Privacy Preservation
in process mining

majid.rafiei at
pads.rwth-aachen.de

Chair of Process and Data Science
RWTH AACHEN UNIVERSITY

© Majid Rafiei (RWTH Aachen University)

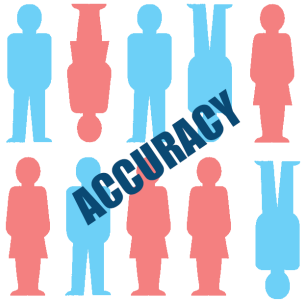
Responsible Process Mining

Making Data/Process Science Green!

© Majid Rafiei (RWTH Aachen University)

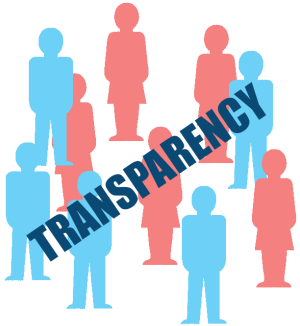
FACT Challenges

- How to avoid unfair conclusions even if they are true?



- How to answer questions with a guaranteed level of accuracy?

- How to answer questions without revealing secrets?



- How to clarify answers such that they become indisputable?

© Majid Rafiei (RWTH Aachen University)

Confidentiality in Process Mining

- Confidentiality in process mining is focused on two important issues:
  - Protecting organizations' sensitive data.
  - Protecting individuals' sensitive data (privacy).

© Majid Rafiei (RWTH Aachen University)

Privacy in Process Mining

- Where can the individual's data be included?
  - Cases: as process instances.
  - Resources: as activity performers.

Case	Activity	Timestamp	Resource
Patient1	Brain Surgery	…	Surgeon1
Patient1	Heart Surgery	…	Surgeon2
Patient3	Kidney Surgery	…	Surgeon3
Patient4	Brain Surgery	…	Surgeon4

© Majid Rafiei (RWTH Aachen University)

Privacy in Process Mining

- The worst case:
  - When the event data include direct personal identifiers.
  - Which is usually not the case.

Could be a real identifier

Case	Activity	Timestamp	Resource
009812	…	…	1230568
009526	…	…	1255980
009823	…	…	1356478
008798	…	…	1352587

Could be a real identifier

© Majid Rafiei (RWTH Aachen University)

Privacy in Process Mining

- Indirect personal data could be included!
  - Which is often the case.
  - Implicitly (based on the context), we know that individuals are doing activities and/or cases are individuals.

Case	Activity	Timestamp	Resource
Patient1	Brain Surgery	…	Surgeon1
Patient1	Heart Surgery	…	Surgeon2
Patient3	Kidney Surgery	…	Surgeon3
Patient4	Brain Surgery	…	Surgeon4

© Majid Rafiei (RWTH Aachen University)

Privacy Spectrum in Process Mining

Different Contexts

Privacy

Case	Activity	Timestamp	Resource
Car1	Welding	2018.01.01:10:00:00	Robot1
Car2	Painting	2018.01.01:10:05:00	Robot2

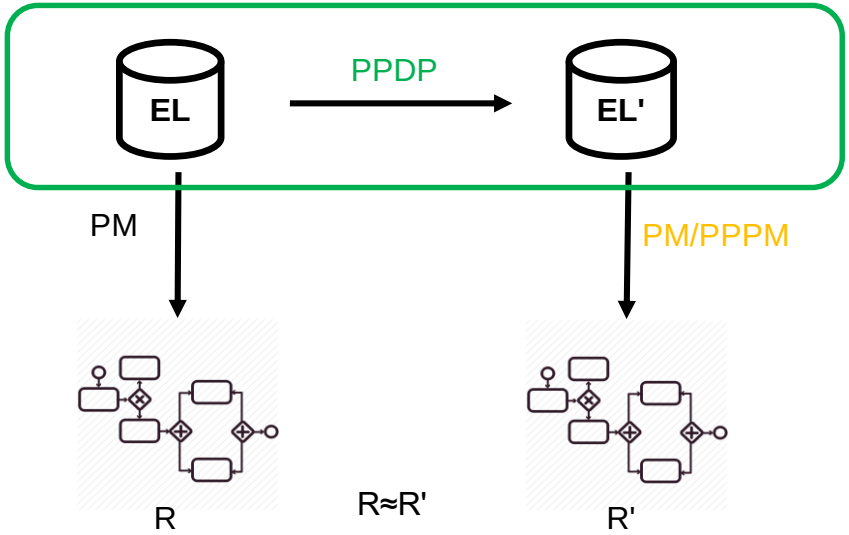
Neither “Case” nor “Resource” is individual.

Case	Activity	Timestamp	Resource
Jack	Brain Surgery	2018.01.01:10:00:00	Dr. Sue
Peter	Heart surgery	2018.01.01:10:05:00	Dr. Frank

Both “Case” and “Resource” are individuals.

© Majid Rafiei (RWTH Aachen University)

PPDP vs PPPM



- Privacy-Preserving Data Publishing (PPDP) aims to hide the identity and/or the sensitive data of record owners, assuming that sensitive data must be retained for data analysis.
- Privacy-Preserving Process Mining (PPPM) aims to extend traditional process mining algorithms to work with the data modified to mask sensitive information.

© Majid Rafiei (RWTH Aachen University)

From RPM to PPDP

RPM → FACT → C →

Responsible Process Mining

Fairness-Accuracy-Confidentiality-Transparency

- Individuals
- Organization

Individuals Privacy → PPDP

- PPDP
- PPPM

© Majid Rafiei (RWTH Aachen University)

PPDP in Databases

- In the most basic form of PPDP, the data holder has a table with the following form:

  D(Explicit_Identifier, Quasi_Identifier (QID), Sensitive_Attributes, Non-Sensitive Attributes)

- Where:
  - ***Explicit_Identifier*** is a set of attributes, containing information that explicitly identifies record owners such as social security number, name, etc.
  - ***Quasi_Identifier*** is a set of attributes that could potentially identify record owners.
  - ***Sensitive_Attributes*** consist of sensitive person-specific information such as disease, salary ,etc.
  - ***Non-Sensitive_Attributes*** contains all attributes that do not fall into the previous three categories.

- PPDP aims to hide the identity and/or the sensitive data of record owners, assuming that sensitive data must be retained for data analysis.

