Privacy-Aware Analysis based on Data Series

Stephan Fahrenkrog-Petersen
Humboldt-Universität zu Berlin
Berlin, Germany
fahrenks@hu-berlin.de

Han van der Aa
University of Mannheim
Mannheim, Germany
han@informatik.uni-mannheim.de

Matthias Weidlich
Humboldt-Universität zu Berlin
Berlin, Germany
matthias.weidlich@hu-berlin.de

***Abstract*—Data that is recorded about the operations of an organization constitutes a valuable source of information for monitoring and improvement. Specific use cases include the assessment of compliance to legal regulations, the analysis of performance bottlenecks, or the optimization of resource utilization. In recent years, a plethora of algorithms for operational analysis using data series, summarized as process mining, have been developed to support these use cases, e.g., by constructing models for simulation and prediction or by comparing the recorded data against a normative specification of a process.**

Data series often contain sensitive information, though, about the individuals that act as service consumers or service providers. Personal information is only partially hidden by obfuscation and pseudonymization and potential privacy breaches need to be prevented for ethical, legal, and economic reasons.

This tutorial is devoted to methods for privacy-aware analysis using data series. It covers essential notions, reviews privacy-disclosure attacks, and outlines techniques to give formal privacy guarantees while largely maintaining the data's utility for operational analysis. The discussion is structured by the adopted perspective on the privacy of individuals, and the degree to which a data series contains contextual information.

I. Overview

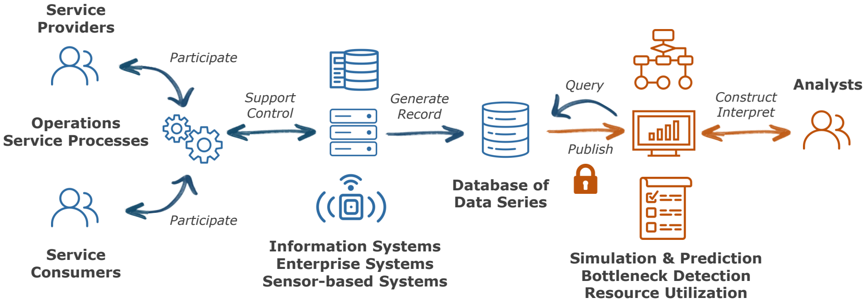
Large organizations produce an ever increasing amount of data that is linked to their operational processes [9]. This phenomenon is observed for traditional business processes such as *order-to-cash* and *procure-to-pay*, as well as for service processes in diverse domains, such as clinical pathways in healthcare and transportation chains in logistics. In these areas, the operations of organizations are widely supported by information systems that record events, messages, transactions, and service calls, which can be linked to the progress of process execution, thereby providing an angle to monitor and improve an organization's operations [40].

We illustrate the essence of operational analysis using data series in Figure 1. That is, operations and service processes, while conducted, involve individuals as service providers and service consumers. For example, in a healthcare environment, a service process may describe the clinical pathway of a patient (i.e., a service consumer) that involves providers such as nurses, clinical assistants, medical doctors, and administrative staff [29]. Various types of systems support and control these service processes, which yields data traces of the operations. In the medical domain, examples for such systems are healthcare information systems, appointment booking systems, or even real-time locating systems that track patients and staff [37]. Through mechanism for data extraction, correlation, and abstraction [8], [35], one then obtains a database of data series, a set of sequences of events that indicate the operational progress for a certain case. For instance, an event may denote the start of a treatment step, while events are further grouped per patient.

A database of data series serves as the starting point for operational analysis. Information is extracted from it by query mechanisms and the results are fed into analysis algorithms. Specific examples for such algorithms include the construction of models to understand, assess, and simulate the service processes [4], [30]; the derivation of predictions about outcomes and performance characteristics [39]; the assessment of changes of a system [1]; and the detection of congestion effects due to competition for scarce resources [36]. Based on data series of a clinical pathway, for instance, statistics on the variants of treatment procedures may be derived and used to construct models of the main patient flow at a particular hospital department. Based thereon, what-if analysis enables an optimization of staffing and the derivation of predictors for the expected wait-time of patients.

