

PLG2: Multiperspective Processes Randomization and Simulation for Online and Offline Settings

Andrea Burattin *

Abstract

Process mining represents an important field in BPM and data mining research. Recently, it has gained importance also for practitioners: more and more companies are creating business process intelligence solutions. The evaluation of process mining algorithms requires, as any other data mining task, the availability of large amount of real-world data. Despite the increasing availability of such datasets, they are affected by many limitations, in primis the absence of a “gold standard” (i.e., the reference model).

This paper extends an approach, already available in the literature, for the generation of random processes. Novelties have been introduced throughout the work and, in particular, they involve the complete support for multiperspective models and logs (i.e., the control-flow perspective is enriched with time and data information) and for online settings (i.e., generation of multiperspective event streams and concept drifts). The proposed new framework is able to almost entirely cover the spectrum of possible scenarios that can be observed in the real-world. The proposed approach is implemented as a publicly available Java application, with a set of APIs for the programmatic execution of experiments.

*A. Burattin is with the University of Innsbruck, Technikerstraße 21a, 6020 Innsbruck, Austria. E-mail: andrea.burattin@uibk.ac.at.

Contents

1	Introduction	2
1.1	Research Challenges	3
2	Related Work	5
3	Process Models in PLG2	6
3.1	Internal Representation of Business Processes	6
3.2	Formal Definition of Business Process	7
4	Random Generation of Business Processes in PLG2	8
4.1	Grammar Capabilities	10
4.2	Grammar Extension with Probabilities	12
5	Process Simulation in PLG2	13
5.1	Multi-Perspective Simulation	14
5.2	Noise Addition	20
6	Stream Simulation in PLG2	21
6.1	Continuous Data Generation	21
6.2	Concept Drifts for Process Models	23
7	Implementation Details of PLG2	26
8	Case Studies	28
8.1	Offline Setting	29
8.2	Online Setting	30
9	Conclusions and Future Work	32

1 Introduction

Process mining [34] gained a lot of attention and is now considered an important field of research, bridging data mining and business process modeling/analysis. In particular, the aim of process mining is to extract useful information from business process executions. Under the umbrella of process mining, different activities could be identified. For example, *control-flow discovery* aims at reconstructing the actual process model starting only from the observations of its executions; *conformance checking* tries to discover discrepancies between the expected (i.e., compliant) executions and the actual ones; *enhancement* extends a process model with additional information obtained from the actual observations.

In data mining, the term *gold standard* (or sometimes also referred as *ground truth*) typically indicates the “correct” answer to a mining task (i.e.,

the reference model). For example, in data clustering, the gold standard may represent the right (i.e., the target) assignment of elements to their corresponding clusters. Many times, referring to a gold standard is fundamental in order to properly evaluate the quality of mining algorithms [9, 26]. Several concepts, like *precision* and *recall* are actually grounded on this idea. In general, it is possible to identify the concept of gold standard for all mining tasks.

As for all other mining challenges, the evaluation of new algorithms is difficult. In order to properly assess the quality of mining algorithms, typically, the evaluation of mining algorithm should be based on real world data. However in the context of business processes, companies are usually reluctant to publicly share their data for analysis purposes. Moreover, detailed information on their running processes (i.e., the reference models) are considered as company assets and, therefore, are kept private.

Since few years, an annual event, called *BPI challenge*¹, releases real world event logs. Despite the importance of this data, the logs are not accompanied with their corresponding gold standards. Moreover, they do not provide examples of all possible real world situations: many times, researchers and practitioners would like to test their algorithms and systems against specific conditions and, to this purpose, those event logs may not be enough. Some other tools, described in the literature (Section 2) can be used to construct business processes or to simulate existing ones. However, they are very difficult to use and limited in several aspects (e.g., they can only generate process models, or can simulate just the control-flow perspective).

1.1 Research Challenges

The final aim of this paper is to support researchers and practitioners in developing new algorithms and techniques for process mining and business process intelligence. Moreover, we put particular emphasis on the online/stream paradigm which, with the advent of the *big data* and *Internet of things*, is rising interest. To achieve our goal we have to face the general data availability problem which, in our context, could be decomposed into several research challenges:

- C1** build large repositories of randomly created process models with control-flow and data perspectives;
- C2** obtain realistic (e.g., noisy) multiperspective event logs, which are referring to a model already known (i.e., the gold standard), to test process mining algorithms;
- C3** generate potentially infinite multiperspective streams of events starting from process models. These streams have to simulate realistic scenar-

¹See <http://www.win.tue.nl/bpi/2015/challenge>.

ios, e.g., they could contain noise, fluctuating event emission rates, and concept drifts.

C1 and C2 are required in order to test an approach against several different datasets, and avoid overfitting phenomena (i.e., tailoring an approach to perform well on particular data, but lacking in abstraction).

C3 is becoming more and more important due to the emerging importance of big data analysis. Big data is typically characterized [14,16] by the *data volume* and *velocity* (a typical way of dealing with such volume and velocity is via unbounded event streams [1,15]); *variety* (for this, we need multiperspective models, not only with the control-flow perspective); *variability* (this led us to properly simulate concept drifts [4]).

In this paper we propose a series of algorithms which can be used to randomly generate multiperspective process models (Section 3). These models can easily be simulated in order to generate multiperspective event logs (Section 5). Moreover, the whole approach is design keeping the simulation of online settings (Section 6) in mind: it is possible to generate drifts on the processes (i.e., local evolutions) and it is possible to simulate multiperspective event streams (which are also replicating the drifts).

Therefore, the aim of this paper is twofold: on one hand, we aim at describing the extensions made with respect to our previous work [5], which constitutes PLG2. On the other hand, we want to highlight the research challenges that need to be solved in order to create realistic and useful test data.

The new approach is implemented in a standalone Java application (Section 7) which is also accompanied by a set of APIs, useful for the programmatic definition of custom experiments.

In summary, this paper extend the work we presented in [5] since we are now able to:

- generate random process models with additional data perspective (or import existing ones);
- have a detailed control over the data attributes (e.g., by controlling their values via scripts);
- have a detailed control over the time perspective (controlled via scripts);
- *evolve* a process model, by randomly changing some of its features (e.g., adding/removing/replacing subprocesses);
- generate a realistic multiperspective event log, with executions of a process models and noise addition (with probabilities for different noise behaviors);

- generate a stream of multiperspective events referring to process models that could change over time with customizable output ratio.

2 Related Work

The idea of generating process models for evaluating process mining algorithms has already been explored.

In particular, van Hee and Liu, in [38], presented an approach to generate random Petri nets representing processes. Specifically, they suggest to use a top-down approach, based on a step-wise refinement of Workflow nets [36], to generate all possible process models belonging to a particular class of Workflow network (also called Jackson nets). This approach has been adopted, for example, in the generation of collections of process model with specific features [37]. A similar and related approach has been reported in [2], where authors propose to generate Petri nets according to a different set of refinement rules.

In both cases, approaches do not address the problem of generating traces from the developed Petri nets. This task, however, has been explored in the past, in particular for the generation of process mining oriented logs [12]. The idea is to decorate a Petri net model, using CPN tool², in order to log executions into event log files. Although the approach is extremely flexible and grounded on a solid tool, it suffers of usability drawbacks. The most important problem consists of the complexity of whole procedure, which is also particularly error prone; the complexity in managing timestamps; the impossibility to simulate data objects (i.e., multiperspective models) in a proper way; and impossibility to simulate streams.

The work reported in [5], extended by this work, provides a first possible complete tool for the random generation of process models and their execution logs.

The approach described in this paper (namely, PLG2) extends previous works in two substantial ways. On one hand it improves the generation of random models and their logs by adding data and time perspectives: the new version of PLG2 is capable of generating random data objects and simulate manually defined ones. Moreover, complete and detailed support for activity and trace timing is provided as well. Secondly, the whole project has been designed with online settings in mind: it is possible to easily (and automatically) generate random new versions of process models, in order to simulate “concept drifts”. Moreover, processes can be simulated to generate multiperspective (i.e., with data and preserving temporal relations) event streams.

²See <http://cpntools.org>.

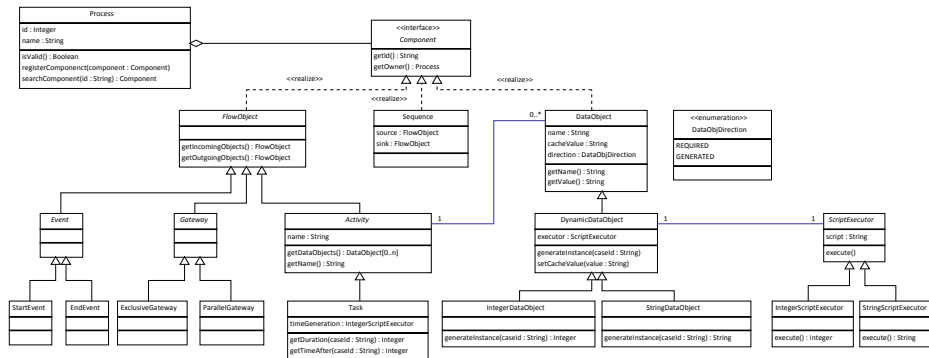


Figure 1: Classes diagram, represented in UML, of the internal structure used for the representation of a process model in PLG2. The structure basically reflects a possible instance of a BPMN model diagram.

3 Process Models in PLG2

This section presents the internal representation used to handle business processes. The generation of random business process is reported as well.

3.1 Internal Representation of Business Processes

In PLG2, the internal structure of a process model is actually rather intuitively derived from the definition of a BPMN process model [28]. Figure 1 depicts the diagram of the classes involved in the modeling. In particular, a process is essentially an aggregation of *components*. Each component can be either a flow object, a sequence or a data object. Flow objects are divided into:

- events (which are divided into *start* or *end*);
- gateways (either *exclusive* or *parallel*);
- tasks (which can be *activities*).