© Majid Rafiei (RWTH Aachen University)

PPDP in Databases

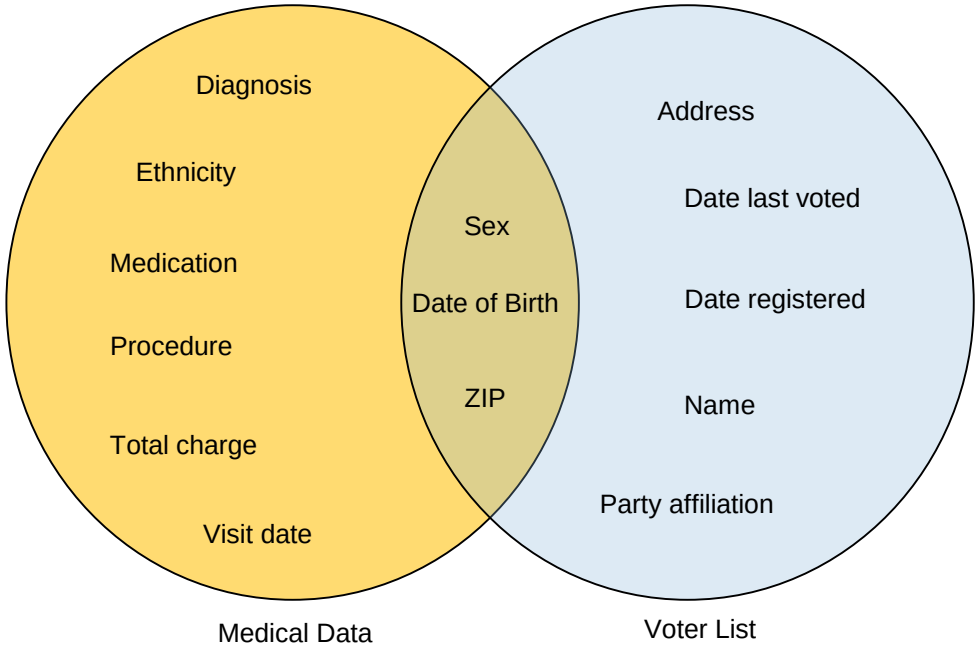
Explicit identifier | Quasi-identifiers | Sensitive attribute

Name	Age	Gender	State of domicile	Religion	Disease
Ramsha	22	Female	Tamil Nadu	Hindu	Gastric ulcer
Yadu	24	Female	Kerala	Hindu	Gastritis
Salima	25	Female	Tamil Nadu	Muslim	Flu
Sunny	25	Male	Karnataka	Muslim	Bronchitis
Joan	24	Female	Kerala	Christian	Heart Disease
Bahuksana	23	Male	Karnataka	Buddhist	Bronchitis
Rambha	19	Male	Kerala	Hindu	Flu
Kishor	24	Male	Karnataka	Hindu	Gastric ulcer
Johnson	17	Male	Kerala	Christian	Cancer
John	19	Male	Kerala	Christian	Flu

© Majid Rafiei (RWTH Aachen University)

PPDP in Databases

- *Linking attack* to re-identify the record owner based on the quasi-identifier



87% of the U.S. population had reported characteristics that made them unique based on only such quasi-identifiers.

Fung, B.C., Wang, K., Fu, A.W.C., Philip, S.Y.: Introduction to privacy-preserving data publishing: Concepts and techniques. Chapman and Hall/CRC (2010)

© Majid Rafiei (RWTH Aachen University)

Record and Attribute Linkage Attacks

- Record Linkage
  - How confidently records can be linked to a target victim, assuming that the victim's record is in the table.
- Attribute Linkage
  - How confidently sensitive attribute values can be linked to a target victim, assuming that the victim's record is in the table.

© Majid Rafiei (RWTH Aachen University)

PPDP in Databases

- To prevent linking attacks, using group-based anonymization approach, the data holder publishes an anonymous table in the following form:

T(QID', Sensitive_Attributes, Non-Sensitive Attributes)

- Where *QID'* is an anonymous version of the original QID obtained by applying anonymization operations such as:
  - Generalization
  - Suppression
  - Anatomization
  - Perturbation
  - …

© Majid Rafiei (RWTH Aachen University)

k-anonymity to Prevent Record Linkage Attack

- We define an equivalence class of an anonymized table to be a set of records that have the same values for the quasi-identifiers.
- k-anonymity requires that each equivalence class contains at least k records.

Quasi-identifiers

Sensitive attribute

	ZIP Code	Age	Disease
1	47677	29	Heart Disease
2	47602	22	Heart Disease
3	47678	27	Heart Disease
4	47905	43	Flu
5	47909	52	Heart Disease
6	47906	47	Cancer
7	47605	30	Heart Disease
8	47673	36	Cancer
9	47607	32	Cancer

3-anonymity

Equivalence classes include 3 records

	ZIP Code	Age	Disease
1	476**	2*	Heart Disease
2	476**	2*	Heart Disease
3	476**	2*	Heart Disease
4	4790*	≥ 40	Flu
5	4790*	≥ 40	Heart Disease
6	4790*	≥ 40	Cancer
7	476**	3*	Heart Disease
8	476**	3*	Cancer
9	476**	3*	Cancer

(partial) suppression

generalization

(partial) suppression

L. Sweeney. k-Anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557–570, 2002.

© Majid Rafiei (RWTH Aachen University)

l-diversity to Prevent Attribute Linkage Attack

- An equivalence class is said to have distinct l-diversity, if there are at least l different values for the sensitive attribute. A table is said to have distinct l-diversity if every equivalence class of the table has distinct l-diversity.

	ZIP Code	Age	Disease	
1	476**	2*	Heart Disease	$l = 1$
2	476**	2*	Heart Disease	
3	476**	2*	Heart Disease	
4	4790*	≥ 40	Flu	$l = 3$
5	4790*	≥ 40	Heart Disease	
6	4790*	≥ 40	Cancer	
7	476**	3*	Heart Disease	$l = 2$
8	476**	3*	Cancer	
9	476**	3*	Cancer	

The table has 1-diversity

A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. ℓ-diversity: Privacy beyond k-anonymity. In *Proc. of the 22nd IEEE International Conference on Data Engineering (ICDE)*, Atlanta, GA, 2006.

© Majid Rafiei (RWTH Aachen University)

C-Confidence Bounding to Prevent Attribute Linkage Attack

- Let $conf(QID \rightarrow s)$ be the percentage of records containing s in an equivalence class with respect to the quasi-identifier QID. C is considered as the maximum value of $conf(QID \rightarrow s)$ in an equivalence class.

	ZIP Code	Age	Disease
1	476**	2*	Heart Disease
2	476**	2*	Heart Disease
3	476**	2*	Heart Disease
4	4790*	≥ 40	Flu
5	4790*	≥ 40	Heart Disease
6	4790*	≥ 40	Cancer
7	476**	3*	Heart Disease
8	476**	3*	Cancer
9	476**	3*	Cancer

$conf(QID \rightarrow Heart\ Disease) = 1$ } $C = 1$

$conf(QID \rightarrow Flu) = 0.33$
$conf(QID \rightarrow Heart\ Disease) = 0.33$
$conf(QID \rightarrow Cancer) = 0.33$ } $C = 0.33$ → The table has 1-confidence bounding

$conf(QID \rightarrow Heart\ Disease) = 0.33$
$conf(QID \rightarrow Cancer) = 0.66$ } $C = 0.66$

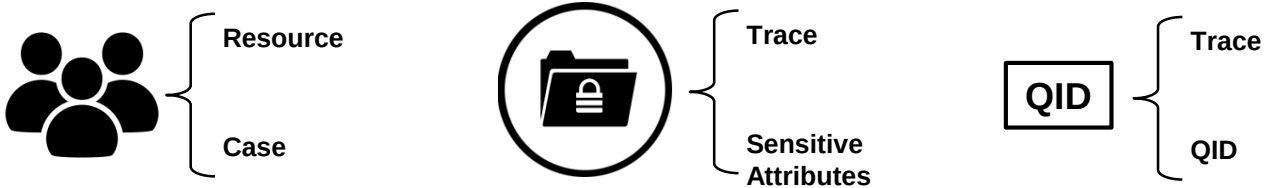
K. Wang, B. C. M. Fung, and P. S. Yu. Handicapping attacker's confidence: An alternative to k-anonymization. *Knowledge and Information Systems (KAIS)*, 11(3):345–368, April 2007.

