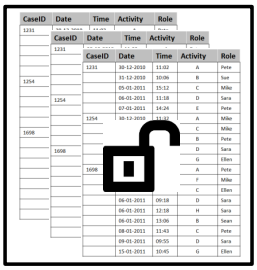
Privacy in Process Mining: Motivation, Method and Research Challenges

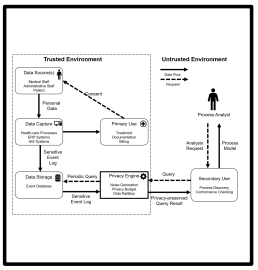
Agnes Koschmider

Group Process Analytics, Kiel University, Germany

Overview



Motivation

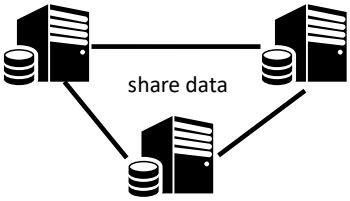


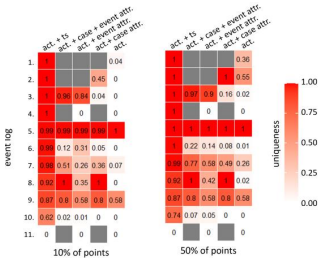
Method

Research Challenges

Privacy Risks

Christian-Albrechts-Universität zu Kiel





S. Nuñez 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. CAiSE 2020: 252-267

Running Example

Christian-Albrechts-Universität zu Kiel

Case attributes | Activity | Timestamp | Event attribute

	Patient	Birth	Gender	Activity	Timestamp	Doctor
Case	104	1935	Male	Blood Test	03/03/19 17:43	Dr. Scott
	104	1935	Male	CT	03/05/19 18:15	Dr. Doe
	104	1935	Male	Surgery	03/07/19 08:23	Dr. Doe
	104	1935	Male	Rehab	03/10/19 09:36	John Brown
Case	105	1968	Male	Blood Test	03/03/19 23:28	Dr. Fox
	105	1968	Male	MRT	03/04/19 23:53	Dr. White
Case	106	1990	Female	Session	03/03/19 12:34	Dr. Black
	106	1990	Female	Abortion	03/08/19 16:23	Dr. Scott
Case	107	1968	Male	Blood Test	03/02/19 18:25	Dr. Scott
	107	1968	Male	MRT	03/06/19 11:32	Dr. Fox

Events

Transform Event Log

Christian-Albrechts-Universität zu Kiel

Case	Birth	Gender	Activity	Timestamp	Doctor
104	1935	Male	[Blood Test, CT, …]	[03/03/19, 03/05/19, …]	[Scott, Doe, …]
105	1968	Male	[Blood Test, MRT, …]	[03/03/19, 03/04/19, …]	[Fox, White, …]
106	1990	Female	[Session, Abortion]	[03/03/19, 03/08/19]	[Black, Scott]
107	1968	Male	[Blood Test, MRT]	[03/02/19, 03/06/19]	[Scott, Fox]

Quantify Uniqueness

Christian-Albrechts-Universität zu Kiel

Considering case atrributes:
given case attribute **Gender**

Case	Birth	Gender	Activity	Timestamp	Doctor
104	1935	Male	[Blood Test, CT, …]	[03/03/19, 03/05/19, …]	[Scott, Doe, …]
105	1968	Male	[Blood Test, MRT, …]	[03/03/19, 03/04/19, …]	[Fox, White, …]
106	1990	Female	[Session, Abortion]	[03/03/19, 03/08/19]	[Black, Scott]
107	1968	Male	[Blood Test, MRT]	[03/02/19, 03/06/19]	[Scott, Fox]