Relevance of Privacy-awareness. In recent years, organizations intensified their efforts to collect fine-granular and accurate operational data. Often, this involves the collection of sensitive data on the involved service providers and service consumers [28] and the presence of respective databases may violate the principle of informational self-determination, meaning the ability of a person to control the access and use of their personal data [3]. Potential breaches of privacy have to be avoided not only for ethical reasons, though. In recent years, various legal measures have been put in place, which prohibit the processing of personal data without prior consent, with the GDPR being a prominent example for such legislation.

Against this background, techniques to privatize data publishing have been developed, based on well-established privacy guarantees: *k-anonymity* [38], which ensures that an individual cannot be distinguished from at least k other individuals; *t-closeness* [24], which ensures that the distribution of sensitive information between equivalence groups of individuals differs by at most a distance of t; *ϵ-differential privacy* [10], which adds noise to the data to ensure that the impact of an individual is probabilistically bound by a threshold based on ϵ. However, to achieve these privacy guarantees, the data needs to be transformed before publishing, which typically induces a trade-off between the strength of privacy guarantee and the loss in utility of the data for some analysis [22].



images: Flaticon.com

Fig. 1: The essence of operational analysis using data series.

The need to investigate attacks on the privacy of individuals is widely acknowledged in the database community. Privacy guarantees have been explored for diverse types of data, starting with traditional relational data [23], through social media data [20], to models that capture the provenance of data after some processing [7]. A similar diversity is observed for the notions of data utility that are induced by specific queries and functions to evaluate over the published data, such as counting queries [19], stochastic gradient descent [44], and complex pattern detection [42]. Overviews of these techniques can be found in earlier tutorials [18], [26].

In this tutorial, we strive for a comprehensive overview of techniques for privacy-aware data processing with a particular focus: We adopt data series as the data model and operational analysis as a use case that induces a particular set of queries and functions that shall be evaluated over the series.

Timeliness of the Tutorial. First ideas on data-driven process analysis have been presented more than twenty years ago [43]. Since then, the field has experienced an enormous development [6], [40]. With analysis techniques becoming more elaborate, the challenge to ensure privacy while maintaining the utility of the data grew as well. Diverse analysis algorithms impose dedicated requirements on the properties to preserve when transforming data series before publishing them.

Privacy-awareness in the analysis based on data series received much attention recently, and a first consolidated set of techniques has been proposed, as outlined in the remainder. At the same time, there is a large number of open scientific problems, reaching from theoretical aspects (bounds for utility loss), through data management topics (impact of data quality issues), to system considerations (performance optimizations).

Goals of the Tutorial. The tutorial exposes the audience to the state of the art for privacy-aware operational analysis using data series. It enables the participants to understand the setting in terms of common notions of data series along with basic queries and functions that are evaluated over them. The tutorial raises awareness for privacy-disclosure attacks on sets of data series and offers a precise, formal characterization of them. It then covers methods to achieve privacy guarantees.

Participants of the tutorial are enabled to conduct research in the field; to enhance the integration of privacy-aware operational analysis with related questions in general data management and engineering; and to explore the transfer of the discussed techniques to other application scenarios.

II. Target Audience & Assumed Background

The tutorial is primarily aimed at researchers and shall assist them in entering the field of privacy-aware analysis using data series. Based on the concepts presented in the tutorial, we highlight a number of open scientific problems, which we expect to be particularly beneficial for young researchers, who are in the course of defining their research focus.

We expect the tutorial to also be relevant to practitioners, who will obtain insights on how operational analysis can be conducted in a privacy-preserving manner.

The research covered in this tutorial is rooted in data engineering, data mining, and information systems analysis. To follow the tutorial, a basic understanding of data structures, graph theory, and probability theory is required. However, we strive for a gentle introduction of the foundations through a large number of illustrative examples. Moreover, some basic understanding of the Python programming language will be useful to follow the hands-on session that illustrates the application of the concepts to real-world data.