© Majid Rafiei (RWTH Aachen University)

PPDP in Process Mining

Case ID	Trace $< e_1, e_2, \dots e_n > = < (r, a, t), (r, a, t), \dots, (r, a, t) >$	Sensitive Attributes	QID	Non-sensitive Attributes
Patient1	<(Surgeon1, Brain Surgery, 2018/10/10-10:30),…>	{Disease,…}	{Age, Gender, …}	…
Patient2	<(Surgeon2, Heart Surgery, 2018/10/11-11:30),…>	{Disease,…}	{Age, Gender, …}	…
Patient3	<(Surgeon3, Kidney Surgery, 2018/10/11-12:30),…>	{Disease,…}	{Age, Gender, …}	…
Patient4	<(Surgeon4, Brain Surgery, 2018/10/12-11:00),…>	{Disease,…}	{Age, Gender, …}	…

- Trace makes this structure complex from privacy point of view.
  - Trace itself can be considered as quasi-identifier and at the same time as sensitive attribute.



© Majid Rafiei (RWTH Aachen University)

Process-mining-specific Challenges

- The quasi-identifier role of traces in process mining causes significant challenges for group-based anonymization techniques because of two specific properties of event data:
  - High variability of traces
  - Pareto distribution of traces

© Majid Rafiei (RWTH Aachen University)

Process-mining-specific Challenges

- High variability of traces:
  - There could be tens of different activities happening in any order.
  - One activity or a bunch of activities could happen repetitively.
  - Some traces could contain a few activities compared to all possible.
- Pareto distribution of traces:
  - Few trace variants are frequent and many trace variants are unique.

- Enforcing k-anonymity on little-overlapping traces in a high-dimensional space is a significant challenge, and the majority part of the data have to be suppressed in order to achieve the desired anonymization.

© Majid Rafiei (RWTH Aachen University)

Types of Background Knowledge

- We consider four main types of background knowledge regarding traces in process mining that can lead to the case (record) and/or attribute linkage attack:
  - Set
    - The adversary knows a subset of activities having been done for the case.
  - Multiset
    - The adversary knows not only a subset of activities having been done for the case, but also the frequency of each activity.
  - Sequence
    - The adversary knows a subsequence of activities having been done for the case.
  - Relative time differences (relative)
    - The adversary knows not only a subsequence of activities, but also the time difference between the activities.

© Majid Rafiei (RWTH Aachen University)

Set Abstraction of an Event Log

Case ID	Activity	Timestamp	Diagnosis	…
1	r	00:00:19 1-11-2019		…
1	b	00:02:52 1-11-2019	AIDS	…
1	d	00:03:33 1-11-2019		…
1	c	00:04:54 1-11-2019		…
1	f	00:06:04 1-11-2019		…
1	c	00:07:19 1-11-2019		…
2	r	00:10:01 2-11-2019		…
2	f	00:16:10 2-11-2019		…
2	c	00:17:30 2-11-2019		…
2	e	00:25:04 2-11-2019	Flu	…
…	…	…	…	…

Case ID	Trace	Diagnosis
1	{d,c,f,r,b}	AIDS
2	{c,e,r,f}	Flu
…	…	…

© Majid Rafiei (RWTH Aachen University)

Multiset Abstraction of an Event Log

Case ID	Activity	Timestamp	Diagnosis	…
1	r	00:00:19 1-11-2019		…
1	b	00:02:52 1-11-2019	AIDS	…
1	d	00:03:33 1-11-2019		…
1	c	00:04:54 1-11-2019		…
1	f	00:06:04 1-11-2019		…
1	c	00:07:19 1-11-2019		…
2	r	00:10:01 2-11-2019		…
2	f	00:16:10 2-11-2019		…
2	c	00:17:30 2-11-2019		…
2	e	00:25:04 2-11-2019	Flu	…
…	…	…	…	…

Case ID	Trace	Diagnosis
1	[r,b,d,c^2,f]	AIDS
2	[r,f,c,e]	Flu
…	…	…

© Majid Rafiei (RWTH Aachen University)

Sequence Abstraction of an Event Log

Case ID	Activity	Timestamp	Diagnosis	…
1	r	00:00:19 1-11-2019		…
1	b	00:02:52 1-11-2019	AIDS	…
1	d	00:03:33 1-11-2019		…
1	c	00:04:54 1-11-2019		…
1	f	00:06:04 1-11-2019		…
1	c	00:07:19 1-11-2019		…
2	r	00:10:01 2-11-2019		…
2	f	00:16:10 2-11-2019		…
2	c	00:17:30 2-11-2019		…
2	e	00:25:04 2-11-2019	Flu	…
…	…	…	…	…

Case ID	Trace	Diagnosis
1	<r,b,d,c,f,c>	AIDS
2	<r,f,c,e>	Flu
…	…	…

© Majid Rafiei (RWTH Aachen University)

Relative Abstraction of an Event Log

Case ID	Activity	Timestamp	Diagnosis	…
1	r	00:00:19 1-11-2019		…
1	b	00:02:52 1-11-2019	AIDS	…
1	d	00:03:33 1-11-2019		…
1	c	00:04:54 1-11-2019		…
1	f	00:06:04 1-11-2019		…
1	c	00:07:19 1-11-2019		…
2	r	00:10:01 2-11-2019		…
2	f	00:16:10 2-11-2019		…
2	c	00:17:30 2-11-2019		…
2	e	00:25:04 2-11-2019	Flu	…
…	…	…	…	…

Case ID	Activity	Timestamp	Diagnosis	…
1	r	00:00:00 1-11-2019		…
1	b	00:02:00 1-11-2019	AIDS	…
1	d	00:03:00 1-11-2019		…
1	c	00:04:00 1-11-2019		…
1	f	00:06:00 1-11-2019		…
1	c	00:07:00 1-11-2019		…
2	r	00:00:00 2-11-2019		…
2	f	00:06:00 2-11-2019		…
2	c	00:07:00 2-11-2019		…
2	e	00:15:00 2-11-2019	Flu	…
…	…	…	…	…

Case ID	Trace	Diagnosis
1	<(r,0),(b,2),(d,3),(c,4),(f,6),(c,7)>	AIDS
2	<(r,0),(f,6),(c,7),(e,15)>	Flu
…	…	…

- Time accuracy = minutes
- Start time in each trace becomes 00:00:00
- All the timestamps in a trace get relative to the start time.

