Continuous Approximation of Collective Systems Behaviour: a Tutorial*

Luca Bortolussi
Dept. of Mathematics and Informatics, University of Trieste, IT.
luca@dmi.units.it

Jane Hillston
Laboratory for Foundations of Computer Science, University of Edinburgh, UK.
jane.hillston@ed.ac.uk

Diego Latella, Mieke Massink
Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, IT.
{Diego.Latella,Mieke.Massink}@isti.cnr.it

ISTI Technical Report n. cnr.isti/2011-TR-021

Abstract

In this paper we will introduce the reader to the field of deterministic approximation of Markov processes, both in discrete and in continuous time. We will discuss fluid approximation of continuous time Markov chains and mean field approximation of discrete time Markov chains, considering the cases in which the deterministic limit process lives in continuous time or in discrete time. We also discuss some more advanced results, especially those concerned with the limit stationary behaviour. We assume a knowledge of modeling with Markov chains, but not on more advanced topics in stochastic processes.

Keywords
deterministic approximation; fluid approximation; mean field approximation; Markov Chains; stochastic process algebras.

1 Introduction

Quantitative modeling in computer science is often based on Markovian stochastic processes [1], both in discrete and in continuous time. This is the case, for instance, in performance analysis [2] and, more recently, in computational systems biology [3]. A common problem faced in these fields is that the number of components constituting a system, as well as the number of local states of components, can be large, leading to a dramatic increase in the size of the state space. The resulting *state space explosion* poses practical limitations on our ability to analyze these systems using standard approaches, like steady state computation and transient analysis of Markov Chains [2], or more sophisticated techniques, like stochastic model-checking [4].

In the last few years, there has been a growing interest in techniques that try to tackle the state space explosion by treating large state spaces in a continuous fashion, approximating the stochastic dynamics with a deterministic dynamical system, usually described by means of a differential equation. This is particularly fruitful for systems which are composed of large clusters of (relatively simple) components.

*Work partially funded by EU Project n. 257414 (ASCENS)

Such techniques have been developed in different contexts and with different flavors, yet their approach is very similar. Broadly speaking, nomenclature is (at least) twofold: continuous approximation is known under the name of both *fluid approximation* and *mean field approximation*. The term *deterministic approximation* is sometimes also used.

Fluid approximation has been introduced and applied in the last five years in order to analyze the collective behavior of stochastic process algebra (SPA) models of large populations [5, 6, 7, 8, 9]. Stochastic process algebras [10, 11, 12] are modeling languages, designed to describe systems of interacting agents, which have Continuous Time Markov Chains (CTMCs) as the relevant semantic domain [1]. Fluid approximation can be applied if in a model there are many instances of a few agent types, and it works by treating the variables counting how many agents of each type are in each state in the system as continuous variables, and treating the global rates of the stochastic transitions as flows, thus obtaining an ordinary differential equation (ODE). After observing that, for large populations, the behavior of the stochastic system and that of the deterministic one were very close [5], it was proved that the solution of the differential equation obtained was the limit of a sequence of CTMC models, for increasing population levels [6], exploiting previous results on deterministic approximation of stochastic processes [13]. Fluid approximation of SPA has been used successfully to describe different kind of systems: computer epidemics [14], biological systems [15, 7, 9, 16], computer networks [17], queues [18], and crowd models [19, 20], just to cite a few.

Another use of fluid techniques can be found in Petri Nets. There have been a variety of different Petri net developments involving *fluid levels*, either as alternatives to, or in addition to, the usual discrete tokens [21, 22, 23]. Perhaps closest the process algebra results outlined above is the work of Silva and Recalde [24] where the motivation is again the state space explosion problem, and the authors present their work in terms of a relaxation of integer token counts for state representation.

Mean field approximation is very similar in spirit to fluid approximation. Mean field techniques have a long history: originally developed in statistical physics (in the context of stochastic models of plasma and dense gases), they have been applied in epidemiology [25, 26, 27], game theory [28, 29], approximation algorithms [30], and in performance modeling of computer networks [31, 32, 33, 34, 35, 36, 37]. Usually, mean field approaches start from a stochastic model expressed directly in terms of a Discrete Time Markov Chain (DTMC), describing a system consisting of a large number of interacting entities, each of which can be in one of a relatively small set of states. Then, one constructs a continuous system, describing the continuous evolution of the number of these entities (more specifically, the variables counting how many entities are in a given state), proving a limit theorem similar to the one for fluid approximation. Many applications of mean field approximation in computer science are concerned with communication networks [31, 32, 34, 37], and the limit theorems are proved just for the specific model in each case. More recently, mean field results for more general frameworks have been presented [38, 39, 40], and applied, for instance, to study properties of gossip protocols [41].