Please note that it is not possible to instantiate general events or gateways or tasks (i.e., those classes are abstract). This technique is used to enforce a proper *typing* of such —otherwise ambiguous— elements. Sequences are used to connect two flow objects. A sequence, clearly, imposes a direction of the flow. Data objects are associated with activities and can be *plain data objects* or *dynamic data objects*. A plain data object is basically a key-value pair. A dynamic data object, instead, has a value that can change every time it is required (i.e., it is dynamically generated by a script). Another characterization of data object is with respect to their “direction”: a data object can be *generated* or *required* by an activity. These two characterizations play

an important role in the process simulation phase. We will get into these details in Section 5.

With respect to the previous version of this work [5], we decided to evolve the internal structure into a more general one. This BPMN standard-oriented representation is fundamental in order to allow much more flexibility. For example, now, it is possible to load BPMN models generated with external tools³, as long as the modeled components are available also in PLG2 (for example, it is not possible to load a BPMN model with inclusive gateways). However, since we are restricting to non-ambiguous components, we also can convert our processes into more formal languages (e.g., it is possible to convert and export the generated models into Petri nets, using the PNML file format⁴).

3.2 Formal Definition of Business Process

The process representation that we just reported can be structured in a more formal definition. Specifically, a process model P can be seen as a graph $P = (V, E)$, where V is the set of nodes, and $E \subseteq V \times V$ is the set of directed edges. However, since in our context not all nodes or edges are equal, we can improve the definition of V and E . Let's then define a process as a tuple $P = ((E_{start}, E_{end}, A, G, D), (S, C))$, where:

- $(E_{start}, E_{end}, A, G, D)$ is a tuple, in which each component is a set of nodes with a specific semantic associated. In particular, E_{start} is the set of starting nodes, E_{end} is the set of end nodes, A is the set of activities, G is the set of gateways, D is the set of data objects;
- (S, C) is another tuple. Each component of this tuple is a set of edges. Specifically, $S \subseteq E_{start} \times A \cup A \times E_{end} \cup A \times A \cup A \times G \cup G \times G \cup A$ is a set of sequences connecting process flow objects.
 $C \subseteq A \times D \cup D \times A$, such that $\forall d \in D \ |\{(\cdot, d) \in C\} \cup \{(d, \cdot) \in C\}| \leq 1$, is a set of associations going from activities to data objects and from data objects to activities. The additional condition guarantees that one data object is connected with at most one activity.

Please note that, the definition just provided partially enforces the semantic correctness of each component involved (for example, it is not possible to connect an end event with a gateway or a data object with an event).

With respect to the data object associations, the C component of a process P permits data objects both incoming and outgoing into and from activities. This behavior is described in the UML classes diagram, reported in Fig. 1, with the `direction` element of a `DataObject`.

³An example of supported tool is Signavio, <http://www.signavio.com>.

⁴See <http://www.pnml.org> for more information on this standard.

4 Random Generation of Business Processes in PLG2

The definition of process just described can be used as general representations for the description of relations between activities, events, gateways, and data objects. In this paper, however, we are also interested in the generation of random process models, in order to be able to create a “process population” capable of describing several behaviors.

In order to generate random processes, we need to combine some well known workflow control-flow patterns [31,35]. The patterns we are interested in are reported in this summarized list:

- *sequence* (WCP-1): direct succession of two activities (i.e., an activity is *enabled* after the completion of the preceding);
- *parallel split* (WCP-2): parallel execution (i.e., once the work reaches the split, it is forked into parallel branches, each executing concurrently);
- *synchronization* (WCP-3): synchronization of parallel branches (i.e., the work is allowed to continue only when all incoming branches are completed);
- *exclusive choice* (WCP-4): mutual execution (i.e., once the work reaches the split, only precisely one of the outgoing branches is allowed to continue);
- *simple merge* (WCP-5): convergence of branches (i.e., each incoming branch results in continuing the work);
- *structured loop* (WCP-21): ability to execute sub-processes repeatedly.

Clearly, these patterns do not describe all the possible behaviors that can be modeled in reality, however we think that most realistic processes are based on them. Actually, we are also going to extend these patterns (with the addition of data-objects), in order to generate multiperspective models.

The way we use these patterns is by progressively combining them in order to build a complete process. The combination of such patterns is performed according to a predefined set of rules. We implement this idea via a Context-Free Grammar (CFG) [22] whose productions are related with the patterns mentioned above. Specifically, we defined the following context-free grammar $G_{Process} = \{V, \Sigma, R, P\}$ where $V = \{P, G, G', G_{\odot}, G_{\oplus}, G_{\otimes}, A, A_{act}, A_{do}, D\}$ is the set of the non-terminal symbols, $\Sigma = \{;, (,), \cup, \otimes, \oplus, \nearrow, \swarrow, e_{start}, e_{end}, a, b, c, \dots, d_1, d_2, d_3, \dots\}$ is the set of all terminals

Symbols	Meaning
e_{start}	The start event of the process
e_{end}	The end event of the process
$()$	Parentheses are used to describe operators precedence
$;$	Operator, it indicates a sequential connection
\circ	Operator, repetition of the first parameter by executing the second
\oplus	Operator, its parameters executed in parallel (“AND”)
\otimes	Operator, its parameters executed in mutual exclusion (“XOR”)
$a \swarrow^d$	Indicates that data object d is required by activity a
$a \nearrow^d$	Indicates that data object d is generated by activity a
a, b, c, \dots	The set of possible activity names
d_1, d_2, d_3, \dots	A set of possible data objects

Table 1: All the terminal symbols of the context-free grammar used for the random generation of business processes and their corresponding meanings.

(their “interpretation” is described in Table 1), R is the set of productions:

$$\begin{aligned}
P &\rightarrow e_{start} ; G ; e_{end} \\
G &\rightarrow G' \mid G_{\circ} \\
G' &\rightarrow A \mid (G;G) \mid (A;G_{\oplus};A) \mid (A;G_{\otimes};A) \mid \epsilon \\
G_{\oplus} &\rightarrow G \oplus G \mid G \oplus G_{\oplus} \\
G_{\otimes} &\rightarrow G \otimes G \mid G \otimes G_{\otimes} \\
G_{\circ} &\rightarrow (G' \circ G) \\
A &\rightarrow A_{act} \mid A_{do} \\
A_{do} &\rightarrow A_{act} \swarrow^D \mid A_{act} \nearrow^D \\
A_{act} &\rightarrow a \mid b \mid c \mid \dots \\
D &\rightarrow d_1 \mid d_2 \mid d_3 \mid \dots
\end{aligned}$$

and P is the starting symbol for the grammar. Using this grammar, a process is described by a string derived from $G_{Process}$.

Analyzing the production rules, it is possible to see that each process requires a starting and a finishing event and, in the middle, there must be a sub-graph G . A sub-graph can be either a “simple sub-graph” (G') or a “repetition of a sub-graph” (G_{\circ}).

Starting from the first case: a sub-graph G' can be a single activity A ; the sequential execution of two sub-graphs ($G;G$); the exclusive or parallel execution of some sub-graphs (respectively, $(A;G_{\otimes};A)$ and $(A;G_{\oplus};A)$); or an “empty” sub-graph ϵ . It is important to note that the generation of parallel and mutual exclusion branches is always “well structured”.

Analyzing the repetition of a sub-graph (G_{\odot}) it should be noticed that, semantically, the repetition of a sub-graph ($G' \circ G$) is described as follows: each time we want to repeat the “main” sub-graph G' , we have to perform another sub-graph G ; the idea is that G (that can even be only a single or empty activity) corresponds to the “roll-back” activities required in order to prepare the system to the repetition of G' (which, also, could be a empty activity).

The structure of G_{\oplus} and G_{\otimes} is simple and expresses the parallel execution or the choice between at least 2 sub-graphs.

A represents the set of possible activities. In this case two productions are possible: A_{act} which generates just an activity, or A_{do} which generates an activity with a data object associated. In this latter case, two more productions are possible: $A_{act} \swarrow^D$ and $A_{act} \nearrow^D$: the first generates an activity with a required data object, the second produces an activity with a generated data object. Finally, the grammar defines activities just as alphabetic identifiers but, actually, the implemented tool “decorates” it with other attributes, such as a unique identifier. The same observation holds for data objects.

Finally, this grammar definition allows for more activities with the same name, however in our implemented generator all the activities are considered to be different.

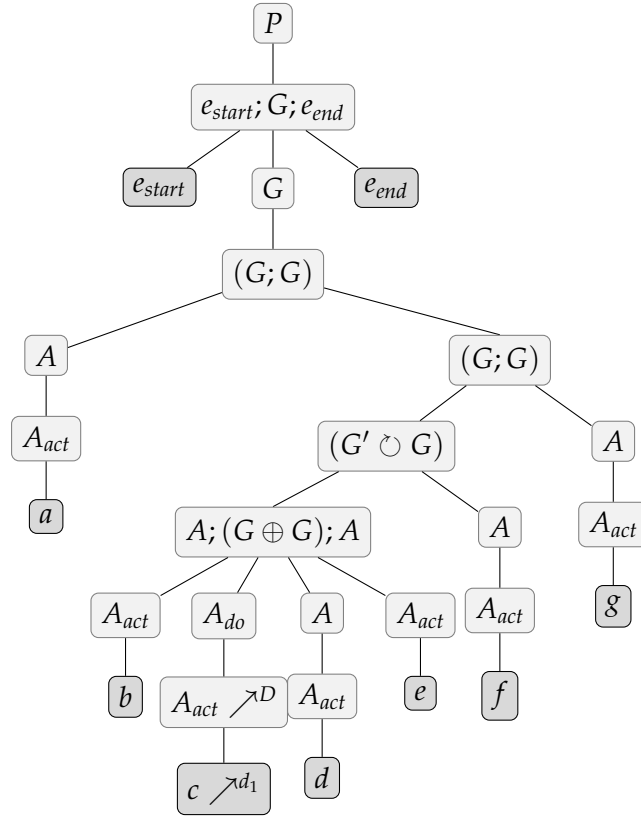
In Fig. 2 an example of all the steps involved in the generation of a process are shown: the derivation tree, the string of terminals, and two graphical representations of the final process (using BPMN and Petri net notations).