© Majid Rafiei (RWTH Aachen University)

An Example

Case Id	Simple Trace	Disease
1	⟨(RE,01-08:30),(V,01-08:45),(RL,01-08:58)⟩	Flu
2	⟨(RE,01-08:46),(HO,01-09:46),(BT,01-10:00),(BT,02-08:00),(V,02-10:00),(RL,02-14:00)⟩	HIV
3	⟨(RE,01-08:50),(HO,01-10:00),(BT,01-10:15),(V,02-10:15),(RL,02-14:15)⟩	Infection
4	⟨(RE,01-08:55),(V,01-09:10),(IN,01-09:30),(RL,01-10:30)⟩	Poisoning
5	⟨(RE,01-09:00),(V,01-09:20),(HO,01-09:55),(BT,01-10:10),(RL,02-16:00)⟩	Cancer
6	⟨(RE,01-09:05),(V,01-10:20),(RL,01-14:20)⟩	Hypotension

- Set
  - $BK = \{\boldsymbol{V}, \boldsymbol{IN}\} \rightarrow Case\ Id = \mathbf{4}$
- Multiset
  - $BK = [HO^1, BT^2] \rightarrow Case\ Id = \mathbf{2}$
- Sequence
  - $BK = < \boldsymbol{RE}, \boldsymbol{V}, \boldsymbol{HO} > \rightarrow Case\ Id = \mathbf{5}$
- Relative
  - Consider case 1 and case 6. If the adversary knows that for a victim case, it took almost four hours to get released (RL) after visiting by a doctor (V), the only matching case is case **6**.

© Majid Rafiei (RWTH Aachen University)

TLKC-Privacy Model for Process Mining

- In reality, it is almost impossible for an adversary to gain all the information of a target victim, and it requires non-trivial effort to gather each piece of background knowledge.
- This realistic limitation is exploited by the TLKC-privacy model for process mining to deal with the challenges appearing because of the quasi-identifier role of traces.
  - It is assumed that the adversary's background knowledge is bounded by at most L values of the quasi-identifier.
  - T $\in \{seconds, minutes, hours, days\}$ refers to the accuracy of timestamps.
  - K refers to the k in the k-anonymity definition.
  - C refers to the bound of confidence regarding the sensitive attribute values in an equivalence class.

Rafiei, M., Wagner, M., van der Aalst, W.M.P.: TLKC-privacy model for process mining. In: 14th International Conference on Research Challenges in Information Science, RCIS 2020 (2020)

© Majid Rafiei (RWTH Aachen University)

TLKC-Privacy Model for Process Mining

- Let EL(T) be an event log with the time accuracy T. It satisfies TLKC-privacy if and only if for any (sub)trace with maximum length L in EL(T), K-anonymity and C-confidence bounding are met with respect to the assumed background knowledge (set, multiset, sequence, relative).
  - Time is only considered when the assumed background knowledge is *relative*.

- *For the precise formal definitions refer to the paper.*

Rafiei, M., Wagner, M., van der Aalst, W.M.P.: TLKC-privacy model for process mining. In: 14th International Conference on Research Challenges in Information Science, RCIS 2020 (2020)

© Majid Rafiei (RWTH Aachen University)

TLKC-Privacy Model (Utility Measure)

- Utility measure:
  - It highly depends on the task which is supposed to be performed on the event log.
  - In process mining, and specifically for process discovery, we want to preserve the maximal frequent traces.
- Maximal Frequent Trace (MFT)
  - Let EL(T) be an event log with the time accuracy T, and Θ be the minimum support threshold. A (sub)trace in EL(T) is maximal in the event log if it is frequent with respect to Θ, and no supertrace of the trace is frequent in the event log.

The Goal is to preserve as many MFT as possible while satisfying TLKC-privacy requirements.

© Majid Rafiei (RWTH Aachen University)

TLKC-Privacy Model (Utility Measure)

- An example:
  - Let $\Theta = \frac{2}{3}$ be the threshold of frequency and EL be the following event log (without timestamps), the set of maximal frequent traces in this event log (MFT_{EL}) is as follows:

Case ID	Trace	Diagnosis
1	<a,b,c,d>	Fever
2	<b,c,e,d>	Flu
3	<b,e,c,d>	Fever

Frequent traces = {<b>,<c>,<d>,<e>,<b,c>,<b,d>,<b,e>,<c,d>,<e,d>,<b,c,d>,<b,e,d>}

MFT_{EL} = {~~<b>,<c>,<d>,<e>,<b,c>,<b,d>,<b,e>,<c,d>,<e,d>,~~<b,c,d>,<b,e,d>}

© Majid Rafiei (RWTH Aachen University)

TLKC-Privacy Model (Violating Trace)

- Violating trace:
  - Let EL(T) be an event log with the time accuracy T. A (sub)trace with maximum length L in EL(T) is violating with respect to the TLKC-privacy requirements if it does not satisfy K-anonymity and C-confidence bounding.

- An event log satisfies TLKC-privacy, if all violating traces w.r.t. the given privacy requirement are removed.
  - There could be numerous number of violating traces.
  - The utility measure needs to be taken into account.

© Majid Rafiei (RWTH Aachen University)

TLKC-Privacy Model (Minimal Violating Trace)

- Minimal Violating Trace (MVT):
  - Let EL(T) be an event log with the time accuracy T. A (sub)trace with maximum length L in EL(T) is minimal violating with respect to the TLKC-privacy requirements if it is violating trace and every proper subtrace of the trace is not a violating trace in the event log.

- Every violating trace in an event log is either an MVT or it contains an MVT. Therefore, if an event log EL contains no MVT, then EL contains no violating trace.
  - The set of minimal violating traces in an event log is much smaller than the set of violating traces.

© Majid Rafiei (RWTH Aachen University)

TLKC-Privacy Model (Minimal Violating Trace)

- An example:
  - Let *sequence* be the type of background knowledge and L=2 be the power of background knowledge (length of sequence), K=2, and C=0.5. For the following event log EL (without timestamps), the set of minimal violating traces (MVT_{EL}) is as follows:

Case ID	Trace	Diagnosis
1	<a,b,c,d>	Fever
2	<b,c,e,d>	Flu
3	<b,e,c,d>	Fever

Violating traces = {<a>,<b>,<c>,<d>,<a,b>,<a,c>,<a,d>,<b,c>,<b,d>,<c,d>,<c,e>,<e,c>}

MVT_{EL} = {<a>,<b>,<c>,<d>,~~<a,b>,<a,c>,<a,d>,<b,c>,<b,d>,<c,d>,<c,e>,<e,c>~~}

© Majid Rafiei (RWTH Aachen University)

TLKC-Privacy Model (Score)

- How to choose events to remove from the minimal violating traces with respect to the utility measure and the privacy requirements?
  - The measure for utility is based on maximal frequent sequences
    - $Utility\ Loss\ (event)\ =\ \#occurrence\ in\ MFT$
  - The measure for privacy is based on minimal violating sequences
    - $Privacy\ Gain\ (event)\ =\ \#occurrence\ in\ MVT$

- Those two measures are combined for obtaining a measure for deleting the events.

$$\text{Score}(event) = \frac{Privacy\ Gain\ (event)}{Utility\ Loss(event) + 1}$$

© Majid Rafiei (RWTH Aachen University)

TLKC-Privacy Model (the Algorithm)

Algorithm 1: TLKC-Privacy Algorithm

Input: Original event log EL
Input: T, L, K, C, and Θ
Input: Sensitive values S
Output: Anonymized event log EL' which satisfies $TLKC$-privacy

generate MFT_{EL} and MVT_{EL};
generate MFT_{tree} and MVT_{tree} as the prefix trees for MFT_{EL} and MVT_{EL};
while *there is node (event) in* MVT_{tree} **do**
    select an event (node) e_w that has the highest score to suppress;
    delete all the MVT and MFT containing the event e_w from MVT_{tree} and MFT_{tree};
    update $Socre(e)$ for all the remaining events (nodes) in MVT_{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';

© Majid Rafiei (RWTH Aachen University)

TLKC-Privacy Model (Example)

- We want to apply TLKC-privacy to the following event log with T=hours, L=2, K=2, C=50%, Θ=25%, and assuming the relative type of background knowledge.