III. Scope and Structure

A. *General Organization*

We partition the tutorial along two dimensions, see Table I.

The *privacy target* dimension partitions the existing approaches into those that deal with privacy considerations related to an intra-case perspective or an inter-case perspective. A data series may be linked to a notion of a case that is defined by a service consumer, i.e., data is grouped per client, customer, or requester. However, individuals are often also involved across cases, so that their sensitive information is distributed over multiple data series. Assuming that the notion of a case is induced by the service consumer, privacy requirements have to be incorporated for the service providers that potentially contribute to many cases. In a healthcare scenario, for instance, a database may record treatment paths per patient. Then, privacy requirements for the intra-case perspective relate to patients, whereas privacy requirements for the inter-case perspective relate to nurses, clinical assistants, medical doctors, and administrative staff of a hospital.

The *richness of the data series*, in turn, characterizes the type of information that is considered in privacy-aware analysis. Data series capture primarily the progress of operations, i.e., they

TABLE I: Overview of techniques for privacy-awareness.

Privacy Target	**Richness of the Data Series**	
	Only Sequence	With Context
Intra-case (Service Consumer)	Noise insertion for queries for groups of series [13], [28], potentially including semantic constraints [14], to achieve differential privacy	Enrich the results of grouping queries with noisy contextual information to achieve local differential privacy [15]
Inter-case (Service Provider)	Strategies to merge data series based on distance metrics to achieve k-anonymity [16], [33]	Combine strategies to merge data series with those that merge attribute values to achieve l-diversity or t-closeness [16], [33]; transformation of the assignment of attribute values [5]

indicate the control-flow according to which operational steps have been conducted. Yet, they may also include contextual information, e.g., on temporal aspects (e.g., start and end times, durations), spatial aspects (e.g., locations, paths), or properties of involved entities (e.g., resources, systems). Turning to the healthcare scenario again, privacy considerations may only relate to the sequences of treatment steps, or also include treatment durations, patient ages, or administered drugs.

We plan for a tutorial of **three hours**:

Part I (0.5 hours)

- Introduction: Background on operational processes; use cases for their data-driven assessment; overview of basic analysis techniques; example scenario.
- Model: Data series and background knowledge; re-identification risks.

Part II (1.5 hours)

- Intra-case, control-flow perspective: Formal definition of privacy-disclosure attacks; techniques to achieve differential privacy for queries that group data series; extensions to incorporate semantic constraints.
- Intra-case, contextual perspective: Attribute-disclosure attacks that include attribute values; techniques to achieve local differential privacy.
- Inter-case, control-flow perspective: Formal definition of identify-disclosure and membership attacks; techniques to achieve k-anonymity through merging of data series.
- Intra-case, contextual perspective: Formal definition of attribute-disclosure attacks; techniques to achieve l-diversity and t-closeness for attributes; decomposition of events to perturb the distributions of contextual information.

Part III (1 hour)

- Hands-on: Jupyter notebook to illustrate basic steps of privacy-aware operational analysis based on real-world data.
- Outlook on the broader field: Techniques for continuous data publishing; privacy-aware computation of performance indicators; secure multi-party computation for privacy-aware operational analysis.
- Summary of current challenges and research directions.

B. Detailed Content

Part I.1: Introduction. This first part discusses the context of operational analysis. We provide background on processes and explain common use cases for their analysis, such as bottleneck detection and the optimization of resource usage. Also, the queries and functions induced by common analysis techniques are reviewed. The concepts will be illustrated with examples from the healthcare domain, where privacy-preserving analysis techniques are of particular importance [31].

Part I.2: Model. To lay the foundations, we introduce common models of data series. Those adopt a relational notion of tuples that denote events, i.e., typed data elements, see Table II. We further highlight the assumptions to be satisfied, e.g., in terms of ordering and data partitioning. In Table II, for instance, the events are ordered by the assigned timestamp, have a type that is derived from the treatment step, and can be partitioned based on a patient identifier.

TABLE II: Healthcare example: Events of a data series indicate the progress of the clinical pathway of a patient.