Broadly speaking, there are three different classes of approaches that have been labeled as mean field. The first one deals with DTMC models that have a deterministic limit in discrete time (i.e. a discrete time dynamical system) [38], the second one considers DTMCs that have a limit in continuous time, described in terms of differential equations [39], and the last one deals with CTMCs which have a limit in continuous time [40]. This last approach is essentially the same as fluid approximation, the only difference being that the authors work directly on a CTMC model, instead of manipulating a SPA model. The second class of mean field techniques, i.e. those concerned with DTMC models that have a deterministic limit in continuous time, is also strongly related to continuous approximations for CTMCs, as will be made clear later in this paper. The first class, instead, is intrinsically different, and can be applied to DTMC models that assume a different mechanism of interaction. In particular, mean field limits in discrete time require us to work with a DTMC in which all entities of the model (try to) perform a move at each step of the process (*clock-synchronous evolution*), while mean field limits in continuous time assume that just one or few entities perform a move in each step (*clock-asynchronous evolution*).

From the previous discussion, it is clear that fluid approximation and mean field techniques

are strongly related, and that there is a certain amount of notational and terminological confusion in the literature. This can create some difficulties to a modeling practitioner or a student wishing to approach the field.

In this paper we seek to overcome these issues by providing a uniform introduction to these techniques. Our intended audience will be computer scientists who have a background in modeling, but we will not assume prior knowledge of continuous approximation. Readers may be motivated by a desire to apply these techniques to a particular problem or by a more general curiosity, but in either case we aim to introduce these methods in a *gentle* way. We will present the methods using a very simple modeling language (basically a direct description of a Markov chain, either in discrete or in continuous time), in order to focus more on aspects related to the dynamics (Section 3). After discussing general issues of the continuous approximation (Section 4), we will first describe the continuous approximation for CTMCs with ODEs (Section 5), and then focus on the ODE, continuous time, approximation of DTMCs (Section 6), followed by a discussion of how these two approaches are related (Section 7). The deterministic approximation of DTMCs in discrete time will be discussed in Section 8. Then we will discuss some general applications, mainly concerned with stationary behavior and independence (Section 9), pointing out various extensions which have recently appeared in the literature. Throughout the paper, we will make use of a running example, a simple model of a computer network epidemic, to illustrate all the approaches (Section 3.1). Notational conventions and basic concepts and definitions are briefly recalled in Section 2. Sections and remarks marked with an asterisk can be safely skipped at a first reading.

2 Preliminaries

In this section we introduce some notation and recall some basic concepts and definitions. For set S, $s \in S$, and binary relation $R \subseteq S \times S$ we let $s\,R$ be defined by $s\,R =_{\text{def}} \{s' \mid s\,R\,s'\}$ and by R^+ we denote the transitive closure of R. We will let $\mathbb{R}_{\geq 0}$ ($\mathbb{R}_{>0}$, respectively) denote the set $\{d \in \mathbb{R} \mid d \geq 0\}$ ($\{d \in \mathbb{R} \mid d > 0\}$, respectively). We let θ denote the (complement of the) test on zero, i.e.

$$\theta(d) =_{\text{def}} \begin{cases} 0, & \text{if } d = 0 \\ 1, & \text{otherwise} \end{cases}$$

For $n > 0$, a *vector* $\mathbf{d}$ in $\mathbb{R}^n$ is a n-tuple $(d_1, d_2, \ldots, d_n)$, where $d_j \in \mathbb{R}$ for $j = 1, \ldots, n$. We let $\mathbf{d}_j$ denote the j-th projection of $\mathbf{d}$, i.e. $\mathbf{d}_j = d_j$ if $\mathbf{d} = (d_1, d_2, \ldots, d_n)$ and $j = 1, \ldots, n$. Vector $\lfloor \mathbf{d} \rfloor$ is defined as $(\lfloor d_1 \rfloor, \lfloor d_2 \rfloor, \ldots, \lfloor d_n \rfloor)$. We let $||\mathbf{d}||$ ($|\mathbf{d}|$, respectively) denote the *2-norm* (*1-norm*, respectively) of vector $\mathbf{d}$, i.e. $||\mathbf{d}|| =_{\text{def}} \sqrt{d_1^2 + d_2^2 + \ldots + d_n^2}$ ($|\mathbf{d}| =_{\text{def}} |d_1| + |d_2| + \ldots + |d_n|$, respectively).

Given a set $A \subseteq \mathbb{R}$, we let $\sup A$ ($\max A$, $\inf A$, and $\min A$, respectively) denote the *least upper bound* of A (*maximum*, *greatest lower bound*, and *minimum*, respectively); furthermore, for a real-valued function f and predicate p, the shorthand $\sup_{p(\mathbf{d})} f(\mathbf{d})$ is often used in place of $\sup\{\mathbf{d}' \mid \exists \mathbf{d}.\, (p(\mathbf{d}) \wedge \mathbf{d}' = f(\mathbf{d}))\}$. For $K \in \mathbb{R}$, $\mathcal{S}^n(A, K)$ will denote the set $\{\mathbf{d} \in A^n \mid |\mathbf{d}| = K\}$. Notice that $\mathcal{S}^n([0,1], 1)$ is the unit simplex in $\mathbb{R}^n$. By $\Theta(f(N))$ we mean asymptotic 'order $f(N)$', by $O(f(N))$ we mean 'order no more than $f(N)$', and by $\Omega(f(N))$ we mean 'order no less than $f(N)$'.[1]

A *vector field* is a function $F : E \to \mathbb{R}^n$, with $E \subseteq \mathbb{R}^n$, which associates a vector in $\mathbb{R}^n$ to each point in E. For vector field $F : E \to \mathbb{R}^n$, we let $F_j : E \to \mathbb{R}$ denote the j-th component of F, i.e.