Quasi-identifier **Sensitive attribute**

Case Id	Trace	Disease
1	$\langle(RE,1),(HO,4),(V,5),(BT,7),(V,8)\rangle$	Cancer
2	$\langle(BT,7),(V,8),(RL,9)\rangle$	Infection
3	$\langle(HO,4),(V,5),(BT,7),(RL,9)\rangle$	Poisoning
4	$\langle(RE,1),(V,6),(V,8),(RL,9)\rangle$	Infection
5	$\langle(HO,4),(V,8),(RL,9)\rangle$	Poisoning
6	$\langle(V,6),(BT,7),(RL,9)\rangle$	Flu
7	$\langle(RE,1),(BT,7),(V,8),(RL,9)\rangle$	Flu
8	$\langle(RE,1),(V,6),(BT,7),(V,8)\rangle$	Cancer

$$\boldsymbol{Trace} = < \boldsymbol{e_1}, \boldsymbol{e_2}, \ldots \boldsymbol{e_n} > = < (\boldsymbol{a},\boldsymbol{t}), (\boldsymbol{a},\boldsymbol{t}), \ldots, (\boldsymbol{a},\boldsymbol{t}) >$$

Timestamps are presented by integer values as hours.

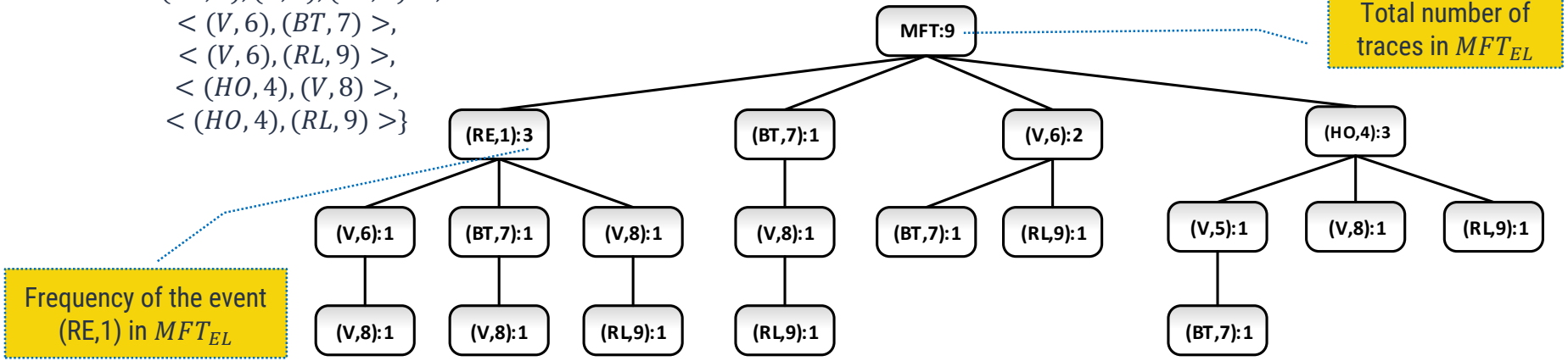
© Majid Rafiei (RWTH Aachen University)

TLKC-Privacy Model (Example)

Generating MFT_{EL} and MFT_{tree}

$$MFT_{EL} = \{< (RE,1),(V,6),(V,8) >, < (RE,1),(BT,7),(V,8) >, < (RE,1),(V,8),(RL,9) >, < (HO,4),V,5),(BT,7) >, < (BT,7),(V,8),(RL,9) >, < (V,6),(BT,7) >, < (V,6),(RL,9) >, < (HO,4),(V,8) >, < (HO,4),(RL,9) >\}$$

MFT_{tree}: Each root-to-leaf path represents one trace in MFT_{EL}



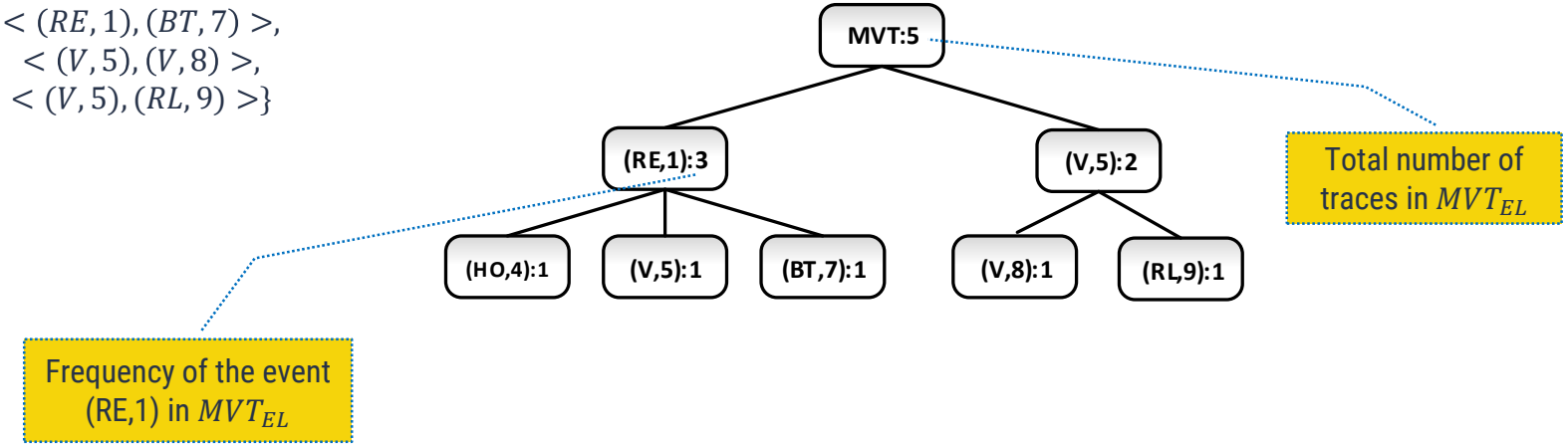
© Majid Rafiei (RWTH Aachen University)

TLKC-Privacy Model (Example)

