Discovering the Influence of Exogenous Data on Decisions in Processes

Adam Banham[1], Yannis Bertrand[2], Robert Andrews[1], Moe T. Wynn[1], and Sander J.J. Leemans[3]

[1] Queensland University of Technology, Brisbane, Australia
adam.banham@hdr.qut.edu.au
[2] KU Leuven, Belgium
[3] Fraunhofer FIT, Germany & RWTH Aachen, Germany

Abstract. Process mining, when applied to data stored in the information systems of businesses, provides insights into the internal performance of their processes. These insights reveal how the behaviour of processes impacts businesses, and can inform planning for various future scenarios by anticipating how processes will perform in these scenarios. However, these scenarios may be influenced by the context of the process, i.e., external data streams (exogenous data) to the process. In these cases, typical process discovery techniques can produce process models that describe what activities could occur next in a given state, but cannot express the effect of external influences on how likely these are to occur. Our contribution presents an extension of stochastic labelled Petri nets and a discovery technique for our new modelling formalism. The proposed formalism can be used to quantify whether the firing likelihood of a transition is influenced by exogenous data when replaying historical process executions over the net. We compare our approach against existing stochastic techniques over several publicly available event logs. Our results show that our approach can outperform existing data-aware techniques in unstructured processes.

Keywords: Process Mining · Process Enhancement · Stochastic Process Mining · Exogenous Data.

1 Introduction

In the field of business process management [8], studying how processes change at runtime [24,10,5,26,29] is important. Studying the dynamics of how processes execute across a collection of varying scenarios enables businesses to plan for future events based on their likely occurrence. Process mining [28] enables businesses to derive the diagnostic information needed to empower such planning by considering historical data about processes. Process discovery techniques [14,17,13] can generate process models that describe when activities occur in a process and how processes are structured [28]. Conformance checking techniques [18,23] can quantify deviations between historical data and a process model to highlight

anomalies at runtime [1,6]. However, to understand the likelihood of an activity at runtime, the mining of stochastic processes is required [9].

Stochastic process mining provides a window into understanding process-level decisions by describing which activities could occur next, and how likely these activities are to occur [9]. Various mechanisms for determining next activity likelihood exist [25,27,5,19,20,4,16,23], with recent extensions considering the dependencies within the history of a trace [16] or case/event attributes [23]. Quantifying the likelihood of an activity may require considering many different dependencies, both internal and external, to the process. Addressing this challenge within stochastic process mining [9,4,16,23] is ongoing. Our contribution is a novel approach for determining whether historical process executions were influenced by external data sources, referred to as *exogenous data* [2].

We investigate how stochastic labelled Petri nets [15] (SLPN) could be used to understand if the behaviour of processes has been influenced by exogenous data. SLPNs are themselves extensions of Petri nets, where transitions are extended with a weight function to represent their chance of firing. Various approaches to deriving these weight functions have been used to investigate several aspects of processes [4,16,23], but all have overlooked exogenous factors.

Behaviour in processes may be informed by internal dependencies, or by exogenous influences, i.e. where processes are influenced by their context [29]. Agriculture supply chains can be influenced by climate variability [30,11] and patient care in hospitals frequently involves the continuous observation of the patient's vital signs [2]. Notably, for these cases, existing discovery techniques [4,16,23] cannot capture exogenous influences on the firing likelihoods. In particular, existing techniques overlook temporal factors in exogenous data, where, between discrete actions, new data becomes available, or no data is available.

This paper shares a common aim with recent work [23] focusing on including endogenous data in the firing likelihoods of activities. However, our work focuses on including exogenous data to cater for contextual factors that may influence the behaviour of processes, rather than solely relying on endogenous data. To do so, we consider behaviour in time series data and introduce a temporal component to analysis. Our approach avoids reducing time series data to a static value through the application of a lossy aggregation [2], and considers the condition of having no exogenous data on firing likelihoods. Therefore, this paper investigates:

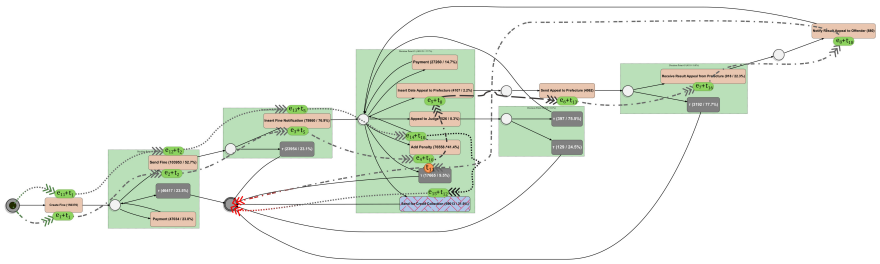
Problem. *How can exogenous influences on the stochastic behaviour of processes be derived and analysed?*

Our answer to this problem is a stochastic discovery technique which discovers weight functions for transitions in labelled Petri nets. The discovery is presented as an optimisation problem, whereby equalities are derived using diagnostic information, and are solved to form weight functions. The discovered weight functions can be used to investigate exogenous influences on processes.

The paper is organised as follows. Section 2 illustrates existing approaches; Section 3 denotes the notation used; Section 4 presents our stochastic modelling formalism; Section 5 defines our discovery method for such a formalism; Sec-

Table 1: An extract from the road fines log [7].

trace	event	event attributes or *endogenous data*				
		process activity	amount	dismissal	pay amount	timestamp
N74775	1	create fine	35.0	NIL		28.05.2005@00:00
	2	send fine				21.09.2005@00:00
	3	insert fine notification				01.10.2005@00:00
	4	add penalty	71.5			30.11.2005@00:00
	5	insert date appeal to prefecture				30.11.2005@00:00
	6	send appeal to prefecture		#		09.01.2006@00:00
	7	receive result appeal from prefecture				17.02.2006@00:00
	8	notify result appeal to offender				31.03.2006@00:00
A8690	9	create fine	36.0	NIL		08.02.2007@00:00
	10	payment			36.0	21.02.2007@00:00
A21188	11	create fine	22.0	NIL		14.07.2008@00:00
	12	send fine				10.12.2008@00:00
	13	insert fine notification				18.12.2008@00:00
	14	add penalty	44.0			16.02.2009@00:00
	15	send for credit collection				15.10.2010@00:00
⋮	⋮	⋮	⋮	⋮	⋮	⋮



Alignment for trace '`N74775`' : $\langle \binom{e_1}{t_1}, \binom{e_2}{t_2}, \binom{e_3}{t_5}, \binom{e_4}{t_{10}}, \binom{e_5}{t_8}, \binom{e_6}{t_{13}}, \binom{e_7}{t_{16}}, \binom{e_8}{t_{18}}, \binom{\gg}{t_{11}} \rangle$.

Alignment for trace '`A21188`' : $\langle \binom{e_{11}}{t_1}, \binom{e_{12}}{t_2}, \binom{e_{13}}{t_5}, \binom{e_{14}}{t_{10}}, \binom{e_{15}}{t_{12}} \rangle$.

Fig. 1: The normative model capturing the decision points (in green) and credit collection is highlighted. Alignments are captured through dashed directed lines.

tion 6 denotes how we quantify our discovered formalism; Section 7 evaluates our contributions; and Section 8 summarises our work.

2 Motivating Example

To motivate stochastic process mining and highlight existing techniques [4,23,16], a running example is introduced. The example uses a snippet of the road fines log [7], which captures the payment and appeal process for fines issued by a police force in Italy (Table 1), and a normative control-flow model [22, Sec. 12.1.3]. Also introduced are optimal alignments [1] for the snippet (Fig. 1). These are important to consider as our approach derives diagnostic information from alignments to inform the firing likelihoods for transitions. This example explores the probability of '*Sent to Credit Collection*' (abbreviated to **credit collection**) occurring. Where the answer for the trace '`A8690`' is straightforwardly 0%, as

the transition for credit collection was never enabled. However, the question is more challenging for the other traces in Table 1.

In stochastic process mining, transitions have weights and the likelihood of a transition firing is a ratio of its weight over the total weight of all enabled transitions. One approach to assigning weights to transitions is to consider the frequency of the fired transitions [4] and estimate a numerical value for a *base* weight, resulting in Fig. 2a. Where for the trace '`N74775`', the discovered net induces a likelihood for credit collection of 31.87% for all three times that it is enabled. The downside of only using base weights is that the likelihood of a transition cannot account for context surrounding the execution.

To contextualise likelihoods for a running execution, base weights can be expanded to *parameterised* weight functions. For example, many parameters on each transition can be introduced allowing for the history of actions in an execution to influence the firing likelihood of the next transition [16] (known as a SLPN-SD). Similarly, parameters can be introduced to account for the value of data attributes as of the last event in an execution [23] (known as a SLDPN). However, these approaches treat executions as discrete changes, and as time passes between actions, the firing likelihoods remain the same.