[1] By a common abuse of notation, with $g(N) = O(f(N))$ we indicate the fact that the function $g(N)$ has asymptotic order $O(f(N))$ (similarly for $\Theta(f(N))$ and $\Omega(f(N))$).

for all $\mathbf{d} \in E$ we have:

$$F(\mathbf{d}) = \begin{pmatrix} F_1(\mathbf{d}) \\ F_2(\mathbf{d}) \\ \vdots \\ F_n(\mathbf{d}) \end{pmatrix} = \begin{pmatrix} F_1(d_1, d_2, \ldots, d_n) \\ F_2(d_1, d_2, \ldots, d_n) \\ \vdots \\ F_n(d_1, d_2, \ldots, d_n) \end{pmatrix}$$

We say that $F : E \to \mathbb{R}^n$ is *Lipschitz*, if and only if there exists $K \in \mathbb{R}_{>0}$ such that for all $\mathbf{d}, \mathbf{d}' \in E$ the following holds:

$$||F(\mathbf{d}') - F(\mathbf{d})|| \leq K \cdot ||\mathbf{d}' - \mathbf{d}||$$

A *trajectory* is a function in $I \to \mathbb{R}^n$, where $I \subseteq \mathbb{R}$ is either an open interval containing 0 or an interval of the form $[0, T[$, for $T > 0$.

We let $\mathbf{x}, \mathbf{y}, \ldots$ denote trajectories; the argument in $\mathbb{R}_{\geq 0}$ can often be interpreted as *time*, so that $\mathbf{x}(t)$ is the point in $\mathbb{R}^n$ at time t on the trajectory $\mathbf{x}$, and $\mathbf{x}_j(t) : \mathbb{R}_{\geq 0} \to \mathbb{R}$ its j-th component. For each vector field $F : \mathbb{R}^n \to \mathbb{R}^n$ and trajectory $\mathbf{x} : \mathbb{R}_{\geq 0} \to \mathbb{R}^n$, the function $F \circ \mathbf{x}$ evaluates the vector field along the trajectory $\mathbf{x}$, i.e. it associates the vector $F(\mathbf{x}(t))$ to each time instant t. With the above definitions, an Ordinary Differential Equation (ODE) is an equation of the form

$$\frac{d\mathbf{x}(t)}{dt} = F(\mathbf{x}(t)) \tag{1}$$

Such an equation may admit a solution, given initial condition $\mathbf{x}(0) = \mathbf{x_0}$ for some vector $\mathbf{x_0} \in \mathbb{R}^n$; the solution is a trajectory $\mathbf{y} : I \to \mathbb{R}^n$ satisfying equation (1) for each $t \in I$. This is also known as the *initial value problem*. It is sometimes convenient to write equations like Equation (1) component-wise, as *coupled systems* of ODEs:

$$\begin{aligned} \tfrac{d\mathbf{x}_1(t)}{dt} &= F_1(\mathbf{x}(t)) \\ \tfrac{d\mathbf{x}_2(t)}{dt} &= F_2(\mathbf{x}(t)) \\ &\vdots \\ \tfrac{d\mathbf{x}_n(t)}{dt} &= F_n(\mathbf{x}(t)) \end{aligned}$$

or, equivalently,:

$$\begin{aligned} \tfrac{dx_1(t)}{dt} &= F_1(x_1(t), x_2(t), \ldots, x_n(t)) \\ \tfrac{dx_2(t)}{dt} &= F_2(x_1(t), x_2(t), \ldots, x_n(t)) \\ &\vdots \\ \tfrac{dx_n(t)}{dt} &= F_n(x_1(t), x_2(t), \ldots, x_n(t)) \end{aligned}$$

A well known result for ODE is the Picard Lindelof theorem, stating that if F is a Lipschitz vector field in E and $\mathbf{x}_0 \in E$, then there are $T_1, T_2 > 0$ such that the initial value problem $\frac{d\mathbf{x}(t)}{dt} = F(\mathbf{x}(t)), \mathbf{x}(0) = \mathbf{x_0}$ has a unique solution $\mathbf{x} : I \to E$, where $I =]-T_1, T_2[$ [42].

The notion of Random Variable (RV) plays a major role in the present paper. Let $\mathcal{D}$ be a *finite* or *countable* set; usually we assume $\mathcal{D} \subset \mathbb{R}^n$.

Definition 2.1. A *$\mathcal{D}$-valued Discrete Random Variable* (DRV) $\mathbf{X}$ is fully characterized by its probability mass function $p_{\mathbf{X}} : \mathcal{D} \to [0, 1]$ which associates a probability value to each element in $\mathcal{D}$, i.e. $p_{\mathbf{X}}(\mathbf{d}) =_{\text{def}} \mathbb{P}\{\mathbf{X} = \mathbf{d}\}$. □

In this paper we will use three classes of $\mathbb{N}$-valued DRVs, namely *Poisson*, *Binomial*, and *Multinomial*.

