisoning |
| 6 | Release (RL) | 01.01.2019-14:20:00 | Employee 6 (E6) | 35 | Corona |
| 2 | Blood Test (BT) | 01.02.2019-08:00:00 | Nurse 1 (N1) | 30 | Infection |
| 2 | Visit (VI) | 01.02.2019-09:30:00 | Doctor 1 (D1) | 30 | Infection |
| 3 | Visit (VI) | 01.02.2019-13:55:00 | Doctor 1 (D1) | 32 | Infection |
| 2 | Release (RL) | 01.02.2019-14:00:00 | Employee 2 (E2) | 30 | Infection |
| 3 | Release (RL) | 01.02.2019-14:15:00 | Employee 2 (E2) | 32 | Infection |
| 5 | Release (RL) | 01.02.2019-16:00:00 | Employee 2 (E2) | 35 | Cancer |

## 2. Motivation and challenges

To motivate the necessity to deal with privacy issues in process mining, we describe the disclosure risks using an example in the health-care context. Consider Table 1 as part of an event log recorded by an information system in a hospital. Note that each case has a sequence of events that are ordered based on the timestamps. This sequence of events is called a *trace* which is a mandatory attribute for a case [1]. For example, case 1, which could be interpreted as patient 1, is first registered by employee 4, then visited by doctor 3, and at the end released from the hospital by employee 6.

Suppose that an adversary knows that a victim patient's data are in the event log (as a *case*), with little information about some event attributes that belongs to the patient, the adversary is able to connect the patient to the corresponding *case id*, so-called *case disclosure* [8]. Consequently, two types of sensitive person-specific information are revealed: (1) the complete sequence of events belonging to the case, and (2) sensitive case attributes. (1) and (2) are generally called *attribute disclosure*. (1) is also called *trace disclosure* that is a specific type of *attribute disclosure* [8]. For example, if the adversary knows that two blood tests were performed for the victim patient, the only matching case is the case with id 2. This attack is called *case linkage* attack. After the case re-identification, the sensitive case attributes are disclosed, e.g., the disease of patient 2 is *infection*. This is called *attribute linkage* attack. Moreover, the complete sequence of events performed for patient 2 is disclosed which contains private information, e.g., the complete sequence of activities performed for the case, the resources who performed the activities for the case, or the exact timestamp of doing a specific activity for the case. We call this attack *trace linkage* which is a specific type of *attribute linkage* attack.

Note that the *attribute linkage* attack does not necessarily need to be launched after the *case linkage*, i.e., if more than one case corresponds to the adversaries knowledge while all the matching cases have the same value for the sensitive case attribute(s) or the same sequence of event attributes (e.g., the same sequence of activities), the *attribute linkage/trace linkage* could happen without a successful *case linkage* attack. For example, if the adversary knows that the activity *visit* has been performed by the resource *doctor* 3 for a victim patient, case 1 and case 6 match this background knowledge. However, they both have the same sequence of activities and resources ($\langle (RE, E4), (VI, D3), (RL, E6) \rangle$). Consequently, the adversary realizes the complete sequence of activities and the resources who performed the activities.

Several group-based privacy preservation techniques, such as $k$-anonymity [3], $l$-diversity [4], and $t$-closeness [9], have been introduced to deal with similar attacks in the context of relational databases. In such techniques, the data attributes are classified into four main categories including; *explicit identifiers*, *quasi-identifiers*, *sensitive attributes*, and *non-sensitive attributes*. The *explicit identifiers* are the attributes that can be used to uniquely identify the data owner, e.g., national id. The *quasi-identifiers* are a set of attributes that could be exploited to uniquely identify the data owner, e.g., $\{age, gender, zipcode\}$. The *sensitive attributes* consist of sensitive person-specific information, e.g., disease or salary, and the *non-sensitive attributes* contain all the attributes that do not fall into the previous three categories [10]. Assuming that *explicit identifiers* suppressed or replaced with dummy identifiers, the group-based privacy preservation techniques aim to perturb potential linkages by generalizing the records into equivalence classes, i.e., groups of records, having the same values on the *quasi-identifier*. These techniques are effective for anonymizing relational data. However, they are not easily applicable to event data due to some specific properties of event data.

In process mining, the *explicit identifiers* (i.e., actual case identifiers) do not need to be stored and processed, and case identifiers are often dummy identifiers, e.g., incremental IDs. As described in the above-mentioned examples, a trace can be considered as a *quasi-identifier* and, at the same time, as a *sensitive attribute*. In other words, a complete sequence of events belonging to a case, is sensitive person-specific information, at the same time, part of a trace, i.e., only some of the event attributes, can be exploited as a *quasi-identifier* to launch *case linkage* and/or *attribute linkage* attacks.

The *quasi-identifier* role of traces in process mining causes significant challenges for group-based privacy preservation techniques because of two specific properties of event data: the *high variability of traces* and the typical *Pareto distribution of traces*. Considering only *activity* as the main event attribute in a trace, the variability of traces in an event log is high because of the following reasons: (1) there could be tens of different activities which could happen in any order, (2) one activity or a bunch of activities could happen repetitively, and (3) traces could contain any non-zero number of activities, i.e., various lengths. Note that this variability becomes even higher when events contain more attributes, e.g., resources. In an event log, trace variants are often distributed similarly to the Pareto distribution, i.e., few trace variants are frequent and many trace variants are unique. Enforcing group-based privacy-preserving approaches on little-overlapping and high-dimensional space is a significant challenge, and often valuable data needs to be suppressed in order to achieve desired privacy requirements [11].

## 3. Preliminaries

In this section, we provide formal definitions for event logs and background knowledge. These formal models will be used in the remainder for describing the attack scenarios and the approach.

### 3.1. Event log

We first introduce some basic notations. For a given set $A$, $A^*$ is the set of all finite sequences over $A$, and $\mathcal{B}(A)$ is the set of all multisets over the set $A$. For $A_1, A_2 \in \mathcal{B}(A)$, $A_1 \subseteq A_2$ if for all $a \in A$, $A_1(a) \leq A_2(a)$. A finite sequence over $A$ of length $n$ is a mapping $\sigma \in \{1, \dots, n\} \to A$, represented as $\sigma = \langle a_1, a_2, \dots, a_n \rangle$ where $a_i = \sigma(i)$ for any $1 \leq i \leq n$. $|\sigma|$ denotes the length of the sequence. For $\sigma_1, \sigma_2 \in A^*$, $\sigma_1 \sqsubseteq \sigma_2$ if $\sigma_1$ is a subsequence of $\sigma_2$, e.g., $\langle a, b, c, x \rangle \sqsubseteq \langle z, x, a, b, b, c, a, b, c, x \rangle$. For $\sigma \in A^*$, $\{a \in \sigma\}$ is the set of elements in $\sigma$, and $[a \in \sigma]$ is the multiset of elements in $\sigma$, e.g., $[a \in \langle x, y, z, x, y \rangle] = [x^2, y^2, z]$. For $x = (a_1, a_2, \dots, a_n) \in A_1 \times A_2 \times \dots \times A_n$, $\pi_{A_i}(x) = a_i$ is the projection of the tuple $x$ on the element from the domain $A_i$, $1 \leq i \leq n$.

**Definition 1** (*Process Instance, Trace*). We define $\mathcal{P} = \mathcal{C} \times \mathcal{E}^* \times \mathcal{S}$ as the universe of all process instances. $\mathcal{C}$ is the universe of case identifiers. $\mathcal{E} = \mathcal{A} \times \mathcal{R} \times \mathcal{T}$ is the universe of main event attributes for process mining where $\mathcal{A}$ is the universe of activities, $\mathcal{R}$ is the universe of resources, and $\mathcal{T}$ is the universe of timestamps. $\mathcal{S} \subseteq \mathcal{D}_1 \cup \dots \cup \mathcal{D}_m$ is the universe of sensitive case attributes where $\mathcal{D}_1$, $\dots$, $\mathcal{D}_m$ are the universes of different case attributes, e.g., disease, salary, age, etc. Given a process instance $p = (c, \sigma, s) \in \mathcal{P}$, $\sigma \in \mathcal{E}^*$ is called the trace attribute of the case $c$.

**Definition 2** (*Event Log*). Let $\mathcal{P} = \mathcal{C} \times \mathcal{E}^* \times \mathcal{S}$ be the universe of process instances. An event log is $EL \subseteq \mathcal{P}$ such that if $(c_1, \sigma_1, s_1) \in EL$, $(c_2, \sigma_2, s_2) \in EL$, and $c_1 = c_2$, then $\sigma_1 = \sigma_2$ and $s_1 = s_2$, i.e., all the case identifiers are unique. Moreover, if $p = (c, \sigma, s) \in EL$, then $\sigma \neq \langle \rangle$.

**Definition 3** (*Perspective, Projection*). Let $\mathcal{P} = \mathcal{C} \times \mathcal{E}^* \times \mathcal{S}$ be the universe of process instances. $ps \in \{\mathcal{A}, \mathcal{R}, \mathcal{A} \times \mathcal{R}, \mathcal{A} \times \mathcal{T}, \mathcal{R} \times \mathcal{T}, \mathcal{A} \times \mathcal{R} \times \mathcal{T}\}$ is a perspective which can be used to project traces of an event log $EL \subseteq \mathcal{P}$. For $\sigma = \langle (a_1, r_1, t_1), \dots, (a_n, r_n, t_n) \rangle \in \mathcal{E}^*$, such that there exists $(c, \sigma, s) \in EL$, $\pi_{ps}(\sigma)$ is the projection of the trace on the given perspective, e.g., for $ps = \mathcal{A} \times \mathcal{R}$, $\pi_{ps}(\sigma) = \langle (a_1, r_1), \dots, (a_n, r_n) \rangle$ is the projection of the trace on the activities and resources. We denote $\mathcal{PS} = \{\mathcal{A}, \mathcal{R}, \mathcal{A} \times \mathcal{R}, \mathcal{A} \times \mathcal{T}, \mathcal{R} \times \mathcal{T}, \mathcal{A} \times \mathcal{R} \times \mathcal{T}\}$ as the universe of perspectives.

**Definition 4** (*Set of Activities/Resources in an Event Log*). Let $\mathcal{P} = \mathcal{C} \times \mathcal{E}^* \times \mathcal{S}$ be the universe of process instances, and $EL \subseteq \mathcal{P}$ be an event log. $A_{EL} = \{a \in \mathcal{A} \mid \exists_{(c,\sigma,s) \in EL} a \in \pi_{\mathcal{A}}(\sigma)\}$ is the set of activities in the event log, and $R_{EL} = \{r \in \mathcal{R} \mid \exists_{(c,\sigma,s) \in EL} a \in \pi_{\mathcal{R}}(\sigma)\}$ is the set of resources in the event log.

**Definition 5** (*Set of Traces/Variants in an Event Log*). Let $\mathcal{P} = \mathcal{C} \times \mathcal{E}^* \times \mathcal{S}$ be the universe of process instances, $EL \subseteq \mathcal{P}$ be an event log, and $ps \in \mathcal{PS}$ be a perspective. $\overline{EL}_{ps} = [\pi_{ps}(\sigma) \mid (c, \sigma, s) \in EL]$ is the multiset of traces in the event log w.r.t. the given perspective. $\widetilde{EL}_{ps} = \{\pi_{ps}(\sigma) \mid (c, \sigma, s) \in EL\}$ is the set of variants, i.e., unique traces, w.r.t. the given perspective, e.g., $\widetilde{EL}_{\mathcal{A}}$ is the set of unique traces w.r.t. the activities.

**Definition 6** (*Directly Follows Relations*). Let $EL \subseteq \mathcal{P}$ be an event log, $ps \in \{\mathcal{R}, \mathcal{A}\}$ be a perspective, $\widetilde{EL}_{ps}$ be the set of variants and $\overline{EL}_{ps}$ be the multiset of traces in the event log $EL$ w.r.t. the given perspective $ps$. $DF_{ps}^{EL} = \{(x, y) \in ps \times ps \mid x >_{ps}^{EL} y\}$ is the set of directly follows relations w.r.t. the given perspective. $x >_{ps}^{EL} y$ iff there exists a trace $\sigma \in \widetilde{EL}_{ps}$ and $1 \leq i < |\sigma|$, s.t., $\sigma(i) = x$ and $\sigma(i+1) = y$. $|x >_{ps}^{EL} y| = \sum_{\sigma \in \widetilde{EL}_{ps}} \overline{EL}_{ps}(\sigma) \times |\{1 \leq i < |\sigma| \mid \sigma(i) = x \wedge \sigma(i+1) = y\}|$ is the number of times $x$ is followed by $y$ in $EL$.

**Definition 7** (*Variant Frequency*). Let $\mathcal{P} = \mathcal{C} \times \mathcal{E}^* \times \mathcal{S}$ be the universe of process instances, and $EL \subseteq \mathcal{P}$ be an event log. Given a perspective $ps \in \mathcal{PS}$, $freq_{ps}^{EL} : \widetilde{EL}_{ps} \to [0, 1]$ is a function that retrieves the relative frequency of the variants in the event log w.r.t. the given perspective. $freq_{ps}^{EL}(\sigma) = \overline{EL}_{ps}(\sigma) / |\overline{EL}_{ps}|$ and $\sum_{\sigma \in \widetilde{EL}_{ps}} freq_{ps}^{EL}(\sigma) = 1$.

Table 2 shows the process instance representation of the event log shown in Table 1, where timestamps are represented as “day-hour:minute". In this event log, *disease* is the attribute which is considered as the sensitive one.

