Granularity in Process Mining: Can we fix it?

Francesca Zerbato[1]*, Ronny Seiger[1], Gemma Di Federico[2], Andrea Burattin[2], and Barbara Weber[1]

[1] University of St. Gallen, Switzerland
francesca.zerbato@unisg.ch
[2] Technical University of Denmark, Kgs. Lyngby, Denmark

Abstract. Process mining techniques rely on the availability of event logs, where events have a certain granularity that is deemed appropriate for representing business activities. In this paper, we discuss why choosing a proper granularity level during preprocessing can be challenging and reflect on the implications that such a "fixed" view over the process bears for the analysis. Then, inspired by use cases in the context of user behavior analysis, we envision possible solutions that allow exploring and mining multiple granularity levels of process activities.

1 Introduction

Process mining enables the analysis of process execution data in the form of events to understand and streamline business processes. Most process mining techniques presume the presence of well-defined processes, whose execution is recorded in event logs. Event logs represent data about the execution of process activities at a certain level of *granularity*, which is often assumed to be the same for all the events in the log. The level of granularity is usually fixated during the preprocessing phase with the help of event abstraction techniques [12], which allow aggregating fine-grained events into coarse-grained events based on numerical factors, such as temporal or geographical proximity, concept hierarchies derived from domain knowledge [2] or common execution patterns [6]. The obtained coarse-grained events describe process activities at the business level, bearing a specific granularity that results from design choices carried out during the preprocessing phase and based on the purpose of the analysis.

However, in some application areas, such as the discovery and analysis of patterns of human [10] and user [1] behavior, fixating the granularity level of process activities during the preprocessing phase can be challenging and the resulting "fixed" view over the process bears implications for the analysis and its outcomes. In this paper, we discuss the problem of *choosing a fixed granularity level for process activities during the preprocessing phase*. In detail, in Sect. 2, we explain why this is a challenging problem and discuss the implications of having a fixed granularity level on the analysis. Then, in Sect. 3, we introduce related work, while in Sect. 4, we elaborate on possible solutions allowing for the representation and mining of multi-granular activities.

2 Fixed Granularity: Challenges and Implications

Process mining has expanded to many application areas, including the discovery and analysis of patterns of human [10] and user [1] behavior from sequences of events, such as sensor data or keystrokes, recording the activities of humans interacting with smart environments or software systems and artifacts over a period of time (hereinafter *interaction traces*). Interaction traces include activities that can be described at many different levels of granularity, i.e., a coarse-grained activity is composed of multiple finer-grained activities, which are sometimes organized in a hierarchy. For example, let us consider a user filtering an event log in a process mining tool and assume that activity A in Fig. 1 corresponds to "filter event log" at level ℓ_1. A can be further detailed into "filter by attribute" (a_1) in ℓ_2, or into "select attribute filter"($a_{1.1}$), "choose attribute"($a_{1.2}$), "choose cases or events"($a_{1.3}$), and "apply attribute filter"($a_{1.4}$) in ℓ_3.

Usually, the granularity level of the activities in an event log is fixated during preprocessing with the help of event abstraction techniques [12]. However, during log preparation, it is not always possible to know (C_1) what is the exact purpose of the analysis, as this often evolves as data are analyzed, and (C_2) what is the most appropriate granularity level of process activities for finding meaningful patterns. Let us consider again the discovery of activity patterns from the interaction traces of users doing process mining and assume that we aim to find common patterns of behavior among users. This goal may change during the analysis, e.g., from analyzing user behavior at a high level to understanding if users implement workarounds to filter the event log. Besides, the expected structure of meaningful patterns is often unknown during preprocessing and emerges after some analysis iterations while exploring the data from different viewpoints

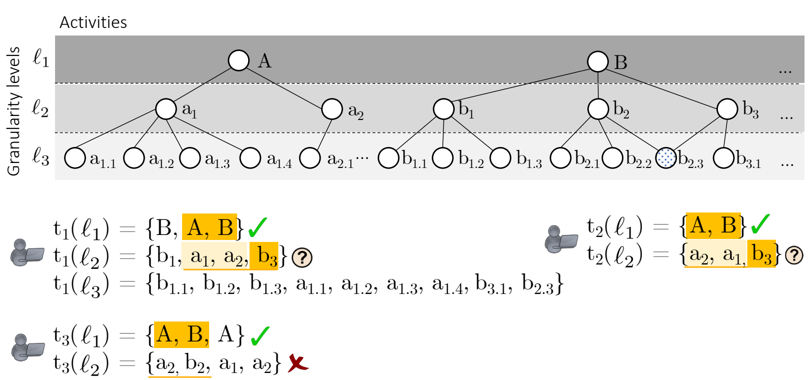

Fig. 1. (Top) Activities (A, B) defined at different levels of granularity (ℓ_1, ℓ_2, ℓ_3). (Bottom) Interaction traces (t_1, t_2, t_3) shown for different levels of granularity. In ℓ_1, pattern $\{A, B\}$ is common to all traces (✓); in ℓ_2 it can be seen in t_1 and t_2 if a_1 and a_2 can be executed in any order (?), while it cannot be observed in t_3 (✗).

and at varying granularity levels. Indeed, patterns can be discovered at all levels of the granularity spectrum and, sometimes, across granularity levels.

