Fixed-delay Events in Generalized Semi-Markov Processes Revisited[⋆]

Tomáš Brázdil, Jan Krčál, Jan Křetínský[⋆⋆], and Vojtěch Řehák

Faculty of Informatics, Masaryk University, Brno, Czech Republic
{brazdil, krcal, jan.kretinsky, rehak}@fi.muni.cz

Abstract. We study long run average behavior of generalized semi-Markov processes with both fixed-delay events as well as variable-delay events. We show that allowing two fixed-delay events and one variable-delay event may cause an unstable behavior of a GSMP. In particular, we show that a frequency of a given state may not be defined for almost all runs (or more generally, an invariant measure may not exist). We use this observation to disprove several results from literature. Next we study GSMP with at most one fixed-delay event combined with an arbitrary number of variable-delay events. We prove that such a GSMP always possesses an invariant measure which means that the frequencies of states are always well defined and we provide algorithms for approximation of these frequencies. Additionally, we show that the positive results remain valid even if we allow an arbitrary number of reasonably restricted fixed-delay events.

1 Introduction

Generalized semi-Markov processes (GSMP), introduced by Matthes in [23], are a standard model for discrete-event stochastic systems. Such a system operates in continuous time and reacts, by changing its state, to occurrences of events. Each event is assigned a random delay after which it occurs; state transitions may be randomized as well. Whenever the system reacts to an event, new events may be scheduled and pending events may be discarded. To get some intuition, imagine a simple communication model in which a server sends messages to several clients asking them to reply. The reaction of each client may be randomly delayed, e.g., due to latency of communication links. Whenever a reply comes from a client, the server changes its state (e.g., by updating its database of alive clients or by sending another message to the client) and then waits for the rest of the replies. Such a model is usually extended by allowing the server to time-out and to take an appropriate action, e.g., demand replies from the remaining clients in a more urgent way. The time-out can be seen as another event which has a fixed delay.

More formally, a GSMP consists of a set S of states and a set $\mathcal{E}$ of events. Each state s is assigned a set $\mathbf{E}(s)$ of events *scheduled* in s. Intuitively, each event in $\mathbf{E}(s)$ is assigned a positive real number representing the amount of time which elapses before

[⋆] The authors are supported by the Institute for Theoretical Computer Science, project No. 1M0545, the Czech Science Foundation, grant No. P202/10/1469 (T. Brázdil, V. Řehák) and No. 102/09/H042 (J. Krčál), and Brno PhD Talent Financial Aid (J. Křetínský).

[⋆⋆] On leave at TU München, Boltzmannstr. 3, Garching, Germany.

the event occurs. Note that several events may occur at the same time. Once a set of events $E \subseteq \mathbf{E}(s)$ occurs, the system makes a *transition* to a new state s'. The state s' is randomly chosen according to a fixed distribution which depends only on the state s and the set E. In s', the *old* events of $\mathbf{E}(s) \smallsetminus \mathbf{E}(s')$ are discarded, each *inherited* event of $(\mathbf{E}(s') \cap \mathbf{E}(s)) \smallsetminus E$ remains scheduled to the same point in the future, and each *new* event of $(\mathbf{E}(s') \smallsetminus \mathbf{E}(s)) \cup (\mathbf{E}(s') \cap E)$ is newly scheduled according to its given probability distribution.

In order to deal with GSMP in a rigorous way, one has to impose some restrictions on the distributions of delays. Standard mathematical literature, such as [15, 16], usually considers GSMP with continuously distributed delays. This is certainly a limitation, as some systems with fixed time delays (such as time-outs or processor ticks) cannot be faithfully modeled using only continuously distributed delays. We show some examples where fixed delays exhibit qualitatively different behavior than any continuously distributed approximation. In this paper we consider the following two types of events:

- *variable-delay*: the delay of the event is randomly distributed according to a probability density function which is continuous and positive either on a bounded interval $[\ell, u]$ or on an unbounded interval $[\ell, \infty)$;
- *fixed-delay*: the delay is set to a fixed value with probability one.

The desired behavior of systems modeled using GSMP can be specified by various means. One is often interested in long-run behavior such as mean response time, frequency of errors, etc. (see, e.g., [1]). For example, in the above communication model, one may be interested in average response time of clients or in average time in which all clients eventually reply. Several model independent formalisms have been devised for expressing such properties of continuous time systems. For example, a well known temporal logic CSL contains a steady state operator expressing frequency of states satisfying a given subformula. In [9], we proposed to specify long-run behavior of a continuous-time process using a timed automaton which observes runs of the process, and measure the frequency of locations of the automaton.

In this paper we consider a standard performance measure, the frequency of states of the GSMP. To be more specific, let us fix a state $\mathring{s} \in S$. We define a random variable $\mathbf{d}$ which to every run assigns the (discrete) frequency of visits to $\mathring{s}$ on the run, i.e. the ratio of the number of transitions entering $\mathring{s}$ to the number of all transitions. We also define a random variable $\mathbf{c}$ which gives timed frequency of $\mathring{s}$, i.e. the ratio of the amount of time spent in $\mathring{s}$ to the amount of time spent in all states. Technically, both variables $\mathbf{d}$ and $\mathbf{c}$ are defined as limits of the corresponding ratios on prefixes of the run that are prolonged ad infinitum. Note that the limits may not be defined for some runs. For example, consider a run which alternates between $\mathring{s}$ and another state s; it spends 2 time unit in $\mathring{s}$, then 4 in s, then 8 in $\mathring{s}$, then 16 in s, etc. Such a run does not have a limit ratio between time spent in $\mathring{s}$ and in s. We say that $\mathbf{d}$ (or $\mathbf{c}$) is well-defined for a run if the limit ratios exist for this run. Our goal is to characterize stable systems that have the variables $\mathbf{d}$ and $\mathbf{c}$ well-defined for almost all runs, and to analyze the probability distributions of $\mathbf{d}$ and $\mathbf{c}$ on these stable systems.

