Federated Stochastic Process Discovery: Definition, Benefits, and Challenges

Hootan Zhian[1]

[1] *The University of Melbourne, Victoria 3010, Australia*

Abstract

A process model is a representation that describes the behavior of a system, constructed by applying process mining algorithms to data found in event logs. The primary advantage of a process model is that it provides a clear understanding of how processes function. The concept of process mining initially emerged from research focused on extracting sequences of activities within workflow environments. Today, modern organizations widely use process mining techniques, including process discovery, to analyze their event logs. These techniques help organizations gain insights into the reality of their operational processes and identify opportunities for improvement. However, in practical scenarios, event log data is often distributed and may contain sensitive information. Traditional process discovery methods rely on event logs stored in centralized repositories. This centralization, however, poses challenges in distributed environments, such as concerns about data availability, privacy, and the high communication and bandwidth demands associated with centralization. I explore how organizations can collaboratively create a process model without sharing their local logs, addressing challenges in distributed environments. Additionally, I discuss how to optimize the quality of extracted models.

Keywords

Stochastic process mining, federated process discovery, cross-silos process discovery, optimization

1. Introduction

Process mining has become an important field that bridges traditional business process management with data-driven analytics. It enables organizations to uncover, monitor, and enhance their actual processes by extracting valuable insights from event logs recorded during business operations. As organizations increasingly digitize their activities, the volume and distribution of process-related data have expanded significantly, presenting both opportunities and challenges for the implementation of process mining techniques. A key problem in process mining is process discovery, which focuses on automatically creating process models from event logs. These models serve as visual representations of actual business processes, helping organizations understand, analyze, and improve their operations [1]. However, traditional process discovery approaches face significant limitations in today's distributed business environment, particularly when dealing with cross-organizational processes and sensitive data [2].

The emergence of federated process discovery addresses the limitations of centralized approaches by enabling organizations to collaborate on process mining initiatives without sharing raw event log data [3]. This approach is particularly crucial in scenarios where data privacy regulations, competitive concerns, or technical constraints prevent the centralization of event logs. Moreover, process discovery is inherently a time-consuming task, which limits the ability to exploit optimization solutions, such as metaheuristics, to a full extent [4]. In contrast, federated stochastic process discovery offers a significant advantage by allowing organizations to avoid the need to aggregate all event logs. This decentralized approach facilitates the application of optimization techniques, enabling a more effective exploration of the solution space to identify optimized solutions.

However, despite these advancements, process discovery still faces challenges related to the evaluation of process models across different and conflicting quality dimensions. Process discovery is a multi-objective optimization problem. Yet, most existing process discovery techniques neglect this fact.

23rd International Conference on Business Process Management

hzhian@student.unimelb.edu.au (H. Zhian)

Traditionally, process discovery techniques aggregate different objectives into a single objective (SO) function. This approach, however, comes with limitations, such as the need for a priori knowledge of objectives' importance and difficulty in evaluating trade-offs between the objectives. Recently, metaheuristic optimization techniques have been used in process discovery [4], particularly when it comes to balancing different quality goals. This success is also seen in the works by Buijs et al. [5, 6], who use multiple quality dimensions to evaluate the quality of discovered process models. Maintaining a diverse list of process models is vital in multi-objective optimization to ensure solutions are uniformly distributed over the search space. Without preventive measures, populations of discovered models can cluster, leading to traps in local maxima.

My PhD project aims to answer these research questions (RQs):
RQ1: How to perform federated stochastic process discovery effectively and efficiently?
RQ2: Can multi-objective metaheuristics improve solutions to the stochastic process discovery problem?
RQ3: How to perform federated stochastic process discovery in online settings?