ID	**Time**	**Patient ID**	**Step**	**Sex**	**Age**	**Drug**	**Staff ID**
e_{42}	10:31	221	Registration	F	54		S26
e_{43}	10:48	221	Vitals				S11
e_{44}	10:50	224	Registration	M	23		S26
e_{45}	10:58	221	Blood Test				S11
e_{46}	11:29	221	Admission Doctor				S05
e_{47}	11:31	224	Vitals				S11
e_{48}	11:49	221	Med. Subscription			Cephalexin	S05

For the introduced model of data series, we then discuss general re-identification risks based on uniqueness measures [41].

Finally, we introduce common types of background knowledge that may be employed in privacy-disclosure attacks, using the classification presented in [33]. It separates such knowledge based on the adopted data types (e.g., sets or sequences) and the semantics of an event (e.g., referring only to an operational step, or also the involved individuals).

Part II.1: Intra-case, control-flow perspective. A series often denotes a service consumer, e.g., a patient, and the control-flow encoded by it may reveal sensitive information, e.g., through treatment steps that relate to an HIV infection. An adversary may link the series with background knowledge. This way, an adversary may re-identify an individual.

A protection against such an attack is offered by injecting differentially private noise to the control-flow data. As a result, the service consumer would be protected through *plausible deniability*. Specifically, common analysis techniques rely on counting queries for data series, so that a differentially private result is achieved by having these queries return a list the noisy counts of series prefixes. State-of-the-art approaches tackle the problem through a step-wise construction of a prefix-tree, where only prefixes with a positive noisy count are expanded [28]. These techniques may be enhanced by using the exponential mechanism to nudge the prefix-tree expansion towards semantically meaningful prefixes [14]. Alternatively, noisy counts can be generated by oversampling of series [13].

Part II.2: Intra-case, contextual perspective. In addition to the control-flow information, contextual information, such as the age and sex of a patient or the subscribed drugs, may be subject to a privacy-disclosure attack. An adversary could link such contextual information either background knowledge.

To prevent such an attack, it is necessary to offer protection for the contextual information of a case. Here, one may resort to differentially private counting queries [15], or to oversampling of series that are at risk, and injecting noise to the timestamps [13].

Part II.3: Inter-case, control-flow perspective. An adversary might also try to gain sensitive information about the individuals that contribute to multiple data series. In practice, this perspective typically relates to service providers, and sensitive information includes details on work habits and performance.

To prevent such attacks, one may minimize the deviation in the recorded process progress to a certain degree, mainly by grouping together similar data series. Here, the similarity of series may be defined based on syntactic edit distances, or also incorporate the semantics of the individual steps. Privacy is then aimed at in terms of group-based notions, such as k-anonymity. Specifically, all differentiating events may be suppressed from a series, until only groups that satisfy the privacy requirements are left [33]. As this may introduce new types of series in the database, one may also merge series into groups of similar series [16].

Part II.4: Inter-case, contextual perspective. It becomes easier to gain information about individuals that participate in multiple data series, i.e., service providers, if contextual information is known. To prevent the respective attacks, the aforementioned strategies for the control-flow perspective need to be enhanced with mechanisms to anonymize the contextual information. Notions such as l-diversity and t-closeness enable us to limit the differences between distributions of the values of context attributes assigned to data series [16], [33]. Then, merging of series is governed not only by their uniqueness, but also by the value distributions for context attributes.

Part III.1: Hands-on. A hands-on session, planned for around 30min, demonstrates the application of some of the privacy-aware analysis techniques. To this end, we rely on a Jupyter notebook and a real-world dataset on the treatment of Sepsis patients at a Dutch hospital [27]. We walk the participants through the following steps: (i) general data exploration; (ii) queries for performance analysis; (iii) a linkage attack on the privacy of the patients; (iv) mitigation of the attack risk through a differentially private version of the queries.

Technically, participants may load the notebook, provided on Github, using Google Colaboratory.[1] This way, participants are able to run and explore the code on their own laptop.