As a working example of GSMP with fixed-delay events, we present a simplified protocol for time synchronization. Using the variable $\mathbf{c}$, we show how to measure reliability of the protocol. Via message exchange, the protocol sets and keeps a client clock

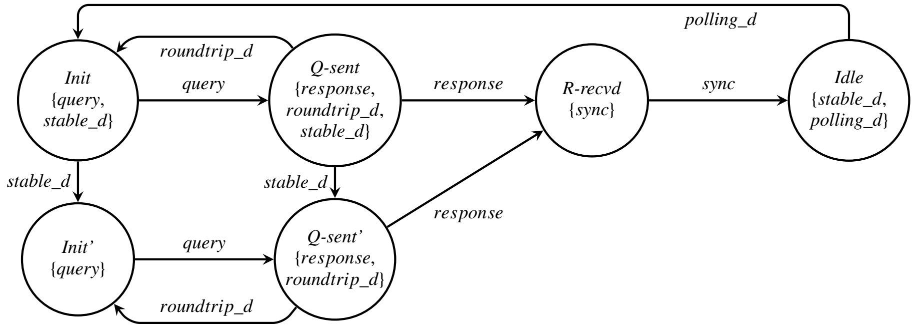

Fig. 1. A GSMP model of a clock synchronization protocol. Below each state label, we list the set of scheduled events. We only display transitions that can take place with non-zero probability.

sufficiently close to a server clock. Each message exchange is initialized by the client asking the server for the current time, i.e. sending a *query* message. The server adds a timestamp into the message and sends it back as a *response*. This query-response exchange provides a reliable data for *synchronization* action if it is realized within a given *round-trip delay*. Otherwise, the client has to repeat the procedure. After a success, the client is considered to be synchronized until a given *stable-time delay* elapses. Since the aim is to keep the clocks synchronized all the time, the client restarts the synchronization process sooner, i.e. after a given *polling delay* that is shorter than the stable-time delay. Notice that the client gets desynchronized whenever several unsuccessful synchronizations occur in a row. Our goal is to measure the portion of the time when the client clock is not synchronized.

Figure 1 shows a GSMP model of this protocol. The delays specified in the protocol are modeled using fixed-delay events *roundtrip_d*, *stable_d*, and *polling_d* while actions are modeled by variable-delay events *query*, *response*, and *sync*. Note that if the stable-time runs out before a fast enough response arrives, the systems moves into primed states denoting it is not synchronized at the moment. Thus, $\mathbf{c}(\mathit{Init'}) + \mathbf{c}(\mathit{Q\text{-}sent'})$ expresses the portion of the time when the client clock is not synchronized.

Our contribution. So far, GSMP were mostly studied with variable-delay events only. There are a few exceptions such as [4, 3, 8, 2] but they often contain erroneous statements due to presence of fixed-delay events. Our goal is to study the effect of mixing a number of fixed-delay events with an arbitrary amount of variable-delay events.

At the beginning we give an example of a GSMP with two fixed-delay events for which it is *not true* that the variables **d** and **c** are well-defined for almost all runs. We also disprove some crucial statements of [3, 4]. In particular, we show an example of a GSMP which reaches one of its states with probability less than one even though the algorithms of [3, 4] return the probability one. The mistake of these algorithms is fundamental as they neglect the possibility of unstable behavior of GSMP.

Concerning positive results, we show that if there is at most one fixed-delay event, then both **d** and **c** are almost surely well-defined. This is true even if we allow an arbitrary number of reasonably restricted fixed-delay events. We also show how to approximate distribution functions of $\mathbf{d}$ and $\mathbf{c}$. To be more specific, we show that for GSMP with at most one unrestricted and an arbitrary number of restricted fixed-delay events, both variables $\mathbf{d}$ and $\mathbf{c}$ have finite ranges $\{d_1, \ldots, d_n\}$ and $\{c_1, \ldots, c_n\}$. Moreover, all values d_i and c_i and probabilities $\mathcal{P}(\mathbf{d} = d_i)$ and $\mathcal{P}(\mathbf{c} = c_i)$ can be effectively approximated.

Related work. There are two main approaches to the analysis of GSMP. One is to restrict the amount of events or types of their distributions and to solve the problems using symbolic methods [8, 2, 21]. The other is to estimate the values of interest using simulation [26, 15, 16]. Concerning the first approach, time-bounded reachability has been studied in [2] where the authors restricted the delays of events to so called expolynomial distributions. The same authors also studied reachability probabilities of GSMP where in each transition at most one event is inherited [8]. Further, the widely studied formalisms of semi-Markov processes (see, e.g., [20, 9]) and continuous-time Markov chains (see, e.g., [6, 7]) are both subclasses of GSMP.

As for the second approach, GSMP are studied by mathematicians as a standard model for discrete event simulation and Markov chains Monte Carlo (see, e.g., [14, 17, 25]). Our work is strongly related to [15, 16] where the long-run average behavior of GSMP with variable-delay events is studied. Under relatively standard assumptions the stochastic process generated by a GSMP is shown to be irreducible and to possess an invariant measure. In such a case, the variables $\mathbf{d}$ and $\mathbf{c}$ are almost surely constant. Beside the theoretical results, there exist tools that employ simulation for model checking (see, e.g., [26, 11]).

In addition, GSMP are a proper subset of stochastic automata, a model of concurrent systems (see, e.g., [12]). Further, as shown in [16], GSMP have the same modeling power as stochastic Petri nets [22]. The formalism of deterministic and stochastic Petri nets (DSPN) introduced by [21] adds deterministic transitions – a counterpart of fixed-delay events. The authors restricted the model to at most one deterministic transition enabled at a time and to exponentially distributed timed transitions. For this restricted model, the authors proved existence of a steady state distribution and provided an algorithm for its computation. However, the methods inherently rely on the properties of the exponential distribution and cannot be extended to our setting with general variable delays. DSPN have been extended by [13, 19] to allow arbitrarily many deterministic transitions. The authors provide algorithms for steady-state analysis of DSPN that were implemented in the tool DSPNExpress [18], but do not discuss under which conditions the steady-state distributions exist.

2 Preliminaries

In this paper, the sets of all positive integers, non-negative integers, real numbers, positive real numbers, and non-negative real numbers are denoted by $\mathbb{N}$, $\mathbb{N}_0$, $\mathbb{R}$, $\mathbb{R}_{>0}$, and $\mathbb{R}_{\geq 0}$, respectively. For a real number $r \in \mathbb{R}$, $\mathrm{int}(r)$ denotes its integral part, i.e. the largest integer smaller than r, and $\mathrm{frac}(r)$ denotes its fractional part, i.e. $r - \mathrm{int}(r)$. Let A be a finite or countably infinite set. A *probability distribution* on A is a function $f : A \to \mathbb{R}_{\geq 0}$ such that $\sum_{a \in A} f(a) = 1$. The set of all distributions on A is denoted by $\mathcal{D}(A)$.

