Chapter 2
Conformance Checking and its Challenges

An analysis of a process is as good as the models used for such analysis. This chapter provides a basic overview on conformance checking. In particular it concentrates on the quality dimensions of a process model with respect to reality, and the challenges that arise when one tries to assess the conformance of a model to reality. This chapter closes with an overview on all the challenges addressed in this book and the respective chapters where each challenge will be covered.

2.1 The Role of Process Models in Conformance Checking

Process models play a crucial role in any process analysis technique. The term process model may refer to any representation, generic or specific, of one or several perspectives of a process. However, one of the most extended meanings of process models are workflow process models, i.e., a process model that captures the order of the actions involved in the process. Workflow process models are the main process modeling type used in this book, and they are referred simply as process models. Figure 2.1 is an example of process model, using an informal modeling notation, for the scholarship process presented in the previous chapter.

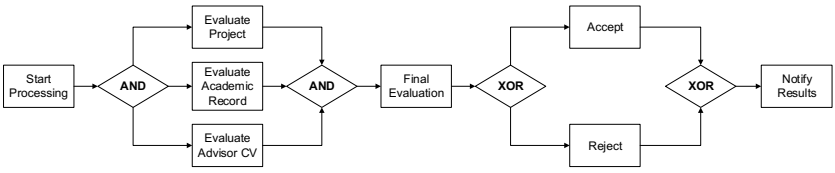

Fig. 2.1: Informal process model of the scholarship process.

J. Munoz-Gama: Conf. Check. ... in Process Mining, LNBIP 270, pp. 11–18, 2016.
DOI: 10.1007/978-3-319-49451-7_2

Different process modeling notations define different types of elements to represent the process. However, there are elements common in most of the notations. These are the *activities* denoting the steps of the process, and usually graphically represented by boxes. Another element common in most process modeling notations are the *flows* between activities, represented as arrows, to denote an ordering relation between activities. Most process models are able to represent the idea of *concurrency* (several activities performed without an specific order), and *choice* (the execution of an activity excludes another). For example, in Figure 2.1 the AND and XOR gateways represent concurrency and choice, respectively. In the context of conformance checking, a process model may either be created by a human modeler or be constructed by an algorithm. Conformance checking then aims to answer how well that model describes reality - as it has been recorded in an event log.

2.2 Dimensions of Conformance Checking

By measuring the conformance between an event log and a process model one is concerned about quantifying if a given model is a valid description of reality. A first naive approach could be to consider that a model and a log conform to each other simply if the model captures all the behavior observed in the log. In other words, a perfect conformance would require that all the traces in the log *fit* in the model. However, there are models that will allow any log to fit, but that have such a trivial structure that they are of little or no use to a process analyst when trying to understand a process. For example, let us consider the model in Figure 2.2 for the scholarship example. The informal semantics behind this model (similar to Petri nets), know as a *flower model*, is that it captures a possible sequence of the activities, in any order and for any length, i.e., the special circle in the middle should be read as the state the process is in and always return to after executing an activity. Therefore, any possible log involving the same activities fit this model. However, as one can see, this model provides absolutely no insight into the process or how the activities are executed. This simple counter-example shows that conformance needs to consider more dimensions than fitness to give a faithful account of how well a model describes a log.

In [77, 73] the multidimensional nature of the conformance is studied, and the authors propose four dimensions – *fitness*, *precision*, *generalization* and *simplicity* – to fully capture the notion of how good a given model is with respect to the reality.

