A Collection of Simulated Event Logs for Fairness Assessment in Process Mining

Timo **Pohl**, Alessandro **Berti**[1,*], Mahnaz Sadat **Qafari** and Wil M.P. van der **Aalst**[1]

[1]*Process and Data Science Chair, RWTH Aachen University, Aachen, Germany*

Abstract

The analysis of fairness in process mining is a significant aspect of data-driven decision-making, yet the advancement in this field is constrained due to the scarcity of event data that incorporates fairness considerations. To bridge this gap, we present a collection of simulated event logs, spanning four critical domains, which encapsulate a variety of discrimination scenarios. By simulating these event logs with CPN Tools, we ensure data with known ground truth, thereby offering a robust foundation for fairness analysis. These logs are made freely available under the CC-BY-4.0 license and adhere to the XES standard, thereby assuring broad compatibility with various process mining tools. This initiative aims to empower researchers with the requisite resources to test and develop fairness techniques within process mining, ultimately contributing to the pursuit of equitable, data-driven decision-making processes.

Keywords

Fairness in Process Mining, Simulated Event Logs, Discrimination Analysis, Data-Driven Decision Making

1. Introduction

The increasing integration of data science and machine learning into decision-making underscores the need for fairness [1, 2]. Despite widespread use, fields like process mining lack robust fairness measures due to a limited number of publicly available event logs [3, 4]. This study seeks to bridge this gap, providing simulated event logs via CPN Tools to aid the evolution of fairness techniques in process mining. The subsequent sections of this paper are organized as follows: Section 2 provides a detailed description of the event logs and their respective attributes. Section 3 offers preliminary insights and analyses derived from these event logs. Section 4 outlines the procedure for downloading and utilizing this resource, supplemented with an illustrative example using the process mining library *pm4py*. The paper culminates in Section 5, wherein we draw conclusions and offer a summation of our work.

Table 1
Basic statistics of the provided collection of event logs.

Event Log	Number of Events	Number of Cases	Number of Variants	Number of Activities
hiring_log_high-xes.gz	63869	10000	386	12
hiring_log_low-xes.gz	72094	10000	296	12
hiring_log_medium-xes.gz	69054	10000	382	12
hospital_log_high-xes.gz	69528	10000	80	10
hospital_log_low-xes.gz	70037	10000	106	10
hospital_log_medium-xes.gz	70124	10000	77	10
lending_log_high-xes.gz	58822	10000	41	12
lending_log_low-xes.gz	60746	10000	31	12
lending_log_medium-xes.gz	58668	10000	33	12
renting_log_high-xes.gz	89972	10000	496	16
renting_log_low-xes.gz	96440	10000	508	16
renting_log_medium-xes.gz	105555	10000	610	16

2. Description of the Resource

We present 12 distinct event logs, split evenly across four domains: hiring, healthcare, lending, and renting. Each log, embedded in the context of German society and containing 10,000 cases, has been meticulously curated and simulated using colored Petri nets[1]. The logs display varying degrees of discrimination within each domain, enabling researchers to delve into real-world scenarios. Table 1 provides some basic statistics on the collection, while Table 2 describes the sensitive attributes. While the chosen attributes related to fairness are debatable, we encourage discourse to refine our understanding of fairness in process mining. Each log's attributes and process are elaborately described to aid in identifying potential sources of discrimination.

All logs comply with the eXtensible Event Stream (XES) [6] standard format, promoting compatibility and interoperability with diverse process mining tools, thus facilitating research across various platforms.

2.1. Hiring Event Logs

Process: The data describes a multifaceted recruitment process with diverse application pathways ranging from minimal processing to extensive multi-step procedures. The variability of these routes, largely dependent on numerous determinants, yields a spectrum of outcomes from instant rejection to successful job offers.
Attributes: The logs include attributes such as age, citizenship, German proficiency, gender, religion, and years of education. While these attributes may inform candidate profiles, their

---

[1]Refer to [5] for a comprehensive explanation of the Petri nets and their parameters.