**Table 2**
The process instance representation of the event log Table 1 (each row is a process instance where timestamps are represented as "day-hour:minute").

| Case Id | Simple trace | Disease |
|---|---|---|
| 1 | <(RE,E4,01-08:30),(VI,D3,01-08:45),(RL,E6,01-08:58)> | Flu |
| 2 | <(RE,E1,01-08:46),(HO,E3,01-09:01),(BT,N1,01-10:02),(BT,N1,02-08:00),(VI,D1,02-09:30),(RL,E2,02-14:00)> | HIV |
| 3 | <(RE,E1,01-08:50),(HO,E3,01-10:00),(BT,N1,01-10:15),(VI,D1,02-13:55),(RL,E2,02-14:15)> | Infection |
| 4 | <(RE,E4,01-08:55),(VI,D2,01-09:10),(IN,N2,01-09:30),(RL,E6,01-10:30)> | Poisoning |
| 5 | <(RE,E1,01-09:00),(VI,D2,01-09:20),(HO,E6,01-09:55),(BT,N2,01-10:10),(RL,E2,02-16:00)> | Cancer |
| 6 | <(RE,E4,01-09:05),(VI,D3,01-10:20),(RL,E6,01-14:20)> | Corona |

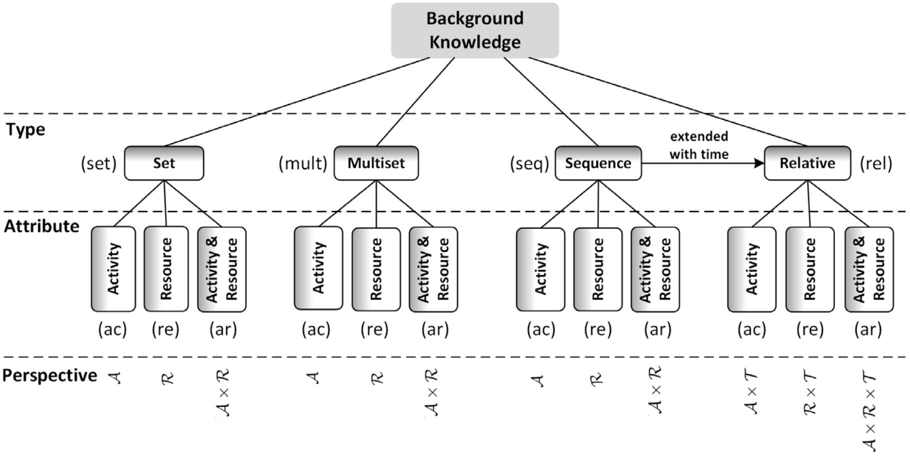


**Fig. 2.** Categorizing background knowledge based on the type and event attributes as well as the corresponding perspectives, e.g., if $type = rel$ and $att = ar$, the corresponding perspective is $ps = \mathcal{A} \times \mathcal{R} \times \mathcal{T}$.

### 3.2. Background knowledge

Regarding the *quasi-identifier* role of traces, we consider four main types of background knowledge including *set*, *multiset* (*mult*), *sequence* (*seq*), and *relative time difference* (*rel*). Using *set* as the type of background knowledge, we assume that an adversary knows a subset of some event attributes contained in the trace attribute of a victim case. In the *multiset* type of background knowledge, the assumption is that an adversary knows a subset of some event attributes included in the trace attribute of a victim case as well as the frequency of the elements. In the *sequence* type of background knowledge, we suppose that an adversary knows a subsequence of some event attributes included in the trace attribute of a victim case.

The exact timestamps of events in an event log impose a high risk regarding the linkage attacks such that little time-related knowledge may easily single out specific events, and consequently the case re-identification. For performance analysis in process mining, we need to have the time-related information. However, the timestamps do not necessarily need to be the actual ones. Therefore, we make all the timestamps relative as defined in Definition 8.

**Definition 8** (*Relative Timestamps*). Let $\sigma = \langle (a_1, t_1), (a_2, t_2), \ldots, (a_n, t_n) \rangle$ be a trace including the time attribute, and $t_0$ be an initial timestamp. $relative(\sigma) = \langle (a_1, t'_1), (a_2, t'_2), \ldots, (a_n, t'_n) \rangle$ is the trace with relative timestamps such that $t'_1 = t_0$ and for each $1 < i \leq n$, $t'_i = t_i - t_1 + t_0$.

Using relative timestamps does not eliminate time-based attacks, since the time differences are real and can be exploited by an adversary. *Relative time difference* type of background knowledge is an extension for the *sequence* type, where the assumption is that an adversary knows a subsequence of some event attributes as well as the relative time differences between the elements. Fig. 2 shows the classification of background knowledge based on the types and event attributes. In the following, we provide formal definitions for different categories of background knowledge based on the main event attributes, i.e., *activity*, *resource*, and *timestamp*. Moreover, one can see that there is a relation between *type*, *attribute*, and *perspective*, i.e., a combination of type and attribute can be mapped to a perspective. For example, if $type = rel$ and $att = ar$, the corresponding perspective is $ps = \mathcal{A} \times \mathcal{R} \times \mathcal{T}$, or if $type \in \{set, mult, seq\}$ and $att = re$, the corresponding perspective is $ps = \mathcal{R}$.

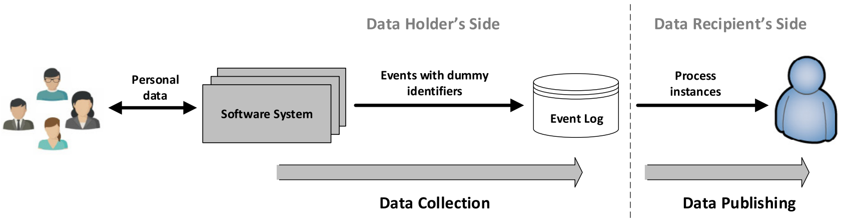


**Fig. 3.** Data collection and data publishing scenario.

**Definition 9** (*Background Knowledge Based on Activities*)**.** Let $EL$ be an event log, and $A_{EL}$ be the set of activities in the event log. $bk_{set,ac}(EL) = 2^{A_{EL}}$, $bk_{mult,ac}(EL) = \mathcal{B}(A_{EL})$, and $bk_{seq,ac}(EL) = A^*_{EL}$ are the sets of candidates of background knowledge based on the activity attribute of the events for the *set*, *multiset*, and *sequence* types of background knowledge. For example, $\{a, b, c\} \in bk_{set,ac}(EL)$, $[a^2, b] \in bk_{mult,ac}(EL)$, and $\langle a, b, c\rangle \in bk_{seq,ac}(EL)$.

**Definition 10** (*Background Knowledge Based on Resources*)**.** Let $EL$ be an event log, and $R_{EL}$ be the set of activities in the event log. $bk_{set,re}(EL) = 2^{R_{EL}}$, $bk_{mult,re}(EL) = \mathcal{B}(R_{EL})$, and $bk_{seq,re}(EL) = R^*_{EL}$ are the sets of candidates of background knowledge based on the resource attribute of the events for the different types of background knowledge.

**Definition 11** (*Background Knowledge Based on Activities&Resources*)**.** Let $EL$ be an event log, $A_{EL}$ be the set of activities in the event log, and $R_{EL}$ be the set of resources in the event log. $bk_{set,ar}(EL) = 2^{A_{EL} \times R_{EL}}$, $bk_{mult,ar}(EL) = \mathcal{B}(A_{EL} \times R_{EL})$, and $bk_{seq,ar}(EL) = (A_{EL} \times R_{EL})^*$ are the sets of candidates of background knowledge based on the activity and resource attribute of the events for the various types of background knowledge.

**Definition 12** (*Background Knowledge Based on Time Differences Between Relative Timestamps*)**.** Let $EL$ be an event log, $A_{EL}$ be the set of activities in the event log, $R_{EL}$ be the set of resources in the event log, and $\mathcal{T}$ be the universe of (relative) timestamps. $bk_{rel,ac}(EL) = (A_{EL} \times \mathcal{T})^*$, $bk_{rel,re}(EL) = (R_{EL} \times \mathcal{T})^*$, and $bk_{rel,ar}(EL) = (A_{EL} \times R_{EL} \times \mathcal{T})^*$ are the sets of candidates of background knowledge based on the relative time differences.

Note that in Definition 12, other attributes are also present. However, our focus is on time differences between relative timestamps. Therefore, we refer to this category of background knowledge as time-based.

## 4. Attack models

Fig. 3 shows our simple scenario of data collection and data publishing. With respect to the types of data holder's models, introduced in [12], we consider a *trusted model*. In the trusted data holder models, the *data holder* is trustworthy, and on the data holder's side, only simple anonymization techniques need to be applied, e.g., suppressing real identifiers. However, the *data recipient*, i.e., a process miner, is not trustworthy and may attempt to identify sensitive information about record owners, i.e., cases. Given a process instance $p = (c, \sigma, s) \in \mathcal{P}$, both $\sigma$ and $s$ are considered as sensitive person-specific information, and part of the trace $\sigma$ can be exploited as the *quasi-identifier* to re-identify the owner of the process instance, i.e., $c$, and/or to learn the sensitive information which belongs to the data owner, i.e., $\sigma$ and/or $s$.

In the following, we provide formal definitions and examples for the attack scenarios based on the main event attributes, i.e., *activity*, *resource*, and *timestamp*. Note that the examples are based on the event log shown in Table 2.

### 4.1. Activity-based attacks

In the activity-based scenarios, we assume that the adversary's knowledge is about the activities performed for a victim case. In the following, we provide formal models based on the introduced types of background knowledge.

- **Based on a set of activities (A1):** In this scenario, we assume that the adversary knows a subset of activities performed for a case, and this information can lead to the *case linkage* and/or *attribute linkage* attacks. Given $EL$ as an event log, we formalize this scenario by a function $match^{EL}_{set,ac} : 2^{A_{EL}} \to 2^{EL}$. For $A \in bk_{set,ac}(EL)$, $match^{EL}_{set,ac}(A) = \{(c, \sigma, s) \in EL \mid A \subseteq \{a \in \pi_{\mathcal{A}}(\sigma)\}\}$. For example, if the adversary knows that $\{VI, IN\}$ is a subset of activities performed for a case, the only matching case is case 4. Therefore, both the sequence of events and the sensitive attribute are disclosed.
- **Based on a multiset of activities (A2):** In this scenario, we assume that the adversary knows a sub-multiset of activities performed for a case, and this information can result in the linkage attacks. Given $EL$ as an event log, we formalize this scenario as follows. $match^{EL}_{mult,ac} : \mathcal{B}(A_{EL}) \to 2^{EL}$. For $B \in bk_{mult,ac}(EL)$, $match^{EL}_{mult,ac}(B) = \{(c, \sigma, s) \in EL \mid B \subseteq [a \in \pi_{\mathcal{A}}(\sigma)]\}$. For example, if the adversary knows that $[HO^1, BT^2]$ is a multiset of activities performed for a case, the only matching case is case 2. Consequently, the complete sequence of events and the disease are disclosed.

- **Based on a sequence of activities (A3):** In this scenario, we assume that the adversary knows a subsequence of activities performed for a case, and this information can lead to the linkage attacks. Given $EL$ as an event log, we formalize this scenario by a function $match^{EL}_{seq,ac} : A^*_{EL} \to 2^{EL}$. For $\sigma \in bk_{seq,ac}(EL)$, $match^{EL}_{seq,ac}(\sigma) = \{(c, \sigma', s) \in EL \mid \sigma \sqsubseteq \pi_{\mathcal{A}}(\sigma')\}$. For example, if the adversary knows that $\langle RE, VI, HO\rangle$ is a subsequence of activities performed for a case, case 5 is the only matching case.

### 4.2. Resource-based attacks

In the resource-based scenarios, we assume that the adversary's knowledge is about the resources who perform activities for a victim case. In the following, we provide formal models based on the main types of background knowledge.

- **Based on a set of resources (R1):** In this scenario, we assume that the adversary knows a subset of resources involved in performing activities for a victim case, and this information can lead to the *case linkage* and/or *attribute linkage* attacks. Given $EL$ as an event log, we formalize this scenario as follows. $match^{EL}_{set,re} : 2^{R_{EL}} \to 2^{EL}$. For $R \in bk_{set,re}(EL)$, $match^{EL}_{set,re}(R) = \{(c, \sigma, s) \in EL \mid R \subseteq \{r \in \pi_{\mathcal{R}}(\sigma)\}\}$. For example, if the adversary knows that $\{E1, D2\}$ is a subset of resources involved in handling a victim case, case 5 is the only matching case. Therefore, both the sequence of events and the sensitive attribute are disclosed.
- **Based on a multiset of resources (R2):** In this scenario, we assume that the adversary knows a sub-multiset of resources involved in performing activities for a victim case, and this information can lead to the linkage attacks. Given $EL$ as an event log, we formalize this scenario as follows. $match^{EL}_{mult,re} : \mathcal{B}(R_{EL}) \to 2^{EL}$. For $S \in bk_{mult,re}(EL)$, $match^{EL}_{mult,re}(S) = \{(c, \sigma, s) \in EL \mid S \subseteq [r \in \pi_{\mathcal{R}}(\sigma)]\}$. For example, if the adversary knows that $[N1^2, E3]$ is a multiset of resources performed activities for a victim case, the only matching case is case 2.
- **Based on a sequence of resources (R3):** In this scenario, we assume that the adversary knows a subsequence of resources who performed activities for a victim case, and this information can result in the linkage attacks. Given $EL$ as an event log, we formalize this scenario by a function $match^{EL}_{seq,re} : R^*_{EL} \to 2^{EL}$. For $\sigma \in bk_{seq,re}(EL)$, $match^{EL}_{seq,re}(\sigma) = \{(c, \sigma', s) \in EL \mid \sigma \sqsubseteq \pi_{\mathcal{R}}(\sigma')\}$. For example, if the adversary knows that $\langle E4, D2\rangle$ is a subsequence of resources who performed activities for a victim case, the only matching case is case 4.

### 4.3. Activity & resource-based attacks

In the activity & resource-based scenarios, we assume that the adversary's knowledge is about activities and the corresponding resources who perform activities for a victim case. In the following, we provide formal models based on the main types of background knowledge.