Fitness As it has been already mentioned, this dimension indicates how much of the observed behavior is captured – *fits* – the process model. For example, the trace ⟨*Start Processing, Evaluate Academic Record, Evaluate Project, Evaluate Advisor CV, Final Evaluation, Accept, Notify Results*⟩ with case id 1 in Table 1.1 perfectly fits the model in Figure 1.2. However, the trace ⟨*Start Processing, Evaluate Academic Record, Evaluate Project, Final Evaluation, Reject, Notify Results*⟩ with case id 3 does not fit the model because evaluate advisor CV is never executed, denoting that the application of the student is rejected

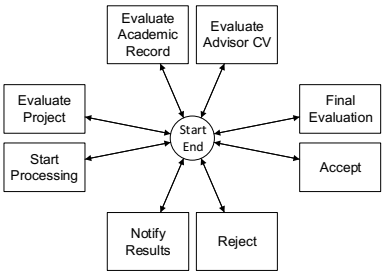

Fig. 2.2: Informal *flower* process model of the scholarship process, modeling any possible sequence of the activities.

without proper evaluation. On the other hand, both traces fit the flower model of Figure 2.2. Part III of this book is devoted to analyze the fitness dimension in a decomposed way, and consequently a more formal presentation of the fitness dimension is presented.

Precision This dimension identifies overly general models: precision penalizes a process model for allowing behavior that is unlikely given the observed behavior in the event log. For example, in the log of Table 1.1 we observe that, although the three documents could be evaluated concurrently, the university employees always first evaluate the academic record. That way, if the student is clearly not suitable for the grant (e.g., the grade does not reach the minimum necessary), the advisor and project evaluation can be done less thoroughly without compromising on the outcome of the evaluation. However, because of that specific order, the model of Figure 2.1 is less precise than reality as it also allows for other unseen execution orders. In contrast, the model shown in Figure 2.3 is a more precise representation of reality than Figure 2.1. The flower model in Figure 2.2 is the perfect example of completely imprecise model. Part II of this book is devoted to the precision dimension, and consequently a more formal presentation of the precision is included in these sections.

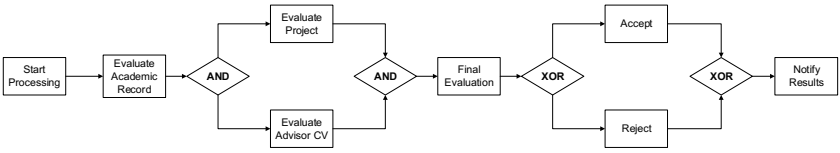

Fig. 2.3: More precise model for the scholarship process.

Generalization This dimension addresses overfitting models: a good model must be able to generalize and reproduce possible future behavior, instead of capturing simply each trace of the observed log. For example, Figure 2.4 shows a model that only captures one possible order for the evaluation of the documents that results necessarily in the acceptance of the application. This model perfectly captures the first trace in the Table 1.1, but it is unable to generalize for any other possible process execution. In [73, 77] the generalization dimension is covered in more detail.

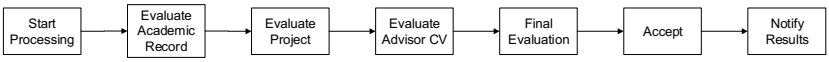

Fig. 2.4: Model that overfits the first trace of the scholarship log, and does not generalize for possible future behavior.

Simplicity This dimension penalizes models that are unnecessarily complex: following the Occam's Razor principle, models that explain the behavior observed in the log in a simple way are preferred than those that use redundant components. Figure 2.5 illustrates an example where explicitly writing out all possible execution sequences of the three evaluate activities complicates the model for the scholarship process unnecessarily. In [73, 77] the simplicity dimension is covered in more detail.

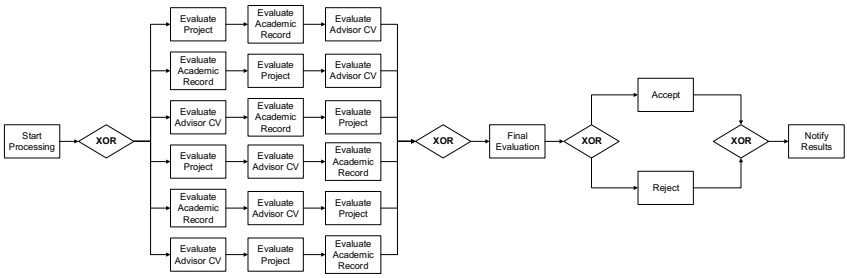

Fig. 2.5: Unnecessary complex model for the scholarship process.