A *σ-field* over a set Ω is a set $\mathcal{F} \subseteq 2^{\Omega}$ that includes Ω and is closed under complement and countable union. A *measurable space* is a pair $(\Omega, \mathcal{F})$ where Ω is a set called *sample space* and $\mathcal{F}$ is a σ-field over Ω whose elements are called *measurable sets*. Given a measurable space $(\Omega, \mathcal{F})$, we say that a function $f : \Omega \to \mathbb{R}$ is a random variable if the inverse image of any real interval is a measurable set. A *probability measure* over a measurable space $(\Omega, \mathcal{F})$ is a function $\mathcal{P} : \mathcal{F} \to \mathbb{R}_{\geq 0}$ such that, for each countable collection $\{X_i\}_{i\in I}$ of pairwise disjoint elements of $\mathcal{F}$, we have $\mathcal{P}(\bigcup_{i\in I} X_i) = \sum_{i\in I} \mathcal{P}(X_i)$ and, moreover, $\mathcal{P}(\Omega) = 1$. A *probability space* is a triple $(\Omega, \mathcal{F}, \mathcal{P})$, where $(\Omega, \mathcal{F})$ is a measurable space and $\mathcal{P}$ is a probability measure over $(\Omega, \mathcal{F})$. We say that a property $A \subseteq \Omega$ holds for *almost all* elements of a measurable set Y if $\mathcal{P}(Y) > 0$, $A \cap Y \in \mathcal{F}$, and $\mathcal{P}(A \cap Y \mid Y) = 1$. Alternatively, we say that A holds *almost surely* for Y.

2.1 Generalized semi-Markov processes

Let $\mathcal{E}$ be a finite set of *events*. To every $e \in \mathcal{E}$ we associate the *lower bound* $\ell_e \in \mathbb{N}_0$ and the *upper bound* $u_e \in \mathbb{N} \cup \{\infty\}$ of its delay. We say that e is a *fixed-delay* event if $\ell_e = u_e$, and a *variable-delay* event if $\ell_e < u_e$. Furthermore, we say that a variable-delay event e is *bounded* if $u_e \neq \infty$, and *unbounded*, otherwise. To each variable-delay event e we assign a *density function* $f_e : \mathbb{R} \to \mathbb{R}$ such that $\int_{\ell_e}^{u_e} f_e(x)\, dx = 1$. We assume f_e to be positive and continuous on the whole $[\ell_e, u_e]$ or $[\ell_e, \infty)$ if e is bounded or unbounded, respectively, and zero elsewhere. We require that f_e have finite expected value, i.e. $\int_{\ell_e}^{u_e} x \cdot f_e(x)\, dx < \infty$.

Definition 1. *A* generalized semi-Markov process *is a tuple* $(S, \mathcal{E}, \mathbf{E}, \mathrm{Succ}, \alpha_0)$ *where*

- S *is a finite set of* states*,*
- $\mathcal{E}$ *is a finite set of* events*,*
- $\mathbf{E} : S \to 2^{\mathcal{E}}$ *assigns to each state s a set of events* $\mathbf{E}(s) \neq \emptyset$ scheduled *to occur in s,*
- $\mathrm{Succ} : S \times 2^{\mathcal{E}} \to \mathcal{D}(S)$ *is the* successor *function, i.e. assigns a probability distribution specifying the successor state to each state and set of events that occur simultaneously in this state, and*
- $\alpha_0 \in \mathcal{D}(S)$ *is the* initial distribution.

A *configuration* is a pair (s, ν) where $s \in S$ and ν is a *valuation* which assigns to every event $e \in \mathbf{E}(s)$ the amount of time that elapsed since the event e was scheduled.[1] For convenience, we define $\nu(e) = \bot$ whenever $e \notin \mathbf{E}(s)$, and we denote by $\nu(\triangle)$ the amount of time spent in the previous configuration (initially, we put $\nu(\triangle) = 0$). When a set of events E occurs and the process moves from s to a state s', the valuation of *old* events of $\mathbf{E}(s) \smallsetminus \mathbf{E}(s')$ is discarded to $\bot$, the valuation of each *inherited* event of $(\mathbf{E}(s') \cap \mathbf{E}(s)) \smallsetminus E$ is increased by the time spent in s, and the valuation of each *new* event of $(\mathbf{E}(s') \smallsetminus \mathbf{E}(s)) \cup (\mathbf{E}(s') \cap E)$ is set to 0.

We illustrate the dynamics of GSMP on the example of Figure 1. Let the bounds of the fixed-delay events *roundtrip_d*, *polling_d*, and *stable_d* be

[1] Usually, the valuation is defined to store the time left before the event appears. However, our definition is equivalent and more convenient for the general setting where both bounded and unbounded events appear.

1, 90, and 100, respectively. We start in the state *Idle*, i.e. in the configuration $(Idle, ((polling\_d, 0), (stable\_d, 0), (\triangle, 0)))$ denoting that $\nu(polling\_d) = 0$, $\nu(stable\_d) = 0$, $\nu(\triangle) = 0$, and $\bot$ is assigned to all other events. After 90 time units, the event *polling_d* occurs and we move to $(Init, ((query, 0), (stable\_d, 90), (\triangle, 90)))$. Assume that the event *query* occurs in the state *Init* after 0.6 time units and we move to $(Q\text{-}sent, ((response, 0), (roundtrip\_d, 0), (stable\_d, 90.6), (\triangle, 0.6)))$ and so forth.

A formal semantics of GSMP is usually defined in terms of general state-space Markov chains (GSSMC, see, e.g., [24]). A GSSMC is a stochastic process Φ over a measurable state-space $(\Gamma, \mathcal{G})$ whose dynamics is determined by an initial measure μ on $(\Gamma, \mathcal{G})$ and a *transition kernel* P which specifies one-step transition probabilities.[2] A given GSMP induces a GSSMC whose state-space consists of all configurations, the initial measure μ is induced by α_0 in a natural way, and the transition kernel is determined by the dynamics of GSMP described above. Formally,