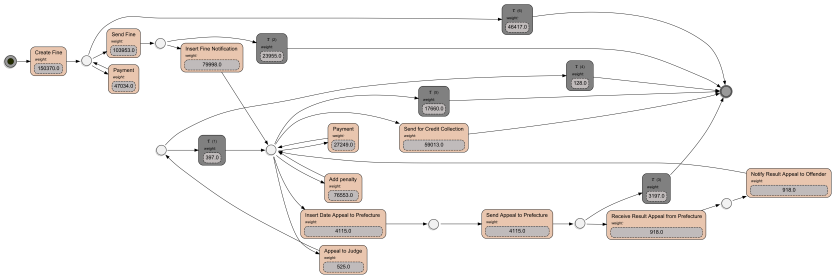
When we consider trace '`N74775`' and the discovered parameterised nets in Fig. 2b and Fig. 2c, then credit collection had firing likelihoods of 20%, 17.3%, and 17.3% or 8.7%, 10.5%, and roughly 0.01% for the respective models. Showing that the dismissal of the fine can be catered for using Fig. 2c. But, if we consider trace '`A21188`' where credit collection did occur, where the likelihoods are 8.1% and 9.2% for Fig. 2c, then the batching nature (seen in Fig. 4) of this action is not captured within parameters, as they only capture discrete changes. Moreover, temporal changes in exogenous data are not supported unless we reduce temporal series into static attributes for events, requiring domain knowledge.

This example shows existing techniques [4,16,23] can cater for endogenous factors that may influence the probability of a transition firing. However, without a transition firing, these techniques are unable to adjust the probability of transitions or account for temporal considerations in exogenous data. Thus, we explore how to consider exogenous influences on the firing likelihood of transitions, and the challenges of including exogenous data for stochastic Petri nets.

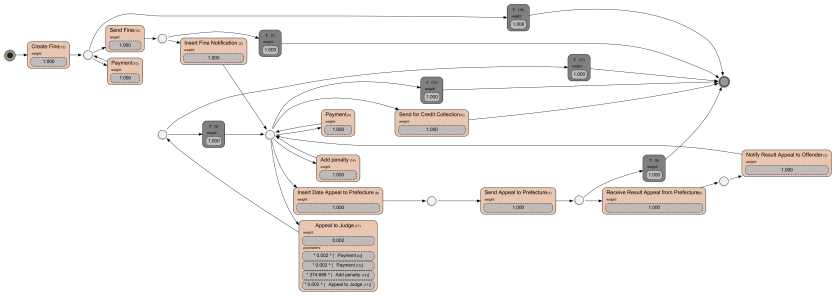
3 Preliminaries

This section introduces the notation used for event logs, exogenous data, process models and alignments.

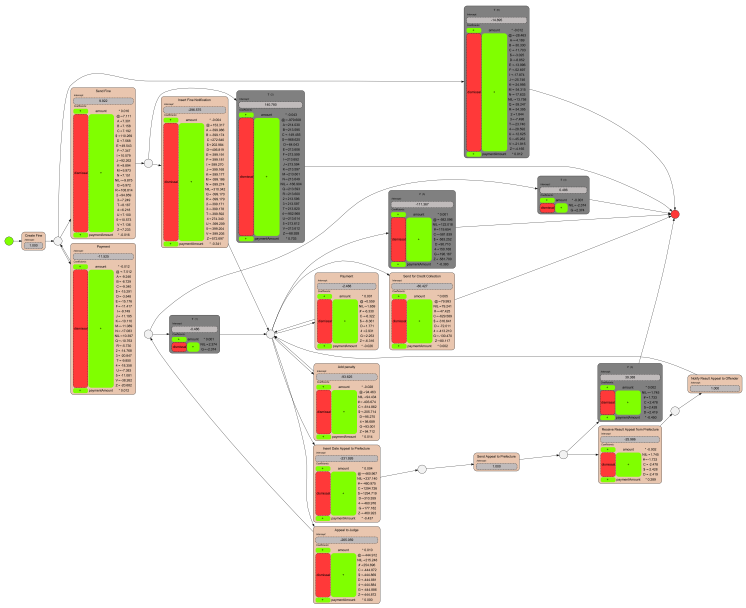
Sets. Curly brackets denote sets, e.g. $\Sigma = \{a, b, c\}$. To denote the number of members in a set Σ, $|\Sigma|$ is written. Given some set, say Σ, a multiset $M \colon \Sigma \mapsto \mathbb{N}_0$ maps the elements of Σ to the natural numbers. Square brackets denote multisets, e.g. $[\,a^2, b^1\,]$ s.t. $a, b \in \Sigma$ and all other elements of Σ are mapped to zero. To denote the infinite set of all multisets over a given set Σ, $\mathcal{M}(\Sigma)$ is used. We write $m \in M$ to access all members of a multiset with non-zero frequencies, i.e. $\{m \in \Sigma \mid M(m) > 0\}$. To state relations between a set S and a multiset



(a) Discovered stochastic labelled Petri net [4, Section 3.6] and only uses base weights.



(b) Discovered SLPN-SD [16, Definition 1].



(c) Discovered SLDPN [23] using event attributes in Table 1.

Fig. 2: Existing stochastic techniques applied to the running example. Weights and parameters are denoted below the transition label.

$$\mathsf{uf} = \langle 1^{01-01-2000}, \ldots, 31188^{14-02-2009}, 31222^{15-02-2009}, \ldots, 44884^{18-06-2013} \rangle$$

(a) Number of unpaid fines (unpaid fines).

$$\mathsf{af} = \langle 62^{01-01-2000}, \ldots, 2.007e6^{14-02-2009}, 2.008e6^{15-02-2009}, \ldots, 2.311e6^{18-06-2013} \rangle$$

(b) Total amount of unpaid fines (amount unpaid).

Fig. 3: Exo–series describing inter-case variables for the road fines log.

M, $S \subseteq M$ denotes that S is a subset of M if and only if : $\forall_{s \in S} M(s) > 0$. To denote the cardinality of a multiset M, $||M|| = \sum_{m \in M} M(m)$ is written.

Sequences. Angled brackets denote sequences, e.g. $\langle a, b, c \rangle$. To concatenate two sequences s, s', $s : s'$ is written. The length of a sequence s is denoted as $|s|$, and if $|s| = 0$ then s is empty. To access an element of a sequence s, s_i is used to access the i-th element of the sequence s.t. $1 \leq i \leq |s|$. To denote a set of possible sequences over a set Σ, Σ^* is used to denote all possible sequences, while Σ^+ is used to denote all possible non-empty sequences.

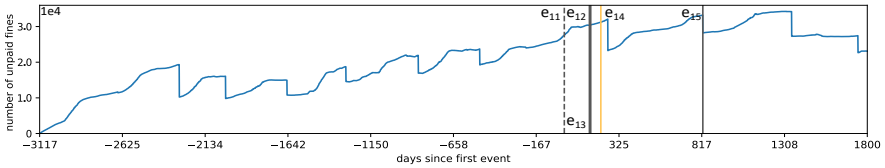
Events. An event describes the execution of a process step (activity). Each *event* consists of the activity and a timestamp of the step [28]. The time of event is accessed through the accessor time. A non-empty sequence of events is a *trace* and describes an execution of a process.

Universes. $\mathcal{U}_{ev}$,$\mathcal{U}_{time}$ denote the sets of identifiers for events, and timestamps. $\mathcal{U}_{att}$,$\mathcal{U}_{val}$ denote the sets of attributes and all possible values that these can take on. $\mathcal{U}_{ev}$,$\mathcal{U}_{time}$,$\mathcal{U}_{att}$,$\mathcal{U}_{val}$ are pairwise disjoint.

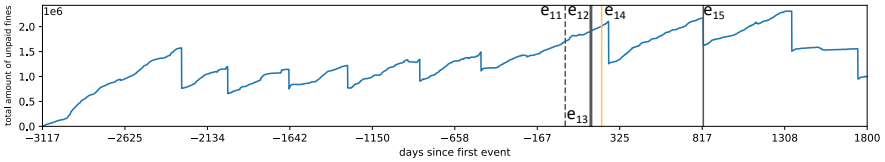
Time Series. Exogenous data will represented be as numerical discrete univariate time series, referred to as an exo–series [2]. Exo–series are sequences of measurements and timestamps, say $x \in (\mathcal{U}_{val} \times \mathcal{U}_{time})^*$, focusing on measuring a single concept. To access the value and time of a measurement, x_i and $\mathsf{time}(x_i)$ is written. $\mathcal{U}_{es} = (\mathcal{U}_{val} \times \mathcal{U}_{time})^*$ denotes the set of all possible exo–series. We call a collection of exo–series measuring the same observation an *exo–panel*.

Example 1. Consider the following exo–series in Fig. 3 for the number of unpaid fines (unpaid fines) and the total amount of unpaid fines (amount unpaid), discussed later in Section 7.1. These exo–series will represent exogenous data that is continuously observed to reflect the amount of work currently active in the information system within the running example. The question we investigate within these examples is: "*does the number of unpaid fines influence whether credit collection occurs within executions?*"

Exogenous Annotated Logs. The source of time series data will come from applying the xPM framework [2, Sec. 4], where an event log is annotated with exogenous data through several determinations [2, Sec. 4.1.], creating an *exogenous annotated log* (xlog). This paper solely focuses on the ($\mathcal{L}$) linked exogenous data [2, Sec. 4.1.2].



(a) exo-series for 'unpaid fines'.



(b) exo-series for 'amount unpaid'.