4.1 Grammar Capabilities

The context free grammar just provided is not capable of generating all the possible business models that could be described using languages such BPMN or Petri net. In particular, we are restricting to block structured ones [23]. Although restricting to block structured processes might seem rough, these processes benefit from very interesting properties [39]. Moreover, recently, the process mining community started to focus on this types of processes [3,24,33], especially for the soundness properties that they can guarantee. With the adoption of the context-free grammar proposed, we decided to stick to this type of language as well.

Please note that the block structure restriction only affects the random process generation part of PLG2: all other components (i.e., process evolution, and simulations for generation of event logs or stream) are still functioning also with imported (and non-block structured) processes.

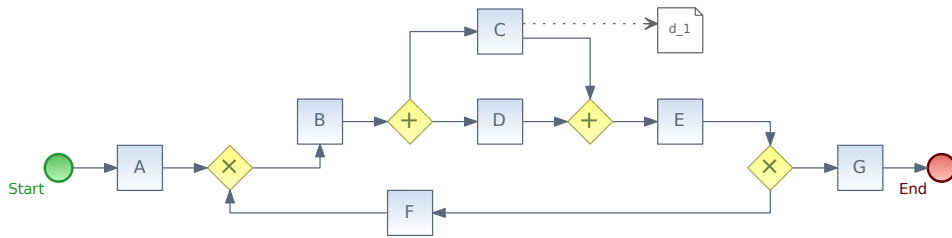
We can also note that there is a straightforward translation of a string produced by the PLG2 grammar into the graph representation introduced in the previous sections. Therefore, the processes generated with PLG2 can



(a) Example of derivation tree. Note that, for space reason, we omitted the explicit representation of some basic productions.

$$e_{start}; (a; ((b; (c \xrightarrow{d_1} \oplus d); e \circ f); g)); e_{end}$$

(b) The string derived from the above tree.



(c) BPMN representation, created by PLG2, for the process generated.

Figure 2: Derivation tree of a process and its string and BPMN representation.

always be expressed as BPMN (with all perspectives) or Petri net (with just the control-flow perspective).

4.2 Grammar Extension with Probabilities

As previously stated, we want to randomly generate strings of terminal using the context-free grammar described earlier. However, in order to provide the user with a deep control over the final structure of the generated processes, we converted the CFG into a stochastic context-free grammar (SCFG) [7,8]. This type of grammars have also been widely used for modeling RNA structures [17,32].

Specifically, to adopt this models, we need to add probabilities associated to each production rule. This allows us to introduce user-defined parameters to control the presence of specific pattern into the generated process. These are the probabilities defined (with indication on whether the user is asked to provide the value):

π_1	for	$G \rightarrow G_{\circ}$	required
π_2	for	$G \rightarrow G'$	as $1 - \pi_1$
π_3	for	$G' \rightarrow A$	required
π_4	for	$G' \rightarrow (G; G)$	required
π_5	for	$G' \rightarrow (A; G_{\oplus}; A)$	required
π_6	for	$G' \rightarrow (A; G_{\otimes}; A)$	required
π_7	for	$G' \rightarrow \epsilon$	required
π_8	for	$G_{\oplus} \rightarrow G \oplus G_{\oplus}$	computed
π_9	for	$G_{\oplus} \rightarrow G \oplus G$	as $1 - \pi_8$
π_{10}	for	$G_{\otimes} \rightarrow G \otimes G_{\otimes}$	computed
π_{11}	for	$G_{\otimes} \rightarrow G \otimes G$	as $1 - \pi_{10}$
π_{12}	for	$A \rightarrow A_{do}$	required
π_{13}	for	$A \rightarrow A_{act}$	as $1 - \pi_{12}$

In order to have a valid grammar the system has to enforce that the probabilities of each production sum to 1. Let's define the groups probabilities as: $G_{Pr} = \{\{\pi_1, \pi_2\}, \{\pi_3, \dots, \pi_7\}, \{\pi_8, \pi_9\}, \{\pi_{10}, \pi_{11}\}, \{\pi_{12}, \pi_{13}\}\}$. Then the following property has to be fulfilled: $\forall Pr \in G_{Pr} \sum_{p \in Pr} p = 1$. As you can see, this property holds by construction for $\{\pi_1, \pi_2\}$ and for $\{\pi_{12}, \pi_{13}\}$. For $\{\pi_3, \dots, \pi_7\}$ it is artificially enforced (the user is required to insert weights, which are then proportionally adapted in order to sum up to 1).

The two remaining sets (i.e., $\{\pi_8, \pi_9\}$ and $\{\pi_{10}, \pi_{11}\}$) are treated slightly differently: in this case the user is required to insert the maximum number

of possible AND/XOR branches. This information let us dynamically compute the probability values. Let's consider the AND case: in the beginning, $\pi_8 = \pi_9 = 0.5$. The system keeps these values unchanged until the maximum number of AND branches are generated (i.e., the number of times that the production rule $G_{\otimes} \rightarrow G \otimes G_{\otimes}$ is consecutively executed). Once the max value is reached, probabilities are changed in order to stop generating more branches: $\pi_8 = 0$ (and therefore $\pi_9 = 1$). Similar approach is adopted for the XOR branches (i.e., $\{\pi_{10}, \pi_{11}\}$). Although this adaptation forces the context-free property of the grammar, we think that, for the final user, it is much more easy to specify the maximum number of branches instead of the actual probabilities.

In order to provide the user with a more detailed control of the grammar, we require to specify an additional parameter, which is called *maximum depth*. This parameter allows the user to control the depth of the derivation tree: once the tree reaches the *maximum depth*, probabilities are artificially changed to these values: $\pi_3 = \pi_7 = 0.5$, $\pi_4 = \dots = \pi_6 = \pi_8 = \pi_{10} = 0$. This probabilities change forces the derivation tree to limit its depth by allowing only new activity or skip patters.

5 Process Simulation in PLG2

In order to evaluate process mining algorithms or, in general, to stress business intelligence systems, we are not only interested in the random generation of processes, but we also need observations of the activities executed for each process instance, i.e. event logs. This section reports details on how we generate multiperspective logs and how to make them more realistic by artificially inserting some noise.

Before getting into the actual simulation algorithms, it is important to define the concept of event log. In order to better understand how an event log is composed, we clarify that an execution of a business process forms a *case*. The sequence of events in a case is called *trace*, and each trace, in turn, consists of the list of *events* which refer to specific activities performed. It is possible to see each event as a set of attributes (i.e., key-value pairs). The fundamental attributes of an event are: (i) the *name* of the executed activity, (ii) the *timestamp* (which reports the execution time of the given event) and (iii) the activity *lifecycle* (whether the event refers to the *beginning* or to the *completion* of an activity). The lifecycle attribute is important when a recorded activity lasted for a certain amount of time (i.e., it is not instantaneous): in this case, two events are recorded, one when the activity begins and another when the activity ends.

More formally, given the set of all possible activity names \mathcal{A} , the set of all possible case identifiers \mathcal{C} , the set of timestamps \mathcal{T} , and the set of lifecycle transitions $\mathcal{L} = \{start, complete\}$, it is possible to define an *event* e

as a tuple, such as $e = (c, a, t, l) \in \mathcal{C} \times \mathcal{A} \times \mathcal{T} \times \mathcal{L}$. In this case, it describes the occurrence of activity a , with lifecycle transition l , for the case c , at time t . Please note that the attributes reported here are just the minimum required ones: other attributes can be added to the event (in general, each data object of the process will generate a new attribute). Given an event $e = (c, a, t, l, a_1, \dots, a_k)$, it is possible to extract each field using a projection operator: $\#_{\text{case}}(e) = c$; $\#_{\text{activity}}(e) = a$; $\#_{\text{time}}(e) = t$; $\#_{\text{lifecycle}}(e) = l$, and so on.

Given a finite set $\mathbb{N}_n^+ = \{1, 2, \dots, n\}$ and a “target” set A , we define a sequence σ as a function $\sigma : \mathbb{N}_n^+ \rightarrow A$. We say that σ maps indexes to the corresponding elements in A . For simplicity, we refer to a sequence using its “string” interpretation: $\sigma = \langle s_1, \dots, s_n \rangle$, where $s_i = \sigma(i)$ and $s_i \in A$. Moreover, we assume to have concatenation and cardinality operators: respectively $\langle e_1^1, \dots, e_n^1 \rangle \cdot \langle e_1^2, \dots, e_m^2 \rangle = \langle e_1^1, \dots, e_n^1, e_1^2, \dots, e_m^2 \rangle$ and $|\langle e_1, \dots, e_n \rangle| = n$.

In our context, we use timestamps to sort the events. Therefore, it is safe to consider a trace just as a sequence of events. In turn, a log is just a set of traces. Therefore, traces are allowed to overlap: given a log l with two traces $t_1 = \langle e_1^1, \dots, e_n^1 \rangle \in l$ and $t_2 = \langle e_1^2, \dots, e_m^2 \rangle \in l$ it is possible to have that $\#_{\text{time}}(e_1^1) \leq \#_{\text{time}}(e_1^2) \leq \#_{\text{time}}(e_n^1)$ or $\#_{\text{time}}(e_1^2) \leq \#_{\text{time}}(e_1^1) \leq \#_{\text{time}}(e_m^2)$.

5.1 Multi-Perspective Simulation

The procedure for the generation of logs out of business process, basically, consists of a simulation engine running a “plain-out activity” [34]. However, in order to properly simulate all the perspectives required, some conventions need to be defined.

The structure of process models that PLG2 can handle is restricted to the family of BPMN models with an unambiguous semantic. Therefore, in PLG2, it is possible to consider a process as its equivalent Petri net representation [27, 29]. The main advantage, in this case, is that it is possible to play the token-game for simulating the process.