Quantify Uniqueness

Christian-Albrechts-Universität zu Kiel

Considering case atrributes:
given case attribute

Case	Birth	Gender	Activity	Timestamp	Doctor
104	1935	Male	[Blood Test, CT, …]	[03/03/19, 03/05/19, …]	[Scott, Doe, …]
105	1968	Male	[Blood Test, MRT, …]	[03/03/19, 03/04/19, …]	[Fox, White, …]
106	1990	Female	[Session, Abortion]	[03/03/19, 03/08/19]	[Black, Scott]
107	1968	Male	[Blood Test, MRT]	[03/02/19, 03/06/19]	[Scott, Fox]

1/4 = 0.25 = 25% re-identification risk

Quantify Uniqueness

Christian-Albrechts-Universität zu Kiel

Considering events as points:
p_1 = (Actvity$_1$, Timestamp$_1$, Doctor$_1$)

Case	Birth	Gender	Activity	Timestamp	Doctor
104	1935	Male	[Blood Test, CT, …]	[03/03/19, 03/05/19, …]	[Scott, Doe, …]
105	1968	Male	[Blood Test, MRT, …]	[03/03/19, 03/04/19, …]	[Fox, White, …]
106	1990	Female	[Session, Abortion]	[03/03/19, 03/08/19]	[Black, Scott]
107	1968	Male	[Blood Test, MRT]	[03/02/19, 03/06/19]	[Scott, Fox]

Quantify Uniqueness

Christian-Albrechts-Universität zu Kiel

Considering events as points:
$p_2 = (Actvity_2, Timestamp_2, Doctor_2)$

Case	Birth	Gender	Activity	Timestamp	Doctor
104	1935	Male	[Blood Test, CT, …]	[03/03/19, 03/05/19 …]	[Scott, Doe, …]
105	1968	Male	[Blood Test, MRT, …]	[03/03/19, 03/04/19, …]	[Fox, White, …]
106	1990	Female	[Session, Abortion]	[03/03/19, 03/08/19]	[Black, Scott]
107	1968	Male	[Blood Test, MRT]	[03/02/19, 03/06/19]	[Scott, Fox]

Quantify Uniqueness

Christian-Albrechts-Universität zu Kiel

Considering events as points:
$p_1 = (Actvity_1)$

Case	Birth	Gender	Activity	Timestamp	Doctor
104	1935	Male	[Blood Test, CT, …]	[03/03/19, 03/05/19, …]	[Scott, Doe, …]
105	1968	Male	[Blood Test, MRT, …]	[03/03/19, 03/04/19, …]	[Fox, White, …]
→ 106	1990	Female	[Session, Abortion]	[03/03/19, 03/08/19]	[Black, Scott]
107	1968	Male	[Blood Test, MRT]	[03/02/19, 03/06/19]	[Scott, Fox]

1/4 = 0.25 = 25% re-identification risk

Quantify Uniqueness

Christian-Albrechts-Universität zu Kiel

Considering events as points:
$p_1 = (Actvity_1, Timestamp_1)$

Case	Birth	Gender	Activity	Timestamp	Doctor
104	1935	Male	[Blood Test, CT, …]	[03/03/19, 03/05/19, …]	[Scott, Doe, …]
105	1968	Male	[Blood Test, MRT, …]	[03/03/19, 03/04/19, …]	[Fox, White, …]
→ 106	1990	Female	[Session, Abortion]	[03/03/19, 03/08/19]	[Black, Scott]
→ 107	1968	Male	[Blood Test, MRT]	[03/02/19, 03/06/19]	[Scott, Fox]