- **Based on a set of (activity,resource) pairs (AR1):** In this scenario, we assume that the adversary knows a subset of (activity,resource) pairs included in the trace attribute of a victim case, and this information can result in the *case linkage* and/or *attribute linkage* attacks. Given $EL$ as an event log, we formalize this scenario as follows. $match^{EL}_{set,ar} : 2^{A_{EL} \times R_{EL}} \to 2^{EL}$. For $AR \in bk_{set,ar}(EL)$, $match^{EL}_{set,ar}(AR) = \{(c, \sigma, s) \in EL \mid AR \subseteq \{(a, r) \in \pi_{\mathcal{A}\times\mathcal{R}}(\sigma)\}\}$. For example, if the adversary knows that $\{(HO, E6)\}$ is a subset of (activity,resource) pairs contained in the trace attribute of a victim case, case 5 is the only matching case, which result is the whole sequence and sensitive attribute disclosure.
- **Based on a multiset of (activity,resource) pairs (AR2):** In this scenario, we assume that the adversary knows a sub-multiset of (activity,resource) pairs included in the trace attribute of a victim case. Given $EL$ as an event log, the scenario can be formalized as follows. $match^{EL}_{mult,ar} : \mathcal{B}(A_{EL} \times R_{EL}) \to 2^{EL}$. For $BS \in bk_{mult,ar}(EL)$, $match^{EL}_{mult,ar}(BS) = \{(c, \sigma, s) \in EL \mid BS \subseteq [(a, r) \in \pi_{\mathcal{A}\times\mathcal{R}}(\sigma)]\}$. For example, if the adversary knows that $[(BT, N1)^2]$ is a multiset of (activity,resource) pairs included in the trace attribute of a victim case, the only matching case is case 2.
- **Based on a sequence of (activity,resource) pairs (AR3):** In this scenario, we assume that the adversary knows a subsequence of (activity,resource) pairs included in the trace attribute of a victim case, and this information can lead to the linkage attacks. Given $EL$ as an event log, we formalize this scenario by a function $match^{EL}_{seq,ar} : (A_{EL} \times E_{EL})^* \to 2^{EL}$. For $\sigma \in bk_{seq,ar}(EL)$, $match^{EL}_{seq,ar}(\sigma) = \{(c, \sigma', s) \in EL \mid \sigma \sqsubseteq \pi_{\mathcal{A}\times\mathcal{R}}(\sigma')\}$. For example, if the adversary knows that $\langle (RE, E4), (VI, D2)\rangle$ is a (activity,resource) pairs included in the trace attribute of a victim case, case 4 is the only matching case.

### 4.4. Time-based attacks

As we discussed in Section 3.2, after making the timestamps relative, the time differences are still real and can be exploited by an adversary. In the following, we extend the attacks of the type *sequence*, i.e., A3, R3, AR3, with the time-related information.

- **Based on relative time differences between activities (AT):** In this scenario, we assume that the adversary knows a subsequence of activities and also the time difference between the activities. Given $EL$ as an event log, the scenario is formalized as follows. $match^{EL}_{rel,ac} : (A_{EL} \times \mathcal{T})^* \to 2^{EL}$. For $\sigma \in bk_{rel,ac}(EL)$, $match^{EL}_{rel,ac}(\sigma) = \{(c, \sigma', s) \in EL \mid \sigma \sqsubseteq relative(\pi_{\mathcal{A}\times\mathcal{T}}(\sigma'))\}$. For example, if an adversary's knowledge is $\langle HO, VI\rangle$, both case 2 and case 3 get matched. However, if the adversary further knows that for a victim case, *visit* performed in the morning of the next day, the only matching case is case 2.

**Table 3**
A simple event log where time difference between relative timestamps are represented by integer values.

| Case Id | Trace | Disease |
|---|---|---|
| 1 | <(RE,E4,1),(HO,E3,4),(VI,D1,5),(BT,N1,7),(VI,D1,8)> | Cancer |
| 2 | <(BT,N1,7),(VI,D1,8),(RL,E2,9)> | Infection |
| 3 | <(HO,E3,4),(VI,D1,5),(BT,N1,7),(RL,E2,9)> | Corona |
| 4 | <(RE,E4,1),(VI,D1,6),(VI,D1,8),(RL,E2,9)> | Infection |
| 5 | <(HO,4),(VI,D1,8),(RL,E2,9)> | Corona |
| 6 | <(VI,D1,6),(BT,N1,7),(RL,E2,9)> | Flu |
| 7 | <(RE,E4,1),(BT,N1,7),(VI,D1,8),(RL,E2,9)> | Flu |
| 8 | <(RE,E4,1),(VI,D1,6),(BT,N1,7),(VI,D1,8)> | Cancer |

**Table 4**
The event log after applying 2-anonymity to Table 3 using *Baseline*-2.

| Case Id | Trace | Disease |
|---|---|---|
| 1 | <(BT,N1,7),(VI,D1,8)> | Cancer |
| 2 | <(BT,N1,7),(VI,D1,8),(RL,E2,9)> | Infection |
| 3 | <(BT,N1,7),(RL,E2,9)> | Corona |
| 4 | <(VI,D1,8),(RL,E2,9)> | Infection |
| 5 | <(VI,D1,8),(RL,E2,9)> | Corona |
| 6 | <(BT,N1,7),(RL,E2,9)> | Flu |
| 7 | <(BT,N1,7),(VI,D1,8),(RL,E2,9)> | Flu |
| 8 | <(BT,N1,7),(VI,D1,8)> | Cancer |

– **Based on relative time differences between resources who performed activities (RT):** According to this scenario, the adversary knows a subsequence of resources and the time difference between the resources involved in handling a case. Given $EL$ as an event log, we formalize this scenario by a function $match^{EL}_{rel,re} : (R_{EL} \times \mathcal{T})^* \rightarrow 2^{EL}$. For $\sigma \in bk_{rel,re}(EL)$, $match^{EL}_{rel,re}(\sigma) = \{(c, \sigma', s) \in EL \mid \sigma \sqsubseteq relative(\pi_{\mathcal{R}\times\mathcal{T}}(\sigma'))\}$. For example, if an adversary's knowledge is $\langle E1, E3\rangle$, both case 2 and case 3 get matched. However, if the adversary further knows that for the victim case, *employee* 3 performed *hospitalization* more than one hour after *registration*, case 3 is the only matching case.

– **Based on relative time differences between (activity,resource) pairs (ART):** In this scenario, the assumption is that the adversary knows a subsequence of (activity,resource) pairs and the time difference between these pairs. Given $EL$ as an event log, we formalize this scenario as follows. $match^{EL}_{rel,ar} : (A_{EL} \times R_{EL})^* \rightarrow 2^{EL}$. For $\sigma \in bk_{rel,ar}(EL)$, $match^{EL}_{rel,ar}(\sigma) = \{(c, \sigma', s) \in EL \mid \sigma \sqsubseteq relative(\sigma')\}$. For example, case 1 and case 6 have the same sequence of (activity,resource) pairs. However, if the adversary knows that for a victim case, it took almost four hours to get released by *employee* 6 after visiting by a doctor, the corresponding possible cases narrow down to only one case, which is case 6.

## 5. Privacy preservation techniques

Traditional $k$-anonymity and its extended privacy preservation techniques assume that an adversary could use all of the quasi-identifier attributes as background knowledge to launch linkage attacks. According to the types of background knowledge introduced in Section 3, this assumption means that the background knowledge of an adversary is $bk_{rel,ar}$ which covers all the information contained in a trace. In the following, we show the results of applying two baseline methods with respect to the aforementioned assumption.

### *5.1. Baseline methods*

In this subsection, we introduce two baseline methods to apply $k$-anonymity on event logs: *Baseline*-1 and *Baseline*-2. *Baseline*-1 is a naïve $k$-anonymity approach where we remove all the trace variants occurring less than $k$ times. *Baseline*-2 maps each violating trace variant, i.e., the variant that does not fulfill the desired $k$-anonymity requirement, to the most similar non-violating subtrace by removing events. In *Baseline*-2, if there exists no non-violating subtrace, the whole trace variant is removed.

Suppose that Table 3 is part of an event log recorded by an information system in a hospital that needs to be published after applying $k$-anonymity. Note that for the sake of simplicity, the time differences between relative timestamps are represented by integers. Since all the traces in this event log are unique if we apply $k$-anonymity with any value greater than 1, using *Baseline*-1, all the traces are removed. If we apply *Baseline*-2 where $k = 2$ then the result is the event log shown in Table 4. One can see that for such a weak privacy requirement 12 events are removed. Now, if we use $k = 4$, Table 5 is the result where 18 events are removed which is more than half of the events.

In [13], the $PRETSA$ method is introduced as a group-based privacy preservation technique for process mining where the authors apply $k$-anonymity and $t$-closeness on event data for privacy-aware process discovery. However, $PRETSA$ focuses on the *resource perspective* of privacy while we focus on the *case perspective*. The $PRETSA$ method assumes a prefix of activity sequences as the background knowledge, and each violating trace is mapped to the most similar non-violating trace. In [5], $PRETSA_{case}$ is introduced as a variant of $PRETSA$ method where only the $k$-anonymity part is considered, and the focus is on the privacy of *cases* rather than *resources*. Therefore, $PRETSA_{case}$ is a specific type of *Baseline*-2 where the background knowledge is a specific type of $bk_{seq,ac}$, i.e., a prefix of activity sequences rather than any subsequence.

### 5.2. $TLKC$-Privacy (extended)

As discussed in [5], it is almost impossible for an adversary to acquire all the information of a target victim, and it requires non-trivial effort to gather each piece of background knowledge. The $TLKC$-privacy exploits this limitation and assumes that the adversary's background knowledge is bounded by at most $L$ values of the quasi-identifier, i.e., the size or power of background knowledge. Based on the types of background knowledge illustrated in Fig. 2, the $TLKC$-privacy considers all the types, i.e., *set*, *multiset*, *sequence*, and *relative*. However, it focuses on the *activity* attribute (ac) and *timestamps* which are included in the *relative* type. In this paper, the technique is extended with the *resource* attribute, i.e., merely *resource* (re) and *activity* along with *resource* (ar) are also considered. In the following, we bound the power of the different types of background knowledge (Definition 9–12) with $L$ as the maximal size of candidates.

**Definition 13** (*Bounded Background Knowledge*). Let $EL$ be an event log, $type \in \{set, mult, seq, rel\}$ be the type of background knowledge, $att \in \{ac, re, ar\}$ be the event attribute of background knowledge, and L be the size of background knowledge. $bk^{L}_{type,att}(EL) = \{cand \in bk_{type,att}(EL) \mid |cand| \le L\}$ are the candidates of the background knowledge whose sizes are bounded by $L$.

In the $TLKC$-privacy, $T \in \{seconds, minutes, hours, days\}$ refers to the accuracy of timestamps, e.g., $T = minutes$ shows that the accuracy of timestamps is limited at *minutes* level, $L$ refers to the power of background knowledge, $K$ refers to the $k$ in the $k$-anonymity definition, and $C$ refers to the bound of confidence regarding the sensitive attribute values in a matching set. We denote $EL(T)$ as the event log with the accuracy of timestamps at the level $T$. The general idea of $TLKC$-privacy is to ensure that the background knowledge of size $L$ in $EL(T)$ is shared by at least $K$ cases, and the confidence of inferring the sensitive value in $S$ is not greater than $C$.

**Definition 14** (*$TLKC$-Privacy*). Let $EL \subseteq \mathcal{P}$ be an event log, $L$ be the maximal size of background knowledge, $T \in \{seconds, minutes, hours, days\}$ be the accuracy of timestamps, $type \in \{set, mult, seq, rel\}$, and $att \in \{ac, re, ar\}$. $EL(T)$ satisfies $TLKC$-privacy if and only if for any $cand \in bk^{L}_{type,att}(EL(T))$ such that $match^{EL(T)}_{type,att}(cand) \neq \emptyset$:

- $|match^{EL(T)}_{type,att}(cand)| \ge K$, where $K \in \mathbb{N}_{>0}$, and
- $Pr(s|cand) = \frac{|\{p \in match^{EL(T)}_{type,att}(cand) | \pi_S(p)=s\}|}{|match^{EL(T)}_{type,att}(cand)|} \le C$ for any $s \in S$, where $0 < C \le 1$ is a real number as the confidence threshold, and $\pi_S(p)$ is the projection of the process instance on the sensitive attribute value.

The $TLKC$-privacy provides a major relaxation from traditional $k$-anonymity based on a reasonable assumption that the adversary has restricted knowledge. It generalizes several privacy preservation techniques including $k$-anonymity, confidence bounding, $(\alpha, k)$-anonymity, and $l$-diversity. It also provides interpretable parameters. Note that the type and attribute of background knowledge implicitly show the perspective (Fig. 2).

#### 5.2.1. Privacy measure

In the subsection, we define *(minimal) violating traces* w.r.t. the privacy requirements of the $TLKC$-privacy.

**Definition 15** (*Violating Trace*). Let $EL \subseteq \mathcal{P}$ be an event log, $L$ be the maximal size of background knowledge, $T \in \{seconds, minutes, hours, days\}$ be the accuracy of timestamps, $type \in \{set, mult, seq, rel\}$, $att \in \{ac, re, ar\}$, $ps \in \mathcal{PS}$ be the corresponding perspective w.r.t. the given $type$ and $att$, and $\sigma \sqsubseteq \pi_{ps}(\sigma')$ such that $(c, \sigma', s) \in EL(T)$. $\sigma$ is a violating (sub)trace with respect to the $TLKC$-privacy requirements if there exists a $cand \in bk^{L}_{type,att}(EL(T))$:

- $cand \sqsubseteq \sigma \vee cand \subseteq \{e \in \sigma\} \vee cand \subseteq [e \in \sigma]$, and
- $|match^{EL(T)}_{type,att}(cand)| < K$ or $Pr(s|cand) > C$ for some $s \in S$.