- Γ is the set of all configurations, and $\mathcal{G}$ is a σ-field over Γ induced by the discrete topology over S and the Borel σ-field over the set of all valuations;
- the initial measure μ allows to start in configurations with zero valuation only, i.e. for $A \in \mathcal{G}$ we have $\mu(A) = \sum_{s \in Zero(A)} \alpha_0(s)$ where $Zero(A) = \{s \in S \mid (s, \mathbf{0}) \in A\}$;
- the transition kernel $P(z, A)$ describing the probability to move in one step from a configuration $z = (s, \nu)$ to any configuration in a set A is defined as follows. It suffices to consider A of the form $\{s'\} \times X$ where X is a measurable set of valuations. Let V and F be the sets of variable-delay and fixed-delay events, respectively, that are scheduled in s. Let $F' \subseteq F$ be the set of fixed-delay events that can occur as first among the fixed-delay event enabled in z, i.e. that have in ν the minimal remaining time u. Note that two variable-delay events occur simultaneously with probability zero. Hence, we consider all combinations of $e \in V$ and $t \in \mathbb{R}_{\geq 0}$ stating that

$$P(z, A) = \begin{cases} \sum_{e \in V} \int_0^\infty \mathrm{Hit}(\{e\}, t) \cdot \mathrm{Win}(\{e\}, t)\, dt & \text{if } F = \emptyset \\ \sum_{e \in V} \int_0^u \mathrm{Hit}(\{e\}, t) \cdot \mathrm{Win}(\{e\}, t)\, dt + \mathrm{Hit}(F', u) \cdot \mathrm{Win}(F', u) & \text{otherwise,} \end{cases}$$

  where the term $\mathrm{Hit}(E, t)$ denotes the conditional probability of hitting A under the condition that E occurs at time t and the term $\mathrm{Win}(E, t)$ denotes the probability (density) of E occurring at time t. Formally,

$$\mathrm{Hit}(E, t) = \mathrm{Succ}(s, E)(s') \cdot \mathbf{1}[\nu' \in X]$$

  where $\mathbf{1}[\nu' \in X]$ is the indicator function and ν' is the valuation after the transition, i.e. $\nu'(e)$ is $\bot$, or $\nu(e) + t$, or 0 for each old, or inherited, or new event e, respectively; and $\nu'(\triangle) = t$. The most complicated part is the definition of $\mathrm{Win}(E, t)$ which intuitively corresponds to the probability that E is the set of events "winning" the competition among the events scheduled in s at time t. First, we define a "shifted" density function $f_{e|\nu(e)}$ that takes into account that the time $\nu(e)$ has already elapsed. Formally, for a variable-delay event e and any elapsed time $\nu(e) < u_e$, we define

$$f_{e|\nu(e)}(x) = \frac{f_e(x + \nu(e))}{\int_{\nu(e)}^\infty f_e(y)\, dy} \qquad \text{if } x \geq 0.$$

[2] Precisely, transition kernel is a function $P : \Gamma \times \mathcal{G} \to [0, 1]$ such that $P(z, \cdot)$ is a probability measure over $(\Gamma, \mathcal{G})$ for each $z \in \Gamma$; and $P(\cdot, A)$ is a measurable function for each $A \in \mathcal{G}$.

Otherwise, we define $f_{e|\nu(e)}(x) = 0$. The denominator scales the function so that $f_{e|\nu(e)}$ is again a density function. Finally,

$$\mathrm{Win}(E,t) = \begin{cases} f_{e|\nu(e)}(t) \cdot \prod_{c\in V\setminus E} \int_t^\infty f_{c|\nu(c)}(y)\, dy & \text{if } E = \{e\} \subseteq V \\ \prod_{c\in V} \int_t^\infty f_{c|\nu(c)}(y)\, dy & \text{if } E = F' \subseteq F \\ 0 & \text{otherwise.} \end{cases}$$

A *run* of the Markov chain is an infinite sequence $\sigma = z_0\ z_1\ z_2 \cdots$ of configurations. The Markov chain is defined on the probability space $(\Omega, \mathcal{F}, \mathcal{P})$ where Ω is the set of all runs, $\mathcal{F}$ is the product σ-field $\bigotimes_{i=0}^{\infty} \mathcal{G}$, and $\mathcal{P}$ is the unique probability measure such that for every finite sequence $A_0, \cdots, A_n \in \mathcal{G}$ we have that

$$\mathcal{P}(\Phi_0{\in}A_0, \cdots, \Phi_n{\in}A_n) = \int\limits_{z_0\in A_0} \cdots \int\limits_{z_{n-1}\in A_{n-1}} \mu(dz_0) \cdot P(z_0, dz_1) \cdots P(z_{n-1}, A_n)$$

where each Φ_i is the i-th projection of an element in Ω (the i-th configuration of a run).

Finally, we define an m-step transition kernel P^m inductively as $P^1(z,A) = P(z,A)$ and $P^{i+1}(z,A) = \int_\Gamma P(z,dy) \cdot P^i(y,A)$.

2.2 Frequency measures

Our attention focuses on frequencies of a fixed state $\mathring{s} \in S$ in the runs of the Markov chain. Let $\sigma = (s_0,\nu_0)\ (s_1,\nu_1) \cdots$ be a run. We define

$$\mathbf{d}(\sigma) = \lim_{n\to\infty} \frac{\sum_{i=0}^{n} \delta(s_i)}{n} \qquad\qquad \mathbf{c}(\sigma) = \lim_{n\to\infty} \frac{\sum_{i=0}^{n} \delta(s_i) \cdot \nu_{i+1}(\triangle)}{\sum_{i=0}^{n} \nu_{i+1}(\triangle)}$$

where $\delta(s_i)$ is equal to 1 when $s_i = \mathring{s}$, and 0 otherwise. We recall that $\nu_{i+1}(\triangle)$ is the time spent in state s_i before moving to s_{i+1}. We say that the random variable $\mathbf{d}$ or $\mathbf{c}$ is *well-defined* for a run σ if the corresponding limit exists for σ. Then, $\mathbf{d}$ corresponds to the frequency of discrete visits to the state $\mathring{s}$ and $\mathbf{c}$ corresponds to the ratio of time spent in the state $\mathring{s}$.

2.3 Region graph

In order to state the results in a simpler way, we introduce the *region graph*, a standard notion from the area of timed automata [5]. It is a finite partition of the uncountable set of configurations. First, we define the region relation $\sim$. For $a, b \in \mathbb{R}$, we say that a and b *agree on integral part* if $\mathrm{int}(a) = \mathrm{int}(b)$ and neither or both a, b are integers. Further, we set the bound $B = \max\left(\{\ell_e, u_e \mid e \in \mathcal{E}\} \setminus \{\infty\}\right)$. Finally, we put $(s_1,\nu_1) \sim (s_2,\nu_2)$ if

- $s_1 = s_2$;
- for all $e \in \mathbf{E}(s_1)$ we have that $\nu_1(e)$ and $\nu_2(e)$ agree on integral parts or are both greater than B;
- for all $e, f \in \mathbf{E}(s_1)$ with $\nu_1(e) \leq B$ and $\nu_1(f) \leq B$ we have that $\mathrm{frac}(\nu_1(e)) \leq \mathrm{frac}(\nu_1(f))$ iff $\mathrm{frac}(\nu_2(e)) \leq \mathrm{frac}(\nu_2(f))$.