Table 2
Sensitive attributes in the provided collection of event logs.

Domain	Age	Citizenship	German Proficiency	Gender	Religion	Years of Education	Underlying Condition	Private Insurance
Hiring	X	X	X	X	X	X		
Hospital	X	X	X	X			X	X
Lending	X	X	X	X		X		
Renting	X	X	X	X	X	X		

misuse could engender discrimination. Variables like age and education may signify experience and skills, citizenship, and German language may address job logistics, but these should not unjustly eliminate applicants. Gender and religion, unrelated to job performance, must not sway hiring. Therefore, the use of these attributes must uphold fairness, avoiding any potential bias.

2.2. Hospital Event Logs

Process: The data depicts a hospital treatment process that commences with registration at an Emergency Room or Family Department and advances through stages of examination, diagnosis, and treatment. Notably, unsuccessful treatments often entail repetitive diagnostic and treatment cycles, underscoring the iterative nature of healthcare provision.

Attributes: The logs incorporate patient attributes such as age, underlying condition, citizenship, German language proficiency, gender, and private insurance. These attributes, influencing the treatment process, may unveil potential discrimination. Factors like age and condition might affect case complexity and treatment path, while citizenship may highlight healthcare access disparities. German proficiency can impact provider-patient communication, thus affecting care quality. Gender could spotlight potential health disparities, while insurance status might indicate socio-economic influences on care quality or timeliness. Therefore, a comprehensive examination of these attributes vis-à-vis the treatment process could shed light on potential biases or disparities, fostering fairness in healthcare delivery.

2.3. Lending Event Logs

Process: This data illustrates the steps within a loan application process. From an initial appointment request, the process navigates various stages, including information verification and underwriting, culminating in loan approval or denial. Additional steps may be required, such as co-signer enlistment or collateral assessment. Some cases experience outright appointment denial, indicating the process's variability, reflecting applicants' differing credit situations.

Attributes: The logs' attributes can aid in identifying influences on outcomes and detecting discrimination. Personal characteristics ('age', 'citizen', 'German speaking', and 'gender') and socio-economic indicators ('YearsOfEducation' and 'CreditScore') can impact the process. While 'YearsOfEducation' and 'CreditScore' can validly inform creditworthiness, 'age', 'citizen', 'language ability', and 'gender' should not bias loan decisions, ensuring these attributes are used responsibly fosters equitable loan processes.

2.4. Renting Event Logs

Process: The data encapsulates a detailed rental process, starting from an application to view a property to potential contract termination. Intermediate steps include initial screening, property viewing, decision-making, and possibly, a thorough screening. If approved, a rental agreement commences, where late payments can lead to eviction, one potential process endpoint. Alternatively, tenants might voluntarily end their contracts. Note that not all applications proceed to the viewing stage.

Attributes: The logs contain attributes that can shed light on potential biases in the process. 'Age', 'citizen', 'German speaking', 'gender', 'religious affiliation', and 'yearsOfEducation' might influence the rental process, leading to potential discrimination. While some attributes may provide useful insights into a potential tenant's reliability, misuse could result in discrimination. Thus, fairness must be observed in utilizing these attributes to avoid potential biases and ensure equitable treatment.

3. Preliminary Analysis

Performing an initial analysis of discrimination across the four contexts - hiring, hospital, lending, and renting - and primarily focusing on the control-flow (the sequence of activities) and time perspectives, notable disparities between the 'protected' and 'non-protected' groups (identified by the values in the cases' attributes) were observed. These discerned discriminations, specifically in the context of high discrimination, can be summarized as follows:

- In the hiring process, the protected group encounters higher rejection rates at the application stage, is provided fewer opportunities for interviews, and receives fewer job offers compared to the non-protected group.
- Within the hospital logs, the protected group tends to follow more complex and longer treatment paths. They undergo 'Thorough Examination' less often and face unsuccessful treatments more frequently than the non-protected group. Additionally, it has been observed that the protected group is predominantly managed by less experienced doctors, potentially contributing to their more complex and lengthy treatment journeys.
- In the lending logs, the protected group experiences higher rates of appointment denials and loan application rejections. They also often go through additional steps such as collateral assessment and co-signer requests, and face a higher rate of loan denials.
- Finally, for renting, the protected group faces higher immediate rejection rates after viewing appointment applications and undergoes 'Extensive Screening' more frequently. Interestingly, they also tend to remain longer in their apartments once accepted as tenants.