Fig. 4: The trace 'A21188' from Table 1 annotated with two exo–series. Verticals denote when events occurred, a dash denotes the first event.

Definition 1 (Exogenous Annotated Logs). *An exogenous-annotated log* $L = (\rho, \Sigma)$ *is a tuple, where* ρ *is the set of exo-panel names and* $\Sigma = \mathcal{M}(\mathcal{U}_{ev}^* \times (\rho \mapsto \mathcal{U}_{es}))$ *is a multiset of traces annotated with a mapping of exo-series.*

To consider exo–series x or a mapping of exo–series X, relative to a given timestamp r, referred to as a *truncated* exo-series, the function trun is introduced. These *truncated* exo-series have a relative timestamp denoting the time away from the truncated timestamp r, e.g.. a relative timestamp of a day, $1d$, denotes that a measurement occurred a day before r. Now, trun is formally denoted as:

$$\mathsf{trun}(X, r) = \bigcup_{X(p)=x} \{p \mapsto \mathsf{trun}(x, r)\} \tag{1}$$

$$\mathsf{trun}(\langle (v,t) \rangle : x, r) = \begin{cases} \langle (v, r - \mathsf{time}(t)) \rangle : \mathsf{trun}(x, r) \text{ if } \mathsf{time}(t) \leq r \\ \langle \rangle \text{ otherwise} \end{cases}$$

Example 2. Consider the following visualisation in Fig. 4 of trace 'A21188' from Table 1 annotated with exo–series for 'unpaid fines' (uf) and 'amount unpaid' (uf) from Example 1. To demonstrate trun, consider applying trun in the context of e_{14} (highlighted vertical in Fig. 4) yields:

$$\begin{aligned}
&\mathsf{trun}(\{\text{unpaid fines } \mapsto \mathsf{uf}, \text{amount unpaid} \mapsto \mathsf{af}\}, \mathsf{time}(e_{14})) \\
&\quad = \{\text{unpaid fines} \mapsto \mathsf{trun}(\mathsf{uf}, \mathsf{time}(e_{14})), \text{amount unpaid} \mapsto \mathsf{trun}(\mathsf{af}, \mathsf{time}(e_{14}))\} \\
&\quad = \{\text{unpaid fines} \mapsto \langle 1^{@4917d}, \ldots, 31188^{@2d}, 31222^{@1d}, 31228^{@0d} \rangle, \\
&\qquad \text{amount unpaid} \mapsto \langle 62^{@4917d}, \ldots, 2\,007\,000^{@2d}, 2\,008\,000^{@1d}, 2\,009\,000^{@0d} \rangle\}
\end{aligned}$$

Petri Nets. Petri nets are used as the modelling notation to describe modelled processes. Two such extensions, labelled Petri nets (LPN) and stochastic Petri nets, are considered in this paper.

Definition 2 (Labelled Petri Nets). *A labelled Petri net (LPN) is a tuple* $(P, T, F, A, \lambda, M_o)$*, where* P *is a set of places,* T *is a set of transitions s.t.* $P \cap T = \emptyset$*,* $F \subseteq ((P \times T) \cup (T \times P))$ *is a set of directed flows,* A *is a set of process activities,* $\lambda : T \mapsto A \cup \{\tau\}$ *is a labelling function, and* $M_o \in \mathcal{M}(P)$ *is an initial marking.*

The preset of a transition t is denoted as ${}^\bullet t = \{p \mid (p,t) \in F\}$, likewise the postset of a transition is $t^\bullet = \{p \mid (t,p) \in F\}$. A transition is enabled in a marking $M \in \mathcal{M}(P)$ if a marking includes all places with an outgoing arc to the transition, i.e. ${}^\bullet t \subseteq M$. A run of an LPN starts from the initial marking M_o and fires transition until no transitions are enabled. The firing of a transition consumes tokens from the preset and generates fresh tokens in the postset of the transition. The set of enabled transitions from a marking M is denoted as $\mathsf{enbl}(M) = \{t \in T \mid {}^\bullet t \subseteq M\}$.

Alignments [1]. An alignment γ is a sequence of moves, where a move is a combination of an event from the log (or a skip $\gg$) and a transition in the model (or a skip $\gg$). However a move in an alignment cannot consist of two skips $\gg$, i.e. $\binom{\gg}{\gg}$. A move $\binom{e}{t}$ is a synchronous move; a move $\binom{e}{\gg}$ is a log move; and a move $\binom{\gg}{t}$ is a model move. Also, given a move m, the following functions, $\mathsf{synm}, \mathsf{logm}, \mathsf{modm}$, identify the type of move using the previous cases. Alignments allow for traces to be mapped to to a run of an LPN.

Sojourn Times. Lastly, we need to consider when a model move could have occurred. To do so, the sojourn times, the waiting time between two fired transitions, are used to inform when model moves could have occurred. In particular, a histogram of the time between synchronous moves in alignments is computed. For instance, say $\gamma = \langle \binom{e}{t}, \binom{\gg}{t''}, \binom{e'}{t'} \rangle$ is an alignment, then the sojourn time of $\mathsf{soj}(t')$ is $\mathsf{time}(e') - \mathsf{time}(e)$, and always is towards the transition t'. Formally the function hist computes all sojourn times, defined as:

Definition 3 (Sojourn times). *Let* $LPN = (P, T, F, A, \lambda, M_o)$ *be a net, let* L *be a log, let* Γ *be a multiset of alignments between LPN and* L *and let* $t \in T$ *be a transition. Then,* hist *is a function that takes a transition* t *to yield a multi-set of sojourn times:*

$$\mathsf{hist}(t) = \bigcup_{\substack{\gamma \in \Gamma, \\ 1 \leq i \leq \Gamma(\gamma)}} \left[\mathsf{time}(e) - \mathsf{time}(e') \mid \forall_{\gamma' : \langle \binom{e'}{t'} \rangle : \gamma'' : \langle \binom{e}{t} \rangle : \gamma''' = \gamma} \nexists_{1 \leq j \leq |\gamma''|} \mathsf{synm}(\gamma''_j) \right]$$

4 Stochastic Petri Nets with Exogenous Dependencies

This section introduces our formalism stochastic labelled Petri nets with exogenous dependencies (Exo-SLPN), where weight functions consider exogenous data. Our modelling formalisation for stochastic process mining builds up existing stochastic labelled Petri nets from process mining [3,16,23].

Definition 4 (Stochastic Labelled Petri Nets With Exogenous Dependencies). *An Exo-SLPN is a tuple* $(P, T, F, A, \lambda, M_o, \rho, \mathsf{trans}, \omega)$*, where* $(P, T, F, A, \lambda, M_o)$ *is an LPN,* ρ *is a set of exo-panels,* $\mathsf{trans} \colon (\rho \mapsto \mathcal{U}_{es}) \mapsto (\rho \mapsto (\mathbb{R} \cup \{\bot\}))$ *transforms a mapping of exo-panels and exo-series into a mapping of exo-panels and numbers or* $\bot$*, and* $w \colon (T \times (\rho \mapsto (\mathbb{R} \cup \{\bot\}))) \mapsto \mathbb{R}$ *is a weight function.*

The control-flow semantics of an Exo-SLPN is those of the corresponding LPN. Additionally, an Exo-SLPN expresses the likelihood of firing an enabled transition as a ratio of the sum of the weights of all enabled transitions. Formally, given a marking M and a mapping of exo-series E, the probability of firing an enabled transition $t \in \mathsf{enbl}(M)$ is:

$$\mathsf{prob}(M, E, t) = \frac{w(t, \mathsf{trans}(E))}{\sum_{t' \in \mathsf{enbl}(M)} w(t', \mathsf{trans}(E))} \tag{2}$$

While Exo-SLPNs support any representation for a weight function w, we consider forms that use three weights for each transition t. That is, a *base* weight ϕ_t and, for each exo-panel p, two influencing parameters, being an *adjustment* parameter $\varphi_{t,p}$ when exogenous data is available and an *alternate* parameter $\psi_{t,p}$ when exogenous data is unavailable. These parameters quantify how the base weight ϕ_t of a transition t is influenced by exogenous data through $\varphi_{t,p}$ and $\psi_{t,p}$. Given a transition t and a mapping of transforms E, this paper considers several forms for a weight function w:

$$w(t, E) = \phi_t \cdot \prod_{E(p)=e} \begin{cases} (\varphi_{t,p})^e & \text{if } e \neq \bot \\ \psi_{t,p} & \text{otherwise} \end{cases} \qquad \textit{with } \phi_t, \varphi_{t,p}, \psi_{t,p} > 0 \tag{3}$$

$$w(t, E) = \phi_t + \sum_{E(p)=e} \begin{cases} e \cdot \varphi_{t,p} & \text{if } e \neq \bot \\ \psi_{t,p} & \text{otherwise} \end{cases} \qquad \textit{with } \phi_t, \varphi_{t,p}, \psi_{t,p} > 0 \tag{4}$$

$$w(t, E) = \phi_t + \sum_{E(p)=e} \begin{cases} e \cdot \varphi_{,p} & \text{if } e \neq \bot \\ \psi_{,p} & \text{otherwise} \end{cases} \qquad \textit{with } \phi_t, \varphi_{,p}, \psi_{,p} > 0 \tag{5}$$

Eq. (3) expresses exogenous influences could be multiplicative with the base weight of a transition, where they can have a dramatic impact on the likelihood of firing. Eq. (4) and Eq. (5) express exogenous influences could be additive with the base weight of a transition, where they can only have a positive impact on the likelihood of firing. However, Eq. (5) only introduces one set of parameters per influence (i.e. exo-panel $p : \varphi_{,p}, \psi_{,p}$), rather than individually $(\varphi_{t,p}, \psi_{t,p})$ on transitions. We explore these options to understand which one may have the best generalisation.