While fixating the granularity level of activities during preprocessing can be challenging, the resulting "fixed view" over the process has implications on what analyses can be done and what patterns can be observed. In the remainder, we discuss some of these important implications (I_1–I_4). To focus on our problem, we set aside the typical challenges of event abstraction and correlation [12] and assume to have the set of all possible activity classes available at different levels of granularity, potentially organized in a hierarchy like the one in Fig. 1.

(I_1) First, fixating the granularity of process activities during preprocessing affects what analyses can be done on the log. Indeed, some techniques do not work well with fine- or mixed-grained events [12] as they generate models that are difficult to comprehend and, thus, hinder the (visual) discovery of patterns.

(I_2) Second, the results of the analysis can be observed only at the chosen level of abstraction. This may lead to the discovery of unrepresentative patterns (e.g., in Fig. 1, pattern $\{A, B\}$ in ℓ_1 is not maintained in ℓ_2) or can prevent the discovery of patterns at other granularity levels. For example, by choosing a too-coarse level of granularity, we lose information about fine-grained events that are mapped into multiple high-level activities. Let us assume that activity B in Fig. 1 corresponds to "discover process", b_2 is "discover process with heuristic miner", b_3 is "discover process with inductive visual miner", and $b_{2.3}$, which is mapped to both b_2 and b_3, is "inspect process model". If we select ℓ_1, we cannot observe that users always inspect the process model ($b_{2.3}$) after having discovered the process with the heuristic (b_2) or inductive visual miner (b_3). Similarly, we lose information about the order of fine-grained activities, including concurrency or interleaving. For example, we cannot detect that activities a_1 (e.g., "filter by attribute") and a_2 (e.g., "filter by variant") are executed interchangeably in t_1 and t_2. Last but not least, when analyzing an event log with a fixed granularity, it is not possible to observe patterns at different levels of abstraction within the same analysis nor patterns that span multiple granularity levels. For example, if we select ℓ_1, we cannot observe that both patterns $\{A, B\}$ and $\{a_1, a_2\}$ occur in t_1 and t_3 within the same analysis, nor can we observe patterns such as $\{A, b_3\}$.

(I_3) Third, fixating the granularity of process activities during preprocessing prevents analysts from controlling it and changing it during the analysis, e.g., to obtain details-on-demand [9] by showing specific parts at a fine-grained level while keeping other less relevant parts more abstract. Some process mining algorithms and tools allow adjusting the mined process model based on certain metrics, but they do not allow users to play with the granularity of activities, e.g., to explore different granularities based on the analysis focus.

(I_4) Also, a fixed granularity hinders the possibility of tracing and explaining the journey from raw data to ready-to-use event logs, as the original events are often "lost" in the abstraction phase. From interviews and conversations with process mining users and experts, we have gathered anecdotal evidence that analysts struggle to explain the models produced by discovery algorithms and often need to access the raw data or re-engage in preprocessing to validate them.

3 Relation between Problem and Existing Work

Regarding activity granularity, the process mining literature has mainly focused on the problem of event abstraction, which has been tackled from two angles. On the one hand, techniques capable of dealing with abstraction to produce a refined version of the log were proposed. On the other hand, research focused on mining algorithms embedding the abstraction directly into the mining phase.

One of the first works in the first category addresses the issue by describing a taxonomy of abstraction patterns [6], showing how different process constructs related to abstraction might be observed and processed. Instead, one of the most recent approaches constructs graph-based models for event data that support the creation and analysis of event logs where events are related to multiple activities [3]. Another recent approach [4] focuses on slicing and dicing event logs where abstraction is handled via process cubes (i.e., where fluid notions of case are possible and each dimension has a *value* and a *granularity*). Many other techniques have been proposed, as thoroughly reviewed in [12]. A result of the literature review is that works can be characterized based on: supervision strategy (unsupervised vs. unsupervised); interleaving of fine-grained events; deterministic nature of the outcome; and whether additional data/perspectives are considered. Approaches in this group focus on raising all the events in the log to the same granularity during preprocessing and, thus, assume that a proper level of granularity is known or discovered before the analysis (cf. C_1 and C_2).