An event log satisfies $TLKC$-privacy, if all violating traces w.r.t. the given privacy requirement are removed. A naïve approach is to determine all violating traces and remove them. However, this approach is inefficient due to the numerous number of violating traces, even for a weak privacy requirement. Moreover, as demonstrated in [5], $TLKC$-privacy is not monotonic w.r.t. $L$. In fact, the anonymity threshold $K$ is monotonic w.r.t. $L$, i.e., if $L' \le L$ and $C = 100\%$, an event log $EL$ which satisfies $TLKC$-privacy must satisfy $TL'KC$-privacy. However, confidence threshold $C$ is not monotonic w.r.t. $L$, i.e., if $\sigma$ is non-violating trace, its subtrace may or may not be non-violating. Therefore, we have to make sure that the conditions should hold for any $L' \le L$. To this end, in the following, we define the extended version of *minimal violating traces* w.r.t. the different perspectives.

**Algorithm 1:** $TLKC$-privacy - extended w.r.t. the different perspectives.

```
Input: Original event log $EL$
Input: $T$, $L$, $K$, $C$, and $\Theta$ (frequency threshold)
Input: Background knowledge type and attribute ($bk_{type,att}$), sensitive attributes $\mathcal{S}$
Output: Anonymized event log $EL'$ which satisfies the desired $TLKC$-privacy requirements
generate $MFT_{ps}^{EL}$ and $MVT_{ps}^{EL}$;
generate $MFT_{ps}^{tree}$ and $MVT_{ps}^{tree}$ as the prefix trees for $MFT_{ps}^{EL}$ and $MVT_{ps}^{EL}$;
while there is node (event) in $MVT_{ps}^{tree}$ do
    select an event (node) $e_w$ that has the highest score to suppress based on $socre(e)_{ps}^{EL}$;
    delete all the MVTs and MFTs containing the event $e_w$ from $MVT_{ps}^{tree}$ and $MFT_{ps}^{tree}$;
    update $socre(e)_{ps}^{EL}$ for all the remaining events (nodes) in $MVT_{ps}^{tree}$;
    add $e_w$ to the suppression set $Sup^{EL}$;
end
foreach $e \in Sup^{EL}$ do
    suppress all instances of $e$ from $EL$;
end
return suppressed $EL$ as $EL'$;
```

**Definition 16** (*Minimal Violating Trace*). Let $EL \subseteq \mathcal{P}$ be an event log, $L$ be the maximal size of background knowledge, $T \in \{seconds, minutes, hours, days\}$ be the accuracy of timestamps, $type \in \{set, mult, seq, rel\}$, $att \in \{ac, re, ar\}$, $ps \in \mathcal{PS}$ be the corresponding perspective w.r.t. the given $type$ and $att$, and $\sigma \sqsubseteq \pi_{ps}(\sigma')$ such that $(c, \sigma', s) \in EL(T)$. $\sigma$ is a minimal violating trace if $\sigma$ is a violating trace (Definition 15) in the $EL$, and every proper subtrace of $\sigma$ is not violating. We denote $MVT_{ps}^{EL}$ as the set of minimal violating traces in the event log $EL$ w.r.t. the perspective $ps$.

Every violating trace in an event log is either a minimal violating trace or it contains a minimal violating trace. Therefore, if an event log contains no minimal violating trace, then it contains no violating trace. Note that the set of minimal violating traces in an event log is much smaller than the set of violating traces in the event log which results in better efficiency for removing violating traces.

### 5.2.2. Utility measure

In the $TLKC$-privacy, the *maximal frequent traces* are defined as a measure for considering data utility, where traces contain *activity* and *timestamp* attributes. Since we extend the $TLKC$-privacy preservation technique to cover all the main perspectives of process mining, the utility measure also needs to be extended. In the following, we provide an extended version of the utility measure considering the perspectives.

**Definition 17** (*Maximal Frequent Trace*). Let $EL$ be an event log, and $ps \in \mathcal{PS}$ be a perspective. For a given minimum support threshold $\Theta$, a non-empty trace $\sigma \sqsubseteq \pi_{ps}(\sigma')$ such that $(c, \sigma', s) \in EL$ is *maximal frequent* in the $EL$ if $\sigma$ is frequent, i.e., the frequency of $\sigma$ is greater than or equal to $\Theta$, and no supertrace of $\sigma$ is frequent in the $EL$. We denote $MFT_{ps}^{EL}$ as the set of maximal frequent traces in the event log $EL$ w.r.t. the perspective $ps$.

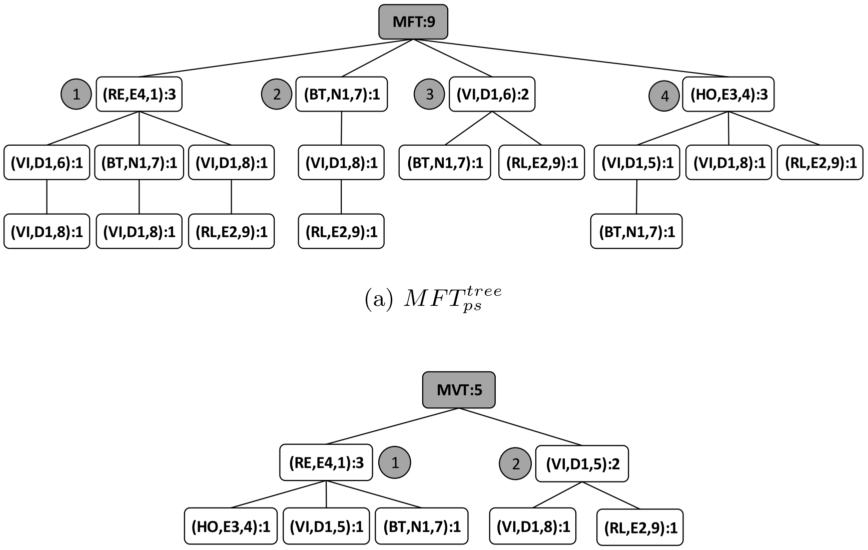
The goal of data utility is to preserve as many MFT as possible w.r.t. the given perspective. For example, in the *control-flow* perspective, i.e., $ps = \mathcal{A}$, the goal in to preserve the maximal frequent traces w.r.t. the activities. Note that in an event log, the set of maximal frequent traces is much smaller than the set of frequent traces. Moreover, any subtrace of a maximal frequent trace is also a frequent trace, and once all the MFTs are discovered, the support counts of any frequent subtrace can be computed by scanning the data once.

### 5.2.3. Balancing privacy and utility

As discussed in the privacy measure section, to provide the desired privacy requirements, all the minimal violating traces need to be removed. However, this should be done w.r.t. the utility measure. According to Definition 16, every proper subtrace of a minimal violating trace is not violating. Therefore, a minimal violating trace can be removed after removing one event of the trace. This event needs to be chosen w.r.t. both utility and privacy measures. To this end, a greedy function is defined to choose an event to remove from the minimal violating traces such that it maximizes the number of removed minimal violating traces, i.e., privacy gain, yet, at the same time, minimizes the number of removed maximal frequent traces, i.e., utility loss.

**Definition 18** (*Score, Privacy Gain, Utility Loss*). Let $EL$ be an event log, $ps \in \mathcal{PS}$ be a perspective, and $events_{ps}(EL) = \{e \in \pi_{ps}(\sigma) \mid (c, \sigma, s) \in EL\}$ be the set of events in the event log w.r.t. the given perspective. $score_{ps}^{EL} : \mathcal{E} \nrightarrow \mathbb{R}_{>0}$ is a function which retrieves the score of the events in the event log w.r.t. the perspective. For $e \in events_{ps}(EL)$, $score_{ps}^{EL}(e) = {}^{PG_{ps}^{EL}(e)}/_{UL_{ps}^{EL}(e)+1}$. $PG_{ps}^{EL}(e)$ is the number of MVTs containing the event $e$, i.e., $PG_{ps}^{EL}(e) = |\{x \in MVT_{ps}^{EL} \mid e \in x\}|$ and $UL_{ps}^{EL}(e)$ is the number of MFTs containing the event $e$, i.e., $UL_{ps}^{EL}(e) = |\{x \in MFT_{ps}^{EL} \mid e \in x\}|$.

Note that in the score (Definition 18), 1 is added to the denominator to avoid diving by zero (when $e$ does not belong to any MFT). The event $e$ with the highest score is called the *winner* event, denoted by $e_w$. Algorithm 1 summarizes all the steps of $TLKC$-privacy. In the following, we show how the algorithm works on the event log Table 3.



(a) $MFT_{ps}^{tree}$

(b) $MVT_{ps}^{tree}$

**Fig. 4.** The $MFT_{ps}^{tree}$ and $MVT_{ps}^{tree}$ generated for the event log Table 3 with $T = hours$, $L = 2$, $K = 2$, $C = 50\%$, $\Theta = 25\%$, $\mathcal{S} = Disease$, and $bk_{rel,ar}^{EL}$.

**Table 5**
The event log after applying 4-anonymity to Table 3 using *Baseline*-2.

| Case Id | Trace | Disease |
|---|---|---|
| 1 | <(BT,N1,7),(VI,D1,8)> | Cancer |
| 2 | <(BT,N1,7),(VI,D1,8)> | Infection |
| 3 | <(RL,E2,9)> | Corona |
| 4 | <(RL,E2,9)> | Infection |
| 5 | <(RL,E2,9)> | Corona |
| 6 | <(RL,E2,9)> | Flu |
| 7 | <(BT,N1,7),(VI,D1,8)> | Flu |
| 8 | <(BT,N1,7),(VI,D1,8)> | Cancer |

Suppose that Table 3 shows a simple event log $EL$ where timestamps are represented by integer values as hours. The first line in Algorithm 1 generates the set of maximal frequent traces ($MFT_{ps}^{EL}$) and the set of minimal violating traces ($MVT_{ps}^{EL}$) from the event log $EL$ with $T = hours$, $L = 2$, $K = 2$, $C = 50\%$, $\Theta = 25\%$, *Disease* as the sensitive attribute $\mathcal{S}$, and $bk_{rel,ar}^{EL}$ as the background knowledge, i.e., $ps = \mathcal{A} \times \mathcal{R} \times \mathcal{T}$. Fig. 4 shows the $MFT_{ps}^{tree}$ and $MVT_{ps}^{tree}$ generated by line 2 in Algorithm 1, where each root-to-leaf path represents one trace, and each node represents an event in a trace with the frequency of occurrence. Table 6 shows the initial score of every event (node) in the $MVT_{ps}^{tree}$ ($score_{ps}^{EL}(e)$). Line 4 determines the winner event $e_w$ which is $(VI, D1, 5)$. Line 5 deletes all the MVTs and MFTs containing the winner event $e_w$, i.e., subtree 2 and the path $\langle (RE, E4, 1), (VI, D1, 5)\rangle$ of subtree 1 in the $MVT_{ps}^{tree}$, and the path $\langle (HO, E3, 4), (VI, D1, 5), (BT, N1, 7)\rangle$ of subtree 4 in the $MFT_{ps}^{tree}$ are removed and frequencies get updated. Line 6 updates the scores based on the new frequencies of events. Table 7 shows the remaining events in $MVT_{ps}^{tree}$ with the updated scores. Line 7 adds the winner event to a suppression set $Sup^{EL}$. Lines 4–7 is repeated until there is no node in $MVT_{ps}^{tree}$. According to Table 7 the next winner event is $(RE, E4, 1)$, and after deleting all the MVTs and MFTs containing this event, $MVT_{ps}^{tree}$ is empty. Therefore, at the end of the ***while*** loop, the suppression set $Sup^{EL} = \{(VI, D1, 5), (RE, E4, 1)\}$. The ***foreach*** loop suppresses all the instances of the events, i.e., *global suppression*, in the $Sup^{EL}$ from the $EL$, and the last line returns the suppressed $EL$ as the anonymized event log $EL'$ which is shown in Table 8.

Compared to Tables 4 and 5 which are the results of applying traditional $k$-anonymity using *Baseline*-2, Table 8 shows that $TLKC$-privacy removes less events (only 6), for the stronger privacy requirements.

### 5.2.4. New utility measure and new score

In this subsection, we first describe the shortcomings of the utility measure and the score introduced in [5] (extended in Definitions 17 and 18), then we introduce a new utility measure and a new score to overcome the drawbacks. According to Definition 18, the score is calculated based on the existence of events in the set of minimal violating traces and the set of maximal

**Table 6**
The initial scores for the events in Fig. 4(b).

| | $(RE, E4, 1)$ | $(HO, E3, 4)$ | $(VI, D1, 5)$ | $(BT, N1, 7)$ | $(VI, D1, 8)$ | $(RL, E2, 9)$ |
|---|---|---|---|---|---|---|
| $PG_{ps}^{EL}(e)$ | 3 | 1 | 3 | 1 | 1 | 1 |
| $UL_{ps}^{EL}(e) + 1$ | 4 | 4 | 2 | 5 | 6 | 5 |
| $score_{ps}^{EL}(e)$ | 0.75 | 0.25 | 1.50 | 0.20 | 0.16 | 0.20 |

**Table 7**
The first updated scores.

| | $(RE, E4, 1)$ | $(HO, E3, 4)$ | $(BT, N1, 7)$ |
|---|---|---|---|
| $PG_{ps}^{EL}(e)$ | 2 | 1 | 1 |
| $UL_{ps}^{EL}(e) + 1$ | 4 | 3 | 4 |
| $score_{ps}^{EL}(e)$ | 0.5 | 0.33 | 0.25 |

frequent traces. However, the sizes of these sets, and consequently the included events, highly depends on the corresponding parameters. The set of MVTs is obtained based on $T$, $L$, $K$, $C$, and $bk_{type,att}$, while the set of MFTs is discovered based on $\Theta$ and the given perspective. Therefore, some of the events included in the set of minimal violating traces may not be included in the set of maximal frequent traces. Consequently, the score of the corresponding events is merely calculated based on the effect on the *privacy gain*. When two or more events have the same score based on the *privacy gain*, the algorithm assumes an equal effect for the data utility aspect and randomly choose one of the events, which is not a valid assumption.