Similarly, Exo-SLPNs support any transformation function trans, and the methods in the remainder of this paper do not pose any restrictions on these functions. But for this paper, a default transformation for numerical time series is used. This transformation returns a weighted average of z-scores (Eq. (7)), where measurements further away from the truncation have a smaller influence on the transformed value. Our intuition is that changes that happened furthest away from the decision point will have less impact on the firing likelihood. Z-scores are used to capture the variability and volatility in exogenous data, quantifying exogenous data for firing likelihoods. However, when there is no variance within a panel, the transformation maps to zero to identify that zero change can be observed from this panel. Similarly, when the truncation results in an empty sequence, the transformation maps to $\bot$ to use of the alternate parameter.

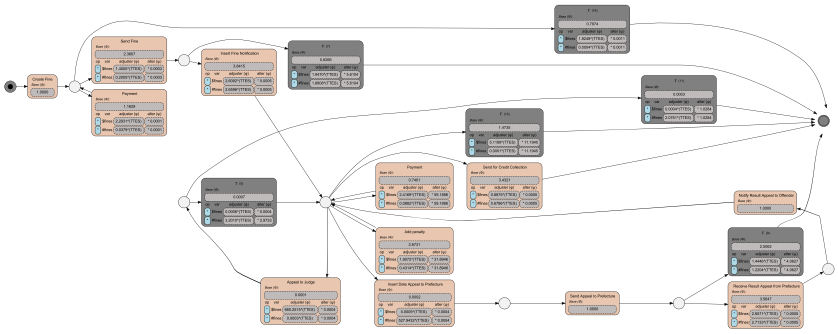


Fig. 5: Exo-SLPN discovered for the road fines log using the normative model.

Definition 5 (Default Numerical Transformation). *Let X be a mapping of panels to truncated exo-series, let $p \in \rho$ be an exo-panel, let $\mu(p)$ be the mean of p, let $\sigma(p)$ be the standard deviation of p, and let $X(p) = x$ be a truncated exo-series for p. Then,* trans' *maps a collection of truncated exo-series into numbers:*

$$\mathsf{trans}'(X)(p) = \begin{cases} \bot & \text{if } X(p) = \langle\rangle \\ 0 & \text{if } \sigma(p) = 0 \wedge X(p) \neq \langle\rangle \\ z & \text{if } \sigma(p) \neq 0 \wedge X(p) = \langle x_1, \ldots, x_n \rangle \end{cases} \tag{6}$$

$$\text{where } z = \frac{\sum_{i=1}^{n} \frac{1}{1+\mathsf{time}(x_i)} \cdot \left| \frac{x_i - \mu(p)}{\sigma(p)} \right|}{\sum_{i=1}^{n} \frac{1}{1+\mathsf{time}(x_i)}} \tag{7}$$

Each transformed-truncated-exo-series is referred to as a TTES.

Example 3. Consider the Exo-SLPN shown in Fig. 5 and trace 'A21188' annotated with exogenous data from Example 2. The goal is to understand the change in probability for credit collection based on exogenous data in the duration between e_{13} and e_{14}. In order to perform transformation, both the mean (mean) and standard deviation (std) of the panels are needed. As such, for unpaid fines these are 20928.975 and 8084.747, for amount unpaid these are 1308182.069 and 479526.257. Therefore, applying trans on the truncated exo-series presented in Example 2, yields a mapping E between panels and TTESes as follows:

$$\begin{aligned} E = \mathsf{trans}'(\{&\text{unpaid fines} \mapsto \langle 1^{@4917d}, \ldots, 31188^{@2d}, 31222^{@1d}, 31228^{@0d} \rangle, \\ &\text{amount unpaid} \mapsto \langle 62^{@4917d}, \ldots, 2\,007\,000^{@2d}, 2\,008\,000^{@1d}, 2\,009\,000^{@0d} \rangle\}) \\ &= \{\text{unpaid fines} \mapsto 0.864, \text{amount unpaid} \mapsto 0.919\}. \end{aligned}$$

Note that a shorthand is used in the following calculation for panels, '$\#F$' for 'unpaid fines' and '$\$F$' for 'amount unpaid'. Then, the probability (using Eq. (3)) of credit collection (**CC**) firing when the trace 'A21188' executed 'add penalty' (**AP**) was:

$$\mathsf{prob}(M, E, \mathbf{CC}) = \frac{w(\mathbf{CC}, E)}{\sum_{t \in \mathsf{enbl}(M)} w(t, E)}$$

$$\approx \frac{3.4321 \cdot \begin{cases} 0.6786^{E(\#F)} & \text{if } E(\#F) \neq \bot \\ 0.0005 & \text{if } E(\#F) = \bot \end{cases} \cdot \begin{cases} 0.8875^{E(\$F)} & \text{if } E(\$F) \neq \bot \\ 0.0005 & \text{if } E(\$F) = \bot \end{cases}}{\cdots + 0.7401 \cdot \begin{cases} 0.0882^{E(\#F)} & \text{if } E(\#F) \neq \bot \\ 59.188 & \text{if } E(\#F) = \bot \end{cases} \cdot \begin{cases} 2.4169^{E(\$F)} & \text{if } E(\$F) \neq \bot \\ 59.188 & \text{if } E(\$F) = \bot \end{cases} + \ldots}$$

$$\approx \frac{3.4321 \cdot 0.6786^{0.864} \cdot 0.8875^{0.919}}{3.4321 \cdot 0.6786^{0.864} \cdot 0.8875^{0.919} + 0.7401 \cdot 0.0882^{0.864} \cdot 2.4169^{0.919} + \ldots}$$

≈ 0.4635 or 46.35%.

5 Discovery

This section presents how to discover an Exo-SLPN, given an event log and an LPN, by casting the discovery as an optimisation problem. Thereby allowing the identification of exogenous influences on processes from historical data. First, we describe how to reconstruct when choices occurred in the process and extract diagnostic information about these moments, referred to as *choice data*. Then, we present the optimisation problem to discover the weights for Exo-SLPN, whereby values for parameters are obtained, i.e. a base weight ϕ, many adjustment φ parameters and many alternative ψ parameters.

5.1 Constructing Choice Data

After alignments have been obtained for traces in an xlog, the next step is to gather model choices and TTES values for these model choices. That is, for each transition fired in the model from replaying the xlog, the following is recorded: (i) the transitions that were enabled, (ii) the TTES values as they were at the time the transition fired, and (iii) the fired transition. For synchronous moves, TTES can be obtained using the event as they have a timestamp (Eq. (9)). For a log move, no transition is recorded so these moves do not introduce any choice data (Eq. (10)). For model moves, no timestamp is available but could be estimated (discussed in Section 5.2) using the function fill (Eq. (11)) to obtain a TTES.

Formally, for a trace, choice data is obtained using Def. 6. Then, choice data for an xlog is then simply the multiset union of its traces.

Definition 6 (Choice data). *Let $(\sigma, \#) \in \Sigma$ be a trace σ annotated with a mapping of exo–series $\#$ from an xlog, let LPN be a net, and let γ be an alignment between σ and LPN. Then,* choice *takes a marking and an alignment, returning a multiset of choice data:*

$$\mathsf{choice}(M, \langle\rangle) = [\,] \tag{8}$$

$$\mathsf{choice}(M, \langle \binom{e}{t} \rangle : \gamma') = [\,(\mathsf{enbl}(M), \mathsf{trans}'(\mathsf{trun}(\#, \mathsf{time}(e))), t)\,] \cup \mathsf{choice}(M \setminus {}^\bullet t \cup t^\bullet, \gamma') \tag{9}$$

$$\mathsf{choice}(M, \langle (\begin{smallmatrix} e \\ \gg \end{smallmatrix}) \rangle : \gamma') = \mathsf{choice}(M, \gamma') \tag{10}$$

$$\mathsf{choice}(M, \langle (\begin{smallmatrix} \gg \\ t \end{smallmatrix}) \rangle : \gamma') = [\,(\mathsf{enbl}(M), \mathsf{fill}(\#, \gamma, |\gamma| - |\gamma'|), t)\,] \cup \mathsf{choice}(M \setminus {}^\bullet t \cup t^\bullet, \gamma') \tag{11}$$

Example 4. Consider the snippet of the road fines log (Table 1) and the alignments in Fig. 1, where the abbreviations are used for activities. Then, the recursion of choice is explored for trace 'A21188' to demonstrate how choice data is derived from the diagnostic information stored in the alignment.