2/4 = 0.50 = 50% re-identification risk

Quantify Uniqueness

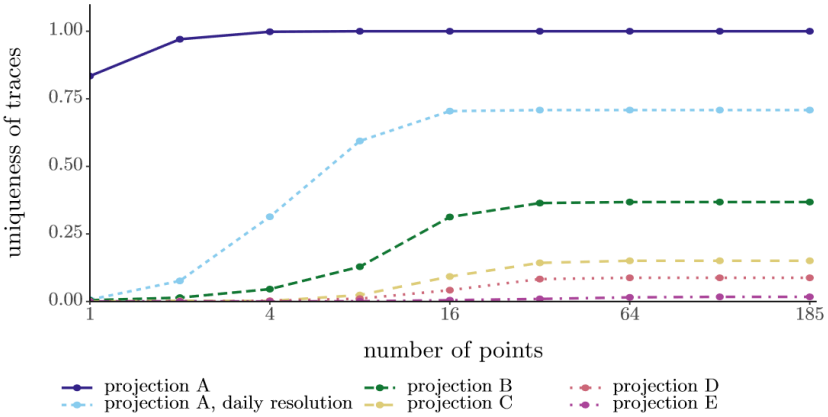
Christian-Albrechts-Universität zu Kiel

Considering events as points:
$p_1 = (Actvity_1, Timestamp_1, Doctor_1)$

Case	Birth	Gender	Activity	Timestamp	Doctor
104	1935	Male	[Blood Test, CT, …]	[03/03/19, 03/05/19, …]	[Scott, Doe, …]
105	1968	Male	[Blood Test, MRT, …]	[03/03/19, 03/04/19, …]	[Fox, White, …]
→ 106	1990	Female	[Session, Abortion]	[03/03/19, 03/08/19]	[Black, Scott]
→ 107	1968	Male	[Blood Test, MRT]	[03/02/19, 03/06/19]	[Scott, Fox]

4/4 = 1.00 = 100% re-identification risk

Uniqueness for Cases of Sepsis Event log



Requirements for Privacy-Preserving Process Mining Techniques

Anonymity

Unlinkability

Transparency

Notice

Accountability

Elkoumy, G., Fahrenkrog-Petersen, S. A., Sani, M. F., Koschmider, A., Mannhardt, F., Voigt, S. N. V., Rafiei, M., & Waldthausen, L. V. Privacy and Confidentiality in Process Mining - Threats and Research Challenges.
ACM Transactions on Management Information Systems, 2021, in press.

State-of-the Art

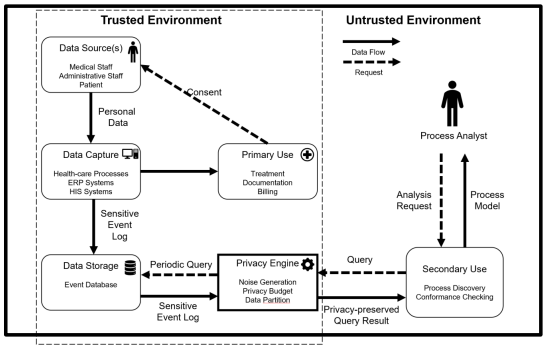
Christian-Albrechts-Universität zu Kiel

	Anonymity	Unlinkability	Notice	Transparency	Accountability
TLKC	x	x			
PRETSA	x	x			
PPPM	x	x			
PRIPEL	x	x			
Multi-party computation	x				

- Beside the requirements of process mining techniques, also data, application and presentation are requirements