Another problem with the current score is that even when there are maximal frequent traces where the event is included, the score does not differentiate the corresponding MFTs based on their frequencies in the event log. For example, suppose that for two events $e_1$ and $e_2$ in the minimal violating traces there are two maximal frequent traces $MFT_1$ and $MFT_2$ such that $e_1$ is only included in $MFT_1$, i.e., $UL(e_1) = 1$, and $e_2$ is only included in $MFT_2$, i.e., $UL(e_2) = 1$. Hence, both events get the same score for the utility aspect. However, the corresponding MFTs may have completely different frequencies in the event log which leads to a different impact on the utility. Particularly, this issue is highlighted when the frequency threshold ($\Theta$) is rather low. For example, if $\Theta = 50\%$, then frequency of $MFT_1$ and $MFT_2$ in the event log could differ up to 50%. Furthermore, the current score is not normalized, and it is not possible for the user to adjust the effect of each aspect on the score. For example, one may want to consider more effect for the data utility aspect compared to the privacy gain aspect.

To overcome the above-mentioned shortcomings, we define a new utility measure that is able to show the impact of every single event on the data utility. We also define a new score based on the new utility measure which provides normalized scores, and the effect of each aspect is adjustable for users. In the new utility measure (Definition 19), we consider the relative frequency of the variants, where the given perspective of the event is included, as the basis of the utility.

**Definition 19** (*New Utility Measure*). Let $EL$ be an event log, $ps \in \mathcal{PS}$ be a perspective, and $events_{ps}(EL) = \{e \in \pi_{ps}(\sigma) \mid (c, \sigma, s) \in EL\}$ be the set of events in the event log w.r.t. the given perspective. For $e \in events_{ps}(EL)$, $nUL_{ps}^{EL}(e) = 1 - \sum_{\{\sigma \in \widetilde{EL} \mid e \in \pi_{ps}(\sigma)\}} freq^{EL}(\sigma)$.

**Definition 20** (*New Score*). Let $EL$ be an event log, $ps \in \mathcal{PS}$ be a perspective, $events_{ps}(EL) = \{e \in \pi_{ps}(\sigma) \mid (c, \sigma, s) \in EL\}$ be the set of events in the event log w.r.t. the given perspective, $\alpha$ be the coefficient of privacy gain ($0 \leq \alpha \leq 1$), $\beta$ be the coefficient of utility loss ($0 \leq \beta \leq 1$), and $\alpha + \beta = 1$. $n\text{-}score_{ps}^{EL} : \mathcal{E} \nrightarrow \mathbb{R}_{>0}$ is a function which retrieves the score of the events in the event log w.r.t. the perspective. For $e \in events_{ps}(EL)$, $n\text{-}score_{ps}^{EL}(e) = \alpha \cdot rPG_{ps}^{EL}(e) + \beta \cdot nUL_{ps}^{EL}(e)$, where $rPG_{ps}^{EL}(e)$ is the relative value of the privacy gain, i.e., $rPG_{ps}^{EL}(e) = {}^{|\{x \in MVT_{ps}^{EL} \mid e \in x\}|}/_{|MVT_{ps}^{EL}|}$.

Algorithm 2 shows the new algorithm based on the new utility measure and new score, where maximal frequent traces are not used anymore, and the score of events included in the minimal violating traces is calculated based on the new score. Note that in both Algorithm 1 and Algorithm 2 the perspective is derived from the background knowledge type and attribute (Fig. 2).

## 6. Experiments

In this section, we evaluate the extended $TLKC$-privacy by applying it to real-life event logs. We explore the effect of applying the technique on both *data utility* and *result utility*. The results are also compared with the baseline methods. The *result utility* analysis evaluates the similarity of the specific results obtained from the privacy-aware event log with the same type of results obtained from the original event log, while the *data utility* analysis compares the privacy-aware event log with the original event log. As discussed in [8], the result utility analysis is highly dependent on the underlying algorithm generating specific results, and the data utility analysis provides a more general evaluation. We perform the evaluation for the three main perspectives including *control-flow*, *organizational*, and *time* perspectives. For the result utility analysis, in each perspective, we focus on a specific type of results. For the control-flow perspective, we focus on *process discovery*, for the organizational perspective, we perform *social network discovery*, and for the time perspective, we perform *bottleneck analysis*. The implementation as a Python program is available on GitHub.[1]

[1] https://github.com/m4jidRafiei/TLKC-Privacy-Ext.

**Algorithm 2:** $TLKC$-privacy - extended w.r.t. the different perspectives, new score, and new utility measure.

```
Input: Original event log $EL$
Input: $T$, $L$, $K$, $C$
Input: Background knowledge type and attribute ($bk_{type,att}$), sensitive attributes $\mathcal{S}$
Output: Anonymized event log $EL'$ which satisfies the desired $TLKC$-privacy requirements
generate $MVT^{EL}_{ps}$ and $MVT^{tree}_{ps}$;
while there is node (event) in $MVT^{tree}_{ps}$ do
    select an event (node) $e_w$ that has the highest score to suppress based on $n\text{-}socre(e)^{EL}_{ps}$;
    delete all the MVTs containing the event $e_w$ from $MVT^{tree}_{ps}$;
    update $n\text{-}socre(e)^{EL}_{ps}$ for all the remaining events (nodes) in $MVT^{tree}_{ps}$;
    add $e_w$ to the suppression set $Sup^{EL}$;
end
foreach $e \in Sup^{EL}$ do
    suppress all instances of $e$ from $EL$;
end
return suppressed $EL$ as $EL'$;
```

**Table 8**
The anonymized event log for Table 3 with $T =$ hours, $L = 2$, $K = 2$, $C = 50\%$, $\Theta = 25\%$, $\mathcal{S} =$ Disease, and $bk^{EL}_{rel,ar}$.

| Case Id | Trace | Disease |
|---|---|---|
| 1 | <(HO,E3,4),(BT,N1,7),(VI,D1,8)> | Cancer |
| 2 | <(BT,N1,7),(VI,D1,8),(RL,E2,9)> | Infection |
| 3 | <(HO,E3,4),(BT,N1,7),(RL,E2,9)> | Corona |
| 4 | <(VI,D1,6),(VI,D1,8),(RL,E2,9)> | Infection |
| 5 | <(HO,E3,4),(VI,D1,8),(RL,E2,9)> | Corona |
| 6 | <(VI,D1,6),(BT,N1,7),(RL,E2,9)> | Flu |
| 7 | <(BT,N1,7),(VI,D1,8),(RL,E2,9)> | Flu |
| 8 | <(VI,D1,6),(BT,N1,7),(VI,D1,8)> | Cancer |

**Table 9**
The general statistics of the event logs used in the experiments.

| Event log | #cases | #events | #unique activity | #unique resource | #unique (activity, resource) |
|---|---|---|---|---|---|
| Sepsis-Cases [14] | 1050 | 15,214 | 16 | – | – |
| BPIC-2012-APP [15] | 13,087 | 60,849 | 10 | 61 | 301 |
| BPIC-2017-APP [17] | 31,509 | 2,39,595 | 10 | 144 | 927 |

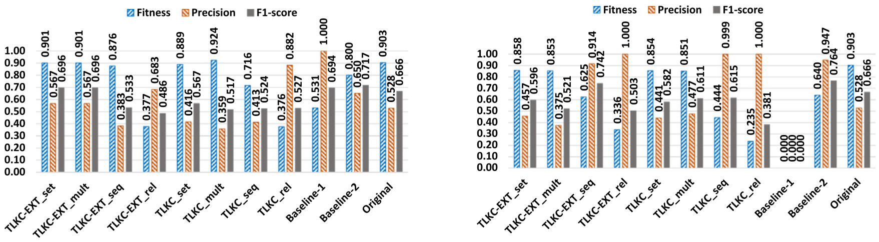
**Table 10**
Some statistics regarding the variants of the event logs used in the experiments w.r.t. the different perspectives.

| Event log | #variants activity perspective | #variants resource perspective | #variants (activity, resource) perspective |
|---|---|---|---|
| Sepsis-Cases [14] | 846 | – | – |
| BPIC-2012-APP [15] | 17 | 2974 | 3872 |
| BPIC-2017-APP [17] | 102 | 24,230 | 24,471 |

### 6.1. *Experimental setup*

For the experiments, we employ two human-centered event logs, where the *case identifiers* refer to individuals: Sepsis-Cases, BPIC-2012-APP, and BPIC-2017-APP. Sepsis-Cases [14] is a real-life event log containing events of sepsis cases from a hospital. BPIC-2012-APP [15] is also a real-life event log about a loan application process taken from a Dutch financial institute. BPIC-2017-APP also pertains to a loan application process of a Dutch financial institute. Table 9 shows the general statistics of these event logs. The Sepsis-Cases event log was included in the experiments because it has some challenging features for privacy preservation techniques, namely, 80% of traces are unique based on the activity perspective which imposes significant challenges for privacy-preserving process discovery algorithms [5,13,16]. BPIC-2017-APP has similar properties w.r.t. the resource perspective, i.e., 76% of traces are unique w.r.t. the resource perspective. Note that Sepsis-Cases does not contain resource information and cannot be used for the *organizational* perspective analysis. We employ BPIC-2012-APP and BPIC-2017-APP for the *organizational* perspective. Table 10 shows some statistics about the variants with respect to different perspectives. For example, as mentioned, in Sepsis-Cases, 80% of traces are unique from the activity perspective, or in BPIC-2017-APP, 76% of traces are unique from the resource perspective.

Overall, we performed more than 1000 experiments for the four different types of background knowledge and different perspectives. 200 different settings are used based on the following values for the main parameters: $L \in \{2,3,4,5,6\}$, $K \in \{20,30,40,50,60\}$, $C \in \{0.2,0.3,0.4,0.5\}$, and $T \in \{hours, minutes\}$. We consider equal weights for the privacy gain and utility loss



(a) The measures with the weak setting. (b) The measures with the strong setting.

**Fig. 5.** The quality measures comparison between the four variants of $TLKC$ and $TLKC$-$EXT$, the original results, and the baseline methods for Sepsis-Cases.

aspects of the score, i.e., $\alpha = 0.5$ and $\beta = 0.5$. In Sepsis-Cases, “diagnose" and “age" are considered as the sensitive case attribute. The numerical attributes are converted to categorical attributes using *boxplots* such that all the values greater than the *upper quartile* are categorized as *high*, the values less than the *lower quartile* are categorized as *low*, and the values in between are categorized as *middle*. Note that the confidence value $C$ should not be greater than 0.5, i.e., there are at least two different sensitive values for a victim case. To show and interpret the results of experiments, we focus on specific *strong* and *weak* settings. We use $T = minutes$, $L = 2$, $K = 20$, and $C = 0.5$ as the weak setting, and $T = minutes$, $L = 6$, $K = 60$, and $C = 0.2$ as the strong setting. Note that in the experiments, $TLKC$ refers to the algorithm presented in [5] which has been extended here w.r.t. the different perspectives, i.e., Algorithm 1, and $TLKC$-$EXT$ refers to Algorithm 2.

### 6.2. Control-flow perspective

In this subsection, we evaluate the effect of applying the extended $TLKC$-privacy on the *result utility* and *data utility* with respect to the control-flow perspective. We perform the control-flow perspective analysis for both event logs.

#### 6.2.1. Result utility

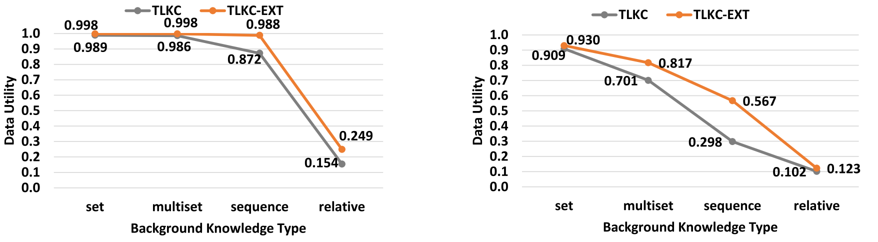
As mentioned, for the result utility analysis of the control-flow perspective, we focus on *process discovery*. The main goal is to find out *how accurately the discovered process model from a privacy-aware event log capture the behavior of the original event log*. To this end, we first discover a process model $M'$ from the privacy-aware event log $EL'$. Then, for $M'$, we calculate *fitness*, *precision*, and *f1-score*, as some model quality measures, w.r.t. the original event log $EL$.

*Fitness* quantifies the extent to which the discovered model can reproduce the traces recorded in the event log [18]. *Precision* quantifies the fraction of the traces allowed by the model which is not seen in the event log [19], and *f1-score* combines the fitness and precision $f1\text{-}score = {}^{2\times precision\times fitness}/_{precision+fitness}$. For process discovery, we use the *inductive miner infrequent* algorithm [20] with the default parameters (noise threshold 0.2). Fig. 5 shows the results of experiments for the quality measures. We consider four variants of our privacy preservation technique based on the introduced types of background knowledge where the attribute is *activity*, i.e., $bk_{set,at}$, $bk_{mult,at}$, $bk_{seq,at}$, and $bk_{rel,at}$. Note that applying privacy preservation techniques may improve some quality measures. However, the aim is to provide as similar results as possible to the original ones and not to improve the quality of discovered models. Therefore, we include the results from the original event log to compare the proximity of the values.