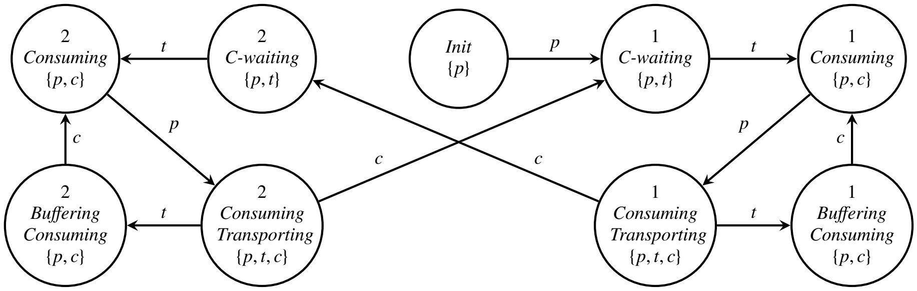

Fig. 2. A GSMP of a producer-consumer system. The events p, t, and c model that a packet production, transport, and consumption is finished, respectively. Below each state label, there is the set of scheduled events. The fixed-delay events p and c have $l_p = u_p = l_c = u_c = 1$ and the uniformly distributed variable-delay event t has $l_t = 0$ and $u_t = 1$.

Note that $\sim$ is an equivalence with finite index. The equivalence classes of $\sim$ are called *regions*. We define a finite *region graph* $G = (V, E)$ where the set of vertices V is the set of regions and for every pair of regions R, R' there is an edge $(R, R') \in E$ iff $P(z, R') > 0$ for some $z \in R$. The construction is correct because all states in the same region have the same one-step qualitative behavior (for details, see [10]).

3 Two fixed-delay events

Now, we explain in more detail what problems can be caused by fixed-delay events. We start with an example of a GSMP with two fixed-delay events for which it is not true that the variables **d** and **c** are well-defined for almost all runs. Then we show some other examples of GSMP with fixed-delay events that disprove some results from literature. In the next section, we provide positive results when the number and type of fixed-delay events are limited.

When the frequencies d and c are not well-defined

In Figure 2, we show an example of a GSMP with two fixed-delay events and one variable-delay event for which it is not true that the variables **d** and **c** are well-defined for almost all runs. It models the following producer-consumer system. We use three components – a producer, a transporter and a consumer of packets. The components work in parallel but each component can process (i.e. produce, transport, or consume) at most one packet at a time.

Consider the following time requirements: each packet production takes *exactly* 1 time unit, each transport takes *at most* 1 time unit, and each consumption takes again *exactly* 1 time unit. As there are no limitations to block the producer, it is working for all the time and new packets are produced precisely each time unit. As the transport takes shorter time than the production, every new packet is immediately taken by the transporter and no buffer is needed at this place. When a packet arrives to the consumer, the consumption is started immediately if the consumer is waiting; otherwise, the packet

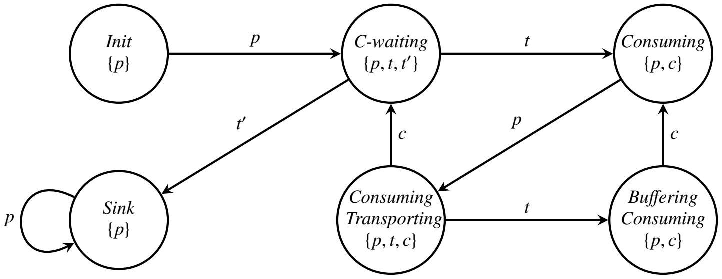

Fig. 3. A GSMP with two fixed-delay events p and c (with $l_p = u_p = l_c = u_c = 1$), a uniformly distributed variable-delay events t, t' (with $l_t = l_{t'} = 0$ and $u_t = u_{t'} = 1$).

is stored into a buffer. When the consumption is finished and the buffer is empty, the consumer waits; otherwise, a new consumption starts immediately.

In the GSMP in Figure 2, the consumer has two modules – one is in operation and the other idles at a time – when the consumer enters the waiting state, it switches the modules. The labels 1 and 2 denote which module of the consumer is in operation.

One can easily observe that the consumer enters the waiting state (and switches the modules) if and only if the current transport takes more time than it has ever taken. As the transport time is bounded by 1, it gets harder and harder to break the record. As a result, the system stays in the current module on average for longer time than in the previous module. Therefore, due to the successively prolonging stays in the modules, the frequencies for 1-states and 2-states oscillate. For precise computations, see [10]. We conclude the above observation by the following theorem.

Theorem 1. *There is a GSMP (with two fixed-delay events and one variable-delay event) for which it is* not *true that the variables* **c** *and* **d** *are almost surely well-defined.*

Counterexamples

In [3, 4] there are algorithms for GSMP model checking based on the region construction. They rely on two crucial statements of the papers:

1. Almost all runs end in some of the bottom strongly connected components (BSCC) of the region graph.
2. Almost all runs entering a BSCC visit all regions of the component infinitely often.

Both of these statements are true for finite state Markov chains. In the following, we show that neither of them has to be valid for region graphs of GSMP.

Let us consider the GSMP depicted in Figure 3. This is a producer-consumer model similar to the previous example but we have only one module of the consumer here. Again, entering the state *C-waiting* indicates that the current transport takes more time than it has ever taken. In the state *C-waiting*, an additional event t' can occur and move the system into a state *Sink*. One can intuitively observe that we enter the state *C-waiting* less and less often and stay there for shorter and shorter time. Hence, the probability that the event t' occurs in the state *C-waiting* is decreasing during the run. For precise computations proving the following claim, see [10].

Claim. The probability to reach *Sink* from *Init* is strictly less than 1.

The above claim directly implies the following theorem thus disproving statement 1.

Theorem 2. *There is a GSMP (with two fixed-delay and two variable delay events) where the probability to reach any BSCC of the region graph is strictly smaller than 1.*

Now consider in Figure 3 a transition under the event p from the state *Sink* to the state *Init* instead of the self-loop. This turns the whole region graph into a single BSCC. We prove that the state *Sink* is almost surely visited only finitely often. Indeed, let $p < 1$ be the original probability to reach *Sink* guaranteed by the claim above. The probability to reach *Sink* from *Sink* again is also p as the only transition leading from *Sink* enters the initial configuration. Therefore, the probability to reach *Sink* infinitely often is $\lim_{n\to\infty} p^n = 0$. This proves the following theorem. Hence, the statement 2 of [3, 4] is disproved, as well.