Given the orthogonal nature of the dimensions, there is no such thing as perfect model, but a set of suitable models. For example, for analyzing the main paths of a organization process the analyst could prioritize fitness over the the other dimensions. On the other hand, if the process model involves critical activities and it is

being used as part of a workflow system, a model with high precision is desired to avoid performing costly actions in the wrong moments.

Next, we outline the basic techniques used in conformance checking, and the challenges addressed in this book - focusing on fitness and precision.

2.3 Replay-based and Align-based Conformance Checking

In early works on conformance, most of the proposed approaches were based on *replaying* the log on the model to detect discrepancies. Some replay-based approaches simply stop at the point where the model is not able to reproduce the trace anymore. Other approaches perform the replay in a non-blocking way, regardless of whether the path of the model is followed or not, like [77]. More sophisticated approaches, such as the approach in [90], include also a *look ahead* function to determine the most promising path. Recently, another family of approaches has appeared, where the conformance check is done in a global manner, by means of *aligning* both the modeled behavior and the behavior observed in the log. Examples of conformance approaches based on alignments are [17, 50]. These approaches handle conformance in a global way, but they are computationally more expensive compared with replay-based approaches. In Part II of this book, both replay-based and alignment-based approaches are explored to check precision. In Part III, a decomposed technique is proposed to alleviate computation time for conformance diagnosis, especially for those analyses based on alignments due their expensive cost.

2.4 Challenges of Conformance Checking

Conformance checking must confront a set of challenges in order to be applied successfully. In particular, we identify five challenges: *four-dimensional conformance*, *big data and real time*, *noise, incompleteness, unfitness, and indeterminism*, *conformance diagnosis* and *modeling notations*. This book addresses all these challenges (see Figure 2.6).

Challenge 1 – *Four-Dimensional Conformance.* Since the multidimensional nature of conformance – *fitness, precision, generalization and simplicity* – has been stated first in [75] and later refined in [78, 77, 73], the relation between the four dimensions and the adequacy of the results has become more and more clear. Works like [35] illustrate the need of metrics for all the dimensions in order to discover good models. However, most of the approaches proposed in conformance, especially on the early days, are focused exclusively on fitness. Conformance checking must provide also measures for other dimensions such as precision, generalization, and simplicity. Hence, the challenge addressed in this book is to provide a versatile, well founded, yet easy to understand way to measure precision. This challenge is addressed in Chapters 4, 5, and 6.

Challenge 2 *– Big Data and Real Time.* The amount of information recorded by the information systems periodically grows exponentially. Event logs become more detailed, complete and large, and with them also the process models. Conformance techniques must evolve accordingly in order to handle this exponential growth, especially those based on the global aligning of behaviors. Moreover, the fast implantation of online and monitoring paradigms in nowadays systems is requiring faster and more fine-grained conformance approaches. In this book, we will address that challenge proposing approaches to measure conformance even on very large models and very large data sets. This challenge is addressed in Chapters 12, 13, 14 and 18.