Figs. 5(a) and 5(b) show how the mentioned quality measures are affected by applying our method with the weak and strong settings (for $TLKC$, we set $\Theta = 0.5$). We compare the measures with the results from the original process model and the introduced baseline methods. If we only consider the quality measures, *Baseline*-2 should be marked as the best one, since it results in better *f1-score* values. However, the baseline methods remove more events from the original event log. Consequently, the corresponding privacy-aware event logs contain significantly less behavior compared to the original event log, and the resulting models have high *precision* and *f1-score*. The result utility analyses show that the extended version of the $TLKC$-privacy leads to the more similar results to the original ones, specifically for the *set* and *multiset* types of background knowledge. However, the results obtained based on the *relative* type of background knowledge have a worse *fitness* value which is not surprising regarding the assumed background knowledge which is considerably strong, at the same time, difficult to achieve in reality. Note that the baseline methods do not protect event data against the *attribute linkage* attack and provide weaker privacy guarantees.

#### 6.2.2. Data utility

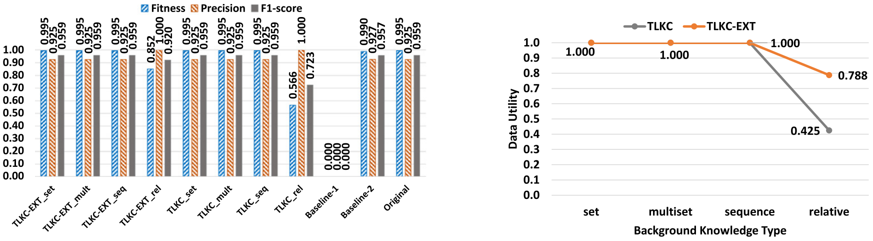
For the data utility analysis, we utilize the *earth mover's distance*, as proposed in [8]. The *earth mover's distance* describes the distance between two distributions [21]. In an analogy, given two piles of earth, it describes the effort required to transform one pile into the other. Assuming $EL$ as the original event log, $EL'$ as a privacy-aware event log, and $ps \in \mathcal{PS}$ as the perspective of analysis. The data utility is calculated as follows: $du(EL, EL') = 1 - \min_{r\in\mathcal{RA}} ul(r, \overline{EL}_{ps}, \overline{EL'}_{ps})$ where $ul(r, \overline{EL}_{ps}, \overline{EL'}_{ps})$ is the distance between the traces of two event logs projected on the given perspective. Note that $r \in \mathcal{RA}$ is used as a reallocation function, and *normalized edit distance* (Levenshtein) [22] is used to calculate the distance between variants. It should also be noted that for the control-flow $ps = \mathcal{A}$.



(a) The data utility with the weak setting.

(b) The data utility with the strong setting.

**Fig. 6.** The data utility comparison between $TLKC$ and $TLKC\text{-}EXT$ which provide the same privacy guarantees for the Sepsis-Cases event log.



(a) The quality measures comparison between the four variants of $TLKC$ and $TLKC\text{-}EXT$, the original results, and the baseline methods.

(b) The data utility comparison between $TLKC$ and $TLKC\text{-}EXT$.

**Fig. 7.** The data and result utility analyses for BPIC-2012-APP considering the strong setting.

Fig. 6 shows the results of data utility analysis where we compare $TLKC$ and $TLKC\text{-}EXT$ which provide the same privacy guarantees. As can be seen, for the weak privacy setting, the data utility results are similar, and $TLKC\text{-}EXT$ performs slightly better for the stronger types of background knowledge. For the strong privacy setting, $TLKC\text{-}EXT$ performs considerably better for the *multiset* and *sequence* types of background knowledge. Comparing the data utility analysis with the result utility analysis shows that the model quality measures alone cannot precisely evaluate the effectiveness of the privacy preservation techniques. For example, in the result utility analysis, for both weak and strong setting, $TLKC\text{-}EXT$ results in an acceptable *f1-score* value. However, the data utility analysis shows that the utility loss is indeed high for this type of background knowledge.

As already mentioned, the Sepsis-Cases event log is a significantly challenging dataset for the privacy preservation techniques due to the high uniqueness of variants. To show the effectiveness of our privacy preservation technique on other event logs, we perform the same type of analyses for BPIC-2012-APP considering only the strong setting. Fig. 7 shows that both data and result utility are high even for the strong types of background knowledge.

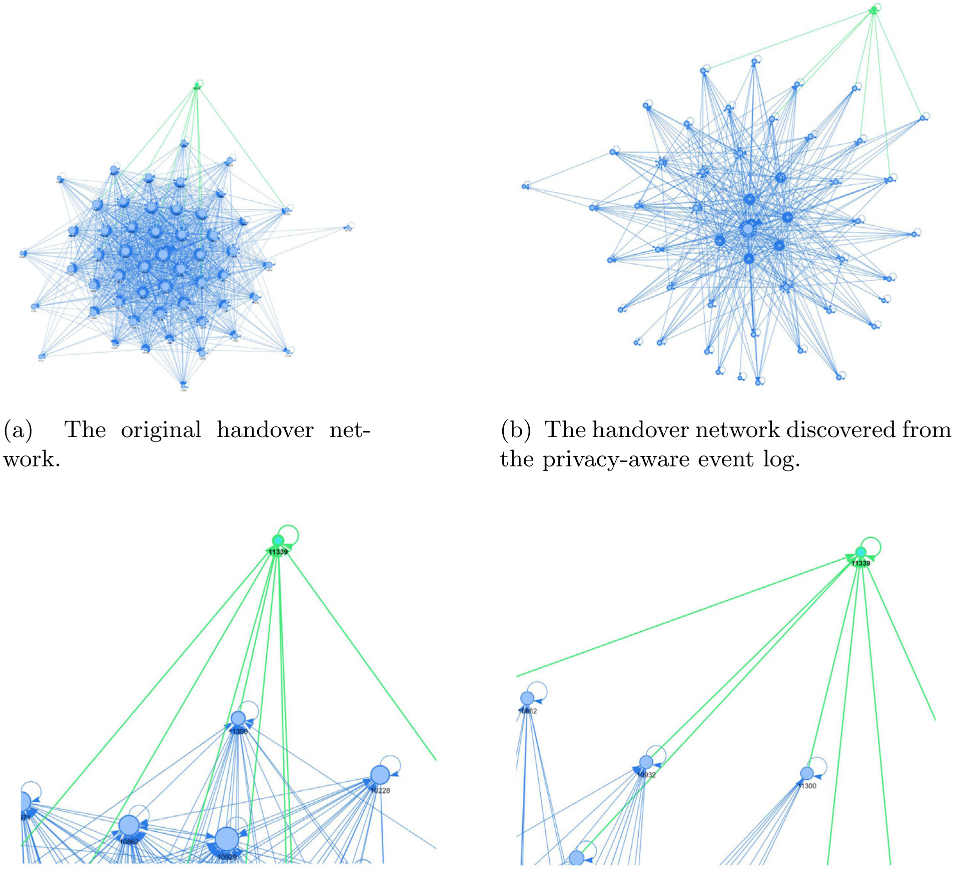
### *6.3. Organizational perspective*

In this subsection, we evaluate the effect of applying the extended $TLKC$-privacy on the *result* and *data* utility of the organizational perspective. The experiments of this perspective are done on BPIC-2012-APP which includes *resource* information.

#### *6.3.1. Result utility*

For the result utility analysis of organizational perspective, we focus on the *social network discovery* techniques. There are different methods for discovering social networks from event logs such as *causality-based*, *joint activities*, *joint cases*, etc [23]. Here, we focus on the *handover* technique which is causality-based. This technique monitors for individual cases how work moves from resource to resource, i.e., there is a *handover* relation from individual $r_1$ to individual $r_2$, if there are two subsequent activities where the first is performed by $r_1$ and the second is performed by $r_2$.

Fig. 8 shows the handover networks discovered from the original event log and a privacy-aware event log when the relation threshold is 0, i.e., all the handovers. The privacy-aware event log was obtained using the $TLKC\text{-}EXT$ privacy preservation technique with the strong setting and *set* as the type of background knowledge. As expected, the density of the network discovered from the privacy-aware event log is less than the original handover network. However, by focusing on some specific nodes, one can



(a) The original handover network.

(b) The handover network discovered from the privacy-aware event log.

(c) The relations of the resource 11339 in the original handover network. The node has 8 input links and no output link, except the self-loop.

(d) The relations of the resource 11339 in the handover network resulting from the privacy-aware event log. The node has 7 input links and no output link, except the self-loop.

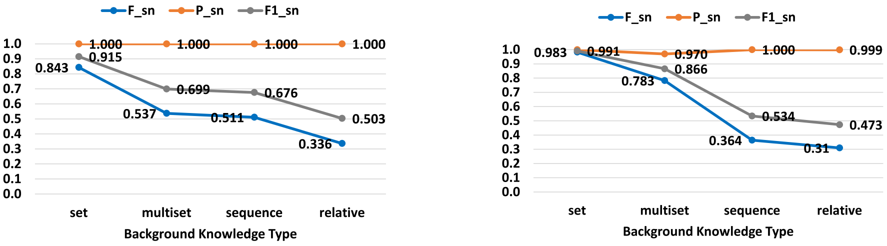
**Fig. 8.** The handover networks discovered from the original event log and a privacy-aware event log for BPIC-2012-APP. The privacy-aware event log was obtained using $TLKC\text{-}EXT$ with the strong setting and *set* as the type of background knowledge.

see that basic concepts are preserved. For example, node 11339 in the original handover network has the following set of input links $\{11302, 11003, 11300, 11121, 11122, 11180, 10932, 10861\}$ and no output link (excluding the self-loop), and in the network discovered from the privacy-aware event log, only the input link from node 11121 is removed.

To quantify the similarity of social networks resulting from an original and a privacy-aware event log, we use a set of measures similar to the quality measure of process models, i.e., *fitness*, *precision*, and *f1-score*. Consider $SN = (R_{EL}, DF_{R}^{EL})$ and $SN' = (R_{EL'}, DF_{R}^{EL'})$ as the handover social networks obtained from an original event log and its corresponding privacy-aware event log, respectively. Since both $TLKC$ and $TLKC\text{-}EXT$ provide privacy guarantees by removing events, the vertices of $SN'$ is a subset of vertices in $SN$, i.e., $R_{EL} \subseteq R_{EL'}$. However, the set of edges in $SN'$ is not necessarily a subset of edges in $SN$, i.e., $SN'$ is not necessarily a subgraph of $SN$. The following equations are used to compute the *fitness* ($F_{sn}$) and the *precision* ($P_{sn}$) for handover networks. The *f1-score* for handover networks ($F1_{sn}$) is the harmonic mean of $F_{sn}$ and $P_{sn}$.

$$F_{sn} = \frac{\sum\limits_{(x,y)\in DF_{R}^{EL} \cap DF_{R}^{EL'}} |x >_{R}^{EL'} y|}{\sum\limits_{(x,y)\in DF_{R}^{EL}} |x >_{R}^{EL} y|}$$

$$P_{sn} = \frac{|(R_{EL} \times R_{EL}) \setminus DF_{R}^{EL} \cap (R_{EL} \times R_{EL}) \setminus DF_{R}^{EL'}|}{|(R_{EL} \times R_{EL}) \setminus DF_{R}^{EL}|}$$



(a) The handover social network comparison for the graphs obtained from the BPIC-2012-APP event log.

(b) The handover social network comparison for the graphs obtained from the BPIC-2017-APP event log.

**Fig. 9.** The social network comparison based on *fitness* ($F_{sn}$), *precision* ($P_{sn}$), and *f1-score* ($F1_{sn}$). The privacy preservation technique is $TLKC\text{-}EXT$ with the strong setting.

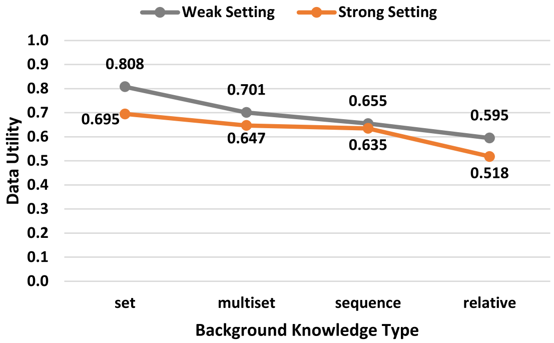


**Fig. 10.** The data utility analysis of organizational perspective for BPIC-2012-APP with the strong and weak settings considering different types of background knowledge, and using $TLKC\text{-}EXT$ as the privacy preservation technique.

Fig. 9 shows the similarity of handover social networks after applying the $TLKC\text{-}EXT$ privacy model with the strong setting to BPIC-2012-APP and BPIC-2017-APP. The *precision* is high for all the types of background knowledge, i.e., the handover social networks obtained from the privacy-aware event logs often do not contain edges that do not exist in the original network. The *fitness* decreases when the background knowledge becomes stronger, i.e., the $SN'$s obtained based on stronger assumptions for the background knowledge have fewer edges in common with the $SN$.