Part III.2: Outlook on the broader field. We provide an outlook on related techniques to achieve privacy-aware processing of data series. That is, privacy considerations may be woven directly into analysis techniques, e.g., the computation of performance indicators [21] or for the analysis of the roles of service providers [32]. We also summarize recent ideas on on the use of Multi-Party-Computation for operational analysis [11] and continuous publishing of data series [34].

Part III.3: Research directions. We conclude with a discussion of challenges for privacy-aware operational analysis [12] and outlined directions for future research.

IV. Related Tutorials

This tutorial has not been presented before. Most closely related are tutorials given by Machanavajjhala, He, and Hay at VLDB 2016 [25] and SIGMOD 2017 [26] on techniques to achieve differential privacy. However, the tutorials focussed on differential privacy for relational data. While they alluded to privatization of network data and trajectories, they did not cover attacks and privacy notions for sequential data. Moreover, Anciaux, Nguyen, and Popa presented a tutorial on managing personal data with strong privacy guarantees [2] at EDBT 2014, which covered platforms for decentralized management of personal data. Such infrastructure considerations are largely orthogonal to our focus on analysis algorithms.

One of the presenters gave a tutorial on temporal analysis of complex systems [17] at ICDE 2018. It focused on sensor data, though, and did not include privacy considerations.

V. Presenters

The three presenters co-authored several research papers on privacy-aware data analysis, including [14]–[16], [28], [41].

Stephan A. Fahrenkrog-Petersen is a research group lead at the Weizenbaum Institute, Germany. He holds a PhD from Humboldt-Universität zu Berlin. His research was published in the proceedings of the premier conferences in the field and in international journals, such as ACM TMIS, DKE, and KAIS. His work received the Distinguished Paper Award at CAiSE 2021 and the Best Student Paper Award at ICPM 2021.

Han van der Aa is a junior professor in the Data and Web Science Group at the University of Mannheim, Germany. He obtained a PhD from the Vrije Universiteit Amsterdam in 2018. His research interests include process modelling, process mining, natural language processing, and complex event processing. His work has been published in journals including IEEE TKDE, Information Systems, and Decision Support Systems and at the BPM, CAISE, ICPM, ICDE, and SIGMOD conferences.

Matthias Weidlich is a full professor and Chair of Databases and Information Systems at Humboldt-Universität zu Berlin, Germany. Matthias' research focuses on process-oriented and event-based information systems. His results appear regularly in premier conferences (SIGMOD, VLDB, ICDE, IJCAI, AAAI, BPM, CAiSE) and journals (TKDE, Information Systems, VLDB Journal) in the field. He serves as Co-Editor-in-Chief for the Information Systems journal and is a member of the steering committee of the ACM DEBS conference series. In the past, he gave tutorials at ICDE, AAMAS, and DEBS.

[1] https://colab.research.google.com/

References

[1] J. N. Adams, C. Pitsch, T. Brockhoff, and W. M. P. van der Aalst. An experimental evaluation of process concept drift detection. *Proc. VLDB Endow.*, 16(8):1856–1869, 2023.

[2] N. Anciaux, B. Nguyen, and I. S. Popa. Tutorial: Managing personal data with strong privacy guarantees. In S. Amer-Yahia, V. Christophides, A. Kementsietsidis, M. N. Garofalakis, S. Idreos, and V. Leroy, editors, *Proceedings of the 17th International Conference on Extending Database Technology, EDBT 2014, Athens, Greece, March 24-28, 2014*, pages 672–673. OpenProceedings.org, 2014.

[3] T. Asikis and E. Pournaras. Optimization of privacy-utility trade-offs under informational self-determination. *Future Gener. Comput. Syst.*, 109:488–499, 2020.

[4] A. Augusto, R. Conforti, M. Dumas, M. L. Rosa, F. M. Maggi, A. Marrella, M. Mecella, and A. Soo. Automated discovery of process models from event logs: Review and benchmark. *IEEE Trans. Knowl. Data Eng.*, 31(4):686–705, 2019.