Privacy Preserving Process Mining

Christian-Albrechts-Universität zu Kiel

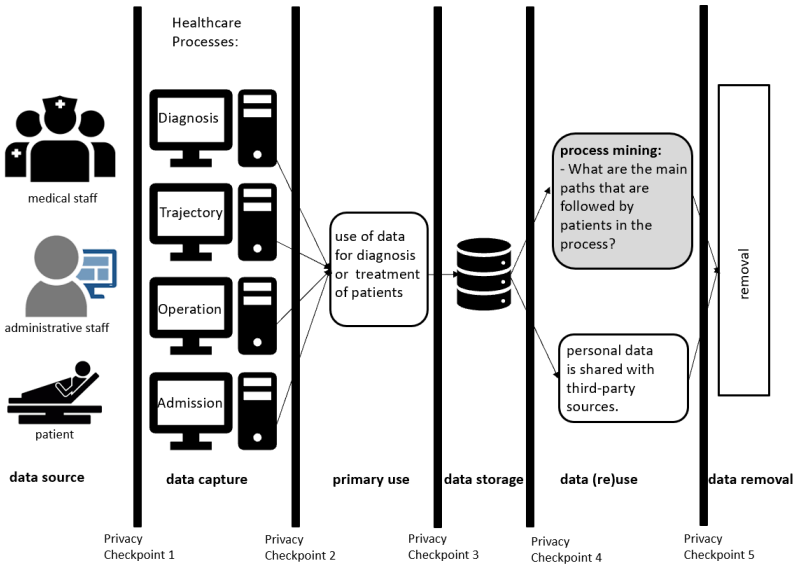


F. Mannhardt, A. Koschmider, N. Baracaldo, M. Weidlich, J. Michael: *Privacy-Preserving Process Mining: Differential Privacy for Event Logs*, Business & Information Systems Engineering 61(5), 2019

J. Michael, A. Koschmider, F. Mannhardt, N. Baracaldo, B. Rumpe: *User-Centered and Privacy-Driven Process Mining System Design for IoT*. CAiSE Forum 2019: 194-206, Springer

Identification of data passes and privacy checkpoints for hospital health processes

Christian-Albrechts-Universität zu Kiel

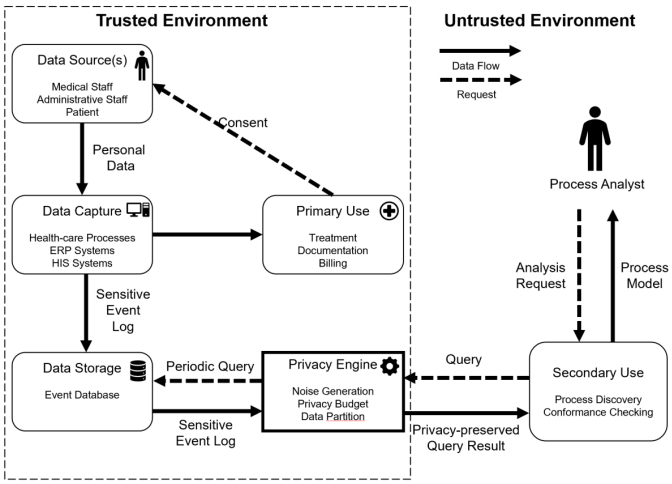


Laplacian mechanism is used to provide differential privacy for counting the number of records in a database

Privacy Model

Christian-Albrechts-Universität zu Kiel

- we assume a **centralized privacy approach**
- sensitive data is stored as an event log in protected data storage
- **privacy engine** acts as the **single point of access** for process mining algorithms and **introduces noise** to each query result
- no difference for data provider between the data used by the process mining algorithm regardless of whether his/her data is included or not



Our Initial Approach – Directly-Follows Relation (DFR)

Christian-Albrechts-Universität zu Kiel

Patient	Source	Target	Further attributes
P1	⊤	A	...
P1	A	B	...
P1	B	C	...
P1	C	⊥	...
P2	⊤	A	...
P2	A	B	...
P2	B	C	...
P2	C	⊥	...
P3	⊤	A	...
P3	A	B	...
P3	B	⊥	...
P4	⊤	A	...
P4	A	D	...
P4	D	⊥	...
P5	⊤	A	...
P5	A	⊥	...

Partition by Relation

Patient	Relation
P1	(A,B)
P2	(A,B)
...	...

Patient	Relation
P1	(B,C)
P2	(B,C)

Laplacian Mechanism

introduce noise to the queries counts according to a user-defined value

Relation	Frequency
$(-, A)$	4
(A, B)	3
(A, D)	1
(B, C)	2
(B, C)	2
$(A, -)$	1
$(B, -)$	1
$(C, -)$	2
$(D, -)$	1

But with noise and possibly previously non-existent relations

Guarantee: Differential Privacy of the DFR query wrt. the input table.

if one would sequentially query information from the same data source, the privacy budget is reduced by the sum of the individual parameters

Our Initial Approach – Activity Sequences

Christian-Albrechts-Universität zu Kiel

Patient	Sequence	Further attributes
P1	$\langle A, B, C, \perp \rangle$	...
P2	$\langle A, B, C, \perp \rangle$	...
P3	$\langle A, B, \perp \rangle$	...
P4	$\langle A, D, \perp \rangle$	...
P5	$\langle A, \perp \rangle$	...