By offering quantifiable instances of discrimination across varying contexts, these logs demonstrate their substantial potential as a resource for furthering the exploration of fairness in process mining, thereby aiding the scientific community in fostering more equitable and just processes.

Listing 1: Example *pm4py* code to use one of the provided event logs.

```python
import pm4py
log = pm4py.read_xes("hiring_log_high.xes.gz") # reads the resource
# splits the log into the protected and non-protected groups.
protected_log = log[log["case:protected"] == True]
nonprotected_log = log[log["case:protected"] == False]
# discovers different process models for the groups using the inductive miner
protected_model = pm4py.discover_process_tree_inductive(protected_log)
nonprotected_model = pm4py.discover_process_tree_inductive(nonprotected_log)
# shows the process models on the screen
pm4py.view_process_tree(protected_model, format="svg")
pm4py.view_process_tree(nonprotected_model, format="svg")
```

4. Downloading and Using the Resource

The event logs [7] are publicly accessible under the CC-BY-4.0 license at https://doi.org/10.5281/zenodo.8059488. These logs can be downloaded and utilized by any process mining tool supporting the XES standard. In Listing 1, we demonstrate how to use the Python library *pm4py* to ingest the hiring (high discrimination) log, split the cases between the protected and unprotected groups, and discover different models for each group.

5. Conclusion

This work is a pivotal step towards embedding fairness in process mining, offering a distinct collection of simulated event logs across diverse domains. These logs fill a critical gap in data resources, which hinders progress in fairness-oriented process mining.

Our initial analysis, though revealing, is rudimentary, highlighting inherent limitations. As simulations, the logs might not encapsulate all real-world process complexities. Yet, they establish a foundation for developing and testing fairness measures.

Future plans include expanding this collection with privacy-preserving, real-world logs, and leveraging insights from this study to advance fairness-oriented process mining techniques. Our resources invite researchers to further explore and compare fairness techniques within process mining.

References

[1] S. Barocas, M. Hardt, A. Narayanan, Fairness in machine learning, Nips tutorial 1 (2017) 2017.

[2] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, D. Roth, A comparative study of fairness-enhancing interventions in machine learning, in: Proceedings of the conference on fairness, accountability, and transparency, 2019, pp. 329–338.

[3] M. S. Qafari, W. M. P. van der Aalst, Fairness-aware process mining, in: On the Move to Meaningful Internet Systems: OTM 2019 Conferences: Confederated International Conferences: CoopIS, ODBASE, C&TC 2019, Rhodes, Greece, October 21–25, 2019, Proceedings, Springer, 2019, pp. 182–192.

[4] T. Pohl, M. S. Qafari, W. M. P. van der Aalst, Discrimination-aware process mining: A discussion, in: International Conference on Process Mining, Springer, 2022, pp. 101–113.

[5] T. Pohl, Fairness in business process analysis, 2023. URL: https://www.pads.rwth-aachen.de/ca/f/bcbspb/.

[6] H. Verbeek, J. C. Buijs, B. F. Van Dongen, W. M. P. van der Aalst, Xes, xesame, and prom 6, in: Information Systems Evolution: CAiSE Forum 2010, Hammamet, Tunisia, June 7-9, 2010, Selected Extended Papers 22, Springer, 2011, pp. 60–75.

[7] T. Pohl, A. Berti, (Un)Fair Process Mining Event Logs, https://doi.org/10.5281/zenodo.8059488, 2023. doi:10.5281/zenodo.8059488, [Online; published on 2023-06-23].