### 6.3.2. Data utility

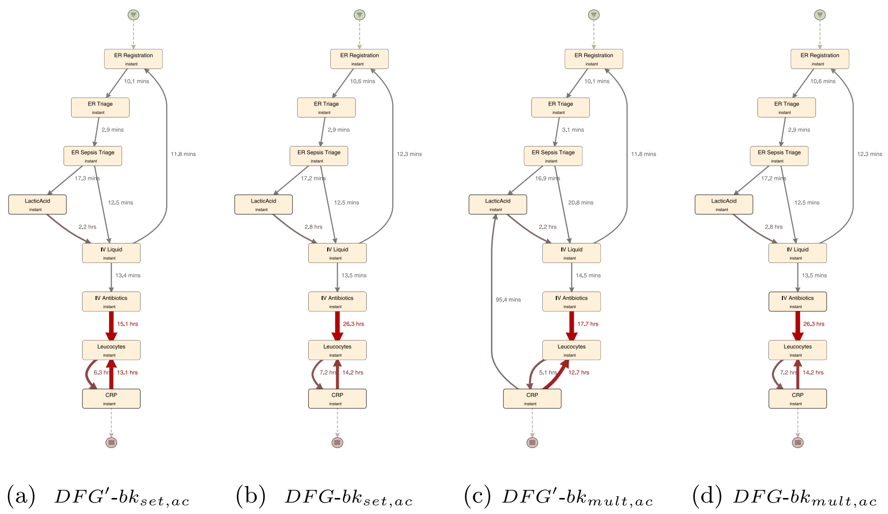
For the data utility analysis of the organizational perspective, we utilize the earth mover's distance, similar to the data utility analysis of the control-flow perspective. Here, the perspective is resource, i.e., $ps = \mathcal{R}$. Fig. 10 shows the results for the data utility analysis for BPIC-2012-APP considering different types of background knowledge using $TLKC\text{-}EXT$ as the privacy preservation technique. As can be seen, the data utility reservation is above 0.5 even for the strong types of background knowledge.

## 6.4. Time perspective

We evaluate the effect on performance analyses by analyzing the bottlenecks w.r.t. the mean duration of cases between activities. Since the privacy preservation techniques may remove some activities, we cannot compare the bottlenecks in the original process model with the bottlenecks in a process model discovered from a privacy-aware event log. Therefore, we first project the original event log on the activities existing in the privacy-aware event log. Then, we discover a performance-annotated directly follows graph $DFG$ from the projected event log and compare it with the performance-annotated directly follows graph $DFG'$ from the privacy-aware event log. A DFG is a graph where the nodes represent activities and the arcs represent causalities. Two activities $a_1$ and $a_2$ are connected by an arrow when $a_1$ is frequently followed by $a_2$ [24].

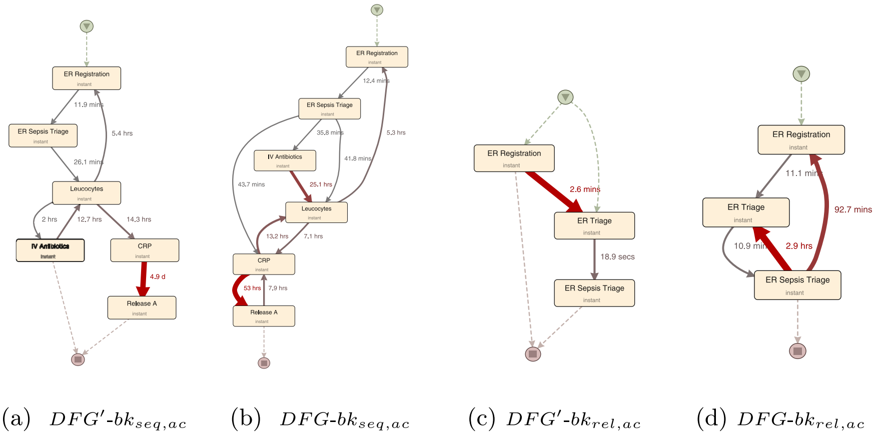
Fig. 11 (*set* and *multiset* as the types of background knowledge) and Fig. 12 (*sequence* and *relative* as the types of background knowledge) show the results for Sepsis-Cases using $TLKC\text{-}EXT$ with the strong setting.[2] As can be seen, the bottlenecks in $DFG$ and $DFG'$ are the same for all the variants, except for DFGs discovered using $bk_{rel,ac}$, where the assumed background knowledge

[2] The results provided by Disco (https://fluxicon.com/disco/) with the sliders set to the maximal number of activities and the minimal paths.



(a) $DFG'$-$bk_{set,ac}$ (b) $DFG$-$bk_{set,ac}$ (c) $DFG'$-$bk_{mult,ac}$ (d) $DFG$-$bk_{mult,ac}$

**Fig. 11.** The performance-annotated DFGs from the projected event log ($DFG$) and an anonymized event log ($DFG'$) for Sepsis-Cases using $TLKC$-$EXT$ with the strong setting and the specified types of background knowledge.



(a) $DFG'$-$bk_{seq,ac}$ (b) $DFG$-$bk_{seq,ac}$ (c) $DFG'$-$bk_{rel,ac}$ (d) $DFG$-$bk_{rel,ac}$

**Fig. 12.** The performance-annotated DFGs from the projected event log ($DFG$) and an anonymized event log ($DFG'$) for Sepsis-Cases using $TLKC$-$EXT$ with the strong setting and the specified types of background knowledge.

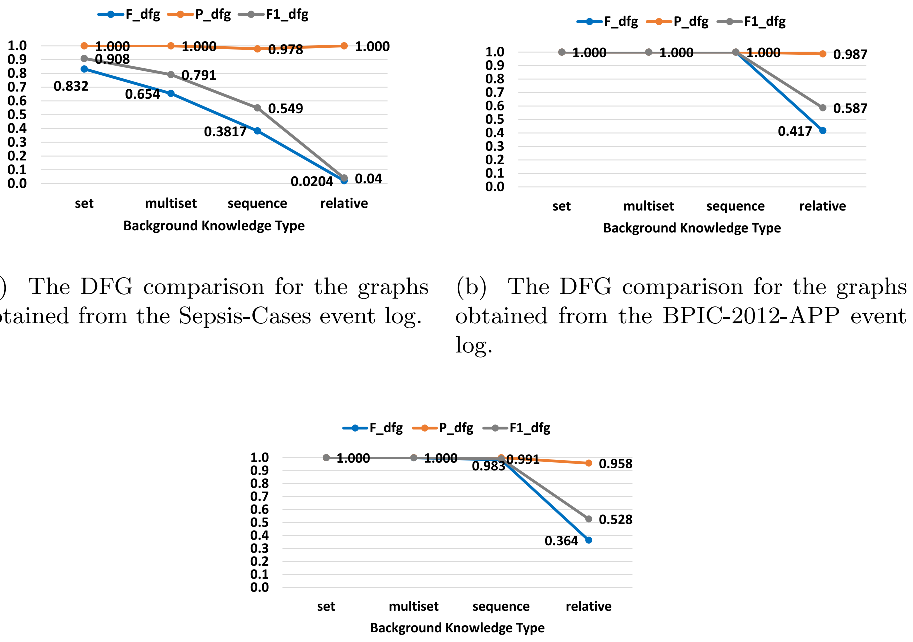
is *relative* which is significantly strong and our data utility analysis in Section 6.2 demonstrated a low data utility preservation for Sepsis-Cases. Note that the mean duration of the cases are different in $DFG$ and $DFG'$ due to the relative timestamps in the privacy-aware event logs.

We also evaluate the similarity of the Directly Follows Graphs (DFGs) resulting from an original event log and its corresponding privacy-aware event log. Let $DFG{=}(A_{EL}, DF^{EL}_{\mathcal{A}})$ and $DFG'{=}(A_{EL'}, DF^{EL'}_{\mathcal{A}})$ be the directly follows graphs obtained from an original and its corresponding privacy-aware event logs, respectively. To compare these graphs, we follow the same approach taken for quantifying the similarity of social networks. The *fitness* ($F_{dfg}$) and *precision* ($P_{dfg}$) for DFGs are calculated as follows:

$$F_{dfg}{=}\frac{\sum\limits_{(x,y)\in DF^{EL}_{\mathcal{A}}\cap DF^{EL'}_{\mathcal{A}}} |x >^{EL'}_{\mathcal{A}} y|}{\sum\limits_{(x,y)\in DF^{EL}_{\mathcal{A}}} |x >^{EL}_{\mathcal{A}} y|}$$

$$P_{dfg}{=}\frac{|(A_{EL}\times A_{EL})\setminus DF^{EL}_{\mathcal{A}} \cap (A_{EL}\times A_{EL})\setminus DF^{EL'}_{\mathcal{A}}|}{|(A_{EL}\times A_{EL})\setminus DF^{EL}_{\mathcal{A}}|}$$

The *f1-score* for DFGs ($F1_{dfg}$) is the harmonic mean of $F_{dfg}$ and $P_{dfg}$. Fig. 13 shows the similarity of DFGs after applying the $TLKC$-$EXT$ privacy model with the strong setting for Sepsis-Cases, BPIC-2012-APP, and BPIC-2017-APP. The *precision* is always high, i.e., the DFGs obtained from the privacy-aware event logs often do not contain directly follows relations that do not exist in the



(a) The DFG comparison for the graphs obtained from the Sepsis-Cases event log.

(b) The DFG comparison for the graphs obtained from the BPIC-2012-APP event log.

(c) The DFG comparison for the graphs obtained from the BPIC-2017-APP event log.

**Fig. 13.** The DFG comparison based on *fitness* ($F_{dfg}$), *precision* ($P_{dfg}$), and *f1-score* ($F1_{dfg}$). The privacy preservation technique is $TLKC\text{-}EXT$ with the strong setting.

original DFG. For the Sepsis-Cases event log, the *fitness* decreases when the background knowledge becomes stronger, i.e., the $DFG'$s obtained based on stronger assumptions for the background knowledge preserve fewer directly follows relations of the original DFG. The *fitness* for the BPIC event logs only drops for the relative type of background knowledge which is considerably strong.

## 7. Related work

In process mining, the research field of confidentiality and privacy is recently receiving more attention. In this section, we list the work that has been done in this research field which is rapidly growing. In [25], *Responsible Process Mining* (RPM) is introduced as the sub-discipline which focuses on possible negative side-effects of applying process mining where *Fairness*, *Accuracy*, *Confidentiality*, and *Transparency* (FACT) are considered as the concerns. In [26], the authors provide an overview of privacy challenges in process mining in human-centered industrial environments. In [27], a method to secure event logs for performing process discovery by the Alpha algorithm is proposed. In [28], the aim is to propose a solution which allows the outsourcing of process mining while ensuring confidentiality. In [29], the goal is to propose a privacy-preserving system design for process mining, where a user-centered view is considered to track personal data. In [30,31], a framework is proposed which provides a generic scheme for confidentiality in process mining. In [32], the authors introduce a privacy-preserving method for discovering roles from event logs. In [33], the authors consider a cross-organizational process discovery context and share public process model fragments as safe intermediates. In [13], the authors apply $k$-anonymity and $t$-closeness on event logs to preserve the privacy of *resources*. In [16,34], the notion of *differential privacy* is employed to preserve the privacy of event logs. In [5], the $TLKC$-privacy is introduced to cope with high variability issues in event logs for applying group-based anonymization techniques. In [35], a uniformization-based approach is proposed to preserve individuals' privacy in process mining. In [36], a secure multi-party computation solution is introduces for preserving privacy in an inter-organizational setting for *process discovery*. In [37], the data privacy and utility requirements for healthcare event data are analyzed. In [38], the authors propose a privacy extension for the XES standard[3] to manage privacy metadata. In [39], the authors propose a measure to evaluate the re-identification risk of event logs. Also, in [8], a general privacy quantification framework, and some measures are introduced to evaluate the effectiveness of privacy preservation techniques. Some tools are also provided for applying the state-of-the-art privacy preservation techniques in the field of process mining such as *PPDP-PM* [40], *ELPaaS* [41], and *Shareprom* [42].

[3] https://xes-standard.org/.

## 8. Conclusion

In this paper, we discussed the challenges regarding directly applying traditional group-based privacy preservation techniques to event logs. We discussed the *linkage attacks* and provided formal models of the possible attacks based on the different types of background knowledge. We extended the $TLKC$-privacy for process mining to cover all the main perspectives of process mining. The data utility preservation aspect of the $TLKC$-privacy was improved by introducing a new utility measure. Moreover, a new score equation was proposed to generate normalized scores for the events that need to be removed. The new equation for the score also provides *privacy gain* and *utility loss* coefficients that can be adjusted by users. Obviously, the extended version of the $TLKC$-privacy inherits all the characteristics of the main approach. Namely, it counteracts both the *case linkage* and the *attribute linkage* attacks. It generalizes several privacy preservation techniques including $k$-anonymity, confidence bounding, ($\alpha$, $k$)-anonymity, and $l$-diversity. It also provides interpretable and tunable parameters.

Similar to the main approach, we implemented four variants of the extended version with respect to the four different types of background knowledge and considering all the main perspectives. The effectiveness of different variants in different perspectives was evaluated based on real-life event logs. Both *data* and *result* utility were analyzed to evaluate the effectiveness. Overall more than 1000 experiments were performed for different types of background knowledge considering different perspectives, and the results were given for a weak and a strong setting. Our experiments showed that the extended $TLKC$-privacy performs better than the previous version considering the data utility preservation aspect. However, in the event logs with the high ratio of unique traces, when the assumed type of background knowledge is very specific, e.g., *relative*, the group-based privacy preservation techniques may not be able to preserve the general data utility, and this negative effect cannot be observed by only result utility analyses.

## CRediT authorship contribution statement

**Majid Rafiei:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Wil M.P. van der Aalst:** Conceptualization, Methodology, Validation, Formal analysis, Resources, Data curation, Writing - review & editing, Visualization, Supervision, Project administration.

## 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.

## Acknowledgment

Funded under the Excellence Strategy of the Federal Government and the Länder, Germany. We also thank the Alexander von Humboldt (AvH) Stiftung for supporting our research.

## References

[1] W.M.P. van der Aalst, Process Mining - Data Science in Action, second ed., Springer, 2016, http://dx.doi.org/10.1007/978-3-662-49851-4.

[2] W.G. Voss, European union data privacy law reform: General data protection regulation, privacy shield, and the right to delisting, Bus. Lawyer 72 (1) (2016).

[3] L. Sweeney, K-anonymity: A model for protecting privacy, Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 10 (05) (2002) 557–570.