Iteratively build Prefix Tree of length i, starting with 1

i = 2

Patient	Prefix
P1	(A,B)
P2	(A,B)
...	...

Patient	Prefix
P4	(A,D)

Iteratively build table with the Laplacian Mechanism

Prune away very low frequency sequences (a-priori inspired)

Sequence	Frequency
$\langle A, B, C \rangle$	2
$\langle A, B \rangle$	1
$\langle A, D \rangle$	2
$\langle A \rangle$	2

i = i +1

Patient	Prefix
P1	(B,C)
P2	(B,C)

Again noisy and non-existent sequences appear

Guarantee: Differential Privacy of the DFR query wrt. the input table.

we treat each trace as a sequence of identifiers

Current Steps

Christian-Albrechts-Universität zu Kiel

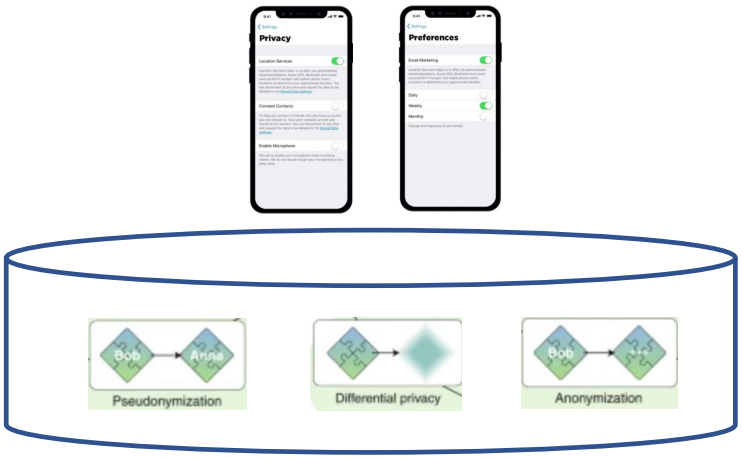
- Development of a log generator for synthetic, privacy-preserving event logs
  - ➢use of Generative Adversarial Network (GAN)