The procedures for the simulation of a process instance are reported in Algorithm 1, 2 and 3. These procedures use the following additional functions: in, out and rnd. Let’s assume a process $P = ((E_{\text{start}}, E_{\text{end}}, A, G, D), (S, C))$, as described in Section 3.2. Given $c \in A \cup G$, we can define $\text{in}(c) = \{c' \mid (c', c) \in S\}$ and $\text{out}(c) = \{c' \mid (c, c') \in S\}$. $\text{rnd}(s)$, instead, given a general set s , returns a randomly selected element e such that $e \in s$.

Algorithm 1 represents the main entry point of the simulation: it expects, as input, a process model and the number of traces to simulate. Then, it basically iterates the generation of single traces in order to populate the log. Line 6 is required in order to properly sort the events, and line 8 introduces, if required, some noise into the trace. The noise generation will be described in Section 5.2.

Algorithm 2 is in charge of the control-flow simulation. The algorithm

Algorithm 1: A general simulation procedure

Input : $P = ((E_{start}, E_{end}, A, G, D), (S, C))$: the process to simulate
 tot : the number of traces to generate

Output: An event log

```
1 log  $\leftarrow \emptyset$ 
2 for  $i = 1$  up to  $tot$  do
3    $t \leftarrow \langle \rangle$  ▷ Generate a new trace
4   SimulateProcess( $P, t, \text{rnd}(E_{start}), \perp$ ) ▷ Algorithm 2
6   sort( $t$ ) ▷ Sort events w.r.t. their times
8   add trace-level noise to  $t$  ▷ See Section 5.2
9   log  $\leftarrow \text{log} \cup \{t\}$  ▷ Add the trace to the log
10 end
11 return log
```

expects as input the process to simulate, the component to analyze, and the sequence (i.e., the edge) that brought the analysis to the current component. First of all, the algorithm requires the definition of a global set t . This set is fundamental for the “token game”: making an analogy with Petri nets, it stores the current marking (i.e., the tokens configuration). In our case, however, the set contains the edges that are “allowed to execute”.

The idea behind Algorithm 2 is to call itself on all elements (events, tasks, and gateways) of the process. Then different behaviors are performed, based on the analyzed element. Specifically, if the element is a task, it is simulated (line 4), and then the algorithm is called on the following component (line 10). If the analyzed element is a XOR gateway, then the call is just passed to one (randomly picked) outgoing element (line 10). If the currently analyzed element is an AND gateway, we made the assumption that it can be either a split or a join (not both at the same time). It is possible to discriminate between split and join by checking the number of outgoing edges (line 13 and 21). If the gateway is an AND split, it is necessary to make one call for each AND branch (line 19). If the gateway is a join, then it is necessary to check whether all the incoming branches are terminated (lines 22-28). If this is the case, then the flow is allowed to continue with the following activities (line 35). Please note that we omitted here the description of the token handling (i.e., insertion, check, and removal) for readability purposes: it is managed in a standard way.

Algorithm 3 is responsible for adding an activity to the provided trace. The algorithm first creates the *start* event for the activity and populates it with the standard fields (lines 1-4). If the activity has a non-instantaneous duration, the algorithm also creates a *complete* event (line 17-20).

In order to determine the activity time and its duration, the system needs to check whether the user specified any of these parameters. If no

Algorithm 2: Simulate Process

Input : $P = ((E_{start}, E_{end}, A, G, D), (S, C))$: the process to simulate
 t : the trace containing the simulated events
 c : process component to simulate
 $s = (i, c)$: incoming sequence

1 tokens \leftarrow globally defined set of tokens (i.e., sequences), initially the empty set
2 **if** c is a Task **or** c is a XOR gateway **then**
3 | **if** c is a Task **then**
4 | | Simulate Activity(P, t, c) ▷ Algorithm 3
5 | | tokens \leftarrow tokens $\setminus \{s\}$
6 | **end**
7 | **if** $|\text{out}(c)| \geq 1$ **then**
8 | | $n \leftarrow \text{rnd}(\text{out}(c))$ ▷ Randomly select the following
9 | | | component
10 | | tokens \leftarrow tokens $\cup \{(c, n)\}$ ▷ Update tokens
11 | | | Simulate Process($P, t, n, (c, n)$) ▷ Recursion
12 | **end**
13 **else if** c is an AND gateway **then** ▷ We treat c as a split
14 | **if** $|\text{out}(c)| > 1$ **then**
15 | | tokens \leftarrow tokens $\setminus \{s\}$
16 | | **forall** the $n \in \text{out}(c)$ **do** ▷ Add all tokens
17 | | | tokens \leftarrow tokens $\cup \{(c, n)\}$
18 | | **end**
19 | | **forall** the $n \in \text{out}(c)$ **do** ▷ Recursive call
20 | | | Simulate Process($P, t, n, (c, n)$)
21 | | **end**
22 **else** ▷ In this case, we treat c as a join
23 | allBranchesSeen \leftarrow **true**
24 | ▷ Check whether all branches (i.e., incoming
25 | | edges) have been executed
26 | **forall** the $p \in \text{in}(c)$ **do**
27 | | **if** $(p, c) \notin t$ **then**
28 | | | allBranchesSeen \leftarrow **false**
29 | | | break
30 | | **end**
31 | **end**
32 | **if** allBranchesSeen is **true** **then** ▷ Remove tokens
33 | | **forall** the $p \in \text{in}(c)$ **do**
34 | | | tokens \leftarrow tokens $\setminus \{(p, c)\}$
35 | | **end**
36 | | $n \leftarrow \text{out}(c)$ ▷ Get the outgoing edge
37 | | tokens \leftarrow tokens $\cup \{(c, n)\}$ ▷ Update tokens
38 | | | Simulate Process($P, t, n, (c, n)$) ▷ Recursive call
39 | | **end**
40 | **end**
41 **end**

Algorithm 3: Simulate Activity

Input : $P = ((E_{start}, E_{end}, A, G, D), (S, C))$: the process
 t : the trace that contain the new events
 a : activity to simulate

▷ Generate the activity start event

- 1 $e_{start} \leftarrow$ new event referring to activity a
- 2 $\#_{activity}(e_{start}) \leftarrow$ the name of activity a
- 3 $\#_{time}(e_{start}) \leftarrow$ activity time ▷ Details in text
- 4 $\#_{lifecycle}(e_{start}) \leftarrow start$

▷ Decorate with all generated data objects

- 5 **forall the** $d \in \{d \mid (a, d) \in C\}$ **do**
- 6 | $\#_d(e_{start}) \leftarrow$ value generated for d
- 7 **end**

▷ Decorate with all required data objects

- 8 **if** $|t| > 1$ **then**
- 9 | **forall the** $d \in \{d \mid (d, a) \in C\}$ **do**
- 10 | | $lastEvent \leftarrow t(|t| - 1)$
- 11 | | $\#_d(lastEvent) \leftarrow$ value generated for d
- 12 | **end**
- 13 **end**

- 14 add event-level noise to e_{start} ▷ See Section 5.2
- 15 $t \leftarrow t \cdot \langle e_{start} \rangle$

▷ Generate the activity completion event

- 16 **if** activity a is not instantaneous **then**
- 17 | $e_{complete} \leftarrow$ new event referring to activity a
- 18 | $\#_{activity}(e_{complete}) \leftarrow$ the name of activity a
- 19 | $\#_{time}(e_{complete}) \leftarrow \#_{time}(e_{start}) +$ activity duration
- 20 | $\#_{lifecycle}(e_{complete}) \leftarrow complete$

▷ Decorate with all generated data objects

- 21 **forall the** $d \in \{d \mid (a, d) \in C\}$ **do**
- 22 | $\#_d(e_{start}) \leftarrow$ value generated for d
- 23 **end**
- 24 add event-level noise to $e_{complete}$ ▷ See Section 5.2
- 25 $t \leftarrow t \cdot \langle e_{complete} \rangle$
- 26 **end**

specifications are reported, then the activity is assumed to be instantaneous and to execute a fixed amount of time after the previous one. However, as said, the user can manually specify these parameters. To do so, the user has to provide two Python [30] functions: `time_after(caseId)` and `time_lasted(caseId)`. Both these functions are called by the simula-

Listing 1: Example of random activity duration (between 5 and 15 minutes) and random time after the execution of an activity (between 1 and 5 minutes).

```
from random import randint
# This Python script is called for the generation of the time
  related features of the activity. Note that the functions
  parameters are the actual case id of the ongoing simulation (
  you can use this value for customize the behavior according to
  the actual instance).

# The time_after(caseid) function is returns the number of second
  to wait before the following activity can start.
def time_after(caseid):
    return randint(60*1, 60*5)

# The time_lasted(caseid) function returns the number of seconds
  the activity is supposed to last
def time_lasted(caseid):
    return randint(60*5, 60*15)
```

tor with the `caseId` parameter valued to the actual case id: this allows the two functions to be case-dependent (for example, it is possible to save files with contextual information). Specifically, `time_after(caseId)` is required to return the number of seconds that have be (virtually) waited before the following activity is allowed to start. `time_lasted(caseId)`, instead, has to return the number of seconds that the activity is supposed to last. This approach is extremely flexible, and allows the user to make very complex simulations. For example, it is possible to define different durations for the same activity depending on which flow the current trace has followed so far, or with respect to the number of iterations on a loop. Examples of such functions are reported in listing 1.

Once the time-related properties of an activity are computed, Algorithm 3, has to deal with the data objects associated with the current activity. In particular, *generated* data objects (see Section 3.1) are supposed to generate values written as the current activity's attribute. *Required* data objects, instead, are written as attributes for the activity which precedes the current one in the trace. The ratio behind this decision is that *generated* data objects are assumed as values written as output of the current activity. *Required* data objects, instead, are variables that has to be observed prior to the execution of the current activity. However, since the simulation is driven by the control-flow, it is necessary to adjust the variable values *a posteriori*.

In order to better understand the utility of *required* data objects, let's consider the process fragment reported in Figure 3. In this case, the simulation will first perform "Activity A" and then either "Activity B" or "Ac-

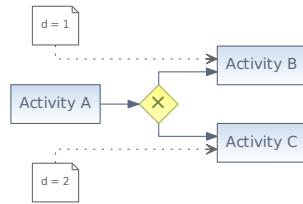


Figure 3: A process fragment of a XOR split gateway with two branches, each of them starting with a different required data object.