Definition 2.2. A DRV X is a *Poisson* RV with parameter μ if and only if $p_{\mathbf{X}}(n) =_{\text{def}} \frac{\mu^n \cdot e^{-\mu}}{n!}$. □

We let $POI[\mu]$ denote the class of RVs with Poisson distribution and parameter $\mu \in \mathbb{R}_{>0}$. Poisson RVs are useful for modeling arrivals of independent events in a given period of time T when the average number of arrivals in T is known and equal to μ.

Definition 2.3. A DRV X is a *Binomial* RV with parameters k and p if and only if

$$p_{\mathbf{X}}(n) =_{\text{def}} \begin{cases} \left(\begin{array}{c} k \\ n \end{array} \right) \cdot p^n \cdot (1-p)^{(k-n)}, & \text{if } 0 \leq n \leq k \\ \\ 0, & \text{otherwise} \end{cases}$$

□

We let $\mathcal{B}[k, p]$ denote the class of RVs with Binomial distribution and parameters k and p. Binomial RVs are useful for studying the number (n) of successful outcomes of k independent trials, where each trial may have a successful outcome with probability p or a failure, with probability $1 - p$. The *Multinomial* class is a generalization of the Binomial one when each trial may have more than just two outcomes.

Definition 2.4. A $\mathbb{R}$-*valued Continuous Random Variable* (CRV) $\mathbf{X}$ is fully characterized by its cumulative distribution function $F_{\mathbf{X}} : \mathcal{D} \to [0, 1]$ such that $F_{\mathbf{X}} =_{\text{def}} \mathbb{P}\{\mathbf{X} \leq t\}$, for each $t \in \mathbb{R}$. □

A class of $\mathbb{R}$-valued CRVs of interest for the present paper is the *Exponential* class.

Definition 2.5. A CRV X is an *Exponential* RV with parameter λ if and only if

$$F_{\mathbf{X}}(t) =_{\text{def}} \begin{cases} 1 - e^{-\lambda \cdot t}, & \text{if } t \geq 0 \\ \\ 0, & \text{otherwise.} \end{cases}$$

□

We let $EXP[\lambda]$ denote the class of RVs with Exponential distribution and parameter $\lambda \in \mathbb{R}_{>0}$. Exponential distributions are useful for modeling the arrival of an event when the average number of (independent) *events per unit time*, i.e. the *rate* λ, is known. It is worth recalling that exponentially distributed RVs are *memoryless*: for all $t, \delta \in \mathbb{R}_{\geq 0}$, $\mathbb{P}\{\mathbf{X} \leq t + \delta \mid \mathbf{X} > t\} = \mathbb{P}\{\mathbf{X} \leq \delta\}$.

The notion of Stochastic Process, as a suitably indexed family of RVs, naturally extends the concept of RV.

Definition 2.6. A *stochastic* process $\mathbf{X}(i)$ is an I-indexed family $\{\mathbf{X}_i \mid i \in I\}$ of RVs, for given index set I. □

Stochastic processes of interest in the context of this paper are *Poisson* processes, *Discrete Time Markov Chains* and *Continuous Time Markov Chains*.

Definition 2.7. A *Poisson* process $\mathcal{N}(t)$ is an $\mathbb{R}_{\geq 0}$-indexed family $\{\mathcal{N}_t \mid t \in \mathbb{R}_{\geq 0}\}$ of DRVs with Poisson distribution functions. Let λ be the average number of events per unit time; then $\mathcal{N}_t \in POI[\lambda \cdot t]$, for all $t \in \mathbb{R}_{\geq 0}$. □

Definition 2.8. A $\mathcal{D}$-*valued (time homogeneous) Discrete Time Markov Chain* (DTMC) $\mathbf{X}(k)$ is an $\mathbb{N}$-indexed family $\{\mathbf{X_k} \mid k \in \mathbb{N}\}$ of $\mathcal{D}$-valued DRVs such that, for all $k, h \in \mathbb{N}$ and $\mathbf{d}, \mathbf{d'}, \mathbf{d_0}, \ldots, \mathbf{d_{k+1}} \in \mathcal{D}$ the following holds:

- $\mathbb{P}\{\mathbf{X_{k+1}} = \mathbf{d_{k+1}} \mid \mathbf{X_0} = \mathbf{d_0}, \ldots, \mathbf{X_k} = \mathbf{d_k}\} = \mathbb{P}\{\mathbf{X_{k+1}} = \mathbf{d_{k+1}} \mid \mathbf{X_k} = \mathbf{d_k}\}$
- $\mathbb{P}\{\mathbf{X_{k+1}} = \mathbf{d'} \mid \mathbf{X_k} = \mathbf{d}\} = \mathbb{P}\{\mathbf{X_{h+1}} = \mathbf{d'} \mid \mathbf{X_h} = \mathbf{d}\}$ □