Generating MVT_{EL} and MVT_{tree}

$$MVT_{EL} = \{< (RE,1), < HO,4) >, < (RE,1), (V,5) >, < (RE,1), (BT,7) >, < (V,5), (V,8) >, < (V,5), (RL,9) >\}$$

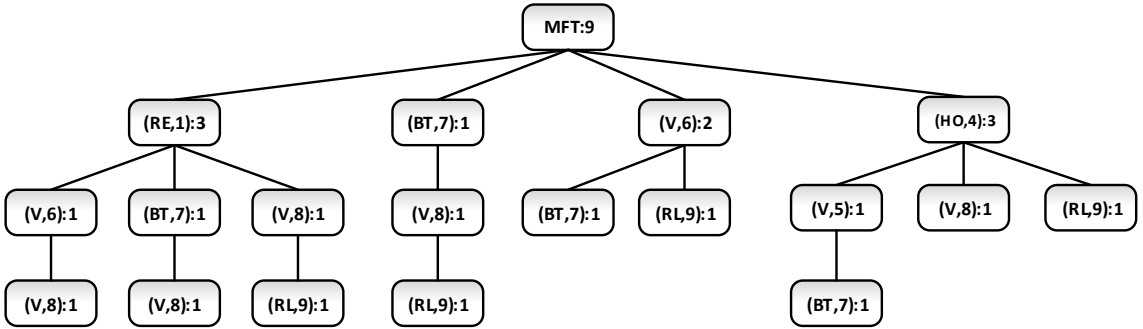
MVT_{tree}: Each root-to-leaf path represents one trace in MVT_{EL}

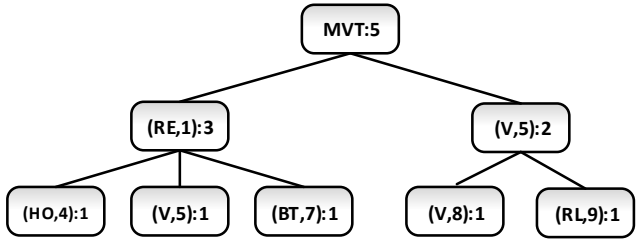


© Majid Rafiei (RWTH Aachen University)

TLKC-Privacy Model (Example)

- Selecting the event with the highest score from MVT_{tree} (the winner event e_w)





The initial scores for the events in MVT_{EL}

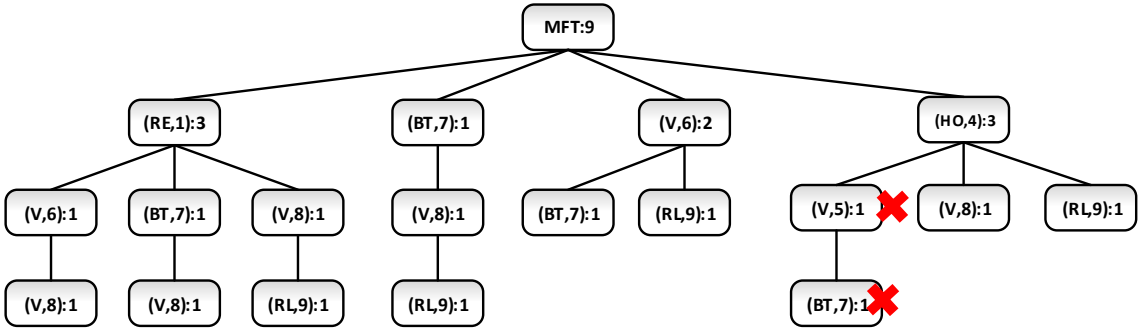
	$(RE,1)$	$(HO,4)$	$(V,5)$	$(BT,7)$	$(V,8)$	$(RL,9)$
$PG(e)$	3	1	3	1	1	1
$UL(e)+1$	4	4	2	5	6	5
$Score(e)$	0.75	0.25	1.50	0.20	0.16	0.20

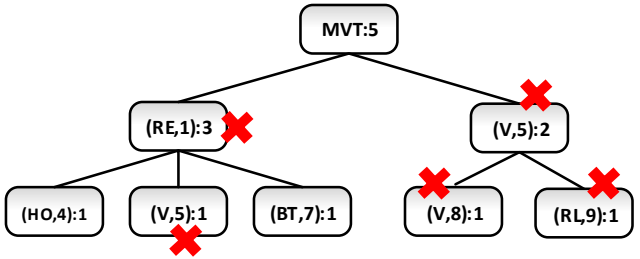
The winner events $e_w = (V,5)$

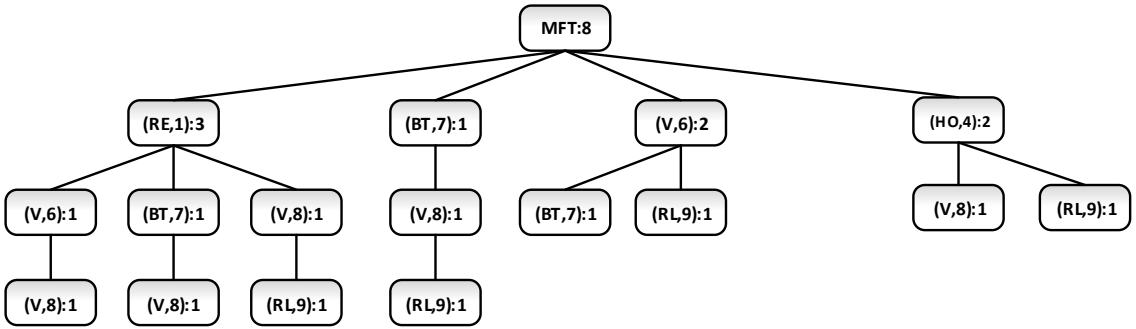
© Majid Rafiei (RWTH Aachen University)

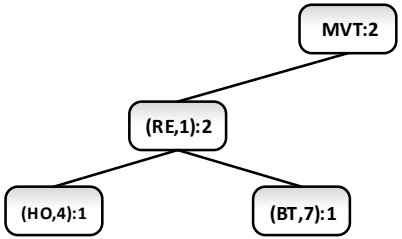
TLKC-Privacy Model (Example)

- Deleting all the traces in MVT_{tree} and MFT_{tree} containing the winner event.





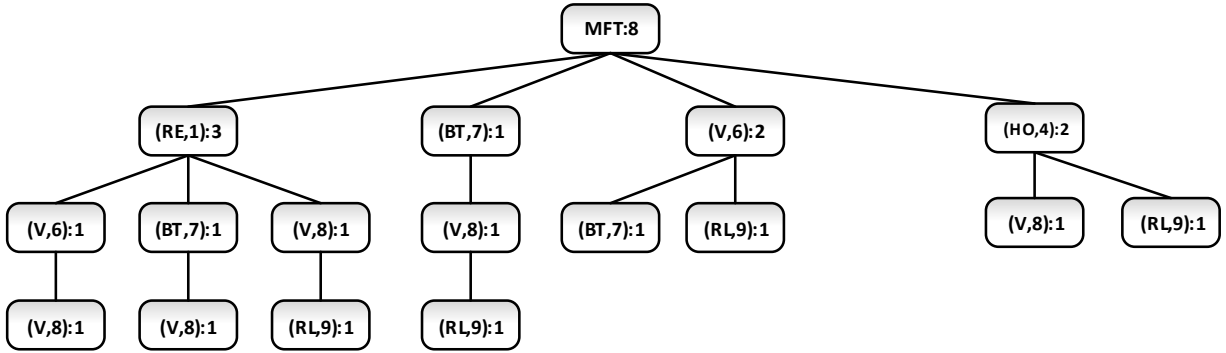


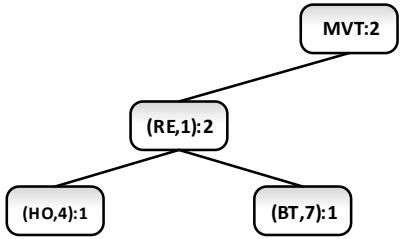


© Majid Rafiei (RWTH Aachen University)

TLKC-Privacy Model (Example)

- Updating scores for the remaining events in MVT_{tree} and e_w to the set of suppressed events (Sup_{EL}).