Listing 2: Example of script for the generation of random integer values (in the range 0, 1000).

```

from random import randint
# This Python script is called for the generation of the integer
  data object. Note that the parameter of this function is the
  actual case id of the ongoing simulation (you can use this
  value to customize your data object). The function name has to
  be "generate".

def generate(caseId):
    return randint(0, 1000)

```

tivity C". However, all the times the simulation engine generates "Activity B", it also decorates the event referring to "Activity A" (belonging to the same trace) with $d = 1$. Instead, all the times the simulation engine generates "Activity C", it also decorates the event referring to "Activity A" (belonging to the same trace) with $d = 2$. Therefore, an analysis system fed with such example trace could infer a correlation between the value of the attribute d of "Activity A", and the following activity.

From the characterization reported in Fig. 1, and described in Section 3, it is possible to distinguish two types of data objects: *plain data objects* and *dynamic data objects*. This distinction is required by the simulation engine in order to properly deal with them: plain data objects are treated as fixed values (i.e., the simulation generates always the same value); dynamic data objects are actually Python scripts whose values are determined by the execution of the script itself. These scripts must implement a `generate(caseId)` function which is supposed to return either an integer or a string value (depending on the type of data object). An example of integer dynamic data object script is reported in listing 2. Please note that, also in this case, the function is called with the `caseId` parameters valued with the actual instance's case id, providing the user with an in-depth, and case depen-

dent, control over the generated values. `ScriptExecutor` component (and its subclasses), reported in Figure 1, are in charge of the execution of the Python scripts.

There is no particular limit on the number of plain and dynamic data objects that a task can have, both required and generated. Clearly, the higher the number of data objects to generate, the longer the simulation will take.

The current random process generator component is only able to generate plain data objects. Specifically, the generated data objects are named `variable_a`, `variable_b`, ... and the values they return are just random strings.

Considering all the simulation aspects described in this section we can conclude that our approach is able to simulate multiperspective models in order to generate multiperspective logs. The “multiperspective” term, in this context, means that the data generate does not only refer to the control-flow perspective, but have also detailed timing properties and the data generate could be extremely articulated and tailored to the actual simulation scenario.

5.2 Noise Addition

In order to generate more realistic data, we introduced a noise component.

The noise component plays a role after the process has been simulated and a trace is available. Specifically, this trace is fed to the noise component which could apply noise at three different “levels”: (i) at the trace level (i.e., noise which involve the trace organization); (ii) at the event level (i.e., noise which involve events on the control-flow perspective); (iii) at the data object level (i.e., noise which involve the data perspective associated to events). The actual noise generation is driven by the parameters set by the user. Such parameters, basically, indicate the probability of applying a particular noise type to the trace. Setting all these values to zero implies having trace with no noise.

The noise details for the trace and -partially- for the event level have already been discussed in the literature and reported in details in [11, 19]. The idea is that the user has to specify the probability of all the different noise events, and the simulator will apply the corresponding effect. Possible trace-level noise phenomena are:

- a trace which is missing its *head* (i.e., its first events). In this case the user has also to specify the maximum size for a head (which will be randomly chosen between 1 and the provided value);
- a trace which is missing its *tail* (i.e., its last events). In this case the user has also to specify the maximum size for the tail (which will be randomly chosen between 1 and the provided value);

- a trace which is missing an *episode* (i.e., a sequence of contiguous events). In this case the user has also to specify the maximum size for an episode (which will be randomly chosen between 1 and the provided value);
- an alien event introduced into the trace, in a random position, with random attributes;
- a doubled event on the trace.

Possible noise effects at the event level are:

- the random change of the activity name of an event;
- the perturbed order between two events of a trace (since the timestamp attributes of two events are involved, we consider this as an event-level noise).

Finally, possible data object-level noises are:

- random modification of an integer dynamic data object. In this case the user has also to specify the maximum value Δ of the change: given the old value v , the new one (i.e., after noise) will be $v + \delta$, with δ random in the closed interval $[-\Delta, +\Delta]$;
- random modification of a string dynamic data object (replacement of the current string with a randomly generated new one).

In order to simplify the noise configuration, we already defined some basic “noise profiles”, such as: (i) complete noise; (ii) noise only on the control-flow; (iii) noise only for data-objects; (iv) noise only on activity names; (v) no noise at all.

6 Stream Simulation in PLG2

As stated before, PLG2 explicitly was design for the simulation of online event streams. Specifically, in this context, we adopted the definition of stream already used in the process mining community [6,25].

6.1 Continuous Data Generation

An event stream differs from an event log in two fundamental aspects. First of all, and event stream has not a predefined end (i.e., the user can generate as many events he wants, so the simulation can last for an unspecified amount of time). The second distinction consists in keeping the events sorted by their time, and not grouped.

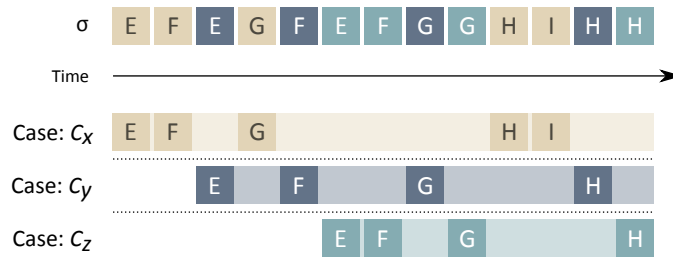


Figure 4: Graphical representation of an event stream. Boxes represent events: their background colors represent the case id, and the letters inside are the activity names. First line reports the stream, following lines are the single cases.

Therefore, differently from an event log (which is a set of sequences, i.e., the traces), an event stream is just a sequence of events. Therefore, the only property that must be enforced is that, given an event stream σ , for all indexes i available, $\#_{time}(\sigma(i)) < \#_{time}(\sigma(i + 1))$. Instead, it will happen, for some indexes i , that $\#_{case}(\sigma(i)) \neq \#_{case}(\sigma(i + 1))$ (i.e., contiguous events refer to different process instances). In this last case, the two events are said to belong to *interleaving traces*. Figure 4 reports a graphical representation of three interleaving traces and how the actual stream looks like.

From an implementation point of view, the idea is to create a socket, which is accepting connections from external clients. PLG2, then, “emits” (i.e., writes on the socket) events that are generated. The challenge, in this case, is to let the system simulate and send events for a potentially infinite amount of time.

In order to generate our continuous stream, we need to ask the user for two parameters: the maximum number of parallel instances running at the same time, and the “time scale”. The first parameter is used to populate the data structures required for the generation of the stream. Then, since the event emission is performed in “the real time” (opposed to the “simulated time”), it might be necessary to scale the simulation time in order to have the desired events emission rate. To this end we need a time multiplier, which is expected to be defined in $(0, \infty]$. This time multiplier is used to transform the duration of a trace (and the time position of all the contained events), from the simulation time to the real time.

The procedure for the generation of streams is reported in Algorithm 4. It starts by allocating as many priority queues as the number of parallel instances of the stream (line 3). These queues are basically used as events buffer. Then, the procedure starts a potentially infinite loop for the events streaming. At the beginning of this loop, the algorithm first needs to check whether the buffer contains enough events. If this is not the case (line 7), then a new process instance is simulated (line 9, using Algorithm 1)

and, after applying the time scale (line 10), all its events are added to the event buffer (line 11). Events are enqueue considering their time order (i.e., events with lower timestamps have higher priority).⁵ Once the algorithm is sure about the availability of events, it extracts (and removes), from the buffer, the event with highest priority (line 13). At this point, it is necessary to make happen the mapping between the simulation and the real time: the algorithm has to wait for a certain amount of time, in order to ensure the correct event distribution in the real time (line 16). After such wait, the event can finally be emitted (line 20), and all connected clients are notified.

Please note that, every time the algorithm has to repopulate the buffer, it asks the framework for the process which has to be simulated (line 8). This is a fundamental point: the user can change the process for the simulation, without stopping the current stream emission, and if such change occurs, a *concept drift* will be observed. Concept drifts [15,18,25,26] represent another important characteristic, which fundamentally differentiate event streams from event logs and, therefore, identify a requirement.

Please note also that, in order to have a more accurate mapping between simulation and real time, the implementation of the buffer population procedure (lines 7-11 of Alg. 4) can be executed in an external thread.⁶

In order to assess the feasibility of this algorithm we run several experiments. In particular, we generated the process model reported in Fig. 5a. This process contains 10 activities, one parallel execution, one loop and one generated data object. Then, we run the streamer of this process for two hours, generating, in total, 4 174 different traces and 67 856 events. The average throughput of the streamer, after an initial configuration stage, was set at 9.4 events per second. Figure 5b reports the memory requirement of the approach: the evolution of the buffer size and the total number of events sent are plotted against the running time of the actual stream. As the plot shows, the average number of stored event is between 300 and 400 events, which represents an affordable memory requirement for any hardware configuration available nowadays.

6.2 Concept Drifts for Process Models

One common characteristic of online settings is the presence of concept drifts. As described in the previous section, the tool is able to dynamically switch the source generating the events. However, in order to change the stream source, a new model is required. To create another model, two op-

⁵Implementation details are skipped here, but some time manipulations are required in order to insert the new trace after all events already enqueued and keeping a certain amount of time from the last event.

⁶This cannot ensure a *completely correct* mapping, however the difference has empirically seen negligible.