$$\begin{aligned}
&\mathsf{choice}(M_o, (\begin{smallmatrix} e_{11} \\ \mathbf{CF} \end{smallmatrix}) : \langle \ldots \rangle) = \\
&\quad [(\{\mathbf{CF}\}, \{\text{unpaid fines} \mapsto 0.551, \text{amount unpaid} \mapsto 0.529\}, \mathbf{CF})] \\
&\quad \cup \mathsf{choice}(M', (\begin{smallmatrix} e_{12} \\ \mathbf{SF} \end{smallmatrix}) : \langle \ldots \rangle) \\
&= [(\{\mathbf{CF}\}, \{\ldots\}, \mathbf{CF})] \\
&\quad \cup [(\{\mathbf{SF}, \tau_{16}, \}, \{\text{unpaid fines} \mapsto 0.797, \text{amount unpaid} \mapsto 0.789\}, \mathbf{SF})] \\
&\ldots \\
&= [(\{\mathbf{CF}\}, \{\ldots\}, \mathbf{CF}), (\{\mathbf{SF}, \tau_{16}\}, \{\ldots\}, \mathbf{SF}), (\{\mathbf{IFN}, \tau_7\}, \{\ldots\}, \mathbf{IFN}), \\
&\quad (\{\mathbf{CC}, \mathbf{AP}, \mathbf{PYb}, \mathbf{AtJ}, \tau_{13}\}, \{\ldots\}, \mathbf{AP})] \\
&\quad \cup [(\{\mathbf{CC}, \mathbf{AP}, \mathbf{PYb}, \mathbf{AtJ}, \tau_{13}\}, \{\text{unpaid fines} \mapsto 1.052, \\
&\quad \text{amount unpaid} \mapsto 1.174\}, \mathbf{CC})]
\end{aligned}$$

Thus for trace 'A21188', five triples of choice data are extracted and are combined with choice data from the other traces in the snippet as shown in Table 2.

5.2 Handling Model Moves

To obtain TTES values from model moves, two types of model moves are distinguished: moves on silent transitions, i.e. $\lambda(t) = \tau$, and moves on labelled transitions, i.e. $\lambda(t) \neq \tau$. To define the fill function, first synchronous (which have TTES values) and log moves (which do not have TTES values, as they do not represent transitions) are considered:

$$\mathsf{fill}(\#, \gamma, i) = \begin{cases} \bot & \text{if } i > |\gamma| & \text{(12a)} \\ \mathsf{trans}'(\mathsf{trun}(\#, \mathsf{time}(e))) & \text{if } \mathsf{synm}(\gamma_i) \mathit{with}\ \gamma_i = (\begin{smallmatrix} e \\ t \end{smallmatrix}) & \text{(12b)} \\ \mathsf{fill}(\#, \gamma, i+1) & \text{if } \mathsf{logm}(\gamma_i) & \text{(12c)} \\ \ldots & \end{cases}$$

Silent Model Moves A silent transition models a change in the state of a system without directly visible external consequences. As such, they have neither timestamps nor TTES. Nevertheless, the decision to fire the silent transition and not a potential competing visible transition depends on the TTES value, and as such must be assigned a TTES. As the firing of a silent transition has no corresponding action in the process (i.e. just in the system), this paper posits

that the moment they happen is the moment the next labelled transition fires:

$$\mathsf{fill}(\#,\gamma,i) = \begin{cases} \dots \\ \mathsf{fill}(\#,\gamma,i+1) & \text{if } \mathsf{modm}(\gamma_i) \wedge \lambda(t) = \tau \textit{ with } \gamma_i = (\begin{smallmatrix} \gg \\ t \end{smallmatrix}) \\ \dots \end{cases} \tag{12d}$$

Labelled Model Moves A labelled transition models the execution of a process step; a labelled model move indicates that this step was executed, but not logged, and hence there is no timestamp available to compute the TTES values. Thus, a TTES is obtained through estimation based on other moves with timestamps around the model move.

For the sake of explanation, assume an alignment with two synchronous moves and a model move in between: $\langle \dots (\begin{smallmatrix} e \\ t \end{smallmatrix}), (\begin{smallmatrix} \gg \\ t' \end{smallmatrix}), (\begin{smallmatrix} e'' \\ t'' \end{smallmatrix}) \dots \rangle$. TTES values depend on when the transition t' was fired, and in this case it was between e and e''. Time wise, this means that the time of e plus the sojourn times of t' and t'' is the time of e'' ($\mathsf{time}(e) + \mathsf{soj}(t') + \mathsf{soj}(e'') = \mathsf{time}(e'')$). Neither $\mathsf{soj}(t')$ or $\mathsf{soj}(t'')$ are known, however a distribution of sojourn times for t' and t'' can be observed using the aligned log. Fig. 6 illustrates this partially: from the moment that $(\begin{smallmatrix} e \\ t \end{smallmatrix})$ happens (vertical line on the left), the distribution of sojourn times of t' (green bars) can be projected. Furthermore, note that $\mathsf{soj}(t')$ cannot have lasted beyond $(\begin{smallmatrix} e'' \\ t'' \end{smallmatrix})$ and that some sojourns times are not appropriate (grayed out bars).

Observe that at each potential timestamp in the green region, the TTES value would be computed using the exo-measures up till the last observed exo-measure. Thus, to estimate the TTES value, the number points in the distribution of $(\begin{smallmatrix} \gg \\ t' \end{smallmatrix})$ (the green region) that occurred before to each exo-measure is used to compute a weighted average accordingly:

$$\mathsf{fill}(\#,\gamma,i) = \begin{cases} \dots \\ \{p \mapsto \begin{cases} \frac{1}{||S||} \sum_{s \in S} E(p)S(s) & \text{if } ||S|| > 0 \\ \bot & \text{otherwise} \end{cases} \mid p \in \#\} \\ \qquad\qquad \text{if } \mathsf{modm}(\gamma_i) \wedge \lambda(t) \neq \tau \end{cases} \tag{12e}$$

with $(\begin{smallmatrix} e \\ t \end{smallmatrix}), (\begin{smallmatrix} e'' \\ t'' \end{smallmatrix})$ being the most recent, next sync moves in γ

and sojourns $S = [s^{\mathsf{hist}(t')(s)} \mid s \in \mathsf{hist}(t') \wedge s \leq \mathsf{time}(e'') - \mathsf{time}(e) \wedge E(p) \neq \bot]$

and transforms of $E = \mathsf{trans}'(\mathsf{trun}(\#, \mathsf{time}(e) + s))$

The edge cases in which no synchronous move precedes or succeeds a model move are handled similarly. Notably two factors are explicitly left out: (i) the distribution over sojourn times of t'' is not leveraged, and (ii) that there may be multiple model moves in between two synchronous moves.

5.3 Solving Equalities

This section outlines how to estimate base weights ϕ, adjustments φ, and alternatives ψ parameters to discover an Exo-SLPN, by casting the problem of

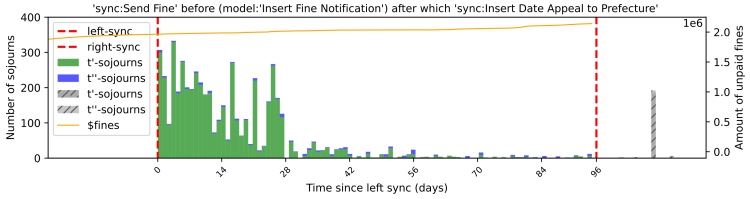


Fig. 6: Illustration of where a model move may have occurred (in green).

estimating these parameters as an linear optimisation problem. To do so, equalities are constructed from the obtained choice data. For each observed choice, the observations in the log are equated with the probability of that choice in the model (expressed in the weight parameters). A standard solver is then applied to minimise the disagreement between the sides of these equations. We considered two solution strategies, one where all the equations are solved in one-shot, and a two-shot approach where base variables are minimised first, then the adjustment and alternative parameters are minimised, both using the following equations.

Recall that the choice data C consists of triples of enabled transitions, the observed TTES values and which transitions fired. For each pair of enabled transitions (T) and TTES values (E) in Eq. (13), an equated fraction is introduced for every observed transition $o \in T$ through Eq. (14). The equated fraction in Eq. (15) states that the relative occurrence of the transition o should match the probability in the model, given the weight parameters denoted in Eq. (16):

$$\forall_{(T,E)\in\{(T,E)|(T,E,t)\in C\}} \tag{13}$$

$$\forall_{o\in\{o|(T,E,o)\in C\}} \tag{14}$$

$$\frac{|||(T,E,o)\in C]||}{|||(T,E,t')\in C]||} \tag{15}$$

$$= \frac{\phi_o \prod_{E(p)=e} \begin{cases} \varphi^e_{o,p} & \text{if } e \neq \bot \\ \psi_{o,p} & \text{otherwise} \end{cases}}{\sum_{t'\in T} \phi_{t'} \prod_{E(p)=e} \begin{cases} \varphi^e_{t',p} & \text{if } e \neq \bot \\ \psi_{t',p} & \text{otherwise} \end{cases}} \tag{16}$$

If all enabled transitions have been observed at least once when constructing equalities, then one of the equations (an observed transition o from Eq. (14)) for every (T, E) (from Eq. (13)) can be removed without loss of information.

Example 5. For instance, the equalities corresponding to our running example are shown in Table 2. By passing these equalities to a solver, the parameters ϕ, φ, ψ for each transition are assigned a value which minimally reduces the difference between these equalities.