[5] E. Batista and A. Solanas. A uniformization-based approach to preserve individuals' privacy during process mining analyses. *Peer-to-Peer Networking and Applications*, 14(3):1500–1519, 2021.

[6] J. Carmona, B. F. van Dongen, A. Solti, and M. Weidlich. *Conformance Checking - Relating Processes and Models*. Springer, 2018.

[7] D. Deutch, A. Frankenthal, A. Gilad, and Y. Moskovitch. On optimizing the trade-off between privacy and utility in data provenance. In G. Li, Z. Li, S. Idreos, and D. Srivastava, editors, *SIGMOD '21: International Conference on Management of Data, Virtual Event, China, June 20-25, 2021*, pages 379–391. ACM, 2021.

[8] K. Diba, K. Batoulis, M. Weidlich, and M. Weske. Extraction, correlation, and abstraction of event data for process mining. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 10(3), 2020.

[9] M. Dumas, M. L. Rosa, J. Mendling, and H. A. Reijers. *Fundamentals of Business Process Management, Second Edition*. Springer, 2018.

[10] C. Dwork. Differential privacy: A survey of results. In M. Agrawal, D. Du, Z. Duan, and A. Li, editors, *Theory and Applications of Models of Computation, 5th International Conference, TAMC 2008, Xi'an, China, April 25-29, 2008. Proceedings*, volume 4978 of *Lecture Notes in Computer Science*, pages 1–19. Springer, 2008.

[11] G. Elkoumy, S. A. Fahrenkrog-Petersen, M. Dumas, P. Laud, A. Pankova, and M. Weidlich. Secure multi-party computation for inter-organizational process mining. In S. Nurcan, I. Reinhartz-Berger, P. Soffer, and J. Zdravkovic, editors, *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*, volume 387 of *Lecture Notes in Business Information Processing*, pages 166–181. Springer, 2020.

[12] G. Elkoumy, S. A. Fahrenkrog-Petersen, M. F. Sani, A. Koschmider, F. Mannhardt, S. N. n. Von Voigt, M. Rafiei, and L. V. Waldthausen. Privacy and confidentiality in process mining: Threats and research challenges. *ACM Trans. Manage. Inf. Syst.*, 13(1), oct 2021.

[13] G. Elkoumy, A. Pankova, and M. Dumas. Mine me but don't single me out: Differentially private event logs for process mining. In C. D. Ciccio, C. D. Francescomarino, and P. Soffer, editors, *3rd International Conference on Process Mining, ICPM 2021, Eindhoven, Netherlands, October 31 - Nov. 4, 2021*, pages 80–87. IEEE, 2021.

[14] S. A. Fahrenkrog-Petersen, M. Kabierski, F. Rösel, H. van der Aa, and M. Weidlich. Sacofa: Semantics-aware control-flow anonymization for process mining. In C. D. Ciccio, C. D. Francescomarino, and P. Soffer, editors, *3rd International Conference on Process Mining, ICPM 2021, Eindhoven, Netherlands, October 31 - Nov. 4, 2021*, pages 72–79. IEEE, 2021.

[15] S. A. Fahrenkrog-Petersen, H. van der Aa, and M. Weidlich. PRIPEL: privacy-preserving event log publishing including contextual information. In D. Fahland, C. Ghidini, J. Becker, and M. Dumas, editors, *Business Process Management - 18th International Conference, BPM 2020, Seville, Spain, September 13-18, 2020, Proceedings*, volume 12168 of *Lecture Notes in Computer Science*, pages 111–128. Springer, 2020.

[16] S. A. Fahrenkrog-Petersen, H. van der Aa, and M. Weidlich. Optimal event log sanitization for privacy-preserving process mining. *Data Knowl. Eng.*, 145:102175, 2023.

[17] A. Gal, A. Senderovich, and M. Weidlich. Online temporal analysis of complex systems using iot data sensing. In *34th IEEE International Conference on Data Engineering, ICDE 2018, Paris, France, April 16-19, 2018*, pages 1727–1730. IEEE Computer Society, 2018.