Challenge 3 *– Noise, Incompleteness, Unfitness, Indeterminism.* Typically, the application of process mining techniques faces some of these four issues: noise, incompleteness, unfitness, and indeterminism. *Noise* in event logs can appear by traces incorrectly recorded (for instance, due to temporary system failure), or traces reflecting exceptional situations not representative of the typical behavior of the process. Noise is a well-known problem in discovery approaches [7], and therefore, conformance approaches proposed should also be noise-aware too. Conformance checking compares reality and model, and therefore, the comparison is only fair if the log really is *complete* regarding what happens in reality, e.g., comparing a small sample from reality to a complex models could lead to incorrect conclusions. However, assuming that a sample log contains all possible behavior is an unrealistic assumption in most of the cases. The number of traces necessary for a complete log grows exponentially when the number of concurrent actions in the model is increased. Moreover, some concurrent actions may look sequentially in the log because performing one action is always much faster than the other. Conformance techniques must include mechanisms to aid the process analyst on deciding whether the problems are real conformance anomalies or result of the incompleteness of the log. *Unfitness* – i.e., situations where the behavior observed in the log cannot be reproduced by the model – is a conformance dimension itself, but it may influence other dimensions: if the model cannot reproduce the observed behavior, it cannot determine the state of the system in that moment. Conformance approaches should try to abstract from how the alignment between observed and modeled behavior is done. This include also the *non-deterministic situations* produced when a trace in the log can be mapped to several sequences in the model. In this book, we will present conformance techniques that mitigate the effects of noise, incompleteness, unfitness, and non-determinism, providing at the same time useful conformance assessment of the process models. This challenge is addressed in Chapters 7, 9, 10, and 11.

Challenge 4 *– Conformance Diagnosis.* The importance of indicating the location of the problems for a proper conformance diagnosis was already emphasized in the seminal work [73]. However, the diagnosis mechanisms cannot be limited to simply locate the possible conformance errors, but they must go a step further: they must provide mechanisms to the analyst to fully understand the causes of the problems. For example, making it possible to dynamically inspect the conformance results at different levels of abstraction, or to group mismatches with

a similar root cause. Diagnosis tools are especially useful for large models or models with a high degree of complexity, where the causes of the problems are difficult to grasp. In this book, we will complement the conformance techniques, with additional approaches to analyze, locate, and rank, the conformance discrepancies detected, aiding on understanding the underlying causes. This challenge is addressed in Chapters 8 and 15.

Challenge 5 – *Modeling Notations.* Most of the approaches presented in conformance so far focus exclusively on the control-flow perspective of the process – i.e., the order of the activities – and to one specific workflow modeling notation, *Petri nets* [65]. Conformance techniques must include other modeling notations, and other perspectives. In addition, there will appear new approaches to check the conformance of multi-perspective models – models capturing more than one perspective – like for example [50], where integer linear programming techniques are used to validate both the control-flow and the data perspectives of a model. In this book, we will go a step in that direction, providing conformance checking techniques adapted for data-aware multi-perspective models, especially suitable for large processes. This challenge is addressed in Chapters 16 and 17.

Figure 2.6 provides an overview of the approach presented on this book, and the techniques proposed to address each one of the challenges aforementioned.

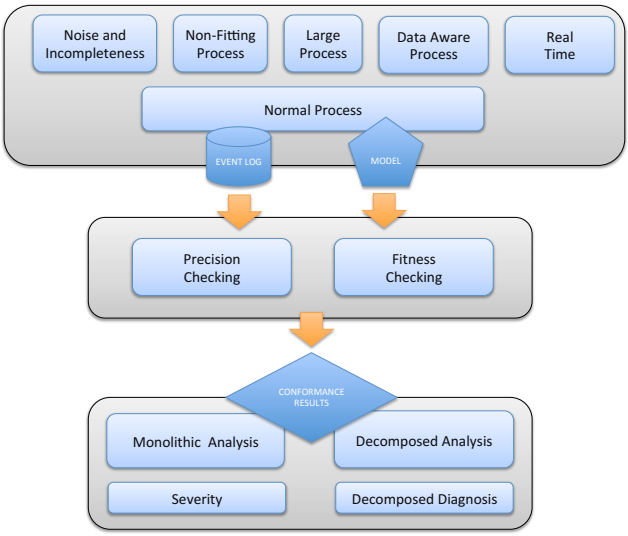

Fig. 2.6: Overview of the conformance analysis and challenges addressed in this book.

Springer

http://www.springer.com/978-3-319-49450-0

Conformance Checking and Diagnosis in Process Mining
Comparing Observed and Modeled Processes
Munoz-Gama, J.
2016, XIV, 202 p. 90 illus., Softcover
ISBN: 978-3-319-49450-0