The first updated scores for the events in MVT_{EL}

	$(RE, 1)$	$(HO, 4)$	$(BT, 7)$
$PG(e)$	2	1	1
$UL(e)+1$	4	3	4
$Score(e)$	0.5	0.33	0.25

$$Sup_{EL} = \{(V, 5)\}$$

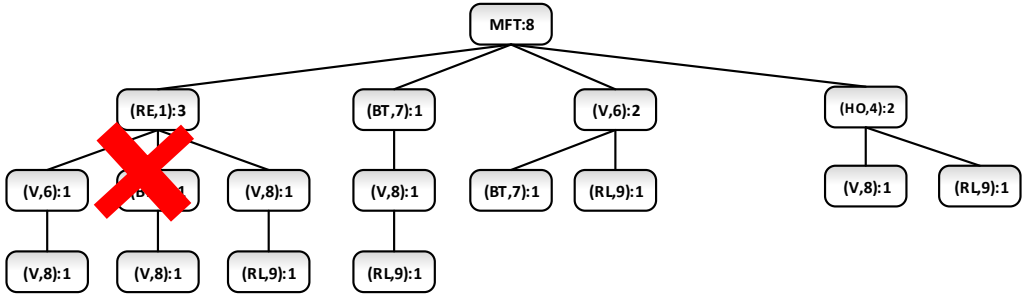
© Majid Rafiei (RWTH Aachen University)

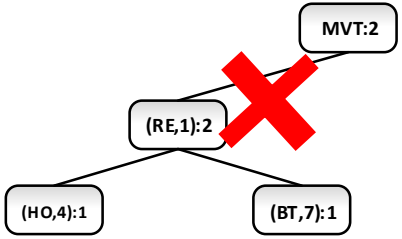
TLKC-Privacy Model (Example)

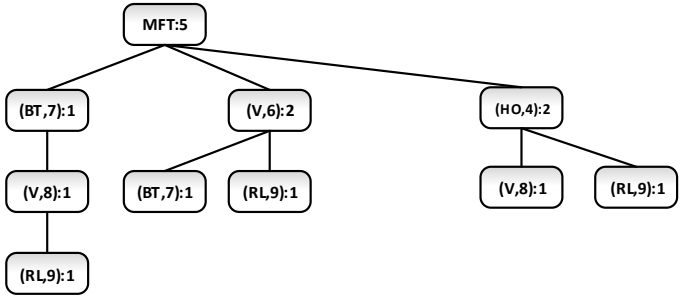
- The previous steps need to be repeated until there is no event in MVT_{tree}.

	$(RE,1)$	$(HO,4)$	$(BT,7)$
$PG(e)$	2	1	1
$UL(e)+1$	4	3	4
$Score(e)$	0.5	0.33	0.25

The winner events $e_w = (RE, 1)$







$\text{Sup}_{\text{EL}} = \{(V, 5), (RE, 1)\}$

© Majid Rafiei (RWTH Aachen University)

TLKC-Privacy Model (Example)

- Removing all the events in Sup_{EL} from the event log.

$$\text{Sup}_{\text{EL}} = \{(\text{V}, 5), (\text{RE}, 1)\}$$

Case Id	Trace	Disease
1	$\langle(RE,1),(HO,4),(V,5),(BT,7),(V,8)\rangle$	Cancer
2	$\langle(BT,7),(V,8),(RL,9)\rangle$	Infection
3	$\langle(HO,4),(V,5),(BT,7),(RL,9)\rangle$	Poisoning
4	$\langle(RE,1),(V,6),(V,8),(RL,9)\rangle$	Infection
5	$\langle(HO,4),(V,8),(RL,9)\rangle$	Poisoning
6	$\langle(V,6),(BT,7),(RL,9)\rangle$	Flu
7	$\langle(RE,1),(BT,7),(V,8),(RL,9)\rangle$	Flu
8	$\langle(RE,1),(V,6),(BT,7),(V,8)\rangle$	Cancer

Removing (V,5) and (RE,1)

Case Id	Trace	Disease
1	$\langle(HO,4),(BT,7),(V,8)\rangle$	Cancer
2	$\langle(BT,7),(V,8),(RL,9)\rangle$	Infection
3	$\langle(HO,4),(BT,7),(RL,9)\rangle$	Poisoning
4	$\langle(V,6),(V,8),(RL,9)\rangle$	Infection
5	$\langle(HO,4),(V,8),(RL,9)\rangle$	Poisoning
6	$\langle(V,6),(BT,7),(RL,9)\rangle$	Flu
7	$\langle(BT,7),(V,8),(RL,9)\rangle$	Flu
8	$\langle(V,6),(BT,7),(V,8)\rangle$	Cancer

T=hours, L=2, K=2, C=0.5

- This event log satisfies TLKC-privacy when the assumed background knowledge is *relative*.
- While it preserves the maximum number of frequent traces when the frequency threshold is Θ=0.25%.

© Majid Rafiei (RWTH Aachen University)

Comparing TLKC-privacy with Traditional k-anonymity

- An abstracted original event log with respect to the relative background knowledge

#events = 30

Case Id	Trace	Disease
1	$\langle(RE,1),(HO,4),(V,5),(BT,7),(V,8)\rangle$	Cancer
2	$\langle(BT,7),(V,8),(RL,9)\rangle$	Infection
3	$\langle(HO,4),(V,5),(BT,7),(RL,9)\rangle$	Poisoning
4	$\langle(RE,1),(V,6),(V,8),(RL,9)\rangle$	Infection
5	$\langle(HO,4),(V,8),(RL,9)\rangle$	Poisoning
6	$\langle(V,6),(BT,7),(RL,9)\rangle$	Flu
7	$\langle(RE,1),(BT,7),(V,8),(RL,9)\rangle$	Flu
8	$\langle(RE,1),(V,6),(BT,7),(V,8)\rangle$	Cancer

#events = 24
Removed events = 6

Case Id	Trace	Disease
1	$\langle(HO,4),(BT,7),(V,8)\rangle$	Cancer
2	$\langle(BT,7),(V,8),(RL,9)\rangle$	Infection
3	$\langle(HO,4),(BT,7),(RL,9)\rangle$	Poisoning
4	$\langle(V,6),(V,8),(RL,9)\rangle$	Infection
5	$\langle(HO,4),(V,8),(RL,9)\rangle$	Poisoning
6	$\langle(V,6),(BT,7),(RL,9)\rangle$	Flu
7	$\langle(BT,7),(V,8),(RL,9)\rangle$	Flu
8	$\langle(V,6),(BT,7),(V,8)\rangle$	Cancer

- Satisfies TLKC-privacy with T=hours, L=2, K=2, C=50%, Θ=25% when the assumed background knowledge is *relative*.