[18] M. Hay, K. Liu, G. Miklau, J. Pei, and E. Terzi. Privacy-aware data management in information networks. In T. K. Sellis, R. J. Miller, A. Kementsietsidis, and Y. Velegrakis, editors, *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2011, Athens, Greece, June 12-16, 2011*, pages 1201–1204. ACM, 2011.

[19] M. Hay, V. Rastogi, G. Miklau, and D. Suciu. Boosting the accuracy of differentially private histograms through consistency. *Proc. VLDB Endow.*, 3(1):1021–1032, 2010.

[20] Z. Jorgensen, T. Yu, and G. Cormode. Publishing attributed social graphs with formal privacy guarantees. In F. Özcan, G. Koutrika, and S. Madden, editors, *Proceedings of the 2016 International Conference on Management of Data, SIGMOD Conference 2016, San Francisco, CA, USA, June 26 - July 01, 2016*, pages 107–122. ACM, 2016.

[21] M. Kabierski, S. A. Fahrenkrog-Petersen, and M. Weidlich. Hiding in the forest: Privacy-preserving process performance indicators. *Inf. Syst.*, 112:102127, 2023.

[22] D. Kifer and A. Machanavajjhala. No free lunch in data privacy. In T. K. Sellis, R. J. Miller, A. Kementsietsidis, and Y. Velegrakis, editors, *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2011, Athens, Greece, June 12-16, 2011*, pages 193–204. ACM, 2011.

[23] C. Li, M. Hay, G. Miklau, and Y. Wang. A data- and workload-aware query answering algorithm for range queries under differential privacy. *Proc. VLDB Endow.*, 7(5):341–352, 2014.

[24] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In R. Chirkova, A. Dogac, M. T. Özsu, and T. K. Sellis, editors, *Proceedings of the 23rd International Conference on Data Engineering, ICDE 2007, The Marmara Hotel, Istanbul, Turkey, April 15-20, 2007*, pages 106–115. IEEE Computer Society, 2007.

[25] A. Machanavajjhala, X. He, and M. Hay. Differential privacy in the wild: A tutorial on current practices & open challenges. *Proc. VLDB Endow.*, 9(13):1611–1614, 2016.

[26] A. Machanavajjhala, X. He, and M. Hay. Differential privacy in the wild: A tutorial on current practices & open challenges. In S. Salihoglu, W. Zhou, R. Chirkova, J. Yang, and D. Suciu, editors, *Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD Conference 2017, Chicago, IL, USA, May 14-19, 2017*, pages 1727–1730. ACM, 2017.

[27] F. Mannhardt and D. Blinde. Analyzing the trajectories of patients with sepsis using process mining. In J. Gulden, S. Nurcan, I. Reinhartz-Berger, W. Guédria, P. Bera, S. Guerreiro, M. Fellmann, and M. Weidlich, editors, *Joint Proceedings of the Radar tracks at the 18th International Working Conference on Business Process Modeling, Development and Support (BPMDS), and the 22nd International Working Conference on Evaluation and Modeling Methods for Systems Analysis and Development (EMMSAD), and the 8th International Workshop on Enterprise Modeling and Information Systems Architectures (EMISA) co-located with the 29th International Conference on Advanced Information Systems Engineering 2017 (CAiSE 2017), Essen, Germany, June 12-13, 2017*, volume 1859 of *CEUR Workshop Proceedings*, pages 72–80. CEUR-WS.org, 2017.

[28] F. Mannhardt, A. Koschmider, N. Baracaldo, M. Weidlich, and J. Michael. Privacy-preserving process mining - differential privacy for event logs. *Bus. Inf. Syst. Eng.*, 61(5):595–614, 2019.

[29] R. Mans, W. M. P. van der Aalst, and R. J. B. Vanwersch. *Process Mining in Healthcare - Evaluating and Exploiting Operational Healthcare Processes*. Springer Briefs in Business Process Management. Springer, 2015.

[30] M. Martini, D. Schuster, and W. M. P. van der Aalst. Mining frequent infix patterns from concurrency-aware process execution variants. *Proc. VLDB Endow.*, 16(10):2666–2678, 2023.