Table 2: The constructed equalities for Example 5. Panel names have been abbreviated. Red equalities are not necessary for the solver.

enabled (T)	TTESes (E)	fired (t)	equality
$\{\mathbf{CF}\}$	$\{\dots\}$	$[\mathbf{CF}]$	$\frac{1}{1} = \frac{\phi_{\mathbf{CF}} \cdot \varphi^{0.282}_{\mathbf{CF},\#\mathrm{F}} \cdot \varphi^{0.357}_{\mathbf{CF},\$\mathrm{F}}}{\phi_{\mathbf{CF}} \cdot \varphi^{0.282}_{\mathbf{CF},\#\mathrm{F}} \cdot \varphi^{0.357}_{\mathbf{CF},\$\mathrm{F}}}$
$\{\mathbf{CF}\}$	$\{\dots\}$	$[\mathbf{CF}]$	$\frac{1}{1} = \frac{\phi_{\mathbf{CF}} \cdot \varphi^{0.530}_{\mathbf{CF},\#\mathrm{F}} \cdot \varphi^{0.552}_{\mathbf{CF},\$\mathrm{F}}}{\phi_{\mathbf{CF}} \cdot \varphi^{0.530}_{\mathbf{CF},\#\mathrm{F}} \cdot \varphi^{0.552}_{\mathbf{CF},\$\mathrm{F}}}$
$\{\mathbf{CF}\}$	$\{\dots\}$	$[\mathbf{CF}]$	$\frac{1}{1} = \frac{\phi_{\mathbf{CF}} \cdot \varphi^{0.633}_{\mathbf{CF},\#\mathrm{F}} \cdot \varphi^{0.598}_{\mathbf{CF},\$\mathrm{F}}}{\phi_{\mathbf{CF}} \cdot \varphi^{0.633}_{\mathbf{CF},\#\mathrm{F}} \cdot \varphi^{0.598}_{\mathbf{CF},\$\mathrm{F}}}$
...			
$\{\mathbf{PYa}, \mathbf{SF}, \tau_{16}\}$	$\{\dots\}$	$[\mathbf{PYa}]$	$\frac{1}{1} = \frac{\phi_{\mathbf{PYa}} \cdot \varphi^{0.803}_{\mathbf{PYa},\#\mathrm{F}} \cdot \varphi^{0.803}_{\mathbf{PYa},\$\mathrm{F}}}{\phi_{\mathbf{PYa}} \cdot \varphi^{0.803}_{\mathbf{PYa},\#\mathrm{F}} \cdot \varphi^{0.803}_{\mathbf{PYa},\$\mathrm{F}} + \dots}$
$\{\mathbf{PYa}, \mathbf{SF}, \tau_{16}\}$	$\{\dots\}$	$[\tau_{16}]$	$\frac{1}{1} = \frac{\phi_{\tau_{16}} \cdot \psi_{\tau_{16},\#\mathrm{F}} \cdot \psi_{\tau_{16},\$\mathrm{F}}}{\phi_{\tau_{16}} \cdot \psi_{\tau_{16},\#\mathrm{F}} \cdot \psi_{\tau_{16},\$\mathrm{F}} + \dots}$
$\{\mathbf{PYa}, \mathbf{SF}, \tau_{16}\}$	$\{\dots\}$	$[\mathbf{SF}]$	$\frac{1}{1} = \frac{\phi_{\mathbf{SF}} \cdot \varphi^{0.439}_{\mathbf{SF},\#\mathrm{F}} \cdot \varphi^{0.380}_{\mathbf{SF},\$\mathrm{F}}}{\phi_{\mathbf{SF}} \cdot \varphi^{0.439}_{\mathbf{SF},\#\mathrm{F}} \cdot \varphi^{0.380}_{\mathbf{SF},\$\mathrm{F}} + \dots}$
$\{\mathbf{PYa}, \mathbf{SF}, \tau_{16}\}$	$\{\dots\}$	$[\mathbf{SF}]$	$\frac{1}{1} = \frac{\phi_{\mathbf{SF}} \cdot \varphi^{0.789}_{\mathbf{SF},\#\mathrm{F}} \cdot \varphi^{0.798}_{\mathbf{SF},\$\mathrm{F}}}{\phi_{\mathbf{SF}} \cdot \varphi^{0.789}_{\mathbf{SF},\#\mathrm{F}} \cdot \varphi^{0.798}_{\mathbf{SF},\$\mathrm{F}} + \dots}$
$\{\mathbf{CC}, \mathbf{PYb}, \mathbf{IAP}, \tau_{13}, \mathbf{AP}, \mathbf{AtJ}\}$	$\{\dots\}$	$[\mathbf{AP}, \mathbf{IAP}]$	$\frac{1}{2} = \frac{\phi_{\mathbf{AP}} \cdot \varphi^{0.310}_{\mathbf{AP},\#\mathrm{F}} \cdot \varphi^{0.313}_{\mathbf{AP},\$\mathrm{F}}}{\phi_{\mathbf{AP}} \cdot \varphi^{0.310}_{\mathbf{AP},\#\mathrm{F}} \cdot \varphi^{0.313}_{\mathbf{AP},\$\mathrm{F}} + \dots}$
			$\frac{1}{2} = \frac{\phi_{\mathbf{IAP}} \cdot \varphi^{0.310}_{\mathbf{IAP},\#\mathrm{F}} \cdot \varphi^{0.313}_{\mathbf{IAP},\$\mathrm{F}}}{\phi_{\mathbf{IAP}} \cdot \varphi^{0.310}_{\mathbf{IAP},\#\mathrm{F}} \cdot \varphi^{0.313}_{\mathbf{IAP},\$\mathrm{F}} + \dots}$
$\{\mathbf{CC}, \mathbf{PYb}, \mathbf{IAP}, \tau_{13}, \mathbf{AP}, \mathbf{AtJ}\}$	$\{\dots\}$	$[\mathbf{AP}]$	$\frac{1}{1} = \frac{\phi_{\mathbf{AP}} \cdot \varphi^{0.919}_{\mathbf{AP},\#\mathrm{F}} \cdot \varphi^{0.865}_{\mathbf{AP},\$\mathrm{F}}}{\phi_{\mathbf{AP}} \cdot \varphi^{0.919}_{\mathbf{AP},\#\mathrm{F}} \cdot \varphi^{0.865}_{\mathbf{AP},\$\mathrm{F}} + \dots}$

6 Conformance Checking

This section outlines how we perform conformance checking on an Exo-SLPN, where the data-aware unit Earth mover's distance [23] (duEMSC) is adapted for our purposes. Data-aware unit Earth movers extends the unit Earth-movers [18] by adding the data perspective as an additional dimension when comparing and calculating probability distributions. The measurement of data-aware unit Earth movers is expressed as follows [23]:

$$\mathrm{duEMSC}(L, M) = 1 - \sum_{D \in L_\Delta} \sum_{A \in L_\Sigma} \max(p_L(A \wedge D) - (p_M(A|D) \cdot p_L(D)), 0)$$

Where L, M denote an event log and a model, and L_Δ, L_Σ are distributions of data sequences and activity sequences observed in the log. In order to apply duEMSC to a given log L and Exo-SLPN M, we need to define data distribution L_Δ and the joint probabilities p_L, p_M for a given activity and data sequence, but more so in our case a sequence of TTES mappings. The first step is to consider our representation of xlog (Def. 1) as a multiset of pairs for activities and data, stepwise walking traces and generating TTESes for each step.

$$L = (\rho, \Sigma) = \bigcup_{(\sigma,E) \in \Sigma} \bigcup_{k=1}^{\Sigma((\sigma,E))} \left[\left(\sigma, \bigcup_{1 \le i \le |\sigma|} \langle \mathsf{trans}'(\mathsf{trun}(E, \mathsf{time}(\sigma_i)))\rangle\right)\right] \quad (17)$$

Where Σ is a multiset of traces annotated with a mapping of exo–series E (Def. 1). Using this form the distributions for sequences over activities A and TTES $\mathcal{E}$ are straightforwardly derived from the multiset representation. Noting that TTES mappings E are continuous, requiring discretisation using a precision reduction factor. Now the probabilities from the log L, given a sequence of

activities A and a sequence of TTES mappings $\mathcal{E}$ can be expressed as:

$$p_L(A \wedge \mathcal{E}) = \frac{L((A,\mathcal{E}))}{|L|} \quad p_L(\mathcal{E}) = \frac{|[\mathcal{E}^{L((A,\mathcal{E}))} \mid (A,\mathcal{E}) \in L]|}{|L|} \tag{18}$$

The last step to consider is the probabilities for all paths T through the Exo-SLPN M that induce the activity sequence A under the assumption of the sequence of mapped TTESes $\mathcal{E}$. Note that the sequence $\mathcal{E}$ must be as long as a path $t \in T$, as such padding is applied to ensure that silent transitions use the next following mapping if available, or the last preceding mapping. Now, the model probability for a given activity sequence A under the assumption of $\mathcal{E}$, is the sum probability of all paths T that induce A, expressed as:

$$p_M(A|\mathcal{E}) = \sum_{\langle t_1,\ldots,t_n\rangle \in T} \prod_{i=1}^{n} \frac{w(t_i,\mathcal{E}_i)}{\sum_{t' \in \mathsf{enbl}(M_i)} w(t',\mathcal{E}_i)} \tag{19}$$

Where M_i is the deterministic marking before firing a transition, e.g. $M_1 = M_o$ and $M_o \xrightarrow{t_1} M' \rightarrow \ldots \xrightarrow{t_{i-1}} M_i$. Now, duEMSC can be denoted using our definition of an xlog L and Exo-SLPN M, expressed as:

$$\text{duEMSC}(L,M) = 1 - \sum_{\mathcal{E} \in L_\Delta} \sum_{A \in L_\Sigma} \max(p_L(A \wedge \mathcal{E}) - (p_M(A|\mathcal{E}) \cdot p_L(\mathcal{E})),\, 0) \tag{20}$$

7 Evaluation

This section compares our Exo-SLPN approach against several existing techniques across several publicly available event logs. The proposed technique has been implemented within the ProM framework[4], and evaluation data has been made publicly available to the best of our ability and inline with ethics[5].