#events = 18
Removed events = 12

Case Id	Trace	Disease
1	$\langle(BT,7),(V,8)\rangle$	Cancer
2	$\langle(BT,7),(V,8),(RL,9)\rangle$	Infection
3	$\langle(BT,7),(RL,9)\rangle$	Poisoning
4	$\langle(V,8),(RL,9)\rangle$	Infection
5	$\langle(V,8),(RL,9)\rangle$	Poisoning
6	$\langle(BT,7),(RL,9)\rangle$	Flu
7	$\langle(BT,7),(V,8),(RL,9)\rangle$	Flu
8	$\langle(BT,7),(V,8)\rangle$	Cancer

- Satisfies 2-anonymity when the assumed background knowledge is *relative*.

© Majid Rafiei (RWTH Aachen University)

Current Research Areas

- Privacy-preserving data publishing techniques for process mining.
- Privacy-preserving algorithms for process mining.
- Measure(s) to quantify confidentiality of event data.
- Confidentiality framework(s) for process mining
- Privacy preservation of the results (e.g., process models, social networks, etc.)
- Cross-organizational process mining (using SMC)
- …

© Majid Rafiei (RWTH Aachen University)

Research Areas (Fairness in Process Mining)

	Creating and managing event data	Process discovery	Conformance checking	Performance analysis	Operational support
Fairness Data Science without prejudice: How to avoid unfair conclusions even if they are true?	The input data may be biased, incomplete or incorrect such that the analysis reconfirms prejudices. By resampling or relabeling the data, undesirable forms of discrimination can be avoided. Note that both cases and resources (used to execute activities) may refer to individuals having sensitive attributes such as race, gender, age, etc.	The discovered model may abstract from paths followed by certain under-represented groups of cases. Discrimination-aware process-discovery algorithms can be used to avoid this. For example, if cases are handled differently based on gender, we may want to ensure that both are equally represented in the model'	Conformance checking can be used to "blame" individuals, groups, or organizations for deviating from some normative model. Discrimination-aware conformance checking (e.g., alignments) needs to separate (1) likelihood, (2) severity and (3) blame. Deviations may need to be interpreted differently for different groups of cases and resources	Straightforward performance measurements may be unfair for certain classes of cases and resources (e.g., not taking into account the context). Discrimination-aware performance analysis detects unfairness and supports process improvements taking into account trade-offs between internal fairness (worker's perspective) and external fairness (citizen/patient/customer's perspective)	Process-related predictions, recommendations and decisions may discriminate (un)intentionally. This problem can be tackled using techniques from discrimination-aware data mining

van der Aalst, W.M.P.: Responsible data science: using event data in a "people friendly" manner. In: International Conference on Enterprise Information Systems. pp. 3–28. Springer (2016)

© Majid Rafiei (RWTH Aachen University)

Research Areas (Accuracy in Process Mining)

	Creating and managing event data	Process discovery	Conformance checking	Performance analysis	Operational support
Accuracy Data Science without guesswork: How to answer questions with a guaranteed level of accuracy?	Event data (e.g., XES files) may have all kinds of quality problems. Attributes may be incorrect, imprecise, or uncertain. For example, timestamps may be too coarse (just the date) or reflect the time of recording rather than the time of the event's occurrence	Process discovery depends on many parameters and characteristics of the event log. Process models should better show the confidence level of the different parts. Moreover, additional information needs to be used better (domain knowledge, uncertainty in event data, etc.)	Often multiple explanations are possible to interpret non-conformance. Just providing one alignment based on a particular cost function may be misleading. How robust are the findings?	In case of fitness problems (process model and event log disagree), performance analysis is based on assumptions and needs to deal with missing values (making results less accurate)	Inaccurate process models may lead to flawed predictions, recommendations and decisions. Moreover, not communicating the (un)certainty of predictions, recommendations and decisions, may negatively impact processes

van der Aalst, W.M.P.: Responsible data science: using event data in a "people friendly" manner. In: International Conference on Enterprise Information Systems. pp. 3–28. Springer (2016)

© Majid Rafiei (RWTH Aachen University)

Research Areas (Confidentiality in Process Mining)

	Creating and managing event data	Process discovery	Conformance checking	Performance analysis	Operational support
Confidentiality Data Science that ensures confidentiality: How to answer questions without revealing secrets?	Event data (e.g., XES files) may reveal sensitive information. Anonymization and de-identification can be used to avoid disclosure. Note that timestamps and paths may be unique and a source for re-identification (e.g., all paths are unique)	The discovered model may reveal sensitive information, especially with respect to infrequent paths or small event logs. Drilling-down from the model may need to be blocked when numbers get too small (cf. k-anonymity)	Conformance checking shows diagnostics for deviating cases and resources. Access-control is important and diagnostics need to be aggregated to avoid revealing compliance problems at the level of individuals	Performance analysis shows bottlenecks and other problems. Linking these problems to cases and resources may disclose sensitive information	Process-related predictions, recommendations and decisions may disclose sensitive information, e.g., based on a rejection other properties can be derived

van der Aalst, W.M.P.: Responsible data science: using event data in a "people friendly" manner. In: International Conference on Enterprise Information Systems. pp. 3–28. Springer (2016)

© Majid Rafiei (RWTH Aachen University)

Research Areas (Transparency in Process Mining)

	Creating and managing event data	Process discovery	Conformance checking	Performance analysis	Operational support
Transparency Data Science that provides transparency: How to clarify answers such that they become indisputable?	Provenance of event data is key. Ideally, process mining insights can be related to the event data they are based on. However, this may conflict with confidentiality concerns	Discovered process models depend on the event data used as input and the parameter settings and choice of discovery algorithm. How to ensure that the process model is interpreted correctly? End-users need to understand the relation between data and model to trust analysis	When modeled and observed behavior disagree there may be multiple explanations. How to ensure that conformance diagnostics are interpreted correctly?	When detecting performance problems, it should be clear how these were detected and what the possible causes are. Animating event logs on models helps to make problems more transparent	Predictions, recommendations and decisions are based on process models. If possible, these models should be transparent. Moreover, explanations should be added to predictions, recommendations and decisions ("We predict that this case be late, because ...")

van der Aalst, W.M.P.: Responsible data science: using event data in a "people friendly" manner. In: International Conference on Enterprise Information Systems. pp. 3–28. Springer (2016)

© Majid Rafiei (RWTH Aachen University)

THANKS!

Any questions?

You can find me at

majid.rafiei@pads.rwth-aachen.de

© Majid Rafiei (RWTH Aachen University)