2. Research Problem

My PhD project addresses three critical challenges in modern process mining. First, organizations increasingly need to collaborate on process mining initiatives while maintaining data privacy. Traditional centralized approaches require sharing raw data, which is often impossible due to privacy regulations, competitive concerns, technical constraints, and high communication and bandwidth costs. Second, current process discovery methods struggle to balance multiple objectives, including model performance metrics like fitness, precision, and simplicity. Third, in spite of federated process discovery potentials, it introduces new challenges, including the complexity of merging distributed process models, maintaining data privacy, and preserving good quality models that reflect the behavior of the whole organization.

3. Related Work

In practice, event logs are often distributed and may contain sensitive data, challenging traditional process mining, which assumes centralized logs. To address this, van der Aalst [3] proposed federated process mining, mapping organization-specific logs into a unified federated log. Two approaches were proposed: sharing filtered event data or abstractions like directly-follows graphs (DFGs) to ensure confidentiality. However, merging these abstractions remains an unresolved challenge. Rojo et al. [7] explored federated mining using distributed devices, such as smartphones, to analyse human actions. Similarly, Khan et al. [2] developed a federated approach for cross-silo mining, using dependency graphs and a privacy-preserving protocol based on the Heuristic Miner algorithm, coordinated by a centralized server. Rafiei and van der Aalst [8] proposed a privacy-aware abstraction-based method for federated process discovery. Their approach uses directly-follows relations to create abstractions, enabling collaboration while protecting sensitive data through mechanisms like handover relations.

Finding the best process model that clearly shows all the details from the event log is a complex problem. It involves trying to meet different objectives that can sometimes conflict with each other, necessitating multi-objective optimization (MOO) techniques [9, 4]. Alkhammash et al. [4] recently demonstrated that process discovery based on the genetic optimization of the ALERGIA grammatical inference algorithm—called GASPD—can construct interesting DFGs in terms of size and accuracy of reflecting the likelihood of traces of the system that generated the event log. Addressing multiple objectives has typically involved consolidating them into a single objective (SO) function. However, this method has several drawbacks, such as the necessity of having prior knowledge about the relative importance of each objective, resulting in only one solution, and complicating the evaluation of trade-offs. The limitations of this approach include the need for an understanding of each objective's significance, the fact that the aggregated function yields only a single solution, the difficulty in assessing trade-offs between objectives, and the potential infeasibility of the solution unless the search space is convex. In contrast, multi-objective optimization (MOO) problems are inherently more complex, as they generate a set of optimal solutions that represent acceptable trade-offs among the various objectives rather than a single solution.

Multi-optimization evolutionary solutions in process discovery began with the work of Van der Aalst et al. [10], utilizing genetic algorithms to extract features from global searches while addressing noise challenges. This concept was further developed through methods like the Evolutionary Tree [11], a genetic approach is applied to identify Pareto frontiers. The evolutionary Miner [12]. Alkhammash et al. [4] have demonstrated the effectiveness of genetic algorithms for optimizing grammatical inference for the discovery of superior stochastic process models by extracting pareto-optimal models.

Maintaining a diverse population in multiobjective optimization (MOO) is crucial to ensure solutions are well-distributed across the Pareto front and to prevent genetic drift [13]. Niching techniques, such as fitness sharing and crowding, help algorithms explore multiple peaks and avoid local optima [14, 15]. These methods have proven effective in enhancing evolutionary algorithms' ability to solve complex, multimodal problems [14, 15].

4. Approach

4.1. Federated Stochastic Process Discovery

I presented an extension of GASPD capable of federated process discovery that operates over a distributed event log in two phases [16]. Unlike existing techniques, which typically rely on projections of complete traces, our approach works directly with multiple collections of complete traces. Each collection captures aspects of the overall process based on a specific data subset. In the first phase, GASPD discovers models with various quality characteristics from each part of the distributed event log, which can be viewed as a collection of local event logs stored on dedicated devices. These superior models, discovered from each local event log, are then sent to a central server. The server aggregates them into a model that describe the overall system.