7.1 Event logs and Exogenous Data

This evaluation uses three event logs, from different domains, paired with exogenous data. The first event log comes from the MIMIC III dataset [12] which contains 'MICU' ward admissions focusing on the first 48 hours of admission and follows the preparation outlined in [2]. For this log, blood pressure measurements collected from nurse observations are used for exogenous data. These measurements were selected as they may inform what medical interventions should or should not occur with an admission.

The second log comes from a smart factory [21], where only the 'WF_101' process is considered and the start events within these executions (as these are associated IoT sensors). This log was paired with IoT sensors used in the factory as exogenous data, where the positioning of a crane and the motor speed of the

[4] Found in the ExogenousData Package: github.com/promworkbench/ExogenousData.
[5] Evaluation data can be found in: github.com/adambanham/exo-slpn-testing

Table 3: Descriptive qualities of event logs used in our evaluation.

log	#traces	#events	#acts	exo-data	μ	σ	#datapoints
MIMIC [12]	4482	28524	13	blood pressure	8	106	976367
S.Factory [21]	34	418	12	crane jib x	139	128	13038
				crane jib y	43	46	13038
				motorspeed	-21	274	1602128
R. Fines [7]	150370	561470	11	unpaid fines	20928	8084	4829
				amount unpaid	1308182	479526	4898

crane are used. These sensors were selected as their readings may influence when human intervention occurs or trigger further actions.

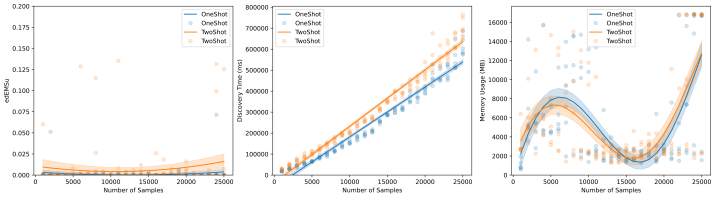
The third log used is the road fines event log [7]. For this log, inter-case variables from the log were used as exogenous data, i.e. the total number of unpaid fines and the amount of unpaid fines seen in the event log. These inter-case time series were created using data points from all events, and then aggregated into daily observations. These variables are used to understand and answer our running example's question Example 1. All three logs are described in terms of control flow, size and exogenous data in Table 3. Our testing was performed using an AMD Ryzen 9 3900XT with 64GB ram, but we saw excessive amounts of system RAM being required (upwards of 256GB) to compute conformance checking, as such both the MIMIC III and road fines log were sampled[6].

7.2 Log Completeness

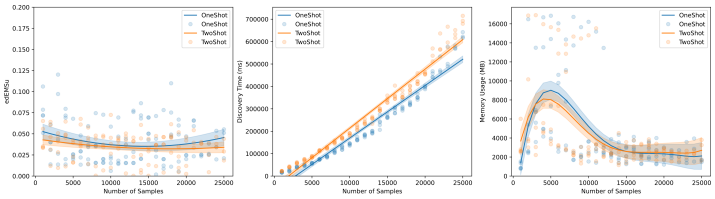
To show that our approach can identify the stochastic nature of a process, we considered how the model quality changes as we apply our approach to progressively more complete samples of a given log. The intuition being that with a more complete understanding of the original log, the discovered stochastic nature should better reflect the original or at least not degrade. We created 25 sample logs of the road fines log, where the n-th sample consists of $1000 \cdot n$ traces, and each progressive larger sample contains all samples from previous sample, i.e. given samples s_i, s_{i+1}, s_{i+j} then $s_i \subset s_{i+1}$ and $(s_i \cup s_{i+1}) \subset s_{i+j}$ where $i \geq 1$ and $j > 1$. Note that our approach considers the temporal nature of each trace, so traces with the same control-flow will have a minor temporal difference and these introduce/expand the choice data used for discovery (Section 5.1). Furthermore, as the sample gets larger, the number of sojourns considered for model moves (Section 5.2) will be increased and will affect the runtime of discovery.

For each sample, we discovered an Exo-SLPN (recording runtime and memory usage) and then using the complete log, we compute duEMSC to quantify the quality of the discovered Exo-SLPN. We sampled the road fines log down to a total of 25,000 traces and used this reduced log as the 'complete log' for testing to avoid excessive system requirements. For a control-flow model the normative model (Fig. 1) was used. Then, our testing was conducted several times for both

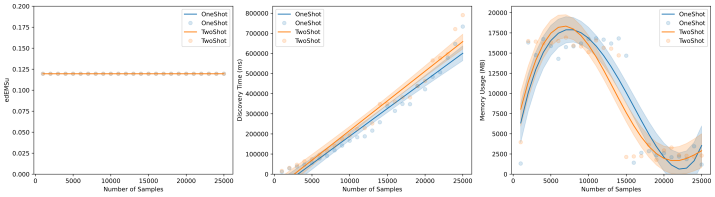
[6] See github.com/adambanham/exo-slpn-testing.



(a) Sampling results when using Eq. (3).



(b) Sampling results when using Eq. (4).



(c) Sampling results when using Eq. (5)

Fig. 7: Results from sampling testing using the road-fines log. The line of best fit and 95% confidence levels are shown to evaluate the trends across runs.

one-shot and two-shot optimisation (Section 5.3). The results of the sampling testing across the alternative weight functions are shown in Fig. 7.

The testing on the multiplicative form (Eq. (3)) in Fig. 7a shows that the equation may overly focus on the exogenous influences and struggle to find the balance between endogenous and exogenous factors. As demonstrated by duEMSC measurements being on average below 0.01 across the sampling with some minor spikes. This struggle may be due to the exogenous influence being modelled as a power of ψ, but interestingly higher quality models were produced by the two-shot approach. Next, the testing for the additive form (Eq. (4)) in Fig. 7b shows that the equation has a higher baseline than Eq. (3) when capturing the stochastic nature, likely due to the exogenous influence being less pronounced. However, Eq. (3) produced a higher quality model than Eq. (4) highest, albeit rather inconsistently, but a large amount of variance can be seen for both equations. The opposite trend occurs for the global additive form (Eq. (5)), which reports no variance at all and a constant reading of 0.119 regardless of sample size. We posit that using Eq. (5) means that exogenous influences are treated as constants and cancel each other out when competition occurs between transitions, which in turn removes all difficulty for optimisation.

Interestingly, the pairwise comparison of two-shot and one-shot optimisation, showed that two-shot has the potential to produce higher quality models, but

the one-shot approach was more consistently better in Fig. 7a and Fig. 7b. Unsurprisingly, the discovery time is linear with the size of the log due to the nature of sojourn calculations, and memory usage was similar between all approaches. Furthermore, the additional runtime needed to perform the two-shot optimisation approach is minor in comparison to the possible quality increase, as such we adopted the two-shot optimisation in the following evaluation.

7.3 Model Quality

To compare our approach against existing ones, we discovered a variety of stochastic extensions of Petri nets and considered how close these nets represent the stochastic nature of a log. As existing techniques and the proposed approach require a control-flow model to discover a stochastic Petri net, the following discovery techniques were used: the inductive miner [14] (IM_f), the directly flows miner [17] (DFM), and the POWL miner [13] (IM_{po}) using the default settings. Additionally the normative model [22, Sec 12.1.3] for the road fines log was included, which contains repeated activities which inductive miners cannot find. These techniques were selected as they discover process trees which ensure that the converted Petri net is sound and alignments can be computed. To quantify the quality of a stochastic Petri net, the unit earth movers distance [18] or the data-aware variant [23] were used (see Section 6). For these measures, a measurement closer to 1.0 reports that the model perfectly describes the stochastic nature of the log. Note that directly comparing between control flow and data aware stochastic conformance measures is discouraged.

The procedure for discovering and quantifying a model consisted of: (i) discovering a control-flow model with the original log, (ii) sampling the original log with replacement, (iii) discovering stochastic weights using the sampled log and control-flow model, and (iv) measure the discovered stochastic model using the original log. The same sampled log is used across techniques. As noted in Section 7.2, the variability of outcomes produced from our approach is high, as such we evaluated them five times and reported on the highest quality outcome.