Theorem 3. *There is a GSMP (with two fixed-delay and two variable delay events) with strongly connected region graph and with a region that is reached infinitely often with probability* 0*.*

4 Single-ticking GSMP

First of all, motivated by the previous counterexamples, we identify the behavior of the fixed-delay events that may cause $\mathbf{d}$ and $\mathbf{c}$ to be undefined. The problem lies in fixed-delay events that can immediately schedule themselves whenever they occur; such an event can occur periodically like ticking of clocks. In the example of Figure 3, there are two such events p and c. The phase difference of their ticking gets smaller and smaller, causing the unstable behavior.

For two fixed-delay events e and e', we say that *e causes e'* if there are states s, s' and a set of events E such that $\mathrm{Succ}(s, E)(s') > 0$, $e \in E$, and e' is newly scheduled in s'.

Definition 2. *A GSMP is called* single-ticking *if either there is no fixed-delay event or there is a strict total order $<$ on fixed-delay events with the least element e (called* ticking *event) such that whenever f causes g then either $f < g$ or $f = g = e$.*

From now on we restrict to single-ticking GSMP and prove our main positive result.

Theorem 4. *In single-ticking GSMP, the random variables* $\mathbf{d}$ *and* $\mathbf{c}$ *are well-defined for almost every run and admit only finitely many values. Precisely, almost every run reaches a BSCC of the region graph and for each BSCC B there are values $d, c \in [0, 1]$ such that* $\mathbf{d}(\sigma) = d$ *and* $\mathbf{c}(\sigma) = c$ *for almost all runs σ that reach the BSCC B.*

The rest of this section is devoted to the proof of Theorem 4. First, we show that almost all runs end up trapped in some BSCC of the region graph. Second, we solve the problem while restricting to runs that *start* in a BSCC (as the initial part of a run outside of any BSCC is not relevant for the long run average behavior). We show that in a BSCC, the variables $\mathbf{d}$ and $\mathbf{c}$ are almost surely constant. The second part of the proof relies on several standard results from the theory of general state space Markov chains. Formally, the proof follows from Propositions 1 and 2 stated below.

4.1 Reaching a BSCC

Proposition 1. *In single-ticking GSMP, almost every run reaches a BSCC of the region graph.*

The proof uses similar methods as the proof in [4]. By definition, the process moves along the edges of the region graph. From every region, there is a minimal path through the region graph into a BSCC, let n be the maximal length of all such paths. Hence, in at most n steps the process reaches a BSCC with positive probability from any configuration. Observe that if this probability was bounded from below, we would eventually reach a BSCC from any configuration almost surely. However, this probability can be arbitrarily small. Consider the following example with event e uniform on $[0, 1]$ and event f uniform on $[2, 3]$. In an intuitive notation, let R be the region $[0 < e < f < 1]$. What is the probability that the event e occurs after the elapsed time of f reaches 1 (i.e. that the region $[e = 0; 1 < f < 2]$ is reached)? For a configuration in R with valuation $((e, 0.2), (f, 0.7))$ the probability is 0.5 but for another configuration in R with $((e, 0.2), (f, 0.21))$ it is only 0.01. Notice that the transition probabilities depend on the difference of the fractional values of the clocks, we call this difference *separation*. Observe that in other situations, the separation of clocks from value 0 also matters.

Definition 3. *Let $\delta > 0$. We say that a configuration (s, ν) is δ-*separated *if for every $x, y \in \{0\} \cup \{\nu(e) \mid e \in \mathbf{E}(s)\}$, we have either $|\mathrm{frac}(x) - \mathrm{frac}(y)| > \delta$ or $\mathrm{frac}(x) = \mathrm{frac}(y)$.*

We fix a $\delta > 0$. To finish the proof using the concept of δ-separation, we need two observations. First, from *any* configuration we reach in m steps a δ-separated configuration with probability at least $q > 0$. Second, the probability to reach a fixed region from *any* δ-separated configuration is bounded from below by some $p > 0$. By repeating the two observations ad infinitum, we reach some BSCC almost surely. Let us state the claims. For proofs, see [10].

Lemma 1. *There is $\delta > 0$, $m \in \mathbb{N}$ and $q > 0$ such that from every configuration we reach a δ-separated configuration in m steps with probability at least q.*

Lemma 2. *For every $\delta > 0$ and $k \in \mathbb{N}$ there is $p > 0$ such that for any pair of regions R, R' connected by a path of length k and for any δ-separated $z \in R$, we have $P^k(z, R') > p$.*

Lemma 2 holds even for unrestricted GSMP. Notice that Lemma 1 does not. As in the example of Figure 3, the separation may be non-increasing for all runs.

4.2 Frequency in a BSCC

From now on, we deal with the bottom strongly connected components that are reached almost surely. Hence, we assume that the region graph G is strongly connected. We have to allow an arbitrary initial configuration $z_0 = (s, \nu)$; in particular, ν does not have to be a zero vector.[3]

[3] Technically, the initial measure is $\mu(A) = 1$ if $z_0 \in A$ and $\mu(A) = 0$, otherwise.

Proposition 2. *In a single-ticking GSMP with strongly connected region graph, there are values $d, c \in [0,1]$ such that for any initial configuration z_0 and for almost all runs σ starting from z_0, we have that* $\mathbf{d}$ *and* $\mathbf{c}$ *are well-defined and* $\mathbf{d}(\sigma) = d$ *and* $\mathbf{c}(\sigma) = c$.

We assume that the region graph is aperiodic in the following sense. A *period* p of a graph G is the greatest common divisor of lengths of all cycles in G. The graph G is *aperiodic* if $p = 1$. Under this assumption[4], the chain $\varPhi$ is in some sense stable. Namely, (i) $\varPhi$ has a unique invariant measure that is independent of the initial measure and (ii) the strong law of large numbers (SLLN) holds for $\varPhi$.

First, we show that (i) and (ii) imply the proposition. Let us recall the notions. We say that a probability measure π on $(\varGamma, \mathcal{G})$ is *invariant* if for all $A \in \mathcal{G}$

$$\pi(A) \quad = \quad \int_{\varGamma} \pi(dx) P(x, A).$$

The SLLN states that if $h : \varGamma \to \mathbb{R}$ satisfies $E_\pi[h] < \infty$, then almost surely

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} h(\varPhi_i)}{n} \quad = \quad E_\pi[h], \tag{1}$$

where $E_\pi[h]$ is the expected value of h according to the invariant measure π.