Set $\mathcal{D}$ is called the *state space* of the DTMC and its elements are the *states* of the DTMC. The first condition in the definition, known as the *memoryless* property, states that the probability that the process is in a certain state at the next step depends only on its state in the current step and not on its past history (i.e. the specific sequence of states through which the current state has been reached); the second condition, known as the *time homogeneity property*, further characterizes this probability, by postulating that it does not depend on the specific step, but only on the current state. From the definition it follows that a DTMC is completely characterized by its *initial probability distribution* $\pi(0) : \mathcal{D} \to [0,1]$ and its *probabilistic transition matrix* $\mathbf{P} : \mathcal{D}^2 \to [0,1]$ such that, for all $\mathbf{d}, \mathbf{d}' \in \mathcal{D}$, $\mathbf{P}(\mathbf{d}, \mathbf{d}') =_{\text{def}} \mathbb{P}\{\mathbf{X_1} = \mathbf{d}' \mid \mathbf{X_0} = \mathbf{d}\}$. In summary: whenever the DTMC is in state $\mathbf{d}$, the probability that it jumps to state $\mathbf{d}'$ is $\mathbf{P}(\mathbf{d}, \mathbf{d}')$; it does not depend on the on the way $\mathbf{d}$ has been reached (memoryless property), nor on the specific step at which $\mathbf{d}$ has been reached (time homogeneity property).

Definition 2.9. A *$\mathcal{D}$-valued Continuous Time Markov Chain* (CTMC) $\mathbf{X}(t)$ is an $\mathbb{R}_{\geq 0}$-indexed family $\{\mathbf{X_t} \mid t \in \mathbb{R}_{\geq 0}\}$ of $\mathcal{D}$-valued DRVs such that, for all $t, t', \delta, t_{k+1}, t_k, \ldots, t_0 \in \mathbb{R}$, with $t_0 < \ldots < t_k < t_{k+1}$ and $\mathbf{d}, \mathbf{d}', \mathbf{d_0}, \ldots, \mathbf{d_{k+1}} \in \mathcal{D}$ the following holds:

- $\mathbb{P}\{\mathbf{X_{t_{k+1}}} = \mathbf{d_{k+1}} \mid \mathbf{X_{t_0}} = \mathbf{d_0}, \ldots, \mathbf{X_{t_k}} = \mathbf{d_k}\} = \mathbb{P}\{\mathbf{X_{t_{k+1}}} = \mathbf{d_{k+1}} \mid \mathbf{X_{t_k}} = \mathbf{d_k}\}$
- $\mathbb{P}\{\mathbf{X_{t'+\delta}} = \mathbf{d}' \mid \mathbf{X_{t'}} = \mathbf{d}\} = \mathbb{P}\{\mathbf{X_{t+\delta}} = \mathbf{d}' \mid \mathbf{X_t} = \mathbf{d}\}$ □

The two conditions are similar to those for DTMCs, but refer to *continuous time* instead of discrete steps. It can be shown that a CTMC can be seen as a DTMC *enriched* with continuous time information. In particular, a *sojourn* time $\mathbf{X_d}$ is associated with each state $\mathbf{d}$ such that $\mathbf{X_d} \in EXP[\lambda_\mathbf{d}]$. The value $\lambda_\mathbf{d}$ is usually called the *exit rate* of $\mathbf{d}$. Let $\mathbf{P}$ be the probabilistic transition matrix of the DTMC associated with a CTMC as above—the so-called *embedded* DTMC. Then the CTMC is fully characterized by its *initial probability distribution* $\pi(0) : \mathcal{D} \to [0,1]$ and its *infinitesimal generator matrix* $\mathbf{Q} : \mathcal{D}^2 \to \mathbb{R}$, such that, for $\mathbf{d} \neq \mathbf{d}'$, $\mathbf{Q}(\mathbf{d}, \mathbf{d}') =_{\text{def}} \lambda_\mathbf{d} \cdot \mathbf{P}(\mathbf{d}, \mathbf{d}')$ for all $\mathbf{d}, \mathbf{d}' \in \mathcal{D}$, with $\mathbf{Q}(\mathbf{d}, \mathbf{d})$ conventionally set to $-\sum_{\mathbf{d}' \neq \mathbf{d}} \mathbf{Q}(\mathbf{d}, \mathbf{d}')$. In summary: when state $\mathbf{d}$ is reached, the CTMC sojourns in $\mathbf{d}$ for a period of time which is a RV in $EXP[\sum_{\mathbf{d}' \neq \mathbf{d}} \mathbf{Q}(\mathbf{d}, \mathbf{d}')]$, after which it jumps to a different[2] state $\mathbf{d}'$ with probability $\frac{\mathbf{Q}(\mathbf{d}, \mathbf{d}')}{\sum_{\mathbf{d}' \neq \mathbf{d}} \mathbf{Q}(\mathbf{d}, \mathbf{d}')}$. Notice that such a probability and the probability of leaving $\mathbf{d}$ at a certain time do not depend on the way $\mathbf{d}$ has been reached or the time already spent in $\mathbf{d}$ (memoryless property), nor on the specific time instant (time homogeneity property).