Results. Table 4 shows the testing results for existing and proposed approaches (far-right columns). Neither control-flow techniques reported 1.0 for any of the used logs, highlighting the need to consider factors outside the control-flow of processes. However, Sign. Deps. [16] was unable to discover a model within 24 hours for the mimic log as it was solving over 50,000 parameters. Surprisingly, neither Alig. Est. [4] or Sign. Deps. reported to capture more than 50% of the stochastic nature of the road fines log using the normative model. All techniques favoured using a more literal and less structured model, i.e. DFM, for the road fines log. Notably, all measurements for the IM_f model and the S. Factory log reported 0.000 due to the model being unable to replay traces.

Moving to the data-aware setting, across the mimic and factory log, our approach outperforms the existing Data Deps. [23] technique in four out of five cases. Showing that Exo-SLPNs can be competitive with other stochastic nets in unstructured processes. However, Data Deps. outperformed in the more structured case, i.e., the road fines log where there is a clear connection

Table 4: Conformance for discovered models with our method highlighted. Highest control-flow scores are bolded. Highest data-aware scores are underlined.

	Disc.	Alig. Est.[4]	Sign. Deps.[16]	Data Deps.[23]	Eq. (3)	Eq. (4)	Eq. (5)
		Unit Earth Movers [18]		Data-Aware Unit Earth Movers [23]			
MIMIC	IM_f	**0.0546**	t/o	<u>0.0020</u>	0.0012	0.0001	0.0001
	IM_{po}	**0.0031**	t/o	0.0006	<u>0.0187</u>	0.0012	0.0000
	DFM	**0.1713**	t/o	0.0064	<u>0.0147</u>	0.0063	0.0067
S. Factory	IM_f	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	IM_{po}	**0.1205**	0.0000	0.0001	<u>0.1176</u>	0.1149	0.0592
	DFM	**0.1282**	0.1258	0.0006	<u>0.1176</u>	0.1150	0.0295
R. Fines	normative	0.2573	**0.2933**	<u>0.1479</u>	0.0253	0.0540	0.1179
	IM_f	0.1003	**0.3865**	<u>0.1654</u>	0.0701	0.0760	0.0413
	IM_{po}	0.2201	**0.3955**	<u>0.3326</u>	0.0437	0.0101	0.0100
	DFM	**0.8178**	**0.8178**	<u>0.5100</u>	0.4266	0.3779	0.2148

between event attributes and outcomes. Comparing the different weight functions, Eq. (3) produced the highest quality model in seven out of nine cases. These results may indicate that the multiplicative form has a greater capacity to capture the stochastic natures of processes, albeit with greater complexity for optimisation.

8 Conclusion

Understanding how likely are individual actions in processes allows businesses to be flexible and adapt to contextual changes. Thus, techniques providing insights into these likelihoods and how systems have acted in the past are important. As without these insights, ad-hoc changes to make actions more likely may be unsuccessful or harmful to the performance of the process. To this end, we explored how the temporal dimension and contextual factors (i.e. exogenous data) surrounding executions of processes can be included within the analysis of firing likelihoods for actions. Extending the state-of-art techniques and evaluating our approach to understand their capacity to capture firing likelihoods.

The contribution of this paper was a novel means to study whether exogenous data influences the behaviour of processes. A new process modelling formalism Exo-SLPN was introduced to enable the quantification of firing likelihoods based on exogenous data. Then, a discovery technique was developed for Exo-SLPNs, which caters for imperfect alignments between the traces and the control-flow model. Lastly, our evaluation of Exo-SLPNs showed our approach could outperform state-of-art techniques in less structured processes.

In future work, we hope to explore how to visualise complex parameterised stochastic Petri nets within the context of a given event log. Exploring alternatives weight forms (Eq. (3)) and transformations (Def. 5) could advance the expressiveness of Exo-SLPN and allow non-numerical signals to be investigated. Also, considering more scalable and faster computations for conformance checking of data-aware stochastic Petri nets would benefit future research.

Acknowledgments. Adam Banham's work was funded through an Australian Government Research Training Program Scholarship, and a QUT, Centre for Data Science Scholarship. He was also supported by RWTH Aachen, through an Advanced Research Opportunities Program Scholarship. Yannis Bertrand's work was supported by the Flemish Fund for Scientific Research (FWO) with grant number G0B6922N.

References

1. Adriansyah, A.: Aligning observed and modeled behavior. Ph.D. thesis, Mathematics and Computer Science, Eindhoven, The Netherlands (2014)
2. Banham, A., Leemans, S.J.J., Wynn, M.T., Andrews, R., Laupland, K.B., Shinners, L.: xPM: Enhancing exogenous data visibility. Artif. Intell. Medicine (2022)
3. Burke, A.: Process mining with labelled stochastic nets. Ph.D. thesis, Information Systems, Brisbane, Australia (2 2024)
4. Burke, A., Leemans, S.J.J., Wynn, M.T.: Stochastic process discovery by weight estimation. In: ICPM Workshops. LNBIP (2020)
5. Camargo, M., Dumas, M., González, O.: Automated discovery of business process simulation models from event logs. Decis. Support Syst. (2020)
6. Carmona, J., van Dongen, B.F., Solti, A., Weidlich, M.: Conformance Checking - Relating Processes and Models (2018)
7. de Leoni, M., Mannhardt, F.: Road traffic fine management process (2015)
8. Dumas, M., La Rosa, M., Mendling, J., Reijers, H.A.: Fundamentals of Business Process Management, Second Edition (2018)
9. Gal, A.: Everything there is to know about stochastically known logs. In: ICPM (2023)
10. Günther, C.W., Rinderle-Ma, S., Reichert, M., van der Aalst, W.M.P., Recker, J.: Using process mining to learn from process changes in evolutionary systems. Int. J. Bus. Process. Integr. Manag. (1) (2008)
11. Janiesch, C., Koschmider, A., Mecella, M., et al.: The internet of things meets business process management: A manifesto. IEEE Systems, Man, and Cybernetics Magazine (4) (2020)
12. Johnson, A.E., Pollard, T.J., Shen, L., wei H. Lehman, L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: MIMIC-III, a freely accessible critical care database. Scientific Data (1) (2016)
13. Kourani, H., van Zelst, S.J.: POWL: partially ordered workflow language. In: BPM. LNCS (2023)
14. Leemans, S.J.J.: Robust Process Mining with Guarantees - Process Discovery, Conformance Checking and Enhancement. LNBIP (2022)
15. Leemans, S.J.J., Maggi, F.M., Montali, M.: Enjoy the silence: Analysis of stochastic petri nets with silent transitions. Inf. Syst. **124**, 102383 (2024)
16. Leemans, S.J.J., Mannel, L.L., Sidorova, N.: Significant stochastic dependencies in process models. Inf. Syst. (2023)
17. Leemans, S.J.J., Poppe, E., Wynn, M.T.: Directly follows-based process mining: Exploration & a case study. In: ICPM (2019)
18. Leemans, S.J.J., van der Aalst, W.M.P., Brockhoff, T., Polyvyanyy, A.: Stochastic process mining: Earth movers' stochastic conformance. Inf. Syst. (2021)
19. Maggi, F.M., Montali, M., Peñaloza, R.: Probabilistic conformance checking based on declarative process models. In: CAiSE Forum. LNBIP, Springer (2020)
20. Maggi, F.M., Montali, M., Peñaloza, R., Alman, A.: Extending temporal business constraints with uncertainty. In: BPM. LNCS, Springer (2020)
21. Malburg, L., Grüger, J., Bergmann, R.: An iot-enriched event log for process mining in smart factories. Zendo (2023)
22. Mannhardt, F.: Multi-perspective process mining. Ph.D. thesis, Mathematics and Computer Science, Eindhoven, The Netherlands (2018)
23. Mannhardt, F., Leemans, S.J.J., Schwanen, C.T., de Leoni, M.: Modelling data-aware stochastic processes - discovery and conformance checking. In: Petri Nets. LNCS (2023)

24. Rinderle, S., Reichert, M., Dadam, P.: Correctness criteria for dynamic changes in workflow systems - a survey. Data Knowl. Eng. (1) (2004)
25. Rogge-Solti, A., van der Aalst, W.M.P., Weske, M.: Discovering stochastic petri nets with arbitrary delay distributions from event logs. In: Business Process Management Workshops. LNBIP (2013)
26. Scheibel, B., Rinderle-Ma, S.: Decision mining with time series data based on automatic feature generation. In: CAiSE. LNCS (2022)
27. Senderovich, A., Shleyfman, A., Weidlich, M., Gal, A., Mandelbaum, A.: To aggregate or to eliminate? optimal model simplification for improved process performance prediction. Inf. Syst. (2018)
28. van der Aalst, W.M.P.: Process Mining - Data Science in Action, Second Edition (2016)
29. vom Brocke, J., Baier, M., Schmiedel, T., Stelzl, K., Röglinger, M., Wehking, C.: Context-aware business process management. Bus. Inf. Syst. Eng. (5) (2021)
30. Yang, J., Ouyang, C., Dik, G., Corry, P., ter Hofstede, A.H.M.: Crop harvest forecast via agronomy-informed process modelling and predictive monitoring. In: CAiSE. LNCS (2022)