[31] A. Pika, M. T. Wynn, S. Budiono, A. H. M. ter Hofstede, W. M. P. van der Aalst, and H. A. Reijers. Towards privacy-preserving process mining in healthcare. In C. D. Francescomarino, R. M. Dijkman, and U. Zdun, editors, *Business Process Management Workshops - BPM 2019 International Workshops, Vienna, Austria, September 1-6, 2019, Revised Selected Papers*, volume 362 of *Lecture Notes in Business Information Processing*, pages 483–495. Springer, 2019.

[32] M. Rafiei and W. M. P. van der Aalst. Mining roles from event logs while preserving privacy. In *Business Process Management Workshops - BPM 2019 International Workshops, Vienna, Austria, September 1-6, 2019, Revised Selected Papers*, pages 676–689, 2019.

[33] M. Rafiei and W. M. P. van der Aalst. Group-based privacy preservation techniques for process mining. *Data Knowl. Eng.*, 134:101908, 2021.

[34] M. Rafiei and W. M. P. van der Aalst. Privacy-preserving continuous event data publishing. In A. Polyvyanyy, M. T. Wynn, A. V. Looy, and M. Reichert, editors, *Business Process Management Forum - BPM Forum 2021, Rome, Italy, September 06-10, 2021, Proceedings*, volume 427 of *Lecture Notes in Business Information Processing*, pages 178–194. Springer, 2021.

[35] A. Rebmann, M. Weidlich, and H. van der Aa. GECCO: constraint-driven abstraction of low-level event logs. In *Proceedings of the 38th International Conference on Data Engineering, ICDE 2022*. IEEE Computer Society, 2022.

[36] A. Senderovich, J. C. Beck, A. Gal, and M. Weidlich. Congestion graphs for automated time predictions. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 4854–4861. AAAI Press, 2019.

[37] A. Senderovich, M. Weidlich, L. Yedidsion, A. Gal, A. Mandelbaum, S. Kadish, and C. A. Bunnell. Conformance checking and performance improvement in scheduled processes: A queueing-network perspective. *Inf. Syst.*, 62:185–206, 2016.

[38] L. Sweeney. k-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.*, 10(5):557–570, 2002.

[39] I. Teinemaa, M. Dumas, M. L. Rosa, and F. M. Maggi. Outcome-oriented predictive process monitoring: Review and benchmark. *TKDD*, 13(2):17:1–17:57, 2019.

[40] W. M. P. van der Aalst. *Process Mining - Data Science in Action, Second Edition*. Springer, 2016.

[41] S. N. von Voigt, S. A. Fahrenkrog-Petersen, D. Janssen, A. Koschmider, F. Tschorsch, F. Mannhardt, O. Landsiedel, and M. Weidlich. Quantifying the re-identification risk of event logs for process mining - empiricial evaluation paper. In S. Dustdar, E. Yu, C. Salinesi, D. Rieu, and V. Pant, editors, *Advanced Information Systems Engineering - 32nd International Conference, CAiSE 2020, Grenoble, France, June 8-12, 2020, Proceedings*, volume 12127 of *Lecture Notes in Computer Science*, pages 252–267. Springer, 2020.

[42] D. Wang, Y. He, E. A. Rundensteiner, and J. F. Naughton. Utility-maximizing event stream suppression. In K. A. Ross, D. Srivastava, and D. Papadias, editors, *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2013, New York, NY, USA, June 22-27, 2013*, pages 589–600. ACM, 2013.

[43] A. J. M. M. Weijters and W. M. P. van der Aalst. Rediscovering workflow models from event-based data using little thumb. *Integr. Comput. Aided Eng.*, 10(2):151–162, 2003.

[44] X. Wu, F. Li, A. Kumar, K. Chaudhuri, S. Jha, and J. F. Naughton. Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. In S. Salihoglu, W. Zhou, R. Chirkova, J. Yang, and D. Suciu, editors, *Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD Conference 2017, Chicago, IL, USA, May 14-19, 2017*, pages 1307–1322. ACM, 2017.