The standard implementation, referred to as FedGASPD, discovers one model from each input event log and merges them, providing a basic baseline for federated process discovery. However, when the discovered models exhibit diverse characteristics, this merging may lead to a loss of distinctiveness and dilute the strong individual features due to the averaging effect.

To address this limitation, I proposed a variation called FedCGASPD, which avoids merging all models into a single output. Instead, it groups similar models into clusters and merges the models within each cluster. This approach yields multiple models that retain the strong characteristics of the merged models, thereby preserving their distinctiveness. This method supports scalable, efficient, and privacy-preserving process discovery across organizational boundaries.

4.2. Stochastic Process Discovery as Multi-Objective Optimization

I conducted a comprehensive review of multi-objective metaheuristics to identify those that efficiently support the discovery of models—based on the ALERGIA grammatical inference algorithm—that are simple (in terms of size), accurate (in terms of Entropic relevance), and diverse [17]. Entropic relevance is chosen as the accuracy measure because it assesses how well a model reflects the likelihood of target traces, balances precision and recall, and is computationally efficient, that is, is computable in time linear to the size of the input event log [18]. I introduced a classification of multi-objective metaheuristics as alternatives to the genetic algorithm used in GASPD, based on their strategies for selecting candidate solutions. Additionally,to ensure the discovered models exhibit a wide range of quality characteristics, I incorporated a niching technique, which allows metaheuristics to explore multiple promising search subspaces simultaneously, reducing the risk of premature convergence to local optima [14]. Finally, I conducted an empirical evaluation using industrial event logs.

5. Research Methodology

My PhD leverages the Design Science research methodology [19], focusing on creating innovative solutions for federated stochastic process discovery and optimization. The approach balances rigor—through literature review, benchmark construction, and simulation-based validation—with practical relevance by testing on real-world datasets and exploring organizational case studies. This dual emphasis ensures both theoretical advancement and actionable outcomes, addressing real implementation challenges. Ultimately, the research aims to bridge academic innovation with practical impact in federated process optimization, aligning with the core principles of Design Science [19].

6. Achieved Results and Future Work

I presented two algorithms for stochastic process discovery in distributed environments [16]. These algorithms are built upon GASPD that operates over centralized event logs. The new algorithms discover process models from several local event logs, possibly scattered across different organizations or silos, and aim to preserve the autonomy and privacy of each party and to decrease data communication requirements and the overall model discovery time. Our experiments demonstrate the effectiveness of FedGASPD and FedCGASPD, providing scalable alternatives for process discovery in distributed environments.

While federated approaches have shown promising results, several avenues remain to further enhance their effectiveness. One key area for improvement lies in systematically exploring the impact of different orders of merging models discovered from local event logs on the quality of the constructed global model. Such an exploration could provide valuable insights into optimizing the merging process.

Another promising direction involves optimizing the discovered models for quality criteria beyond size and Entropic relevance. Additionally, exploring alternative policies for selecting superior local models represents an intriguing avenue for future research. One critical challenge identified is that the process of merging models can introduce features into the global model that do not reflect any behavior present in the local logs. To address this, performing optimization directly on the global model could help achieve a more accurate representation of the overall behavior of the participating organizations.

To improve the quality and diversity of discovered models, I implemented and evaluated nine metaheuristics as alternatives to the Genetic Algorithm used in the GASPD stochastic process discovery algorithms, enhanced by niching techniques to ensure the diversity of the discovered models. For performance evaluation, I used two metrics: dominance count and Diversity Comparison Indicator (DCI) [20]. Dominance count evaluates how many solutions from one algorithm dominate others on the global Pareto front, while DCI assesses diversity by examining solution spread in objective space. Experiments on real event logs reveal that niching techniques enhance model diversity and help avoid local maxima. Empirical results across twelve logs show Differential Evolution (DE) consistently surpasses other metaheuristics, including GASPD. Future work may refine niching methods and improve DE and other metaheuristics for stronger process discovery algorithms.