[4] A. Machanavajjhala, J. Gehrke, D. Kifer, M. Venkitasubramaniam, L-diversity: Privacy beyond k-anonymity, in: 22nd International Conference on Data Engineering, ICDE'06, IEEE, 2006, p. 24.

[5] M. Rafiei, M. Wagner, W.M.P. van der Aalst, TLKC-privacy model for process mining, in: F. Dalpiaz, J. Zdravkovic, P. Loucopoulos (Eds.), Research Challenges in Information Science - 14th International Conference, RCIS 2020, Limassol, Cyprus, September 23–25, 2020, Proceedings, in: Lecture Notes in Business Information Processing, vol. 385, Springer, 2020, pp. 398–416, http://dx.doi.org/10.1007/978-3-030-50316-1_24.

[6] K. Wang, B.C.M. Fung, P.S. Yu, Handicapping attacker's confidence: an alternative to $k$ -anonymization, Knowl. Inf. Syst. 11 (3) (2007) 345–368, http://dx.doi.org/10.1007/s10115-006-0035-5.

[7] R.C. Wong, J. Li, A.W. Fu, K. Wang, (Alpha, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing, in: T. Eliassi-Rad, L.H. Ungar, M. Craven, D. Gunopulos (Eds.), Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20–23, 2006, ACM, 2006, pp. 754–759, http://dx.doi.org/10.1145/1150402.1150499.

[8] M. Rafiei, W.M.P. van der Aalst, Towards quantifying privacy in process mining, in: International Conference on Process Mining - ICPM 2020 International Workshops, Padua, Italy, October 4–9, 2020, 2020.

[9] N. Li, T. Li, S. Venkatasubramanian, T-closeness: Privacy beyond k-anonymity and l-diversity, in: R. Chirkova, A. Dogac, M.T. Özsu, T.K. Sellis (Eds.), Proceedings of the 23rd International Conference on Data Engineering, ICDE 2007, the Marmara Hotel, Istanbul, Turkey, April 15–20, 2007, IEEE Computer Society, 2007, pp. 106–115, http://dx.doi.org/10.1109/ICDE.2007.367856.

[10] C.C. Aggarwal, S.Y. Philip, Privacy-Preserving Data Mining: Models and Algorithms, Springer Science & Business Media, 2008.

[11] C.C. Aggarwal, On k-anonymity and the curse of dimensionality, in: K. Böhm, C.S. Jensen, L.M. Haas, M.L. Kersten, P. Larson, B.C. Ooi (Eds.), Proceedings of the 31st International Conference on Very Large Data Bases, Trondheim, Norway, August 30 - September 2, 2005, ACM, 2005, pp. 901–909.

[12] J. Gehrke, Models and methods for privacy-preserving data analysis and publishing, in: L. Liu, A. Reuter, K. Whang, J. Zhang (Eds.), Proceedings of the 22nd International Conference on Data Engineering, ICDE 2006, 3–8 April 2006, Atlanta, GA, USA, IEEE Computer Society, 2006, p. 105, http://dx.doi.org/10.1109/ICDE.2006.100.

[13] S.A. Fahrenkrog-Petersen, H. van der Aa, M. Weidlich, PRETSA: event log sanitization for privacy-aware process discovery, in: International Conference on Process Mining, ICPM 2019, Aachen, Germany, June 24–26, 2019, IEEE, 2019, pp. 1–8, http://dx.doi.org/10.1109/ICPM.2019.00012.

[14] F. Mannhardt, Sepsis cases-event log. Eindhoven university of technology, 2016, https://doi.org/10.4121/uuid:915d2bfb-7e84-49ad-a286-dc35f063a460.

[15] B.F. Van Dongen, BPIC 2012. Eindhoven university of technology, 2012, http://dx.doi.org/10.4121/uuid:3926db30-f712-4394-aebc-75976070e91f.

[16] F. Mannhardt, A. Koschmider, N. Baracaldo, M. Weidlich, J. Michael, Privacy-preserving process mining - differential privacy for event logs, Bus. Inf. Syst. Eng. 61 (5) (2019) 595–614, http://dx.doi.org/10.1007/s12599-019-00613-3.

[17] B.F. Van Dongen, BPIC 2017. Eindhoven university of technology, 2017, https://doi.org/10.4121/uuid:5f3067df-f10b-45da-b98b-86ae4c7a310b.

[18] A. Adriansyah, B.F. van Dongen, W.M.P. van der Aalst, Conformance checking using cost-based fitness analysis, in: Proceedings of the 15th IEEE International Enterprise Distributed Object Computing Conference, EDOC, 2011, pp. 55–64.

[19] A. Adriansyah, J. Munoz-Gama, J. Carmona, B.F. van Dongen, W.M.P. van der Aalst, Measuring precision of modeled behavior, Inf. Syst. E-Business Management 13 (1) (2015) 37–67.

[20] S.J.J. Leemans, D. Fahland, W.M.P. van der Aalst, Discovering block-structured process models from event logs containing infrequent behaviour, in: Business Process Management Workshops - BPM International Workshops, 2013, pp. 66–78.

[21] L. Rüschendorf, The Wasserstein distance and approximation theorems, Probab. Theory Related Fields 70 (1) (1985) 117–129.

[22] V.I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals, in: Soviet Physics Doklady, vol. 10, 1966, pp. 707–710.

[23] W.M.P. van der Aalst, H.A. Reijers, M. Song, Discovering social networks from event logs, Comput. Support. Coop. Work (CSCW) 14 (6) (2005) 549–593.

[24] S.J. Leemans, D. Fahland, W.M.P. Aalstvan der Aalst, Scalable process discovery and conformance checking, Softw. Syst. Model. 17 (2) (2018) 599–631.

[25] W.M.P. van der Aalst, Responsible data science: Using event data in a "people friendly" manner, in: S. Hammoudi, L.A. Maciaszek, M. Missikoff, O. Camp, J. Cordeiro (Eds.), Enterprise Information Systems - 18th International Conference, ICEIS 2016, Rome, Italy, April 25–28, 2016, Revised Selected Papers, in: Lecture Notes in Business Information Processing, vol .291, Springer, 2016, pp. 3–28, http://dx.doi.org/10.1007/978-3-319-62386-3_1.

[26] F. Mannhardt, S.A. Petersen, M.F. Oliveira, Privacy challenges for process mining in human-centered industrial environments, in: 14th International Conference on Intelligent Environments, IE 2018, Roma, Italy, June 25–28, 2018, IEEE, 2018, pp. 64–71, http://dx.doi.org/10.1109/IE.2018.00017.

[27] G. Tillem, Z. Erkin, R.L. Lagendijk, Privacy-preserving alpha algorithm for software analysis, in: 37th WIC Symposium on Information Theory in the Benelux/6th WIC/IEEE SP, 2016.

[28] A. Burattin, M. Conti, D. Turato, Toward an anonymous process mining, in: Future Internet of Things and Cloud (FiCloud), 2015 3rd International Conference on, IEEE, 2015, pp. 58–63.

[29] J. Michael, A. Koschmider, F. Mannhardt, N. Baracaldo, B. Rumpe, User-centered and privacy-driven process mining system design for IoT, in: C. Cappiello, M. Ruiz (Eds.), Information Systems Engineering in Responsible Information Systems - CAiSE Forum 2019, Rome, Italy, June 3–7, 2019, Proceedings, in: Lecture Notes in Business Information Processing, vol. 350, Springer, 2019, pp. 194–206, http://dx.doi.org/10.1007/978-3-030-21297-1_17.

[30] M. Rafiei, L. von Waldthausen, W.M.P. van der Aalst, Ensuring confidentiality in process mining, in: P. Ceravolo, M.T.G. López, M. van Keulen (Eds.), Proceedings of the 8th International Symposium on Data-Driven Process Discovery and Analysis (SIMPDA 2018), Seville, Spain, December 13–14, 2018, in: CEUR Workshop Proceedings, vol. 2270, CEUR-WS.org, 2018, pp. 3–17.

[31] M. Rafiei, L. von Waldthausen, W.M.P. van der Aalst, Supporting confidentiality in process mining using abstraction and encryption, in: P. Ceravolo, M. van Keulen, M.T.G. López (Eds.), Data-Driven Process Discovery and Analysis - 8th IFIP WG 2.6 International Symposium, SIMPDA 2018, Seville, Spain, December 13–14, 2018, and 9th International Symposium, SIMPDA 2019, Bled, Slovenia, September 8, 2019, Revised Selected Papers, in: Lecture Notes in Business Information Processing, vol. 379, Springer, 2019, pp. 101–123, http://dx.doi.org/10.1007/978-3-030-46633-6_6.

[32] M. Rafiei, W.M.P. van der Aalst, Mining roles from event logs while preserving privacy, in: C.D. Francescomarino, R.M. Dijkman, U. Zdun (Eds.), Business Process Management Workshops - BPM 2019 International Workshops, Vienna, Austria, September 1–6, 2019, Revised Selected Papers, in: Lecture Notes in Business Information Processing, vol. 362, Springer, 2019, pp. 676–689, http://dx.doi.org/10.1007/978-3-030-37453-2_54.

[33] C. Liu, H. Duan, Q. Zeng, M. Zhou, F. Lu, J. Cheng, Towards comprehensive support for privacy preservation cross-organization business process mining, IEEE Trans. Serv. Comput. 12 (4) (2019) 639–653, http://dx.doi.org/10.1109/TSC.2016.2617331.

[34] S.A. Fahrenkrog-Petersen, H. van der Aa, M. Weidlich, PRIPEL: privacy-preserving event log publishing including contextual information, in: D. Fahland, C. Ghidini, J. Becker, M. Dumas (Eds.), Business Process Management - 18th International Conference, BPM 2020, Seville, Spain, September 13–18, 2020, Proceedings, in: Lecture Notes in Computer Science, vol. 12168, Springer, 2020, pp. 111–128, http://dx.doi.org/10.1007/978-3-030-58666-9_7.

[35] E. Batista, A. Solanas, A uniformization-based approach to preserve individuals' privacy during process mining analyses, Peer-To-Peer Netw. Appl. (2021) 1–20.

[36] G. Elkoumy, S.A. Fahrenkrog-Petersen, M. Dumas, P. Laud, A. Pankova, M. Weidlich, Secure multi-party computation for inter-organizational process mining, in: S. Nurcan, I. Reinhartz-Berger, P. Soffer, J. Zdravkovic (Eds.), Enterprise, Business-Process and Information Systems Modeling - 21st International Conference, BPMDS 2020, 25th International Conference, EMMSAD 2020, Held At CAiSE 2020, Grenoble, France, June 8–9, 2020, Proceedings, in: Lecture Notes in Business Information Processing, vol. 387, Springer, 2020, pp. 166–181, http://dx.doi.org/10.1007/978-3-030-49418-6_11.

[37] A. Pika, M.T. Wynn, S. Budiono, A.H.M. ter Hofstede, W.M.P. van der Aalst, H.A. Reijers, Towards privacy-preserving process mining in healthcare, in: C.D. Francescomarino, R.M. Dijkman, U. Zdun (Eds.), Business Process Management Workshops - BPM 2019 International Workshops, Vienna, Austria, September 1–6, 2019, Revised Selected Papers, in: Lecture Notes in Business Information Processing, vol. 362, Springer, 2019, pp. 483–495, http://dx.doi.org/10.1007/978-3-030-37453-2_39.

[38] M. Rafiei, W.M.P. van der Aalst, Privacy-preserving data publishing in process mining, 2021, CoRR abs/2101.02627, https://arxiv.org/abs/2101.02627.

[39] S.N. von Voigt, S.A. Fahrenkrog-Petersen, D. Janssen, A. Koschmider, F. Tschorsch, F. Mannhardt, O. Landsiedel, M. Weidlich, Quantifying the re-identification risk of event logs for process mining - empiricial evaluation paper, in: Advanced Information Systems Engineering, CAiSE, 2020.

[40] M. Rafiei, W.M.P. van der Aalst, Practical aspect of privacy-preserving data publishing in process mining, 2020, CoRR abs/2009.11542, https://arxiv.org/abs/2009.11542.

[41] M. Bauer, S.A. Fahrenkrog-Petersen, A. Koschmider, F. Mannhardt, H. van der Aa, M. Weidlich, ELPaaS: Event log privacy as a service, in: Proceedings of the Dissertation Award, Doctoral Consortium, and Demonstration Track At BPM 2019, 2019.

[42] G. Elkoumy, S.A. Fahrenkrog-Petersen, M. Dumas, P. Laud, A. Pankova, M. Weidlich, Shareprom: A tool for privacy-preserving inter-organizational process mining, in: Proceedings of the Best Dissertation Award, Doctoral Consortium, and Demonstration & Resources Track At BPM 2020 Co-Located with the 18th International Conference on Business Process Management (BPM 2020), Sevilla, Spain, September 13–18, 2020, in: CEUR Workshop Proceedings, vol. 2673, CEUR-WS.org, 2020, pp. 72–76.

**Majid Rafiei** is a Scientific Assistant (Ph.D. candidate) at the Chair of Process and Data Science (PADS) - RWTH Aachen University. He is graduated as master of engineering in Electronic Commerce from Amirkabir University of Technology, Tehran. Currently he is working on Process Mining and specifically on Responsible Process Mining (RPM), where the aim is to use process mining with respect to Fairness, Accuracy, Confidentiality, and Transparency (FACT).

**Prof.dr.ir. Wil van der Aalst** is a full professor at RWTH Aachen University leading the Process and Data Science (PADS) group. He is also part-time affiliated with the Technische Universiteit Eindhoven (TU/e). Until December 2017, he was the scientific director of the Data Science Center Eindhoven (DSC/e) and led the Architecture of Information Systems group at TU/e. Since 2003, he holds a parttime position at Queensland University of Technology (QUT). Currently, he is also a distinguished fellow at Fondazione Bruno Kessler (FBK) in Trento and a member of the Board of Governors of Tilburg University.