Algorithm 4: Stream

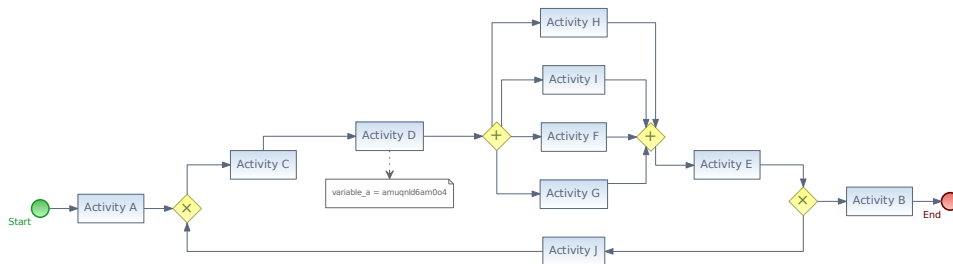
Input : p : the number of parallel instances
 m : time multiplier

▷ Initialization of the data structures

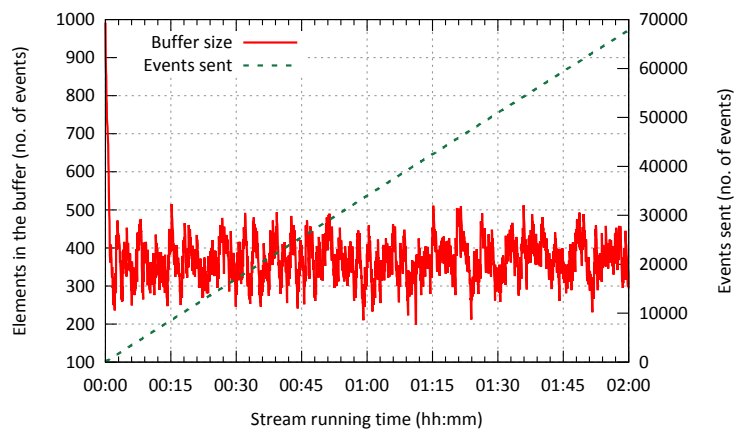
```
1 queues  $\leftarrow \emptyset$  ▷ This is an event buffer
2 for  $i = 1$  up to  $p$  do
3   | queues  $\leftarrow$  queue  $\cup$  {a new priority queue}
4 end
5  $l \leftarrow \perp$ 
6 forever do
7   | ▷ Populate the event buffer
8   | if  $|\text{queues}| < 2p$  then ▷ Here  $|\text{queues}|$  is the sum of sizes
9   |   | of all queues contained. Although the inequality
10  |   | could be  $< 1$ , we prefer to use  $2p$  since these
11  |   | operations could be performed in a different
12  |   | concurred thread
13  |   |  $proc \leftarrow$  the process to simulate
14  |   |  $t \leftarrow$  simulate a new trace for  $proc$  ▷ Alg. 1
15  |   | scale the trace duration (and events times) according to  $m$ 
16  |   | distribute the events of  $t$  (sorted by their time) to the queue
17  |   | with the highest priority of the last event
18  |   | end
19  |   | ▷ The actual streaming
20  |   |  $e \leftarrow$  extract (and remove) the event with highest priority from all
21  |   | queues ▷ From queues
22  |   | if  $l \neq \perp$  then
23  |   |   |  $w \leftarrow \#_{time}(e) - \#_{time}(l)$ 
24  |   |   | wait for  $w$  time units
25  |   |   | end
26  |   |  $l \leftarrow e$ 
27  |   |  $\#_{time}(e) \leftarrow$  now
28  |   | emit  $e$  ▷ To all connected clients
29 end
```

tions are available: one is to load or generate from scratch a model; the other is to “evolve” an existing one: this is an important feature of PLG2.

To evolve an existing model, PLG2 replaces an activity with a subprocess generated using the context-free grammar described in Section 4. This operation, which takes place randomly, and with a probability provided by the user, is repeated for each activity of the process. The new process could be very similar to the originating one, or very different, and this basically depends on the probability configured by the user.



(a) Process model used for performance computation of the stream reported in this section.

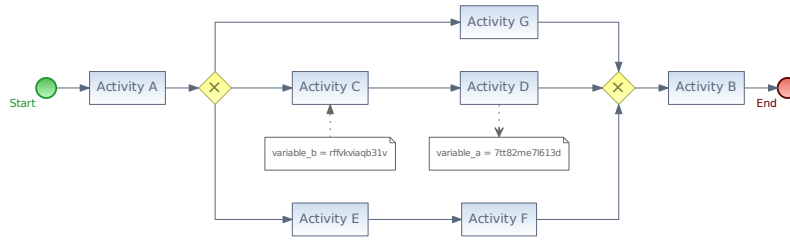


(b) Size of the buffer and total number of events sent for the process reported in Fig. 5a.

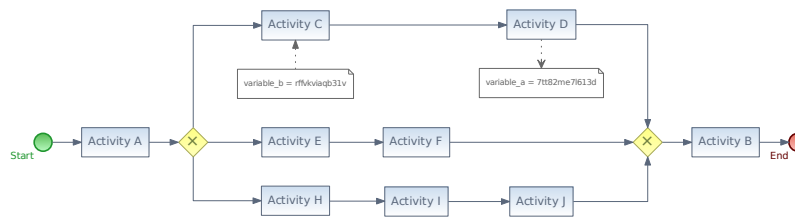
Figure 5: Simulations details: the process model used and the size of the buffer. The entire simulation lasted for two hours.

For example, Figure 6 reports two evolutions of the process model, which has been randomly generated and which is reported in (a). In (b) the procedure applies the evolution by replacing “Activity G” with the sequence of three activities (“Activity H”, “Activity I” and “Activity J”). In (c), the evolution involves “Activity D” (and the associated data object) which is replaced with a skip (i.e., it is removed).

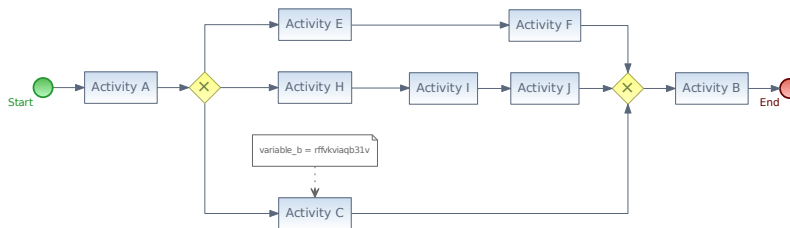
Please note that an evolution could involve the creation or the deletion of data objects as well. Process evolution, therefore, can be used for the definition of particular experiments (e.g., a stream with random concept drifts occurring every 1000 events).



(a) Starting process model, randomly generated.



(b) First evolution of the process model. In this case, activity G has been replaced by a sequence of three activities (H, I, J).



(c) Second evolution of the process model. In this case activity D — and the associated data object 'variable_a' — have been removed.

Figure 6: A process model randomly generated with two sequential evolutions. Please note that new activities can be introduced or removed (with the associated data objects).

7 Implementation Details of PLG2

PLG2 has been implemented in a Java application. It is available as open source project and also binary files are provided for convenience.⁷ The project APIs can also be easily used to randomly generate processes or logs. Listing 3 reports the Java code required to generate a random process model; to simulate it in order to create 1000 traces; and to export the

⁷See <http://plg.processmining.it> and <https://github.com/delas/plg>.

Listing 3: Java fragment for the creation of a new process, its simulation (to generate 1000 traces), and its export as Petri net.

```
// process randomization
Process p = new Process("test");
ProcessGenerator.randomizeProcess(p,
    RandomizationConfiguration.BASIC_VALUES);

// log simulation to generate 1000 traces
XLog l = new LogGenerator(p,
    new SimulationConfiguration(1000)).generateLog();

// export as pnml
new PNMLExporter().exportModel(p, "p.pnml");
```

process as a Petri net (using the PNML standard).

The current implementation is able to load BPMN files generated with Signavio or PLG2. It is also possible to export a model as PNML [21] or PLG2 file. Moreover, it is possible to export graphical representation of the model (both in terms of BPMN and Petri net) using the Graphviz file format [13]. The simulation of log files generates a XES⁸-compliant [20] objects, which can be exported both as (compressed) XES or (compressed) MXML. These formats are widely used by most process mining tools.

Figure 7 reports a screenshot of the current implementation of PLG2. From the picture it is possible to see the main structure of the GUI: there is a “workspace” list of generated processes on the left. The selected process is shown on the main area. Right clicking on activities allows the user to set up activity-specific properties (such as times, or data objects). On the bottom part of the main application it is possible to see the PLG2 console. Here the application reports all the log information, useful for debugging purposes. The application dialog in the foreground is used for the configuration of the Python script which will be used to determine the time properties. As shown, specific syntax highlighting and other typing hints (such as automatic indentation) helps the user in writing Python code. The stream dialog is also displayed in foreground. As can be seen, in this case, it is possible to dynamically change the streamed process and the time multiplier. The right hand side of such dialog (in the rectangle with black background), moreover, reports “a preview” of the stream: 30 seconds of the stream are reported (each round dot represents, in this case, up to 3 events).

As stated previously, some components of PLG2 require the execution of Python scripts. To deal with that we used the Jython framework⁹ which,

⁸See <http://www.xes-standard.org>.

⁹See <http://www.jython.org>.

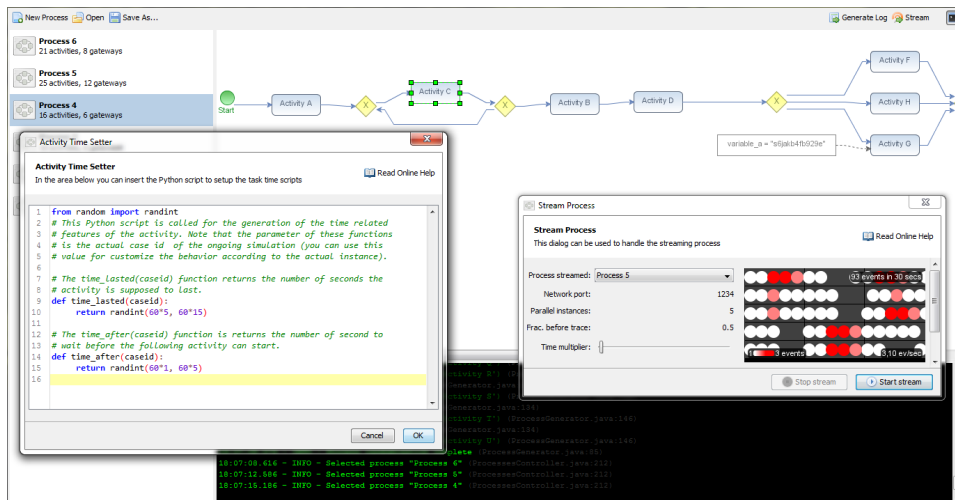


Figure 7: Screenshot of PLG2. From the current visualization it is possible to see that several models have been created in the workspace, the dialog for the time rules configuration for “Activity C”, and the console, which shows general information on what is going on. The stream dialog is reported as well.

basically, is an implementation of Python which can run in Java. The interaction between Java and Python objects is encapsulated in the `ScriptExecutor` hierarchy, reported in Fig. 1.