We set h as follows. For a run $(s_0, \nu_0)(s_1, \nu_1)\cdots$, let $h(\varPhi_i) = 1$ if $s_i = \mathring{s}$ and 0, otherwise. We have $E_\pi[h] < \infty$ since $h \leq 1$. From (1) we obtain that almost surely

$$\mathbf{d} \quad = \quad \lim_{n\to\infty} \frac{\sum_{i=1}^{n} h(\varPhi_i)}{n} \quad = \quad E_\pi[h].$$

As a result, $\mathbf{d}$ is well-defined and equals the constant value $E_\pi[h]$ for almost all runs. We treat the variable $\mathbf{c}$ similarly. Let $W((s, \nu))$ denote the expected waiting time of the GSMP in the configuration (s, ν). We use a function $\tau((s, \nu)) = W((s, \nu))$ if $s = \mathring{s}$ and 0, otherwise. Since all the events have finite expectation, the functions W and τ are bounded and we have $E_\pi[W] < \infty$ and $E_\pi[\tau] < \infty$. We show in [10] that almost surely

$$\mathbf{c} \quad = \quad \lim_{n\to\infty} \frac{\sum_{i=1}^{n} \tau(\varPhi_i)}{\sum_{i=1}^{n} W(\varPhi_i)} \quad = \quad \frac{E_\pi[\tau]}{E_\pi[W]}.$$

Therefore, $\mathbf{c}$ is well-defined and equals the constant $E_\pi[\tau]/E_\pi[W]$ for almost all runs.

Second, we prove (i) and (ii). A standard technique of general state space Markov chains (see, e.g., [24]) yields (i) and (ii) for chains that satisfy the following condition. Roughly speaking, we search for a set of configurations C that is visited infinitely often and for some ℓ the measures $P^\ell(x, \cdot)$ and $P^\ell(y, \cdot)$ are very similar for any $x, y \in C$. This is formalized by the following lemma.

Lemma 3. *There is a measurable set of configurations* C *such that*

1. *there is* $k \in \mathbb{N}$ *and* $\alpha > 0$ *such that for every* $z \in \varGamma$ *we have* $P^k(z, C) \geq \alpha$, *and*

[4] If the region graph has period $p > 1$, we can employ the standard technique and decompose the region graph (and the Markov chain) into p aperiodic components. The results for individual components yield straightforwardly the results for the whole Markov chain, see, e.g., [9].

2. *there is $\ell \in \mathbb{N}$, $\beta > 0$, and a probability measure κ such that for every $z \in C$ and $A \in \mathcal{G}$ we have $P^{\ell}(z, A) \geq \beta \cdot \kappa(A)$.*

Proof (Sketch). Let e be the ticking event and R some reachable region where e is the event closest to its upper bound. We fix a sufficiently small $\delta > 0$ and choose C to be the set of δ-separated configurations of R. We prove the first part of the lemma similarly to Lemmata 1 and 2. As regards the second part, we define the measure κ uniformly on a hypercube X of configurations (s, ν) that have $\nu(e) = 0$ and $\nu(f) \in (0, \delta)$, for $f \neq e$. First, assume that e is the only fixed-delay event. We fix $z = (s', \nu')$ in R; let $d = u_e - \nu'(e) > \delta$ be the time left in z before e occurs. For simplicity, we assume that each variable-delay events can occur after an arbitrary delay $x \in (d - \delta, d)$. Precisely, that it can occur in an ε-neighborhood of x with probability bounded from below by $\beta \cdot \varepsilon$ where β is the minimal density value of all $\mathcal{E}$. Note that the variable-delay events can be "placed" this way arbitrarily in $(0, \delta)$. Therefore, when e occurs, it has value 0 and all variable-delay events can be in interval $(0, \delta)$. In other words, we have $P^{\ell}(z, A) \geq \beta \cdot \kappa(A)$ for any measurable $A \subseteq X$ and for $\ell = |\mathcal{E}|$.

Allowing other fixed-delay events causes some trouble because a fixed-delay event $f \neq e$ cannot be "placed" arbitrarily. In the total order $<$, the event f can cause only strictly greater fixed-delay events. The greatest fixed-delay event can cause only variable-delay events that can be finally "placed" arbitrarily as described above. □

5 Approximations

In the previous section we have proved that in single-ticking GSMP, $\mathbf{d}$ and $\mathbf{c}$ are almost surely well-defined and for almost all runs they attain only finitely many values $d_1 \ldots, d_k$ and $c_1, \ldots, c_k$, respectively. In this section we show how to approximate d_i's and c_i's and the probabilities that $\mathbf{d}$ and $\mathbf{c}$ attain these values, respectively.

Theorem 5. *In a single-ticking GSMP, let $d_1, \ldots, d_k$ and $c_1, \ldots, c_k$ be the discrete and timed frequencies, respectively, corresponding to BSCCs of the region graph. For all $1 \leq i \leq k$, the numbers d_i and c_i as well as the probabilities $\mathcal{P}(\mathbf{d} = d_i)$ and $\mathcal{P}(\mathbf{c} = c_i)$ can be approximated up to any $\varepsilon > 0$.*

Proof. Let $X_1, \ldots, X_k$ denote the sets of configurations in individual BSCCs and d_i and c_i correspond to X_i. Since we reach a BSCC almost surely, we have

$$\mathcal{P}(\mathbf{d} = d_i) = \sum_{j=1}^{k} \mathcal{P}(\mathbf{d} = d_i \mid \mathit{Reach}(X_j)) \cdot \mathcal{P}(\mathit{Reach}(X_j)) = \sum_{j=1}^{k} \mathbf{1}[d_j = d_i] \cdot \mathcal{P}(\mathit{Reach}(X_j))$$

where the second equality follows from the fact that almost all runs in the j-th BSCC yield the discrete frequency d_j. Therefore, $\mathcal{P}(\mathbf{d} = d_i)$ and d_i can be approximated as follows using the methods of [25].

Claim. Let X be a set of all configurations in a BSCC $\mathcal{B}$, $X_{\mathring{s}} \subseteq X$ the set of configurations with state $\mathring{s}$, and d the frequency corresponding to $\mathcal{B}$. There are computable constants $n_1, n_2 \in \mathbb{N}$ and $p_1, p_2 > 0$ such that for every $i \in \mathbb{N}$ and $z_X \in X$ we have

$$\begin{aligned} |\mathcal{P}(\mathit{Reach}(X)) - P^i(z_0, X)| &\leq (1 - p_1)^{\lfloor i/n_1 \rfloor} \\ |d - P^i(z_X, X_{\mathring{s}})| &\leq (1 - p_2)^{\lfloor i/n_2 \rfloor} \end{aligned}$$

Further, we want to approximate $c_i = E_\pi[\tau]/E_\pi[W]$, where π is the invariant measure on X_i. In other words, we need to approximate $\int_{X_i} \tau(x)\pi(dx)$ and $\int_{X_i} W(x)\pi(dx)$. An n-th approximation $w_n(x)$ of $W(x)$ can be gained by discretizing the regions into, e.g., $1/n$-large hypercubes. If W is continuous, then $(w_n)_{n=1}^{\infty}$ is its pointwise approximation. Moreover, if W is bounded, then it is dominated by the approximation function w_n. Hence the approximation is correct by the dominated convergence theorem. Note that τ only differs from W in being identically zero on some regions. Therefore, the following claim concludes the proof. For details, see [10].