On the second family of approaches, one of the first works is the Fuzzy Miner [5], which introduces the metaphor of a process model as a road map, where relevant information is shown, whereas less significant but highly correlated behavior is aggregated, and less significant and lowly correlated behavior is hidden. These notions of significance and correlations are parametric (and the user can control them) yet are based on predefined heuristics. Other approaches focused on subsets of activities [11] to create and analyze "local models", while domain-specific techniques based on well-defined execution subtraces have also been proposed, e.g., to focus on software system processes [7]. A recent approach allows to iteratively project and abstract the event log and discover a multi-level process model starting from activity trees [8], i.e., hierarchical clusters of activities. While these approaches allow the user to adjust the abstraction level during the mining phase, i.e., partially addressing challenges C_1 and C_2 and implications I_1 and I_2, they focus on reducing the complexity of the mined models based on predefined metrics or subprocesses and do not support the interactive exploration of activity granularities (I_3), e.g., based on semantic relationships among activity classes, nor they explicitly address traceability and explainability (I_4).

4 Solving the Fixed Granularity Problem: Initial Ideas

How to best tackle the introduced challenges (C_1, C_2) and the implications (I_1–I_4) is still not clear. Currently, we are thinking about two possible directions, envisioning a scenario where the choice of the granularity level of activities is deferred from the traditional preprocessing phase to the analysis, where users can explore multiple granularity levels and select the desired one interactively.

(i) A first approach could consist of changing the typical process mining workflow by introducing a new step in charge of preprocessing the log to be consumed by mining algorithms. Envisioning an interactive system, we should be able to repeat the preprocessing on-the-fly, thus allowing users to quickly explore scenarios where activities have different granularity levels while leveraging existing process mining algorithms and process modeling languages.

(ii) Another approach would comprise the construction of novel process mining algorithms capable of mining models from event logs that allow the specification of multiple granularity levels for process activities within the same model (e.g., along the lines of subprocesses in BPMN or activity trees [8]).

Acknowledgment. This work is part of the ProMiSE project, funded by the Swiss National Science Foundation (SNSF) under Grant No.: 200021_197032.

References

1. Abbad Andaloussi, A., Zerbato, F., Burattin, A., Slaats, T., Hildebrandt, T.T., Weber, B.: Exploring how users engage with hybrid process artifacts based on declarative process models: a behavioral analysis based on eye-tracking and think-aloud. Software and Systems Modeling (2020)
2. Baier, T., Mendling, J., Weske, M.: Bridging abstraction layers in process mining. Information Systems **46**, 123–139 (2014)
3. Esser, S., Fahland, D.: Multi-dimensional event data in graph databases. Journal on Data Semantics (2021)
4. Ghahfarokhi, A.F., Berti, A., van der Aalst, W.M.: Process comparison using object-centric process cubes. arXiv preprint arXiv:2103.07184 (2021)
5. Günther, C.W., Van Der Aalst, W.M.: Fuzzy mining–adaptive process simplification based on multi-perspective metrics. In: Int. Conf. on Business Process Management. pp. 328–343. Springer (2007)
6. Jagadeesh Chandra Bose, R.P., van der Aalst, W.M.P.: Abstractions in process mining: A taxonomy of patterns. In: BPM. pp. 159–175. Springer (2009)
7. Leemans, M., van der Aalst, W.M.P., van den Brand, M.G.J.: Hierarchical performance analysis for process mining. In: Int. Conf. on Software and System Process. p. 96–105. ICSSP '18, ACM (2018)
8. Lu, X., Gal, A., Reijers, H.A.: Discovering hierarchical processes using flexible activity trees for event abstraction. In: ICPM. pp. 145–152 (2020)
9. Shneiderman, B.: The eyes have it: A task by data type taxonomy for information visualizations. In: IEEE Symp. on Visual Languages. pp. 336–343. IEEE (1996)
10. Tax, N., Sidorova, N., Haakma, R., van der Aalst, W.: Mining process model descriptions of daily life through event abstraction. In: Intelligent Systems and Applications. pp. 83–104. Springer (2018)
11. Tax, N., Sidorova, N., van der Aalst, W.M.P., Haakma, R.: Heuristic approaches for generating local process models through log projections. In: IEEE Symp. Series on Computational Intelligence (SSCI). pp. 1–8. IEEE (2016)
12. van Zelst, S.J., Mannhardt, F., de Leoni, M., Koschmider, A.: Event abstraction in process mining: literature review and taxonomy. Granular Computing pp. 1–18 (2020)