Since it is possible to repeat the code fragment reported in Listing 3 as many times as required, we are able to fulfill C1 (Section 1.1). The detailed process simulation, the advanced data values generation and the noise configuration are necessary to create realistic multiperspective event logs and therefore to accomplish C2. Finally, the feasibility of the stream procedure reported, together with its main features (such as the possibility to generate multiperspective streams, the dynamic change of the originating process model and the possibility to adapt the time between events emitted) makes possible to successfully cope with C3.

8 Case Studies

In this section we would like to propose two possible scenarios in which PLG2 could easily be applied. In particular we will show a multiperspective analysis, performed in offline setting; and a control-flow discovery activity in online scenario.

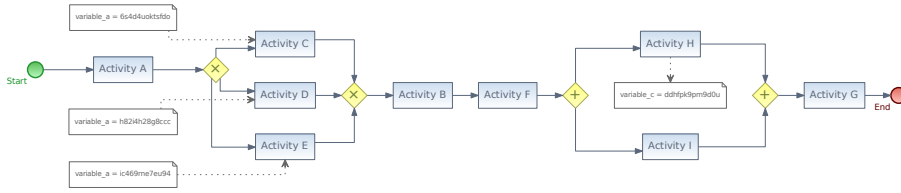


Figure 8: Process model used for the offline simulation.

8.1 Offline Setting

On the first case study, we would like to analyze both the control-flow and the data perspectives of a log files.

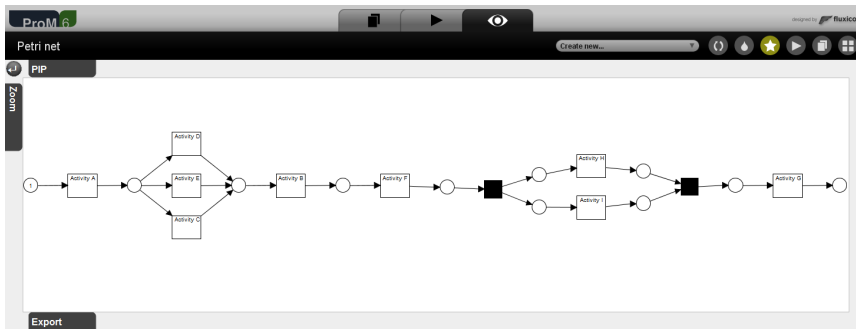
To perform our test, we generated a random process model, and then we manually slightly modified it, in order to fit our goals. Specifically, we added required data objects to activities C, D, and E. These data objects are all named `variable_a`, but each of them has a different value. The generated model is reported in Figure 8 and represents our gold standard. Therefore, when we perform the data analysis, we expect the presence of a variable influencing the control-flow for those activities.

To perform our simulation, we generated a log with 2000 traces and then we analyzed it using ProM¹⁰ [40]. For the control-flow discovery analysis we run the Inductive Miner [24] algorithm. Then, we converted the generated model into a Petri net. The result is reported in Figure 9a. As we can see, from the behavioral point of view, the mined model reflects the original one, except for the data perspective (which cannot be extracted with Inductive Miner). Starting from the Petri net mined, we run the Data-flow Discovery plugin [10] in order to add data variables governing the control-flow. The result, which is reported in Figure 9b, shows the presence of a variable named `variable_a` which is written by activity A, and read by activities C, D, and E. The screenshot also reports the actual guard for activity E (i.e., the value that is required in order to execute that activity). Both the control-flow and data flow mined reflect the expected ones.

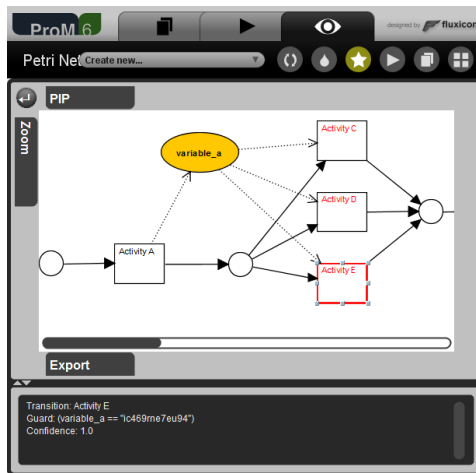
As a second test, we mined the control-flow using the tool Disco.¹¹ The control-flow discovered by the tool is shown in Figure 10. The formalism, adopted by the tool for the representation of business processes, allows us to see, basically, only direct following relationships. As we can note, activities C, D and E are executed, respectively 676, 622, and 702 times. Since in total we have 2000 traces, this is an indication that maybe those activities are mutually exclusive (although this is not necessary). Instead, activities H and I are both executed 2000 times but we see there are connections be-

¹⁰See <http://www.promtools.org>.

¹¹See <http://fluxicon.com/disco/>.



(a) Control-flow extracted using the Inductive Miner algorithm and then converted into a Petri net.



(b) Result of mining the data flow, which decorates the mined Petri net.

Figure 9: Results of mining activities (control-flow and data flow) performed using ProM plugins.

tween them. These connections indicate that the activities are not executed in a specific order (i.e., they are parallel). These behavioral characteristics reflect the gold standard.

8.2 Online Setting

For the second case study, we decided to analyze the online scenario with concept drifts. To achieve this goal, we created a second model (M_2), different from the previous one (M_1). Then we started streaming events referring to M_1 .

In the meanwhile, we configured the stream mining plugin implemented in ProM and described in [6]. Specifically, we used the mining approach based on Lossy Counting, with parameter $\epsilon = 0.032$. We also configured

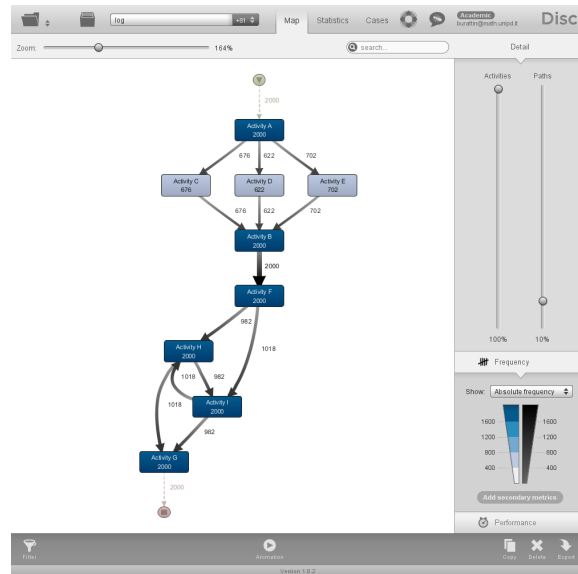


Figure 10: Result of mining the log using the tool Disco.

the miner to update the graphical representation of the process model every 500 events received. The sequence of models extracted is reported in Figure 11.

The first two models extracted are not equivalent to the expected one, since the miner needs several observations in order to reinforce and accept the patterns. Figure 11c shows the model which is equivalent to the gold standard (in this picture split/join semantics are not reported for readability purposes). At this point, we decided to change the stream, and emit events referring to M_2 (i.e., we simulated the occurrence of a concept drift). The first model extracted after such concept drift is reported in Figure 11d and shows both process models M_1 and M_2 embedded into the same representation. This is a known phenomenon, and is due to the *inertia* of the stream-based approaches. After some events, since the miner is not receiving anymore observations from M_1 , it starts to forget its structures. After some more events, the second model M_2 is definitely discovered, as shown in Figure 11g, and no traces of M_1 are left.

With these two case studies, we tried to show some of the possible usages of the described approaches. In these tests, we just used algorithms already available in the literature. However, the primary goal should be testing new ones. Moreover, in the described cases, we just manually compared the mined models and the expected ones but this could be done automatically. Finally, since we provide libraries to perform all functionalities via Java code, batch approaches could be designed, in order to perform the same operations against large repositories with models expressing very dif-

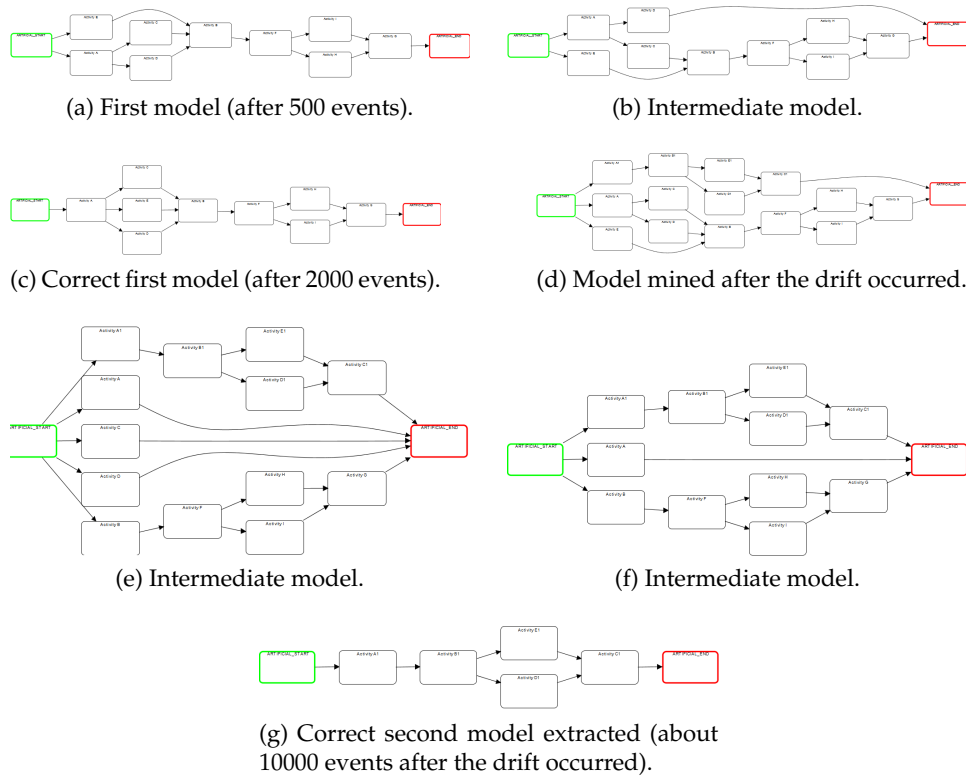


Figure 11: Evolution of discovered models during a process stream simulation with a concept drift occurred.

ferent behaviors.