Acknowledgments

My PhD program has been funded by the Melbourne Research Scholarship from the University of Melbourne. I want to thank Prof. Artem Polyvyanyy and Prof. Rajkumar Buyya for supporting and inspiring my PhD research.

Declaration on Generative AI

The author has not employed any Generative AI tools.

References

[1] W. van der Aalst, T. Weijters, L. Maruster, Workflow mining: Discovering process models from event logs, Transactions on Knowledge and Data Engineering 16 (2004) 1128–1142.

[2] A. Khan, A. Ghose, H. Dam, Cross-silo process mining with federated learning, Journal 13121 (2021) 612–626.

[3] W. van der Aalst, Federated process mining: Exploiting event data across organizational boundaries, 2021, pp. 1–7.

[4] H. Alkhammash, A. Polyvyanyy, A. Moffat, Stochastic directly-follows process discovery using grammatical inference, in: CAiSE, 2024, pp. 87–103.

[5] J. Buijs, B. van Dongen, W. van der Aalst, A genetic algorithm for discovering process trees, in: Proceedings of the IEEE Congress on Evolutionary Computation, 2012, pp. 1–8.

[6] J. Buijs, B. van Dongen, W. van der Aalst, Quality dimensions in process discovery: The importance of fitness, precision, generalization and simplicity, International Systems 23 (2014).

[7] J. Rojo, J. Garcia-Alonso, J. Berrocal, J. Hernández, J. Murillo, C. Canal, Sowcompact: A federated process mining method for social workflows, Information Sciences 595 (2022) 18–37.

[8] M. Rafiei, W. van der Aalst, An abstraction-based approach for privacy-aware federated process mining, IEEE Access 11 (2023) 33697–33714.

[9] A. Augusto, M. Dumas, M. Rosa, Metaheuristic optimization for automated business process discovery, in: Business Process Management, 2019, pp. 268–285.

[10] W. Van der Aalst, A. de Medeiros, A. Weijters, Genetic process mining, in: Applications and Theory of Petri Nets 2005, 2005.

[11] W. Van der Aalst, T. Weijters, L. Maruster, Workflow mining: Discovering process models from event logs, Transactions on Knowledge and Data Engineering 16 (2004) 1128–1142.

[12] T. Molka, D. Redlich, W. Gilani, X. Zeng, M. Drobek, Evolutionary computation-based discovery of hierarchical business process models, Springer, 2015.

[13] A. Konaka, D. Coit, A. Smith, Multi-objective optimization using genetic algorithms: A tutorial, Reliability Engineering and System Safety (2006) 992–1007.

[14] B. Sareni, L. Krahenbuhl, Fitness sharing and niching methods revisited, IEEE Transactions on Evolutionary Computation 2 (1998).

[15] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, A fast and elitist multiobjective genetic algorithm: Nsga-ii, IEEE Transactions on Evolutionary Computation 6 (2002) 182–197.

[16] H. Zhian, R. Buyya, A. Polyvyanyy, Federated stochastic process discovery using grammatical inference, in: Proceedings of the CAiSE 2025 Conference, 2025. Accepted for publication.

[17] H. Zhian, R. Buyya, Artem, Multi-objective metaheuristics for effective and efficient stochastic process discovery, in: Proceedings of the BPM 2025 Conference, 2025. Accepted for publication.

[18] H. Alkhammash, A. Polyvyanyy, A. Moffat, L. García-Bañuelos, Entropic relevance: A mechanism for measuring stochastic process models discovered from event data, Information Systems 107 (2022) 101922.

[19] R. Hevner, S. March, J. Park, R. Sudha, Design science in information systems research, Management Information Systems Quarterly 28 (2008) 6.

[20] J. Doe, J. Smith, Diversity comparison of pareto front approximations in many-objective optimization, Journal of Optimization Research (2023) 123–145.

A. Online Resources

Our implemented algorithms are publicly available via the following links.

- Federated Process Discovery.
- Process Discovery Optimization.