Remark 2.1. It is worth pointing out here that, although there is no intrinsic notion of time in DTMCs, in the literature quite often they are regarded as models for systems with a central clock. Each tick of such a clock is considered as corresponding to a step of the DTMC. In the present tutorial we will adhere to such an interpretation of DTMCs unless otherwise stated.

For a detailed introduction to DTMCs and CTMCs, the reader is referred to [1].

3 A low-level language

In this tutorial we will present various aspects of deterministic approximation, which will be illustrated by means of a basic description language. This low-level language is close to the descriptive level of Markov Chains (MCs) and facilitates their compact description, by representing the different possible transitions parametrically.

The language is similar to the one of the PRISM model checker [44]; it resembles also the language used in [45]. The main difference is that our language lacks an explicit treatment of guards, which can be incorporated into the functions associated with transitions (cf. Section 5.6).

[2]Notice that, traditionally, self-loops are not allowed in CTMCs. On the other hand, self-loops naturally arise when higher level modeling languages are used, such as SPAs, for example. It can be shown that self-loops have no impact as far as standard CTMC analysis techniques are concerned, e.g. transient or steady-state analysis techniques, but they play a role when more sophisticated notions and analyses techniques are concerned, e.g. Strong Markovian Bisimulation or satisfaction of *next* (Stochastic) Temporal Logic formulae (see, e.g. [43]).

We chose to use this language rather than SPA because most of the approaches to deterministic approximation are formulated directly in terms of MCs, and their relevant ideas are better captured at this level. This language, in addition, has many traits in common with Stochastic Petri Nets, a modeling formalism in widespread use. Other, higher level languages, like SPAs, can be mapped to this language in a more or less straighforward way. For instance, SPAs with a fluid semantics in terms of ODEs can be easily mapped to this language along the lines of the ODE derivation, also via a formal Structured Operational Semantics style definition (see, for instance, [46, 47]). This is illustrated in Appendix A; similar translations can be defined for probabilistic process algebras and DTMCs.

We first define *qualitative* models, as opposed to DTMCs and CTMCs ones. The latter, in the context of this paper, will be called *quantitative* models, for obvious reasons. They are obtained by refinement of qualitative models.

Definition 3.1. A *qualitative* model is a tuple $\mathcal{X} = (\mathbf{X}, \mathcal{D}, \mathcal{T}, \mathbf{d_0})$, where:

1. $\mathbf{X} = (X_1, \ldots, X_n)$ is a tuple of *variables*.
2. Each X_i takes values in a *finite* or *countable* domain $\mathcal{D}_i$. We usually assume that $\mathcal{D}_i \subset \mathbb{R}$. Hence, $\mathcal{D} = \prod_i \mathcal{D}_i$ is the *state space* of the model.
3. $\mathbf{d_0} \in \mathcal{D}$ is the *initial state* of the model.
4. $\mathcal{T} = \{\tau_1, \ldots, \tau_m\}$ is the set of *transitions*, of the form $\tau_j = (a, \mathbf{s}, \mathbf{t}, f)$, where:
   (a) a is the *label* of the transition;
   (b) $\mathbf{s} \in \mathbb{R}^n$, $\mathbf{s} \geq \mathbf{0}$ is the *pre-vector*, i.e. a vector of non-negative components specifying how many units of each variable are consumed by the transition.
   (c) $\mathbf{t} \in \mathbb{R}^n$, $\mathbf{t} \geq \mathbf{0}$ is the *post-vector*, i.e. a vector of non-negative components specifying how many units of each variable are created by the transition.
   (d) $f : \mathcal{D} \to \mathbb{R}_{\geq 0}$ is the *enabling function* of the transition, such that $f(\mathbf{d}) = 0$ whenever $\mathbf{d} + \mathbf{t} - \mathbf{s} \notin \mathcal{D}$

   The pre-vectors and post-vectors are combined in the *state-change* vector $\mathbf{v} = \mathbf{t} - \mathbf{s}$, giving the net change on each variable due to the transition. Function f gives the enabling status of the transition: τ_j is *enabled* in state $\mathbf{d}$ if and only if $f(\mathbf{d}) > 0$. □

For transition $\tau = (a, \mathbf{s}, \mathbf{t}, f)$ we let a_τ denote the label of τ and similarly for the other components. In the context of this tutorial, it is convenient to make models depend on an index N, which is a parameter associated with a notion of "size" of the model (cf. Section 4). This is used to index a sequence of models of increasing sizes. We denote this by $\mathcal{X}^{(N)}$, where it is intended that each component of $\mathcal{X}^{(N)}$ may depend on N, thus we get $\mathcal{D}^{(N)}$, $f_\tau^{(N)}$, etc.; for the sake of readability, we will often refrain from indicating the parameter N, when there is not potential for confusion. The enabling functions are left under-specified on purpose. Qualitative models give only *behavioral* information; consequently, the only thing which matters is whether $f_\tau(\mathbf{d})$ is positive, in which case transition τ is enabled in state $\mathbf{d}$, the specific value being irrelevant. When *qualitative* models are refined into *quantitative* ones, namely DTMCs or CTMCs, each such function is refined so that it will specify the *probability* of τ, in the DTMC case, or its *rate*, in the CTMC case.