Claim. On each region, W is continuous and bounded and can be approximated.

6 Conclusions, future work

We have studied long run average properties of generalized semi-Markov processes with both fixed-delay and variable-delay events. We have shown that two or more (unrestricted) fixed-delay events lead to considerable complications regarding stability of GSMP. In particular, we have shown that the frequency of states of a GSMP may not be well-defined and that bottom strongly connected components of the region graph may not be reachable with probability one. This leads to counterexamples disproving several results from literature. On the other hand, for single-ticking GSMP we have proved that the frequencies of states are well-defined for almost all runs. Moreover, we have shown that almost every run has one of finitely many possible frequencies that can be effectively approximated (together with their probabilities) up to a given error tolerance.

In addition, the frequency measures can be easily extended into the mean payoff setting. Consider assigning real rewards to states. The mean payoff then corresponds to the frequency weighted by the rewards.

Concerning future work, the main issue is efficiency of algorithms for computing performance measures for GSMP. We plan to work on both better analytical methods as well as practicable approaches to Monte Carlo simulation. One may also consider extensions of our positive results to controlled GSMP and games on GSMP.

References

1. de Alfaro, L.: How to specify and verify the long-run average behavior of probabilistic systems. In: Proceedings of LICS'98. pp. 454–465. IEEE Computer Society Press (1998)
2. Alur, R., Bernadsky, M.: Bounded model checking for GSMP models of stochastic real-time systems. In: Proceedings of 9th International Workshop Hybrid Systems: Computation and Control (HSCC). Lecture Notes in Computer Science, vol. 3927, pp. 19–33. Springer (2006)
3. Alur, R., Courcoubetis, C., Dill, D.: Model-checking for probabilistic real-time systems. In: Proceedings of ICALP'91. Lecture Notes in Computer Science, vol. 510, pp. 115–136. Springer (1991)
4. Alur, R., Courcoubetis, C., Dill, D.: Verifying automata specifications of probabilistic real-time systems. In: Real-Time: Theory in Practice. Lecture Notes in Computer Science, vol. 600, pp. 28–44. Springer (1992)
5. Alur, R., Dill, D.: A theory of timed automata. Theoretical Computer Science 126(2), 183–235 (1994)

6. Baier, C., Haverkort, B., Hermanns, H., Katoen, J.P.: Model-checking algorithms for continuous-time Markov chains. IEEE Transactions on Software Engineering 29(6), 524–541 (2003)
7. Barbot, B., Chen, T., Han, T., Katoen, J., Mereacre, A.: Efficient CTMC model checking of linear real-time objectives. Tools and Algorithms for the Construction and Analysis of Systems pp. 128–142 (2011)
8. Bernadsky, M., Alur, R.: Symbolic analysis for GSMP models with one stateful clock. In: Proceedings of 10th International Workshop Hybrid Systems: Computation and Control (HSCC). Lecture Notes in Computer Science, vol. 4416, pp. 90–103. Springer (2007)
9. Brázdil, T., Krčál, J., Křetínský, J., Kučera, A., Řehák, V.: Measuring performance of continuous-time stochastic processes using timed automata. In: Proceedings of 14th International Conference on Hybrid Systems: Computation and Control (HSCC'11). pp. 33–42. ACM Press (2011)
10. Brázdil, T., Krčál, J., Křetínský, J., Řehák, V.: Fixed-delay Events in Generalized Semi-Markov Processes Revisited. ArXiv e-prints (Sep 2011)
11. Ciardo, G., Jones III, R., Miner, A., Siminiceanu, R.: Logic and stochastic modeling with SMART. Performance Evaluation 63(6), 578–608 (2006)
12. D'Argenio, P., Katoen, J.: A theory of stochastic systems Part I: Stochastic automata. Information and computation 203(1), 1–38 (2005)
13. German, R., Lindemann, C.: Analysis of stochastic Petri nets by the method of supplementary variables. Performance Evaluation 20(1-3), 317–335 (1994)
14. Glynn, P.: A GSMP formalism for discrete event systems. Proceedings of the IEEE 77, 14–23 (1989)
15. Haas, P.: On simulation output analysis for generalized semi-markov processes. Commun. Statist. Stochastic Models 15, 53–80 (1999)
16. Haas, P.: Stochastic Petri Nets: Modelling, Stability, Simulation. Springer Series in Operations Research and Financial Engineering, Springer (2010)
17. Haas, P., Shedler, G.: Regenerative generalized semi-Markov processes. Stochastic Models 3(3), 409–438 (1987)
18. Lindemann, C., Reuys, A., Thummler, A.: The DSPNexpress 2.000 performance and dependability modeling environment. In: Fault-Tolerant Computing, 1999. Digest of Papers. Twenty-Ninth Annual International Symposium on. pp. 228–231. IEEE Computer Society Press (1999)
19. Lindemann, C., Shedler, G.: Numerical analysis of deterministic and stochastic Petri nets with concurrent deterministic transitions. Performance Evaluation 27, 565–582 (1996)
20. López, G., Hermanns, H., Katoen, J.: Beyond memoryless distributions: Model checking semi-Markov chains. Process Algebra and Probabilistic Methods. Performance Modelling and Verification pp. 57–70 (2001)
21. Marsan, M., Chiola, G.: On Petri nets with deterministic and exponentially distributed firing times. Advances in Petri Nets 1987 pp. 132–145 (1987)
22. Marsan, M., Balbo, G., Conte, G., Donatelli, S., Franceschinis, G.: Modelling with Generalized Stochastic Petri Nets. Wiley (1995)
23. Matthes, K.: Zur Theorie der Bedienungsprozesse. Transactions of the Third Prague Conference on Information Theory, Statistical Decision Functions, Random Processes pp. 513–528 (1962)
24. Meyn, S., Tweedie, R.: Markov Chains and Stochastic Stability. Cambridge University Press (2009)
25. Roberts, G., Rosenthal, J.: General state space Markov chains and MCMC algorithms. Probability Surveys 1, 20–71 (2004)
26. Younes, H., Simmons, R.: Probabilistic verification of discrete event systems using acceptance sampling. In: Computer Aided Verification. pp. 23–39. Springer (2002)