9 Conclusions and Future Work

This paper describes PLG2, which is the evolution of an already available tool. The old tool was able to randomly generate process models and simulate them. The new tool introduces updates on two sides: on one hand it extends the support to multiperspective models (by adding detailed control of time perspective and introducing data objects); on the other hand, full support for the simulation of online settings (generating drifting models and simulating event streams) is provided.

We believe, that the combination of the two newly introduced aspects allows the tool to be a valid instrument for the data mining, information systems, and process mining community, since it allows the simulation of very complex scenarios. As the predecessor of this tool has proven, by its wide adoption, we think that the new features of PLG2 are important in order to push and help researchers to tackle the new challenges that up-

coming settings propose (for example, *big data* requires to handle streams of multiperspective data).

We think that a lot of work is necessary in this field: the simulation of real scenarios is a very tough and broad task. In particular, it is important to investigate how to generate even more realistic scenarios. To achieve such realism, it is necessary to work both on the model generation (control-flow, time and data perspective) and on the simulation (for example identifying new types of noise). For example, the introduction of noise on the modeling could be considered (e.g., inserting or removing edges randomly, or in specific contexts).

An example of possible future work consists in the ad hoc simulation of the social perspective (identifying common patterns and possible behaviors) which, right now, is already possible, but just through the data perspective (e.g., generating data that describe the originators). Another future work, on the simulation part consists in introducing noise referring not to the trace/event modification, but to the distribution of the cases (i.e., not all control-flow paths are equally probable).

References

- [1] C. Aggarwal. *Data Streams: Models and Algorithms*, volume 31 of *Advances in Database Systems*. Springer US, Boston, MA, 2007.
- [2] G. Bergmann, A. Horváth, I. Ráth, and D. Varró. A Benchmark Evaluation of Incremental Pattern Matching in Graph Transformation. In *ICGT International conference on Graph Transformations*, number i, pages 396–410, Berlin, Heidelberg, 2008. Springer-Verlag.
- [3] J. Buijs, B. van Dongen, and W. M. P. van der Aalst. A Genetic Algorithm for Discovering Process Trees. In *Proceedings of WCCI 2012 IEEE World Congress on Computational Intelligence*, pages 925–932, Brisbane, Australia, 2012.
- [4] A. Burattin. *Process Mining Techniques in Business Environments*. Springer International Publishing, 2015.
- [5] A. Burattin and A. Sperduti. PLG: a Framework for the Generation of Business Process Models and their Execution Logs. In *Business Process Management Workshops (BPI)*, pages 214–219, Hoboken, New Jersey, USA, 2010. Springer Berlin Heidelberg.
- [6] A. Burattin, A. Sperduti, and W. M. P. van der Aalst. Control-flow Discovery from Event Streams. In *Proceedings of the IEEE Congress on Evolutionary Computation*, pages 2420–2427. IEEE, 2014.

- [7] N. Chomsky. Three models for the description of language. *IRE Transactions on Information Theory*, 2(3):113–124, 1956.
- [8] N. Chomsky and G. A. Miller. *Introduction to the formal analysis of natural languages*. Wiley, 1963.
- [9] K. J. Cios, R. W. Swiniarski, W. Pedrycz, and L. A. Kurgan. *Data Mining - A Knowledge Discovery Approach*. Springer US, 2007.
- [10] M. de Leoni and W. M. P. van der Aalst. Data-aware process mining: discovering decisions in processes using alignments. *Proceedings of the 28th Annual ACM Symposium on Applied Computing - SAC '13*, page 1454, 2013.
- [11] A. K. A. de Medeiros. *Genetic Process Mining*. Phd thesis, Technische Universiteit Eindhoven, 2006.
- [12] A. K. A. de Medeiros and C. W. Günther. Process Mining: Using CPN Tools to Create Test Logs for Mining Algorithms. In *Proceedings of the Sixth Workshop and Tutorial on Practical Use of Coloured Petri Nets and the CPN Tools*, pages 177–190, 2005.
- [13] J. Ellson, E. R. Gansner, E. Koutsofios, S. C. North, and G. Woodhull. Graphviz and Dynagraph Static and Dynamic Graph Drawing Tools. Technical report, AT&T Labs - Research, Florham Park NJ 07932, USA, 2004.
- [14] W. Fan and A. Bifet. Mining Big Data : Current Status , and Forecast to the Future. *ACM SIGKDD Explorations Newsletter*, 14(2):1–5, 2013.
- [15] J. a. Gama. *Knowledge Discovery from Data Streams*, volume 20103856 of *Chapman & Hall/CRC Data Mining and Knowledge Discovery Series*. Chapman and Hall/CRC, May 2010.
- [16] Gartner. Big Data, IT Glossary.
- [17] R. Giegerich. Introduction to Stochastic Context Free Grammars. In *RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods*, volume 1097, pages 85–106. Springer Science+Business Media New York, 2014.
- [18] V. Grossi and F. Turini. Stream mining: a novel architecture for ensemble-based classification. *Knowledge and Information Systems*, Feb. 2011.
- [19] C. W. Günther. *Process mining in Flexible Environments*. Phd thesis, Technische Universiteit Eindhoven, Eindhoven, 2009.

- [20] C. W. Günther and E. H. M. W. Verbeek. XES Standard Definition. www.xes-standard.org, 2009.
- [21] L. Hillah, E. Kindler, F. Kordon, L. Petrucci, and N. Trèves. A primer on the Petri net markup language and ISO/IEC 15909-2. *Petri Net Newsletter*, (October):101–120, 2009.
- [22] J. E. Hopcroft, R. Motwani, and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation*. Pearson, 3rd editio edition, 2006.
- [23] O. Kopp, D. Martin, D. Wutke, and F. Leymann. The Difference Between Graph-Based and Block-Structured Business Process Modelling Languages. *Enterprise Modelling and Information Systems*, 4(1):3–13, 2009.
- [24] S. J. J. Leemans, D. Fahland, and W. M. P. van der Aalst. Discovering Block-Structured Process Models from Event Logs - A Constructive Approach. In *Proceedings of Petri Nets*, pages 311–329. Springer Berlin Heidelberg, 2013.
- [25] F. M. Maggi, A. Burattin, M. Cimitile, and A. Sperduti. Online Process Discovery to Detect Concept Drifts in LTL-Based Declarative Process Models. In *On the Move to Meaningful Internet Systems: OTM 2013 Conferences*, pages 94–111. Springer Berlin Heidelberg, 2013.
- [26] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*, volume 35. Cambridge University Press, 1st edition, June 2008.
- [27] T. Murata. Petri nets: Properties, analysis and applications. *Proceedings of the IEEE*, 77(4):541–580, 1989.
- [28] OMG. *Business Process Model and Notation (BPMN) - Version 2.0, Beta 1*. 2009.
- [29] J. L. Peterson. Petri Nets. *ACM Computing Surveys (CSUR)*, 9(3):223–252, 1977.
- [30] Python Software Foundation. Python Language Reference, version 2.x.
- [31] N. Russell, A. H. ter Hofstede, W. M. P. van der Aalst, and N. Mulyar. Workflow Control-flow Patterns: A Revised View. *BPM Center Report BPM-06-22*, BPMcenter.org, 2006.
- [32] Y. Sakakibara, M. Brown, R. Hughey, I. S. Mian, K. Sjölander, R. C. Underwood, and D. Haussler. Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Research*, 22(23):5112–5120, 1994.

- [33] W. M. van der Aalst, J. Buijs, and B. van Dongen. Towards improving the representational bias of process mining. In *Lecture Notes in Business Information Processing*, volume 116 LNBIP, pages 39–54, 2012.
- [34] W. M. P. van der Aalst. *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Springer Berlin / Heidelberg, 2011.
- [35] W. M. P. van der Aalst, A. H. ter Hofstede, B. Kiepuszewski, and A. P. Barros. Workflow Patterns. *Distributed and Parallel Databases*, 14(1):5–51, 2003.
- [36] W. M. P. van der Aalst and K. M. van Hee. *Workflow management: models, methods, and systems*. The MIT press, 2004.
- [37] K. Van Hee, M. La Rosa, Z. Liu, and N. Sidorova. Discovering characteristics of stochastic collections of process models. In *9th International Conference, BPM 2011, Clermont-Ferrand, France, August 30 - September 2, 2011. Proceedings*, volume 6896 LNCS, pages 298–312, 2011.
- [38] K. M. van Hee and Z. Liu. Generating Benchmarks by Random Stepwise Refinement of Petri Nets. In *Proceedings of workshop APNOC/-SUMo*, 2010.
- [39] J. Vanhatalo, J. Vanhatalo, V. Hagen, and J. Koehler. The Refined Process Structure Tree. In *Proceedings of the 6th International Conference, BPM 2008*, pages 100–115. Springer Berlin Heidelberg, 2008.
- [40] E. H. M. W. Verbeek, J. Buijs, B. van Dongen, and W. M. P. van der Aalst. ProM 6: The Process Mining Toolkit. In *BPM 2010 Demo*, pages 34–39, 2010.