- Noise/Outlier quantification model

Challenges: Interpretable Quantification of Privacy Disclosure

Christian-Albrechts-Universität zu Kiel

- more reliable and interpretable metrics of privacy disclosure



Challenges: Balancing Risk and Utility

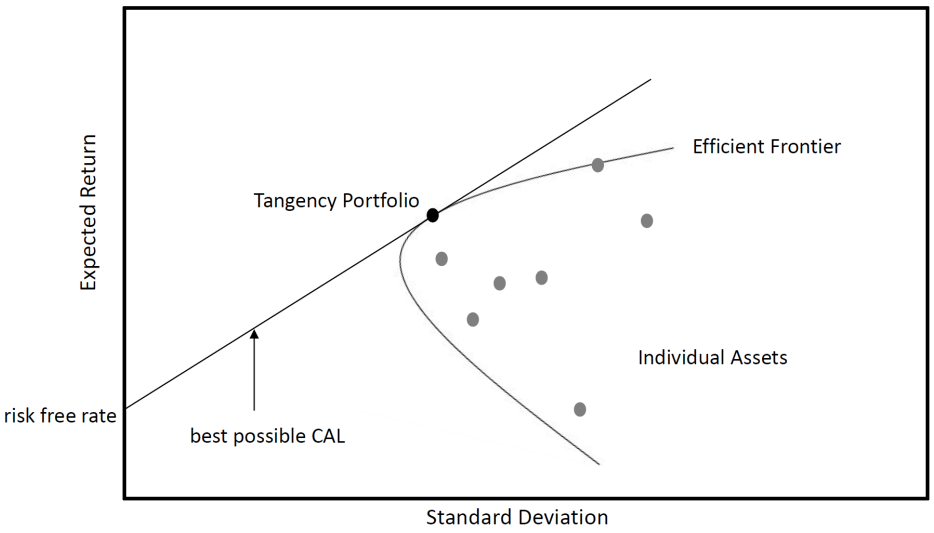
Christian-Albrechts-Universität zu Kiel

- Trade-off between disclosure risk and utility

Elkoumy, G., Fahrenkrog-Petersen, S. A., Sani, M. F., Koschmider, A., Mannhardt, F., Voigt, S. N. V., Rafiei, M., & Waldthausen, L. V. Privacy and Confidentiality in Process Mining - Threats and Research Challenges.
ACM Transactions on Management Information Systems, 2021, in press.

Challenges: Balancing Risk and Utility

Christian-Albrechts-Universität zu Kiel

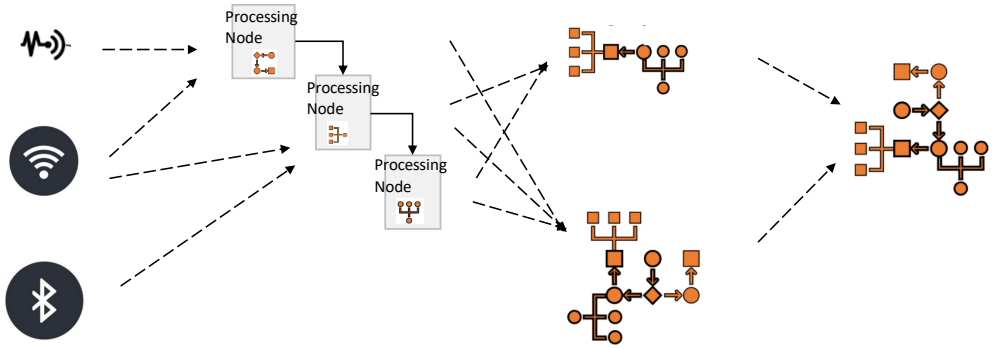


source: https://en.wikipedia.org/wiki/Efficient_frontier

Challenges: Distributed Privacy

Christian-Albrechts-Universität zu Kiel

- Distributed Privacy



Challenges: Distributed Privacy

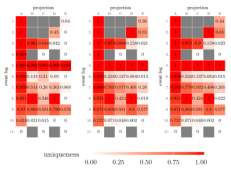
Christian-Albrechts-Universität zu Kiel

- Computational Challenges:
  - ➢ with increasing dimensions of attributes, it becomes more unpractical to achieve privacy-preserving process mining
- Traceability Challenge:
  - ➢ trace data-life cycle and ensure consent, right to be forgotten
- Transparency Challenge
  - ➢ notify who is using the data

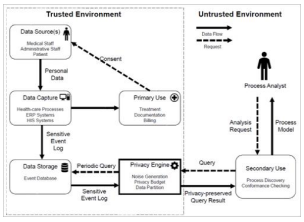
Elkoumy, G., Fahrenkrog-Petersen, S. A., Sani, M. F., Koschmider, A., Mannhardt, F., Voigt, S. N. V., Rafiei, M., & Waldthausen, L. V. Privacy and Confidentiality in Process Mining - Threats and Research Challenges. *ACM Transactions on Management Information Systems*, 2021, in press.

Summary and Outlook

Christian-Albrechts-Universität zu Kiel



S. Nuñez 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. CAiSE 2020: 252-267



F. Mannhardt, A. Koschmider, N. Baracaldo, M. Weidlich, J. Michael: Privacy-Preserving Process Mining - Differential Privacy for Event Logs. Business & Information Systems Engineering 61(5): 595-614 (2019)

G. Elkoumy, S.A. Fahrenkrog-Petersen, M. Fani Sani, A. Koschmider, F. Mannhardt, S. Nuñez von Voigt, M. Rafiei, L. von Waldthausen: Privacy and Confidentiality in Process Mining - Threats and Research Challenges, ACM Transactions of Management Information Systems, 2021