The semantics of the language is straightforward. A *transition system* (TS) is associated with each model as described below.

Definition 3.2. Given model $\mathcal{X} = (\mathbf{X}, \mathcal{D}, \mathcal{T}, \mathbf{d_0})$, let relation $R \subseteq \mathcal{D} \times \mathcal{D}$ be defined by $R =_{\text{def}} \sum_{\tau \in \mathcal{T}} \{(\mathbf{d}, \mathbf{d}') \mid \mathbf{d}, \mathbf{d}' \in \mathcal{D},\, f_\tau(\mathbf{d}) > 0, \mathbf{d}' = \mathbf{d} + \mathbf{v}_\tau\}$. The TS of $\mathcal{X}$, $TS_\mathcal{X}$, is the triple $(S, \to, s_0)$ where $S =_{\text{def}} \{\mathbf{d_0}\} \cup (\mathbf{d_0}\, R^+)$ is the set of states, $\to =_{\text{def}} R \cap S \times S$ is the transition relation, and $s_0 =_{\text{def}} \mathbf{d_0}$ is the initial state. □

When we model a concrete scenario, we are usually describing a certain number of different types of entities or objects, which are present in one or more instances in the model and which interact to produce the model's behavior. For instance, in an ecological model we can have many instances of prey and predators (the two entity types) interacting with one another. Each instance of an entity type is usually in one of its many internal states, determining which interactions the entity can take part in. In our setting, such models are constructed using variables to count how many instances of a given entity type are in a given state,[3] while transitions describe the different interactions and events changing the state of one or more entities (for instance, a predator eating a prey in an ecological network).

It is worth pointing out that the models of the low level language introduced in this section and used in this paper are *system level* models: the global features of the system of interest are explicitly specified in the model. This is different from *compositional* model construction, as, for instance, is the case with SPA, where the behavior of each system component is usually specified separately and the system level description is obtained by composing system components via suitable composition operators. As we already mentioned, models written in high level languages offering compositional features, can be easily translated into the modeling language used here. We finally note that the language is well suited for *clock-asynchronous* or *interleaving* models of system behavior, which are the kind of models we are mainly concerned with in this paper.

Consider a system composed of several entities. The behavior of each entity is characterized by the execution of specific actions. Each such execution, in turn, can be thought of as a state transition of the entity. For the sake of clarity, we call such transitions (entity) *local* transitions. In a *clock-asynchronous* model of such a system, each entity behaves independently from the others, *except* when it *cooperates* with one or more of them, e.g. by means of communication or rendez-vous, as in many process algebras. In a model of such a system written in our language, in most cases, a transition τ, which describes a *global*, system level, transition, corresponds to the execution of *a local transition of a single entity*. The exception are cooperation transitions (or synchronization transitions, as they are also called), which instead involve *all (but only) the cooperating entities*, of which there are usually just a few; all other entities are *not directly affected* by the execution of τ (i.e. they do not change their state). The number of entities participating in the execution of τ is $|\mathbf{s}_\tau|$, which is typically a small number (1 in the case in which no cooperation takes place). However, in Section 8 we present a discussion of clock-synchronous semantics, based on the maximum-parallelism assumption [38, 48]: all entities of the model move in a single step, meaning that a maximal set of transitions is fired.

3.1 Main Example

We introduce an example that will be used throughout the paper to present the various aspects of the techniques involved in continuous approximation. We present a simple model of a worm epidemic in a network of computers. The model describes a network of computers in which each node can be infected by a worm. Once this occurs, the worm remains latent for a while, and then activates. When it is active, it tries to propagate over the network by sending messages to other nodes. After some time, an infected computer can be patched, so that the infection is recovered. We assume that recovered computers can become susceptible of infection again after a while, hence modeling the appearance of a new version of the worm. Non-infected computers may also be patched, but this event happens less frequently. Each node in the network can specifically be infected from two sources, i.e. by the activity of a worm of an infected node or by an external source (for instance, by an email attachment received from outside the network).

In this scenario, we have one single entity class (a computer in the network), the elements of which can be in one of several local states: susceptible (S), exposed (E —this is the latent infection period), actively infected (I), and recovered (R). Consequently, the variables of the model, counting the number of entities in each local state, are collected in the tuple (S, E, I, R),

[3]For instance, in the predator-prey model, suppose prey have three internal states and predators have four internal states, then we will use seven variables: X_1, X_2, X_3 for the internal state of the prey and X_4, X_5, X_6, X_7 for the internal states of the predators.

the global state. Let N denote the number of nodes originally in the network, and assume that no nodes can be added or removed from the network. Then each variable takes values in $\{0, 1, \ldots, N\}$, with the further restriction that $S+E+I+R = N$ in all states; thus: $\mathcal{D}^{(N)} =_{\text{def}} \mathcal{S}^4(\{0, \ldots, N\}, N)$. We turn now to describe the transitions. In this specific model, for the sake of simplicity, we do not explicitly represent pairwise interactions between entities. Instead we give a more abstract system level description. Consequently, the list of transitions is given below, where pre-vectors and post-vectors are defined using the unit vectors $\mathbf{e_S} =_{\text{def}} (1,0,0,0), \mathbf{e_E} =_{\text{def}} (0,1,0,0), \mathbf{e_I} =_{\text{def}} (0,0,1,0)$, and $\mathbf{e_R} =_{\text{def}} (0,0,0,1)$:

- Infection of a susceptible node from an external source:
  $\tau_e = (\text{ext}, \mathbf{e_S}, \mathbf{e_E}, f_e)$, with $f_e(S,E,I,R) =_{\text{def}} \theta(S)$.
- Infection of a susceptible node from a malicious contact with an infected node:
  $\tau_i = (\text{inf}, \mathbf{e_S}, \mathbf{e_E}, f_i)$, with $f_i(S,E,I,R) =_{\text{def}} \theta(S \cdot I)$.
- Activation of the infection in an exposed node:
  $\tau_a = (\text{act}, \mathbf{e_E}, \mathbf{e_I}, f_a)$, with $f_a(S,E,I,R) =_{\text{def}} \theta(E)$.
- Patching of an infected node:
  $\tau_r = (\text{rec}, \mathbf{e_I}, \mathbf{e_R}, f_r)$, with $f_r(S,E,I,R) =_{\text{def}} \theta(I)$.
- Patching of a non-infected node:
  $\tau_p = (\text{rec}, \mathbf{e_S}, \mathbf{e_R}, f_p)$, with $f_p(S,E,I,R) =_{\text{def}} \theta(S)$, and
  $\tau_q = (\text{rec}, \mathbf{e_E}, \mathbf{e_R}, f_q)$, with $f_q(S,E,I,R) =_{\text{def}} \theta(E)$.
- Loss of immunity of a recovered node:
  $\tau_s = (\text{rec}, \mathbf{e_R}, \mathbf{e_S}, f_s)$, with $f_s(S,E,I,R) =_{\text{def}} \theta(R)$.

As enabling functions we chose the expected ones: for example, in a given state (s,e,i,r), transition τ_e is enabled if and only if there are susceptible nodes to get infected, i.e. $s > 0$. The qualitative model of the epidemic example is thus completely specified below, where $(s_0, e_0, i_0, r_0)^{(N)} \in \mathcal{S}^4(\{0, \ldots, N\}, N)$ is a given initial state:

$$\mathcal{E}^{(N)} =_{\text{def}} ((S,E,I,R), \mathcal{S}^4(\{0,\ldots,N\},N), \{\tau_e, \tau_i, \tau_a, \tau_r, \tau_p, \tau_q, \tau_s\}, (s_0, e_0, i_0, r_0)^{(N)})$$

3.2 DTMC and CTMC models and their semantics

As we anticipated above, in order to specify a quantitative model in our language, we start from a qualitative one and we refine transition functions. More specifically, given qualitative model $\mathcal{X}^{(N)} = (\mathbf{X}^{(N)}, \mathcal{D}^{(N)}, \mathcal{T}^{(N)}, \mathbf{d_0}^{(N)})$

- a DTMC model $\mathcal{X}_D^{(N)}$ is obtained by replacing, in each transition $\tau \in \mathcal{T}^{(N)}$, the transition function $f_\tau^{(N)}$ with a transition *probability* function $p_\tau^{(N)} : \mathcal{D}^{(N)} \to [0,1]$, which must define a sub-probability distribution, i.e. for each $\mathbf{d} \in \mathcal{D}^{(N)}$, it is required that: $\sum_{\tau \in \mathcal{T}^{(N)}} p_\tau^{(N)}(\mathbf{d}) \leq 1$.
- a CTMC model $\mathcal{X}_C^{(N)}$ is obtained by replacing, in each transition $\tau \in \mathcal{T}^{(N)}$, the transition function $f_\tau^{(N)}$ with a transition *rate* function $r_\tau^{(N)} : \mathcal{D}^{(N)} \to \mathbb{R}_{\geq 0}$. For each $\mathbf{d} \in \mathcal{D}^{(N)}$ and $\tau \in \mathcal{T}^{(N)}$ such that $r_\tau^{(N)}(\mathbf{d}) > 0$ (i.e. τ is enabled), the rate of τ is $r_\tau^{(N)}(\mathbf{d})$.

DTMC semantics. The definition of a DTMC from a given DTMC model is straightforward. The state space of the DTMC is $\mathcal{D}^{(N)}$. The probabilistic transition matrix $\mathbf{P}$ is defined using the transition probability functions, as follows:

$$\mathbf{P}(\mathbf{d}, \mathbf{d}') = \sum_{\tau \in \mathcal{T}^{(N)} \mid \mathbf{v}_\tau = \mathbf{d}' - \mathbf{d}} p_\tau^{(N)}(\mathbf{d}), \quad \mathbf{d} \neq \mathbf{d}'$$

i.e. we add the probability of all transitions changing state from $\mathbf{d}$ to $\mathbf{d}'$. This is expressed in the summation above by requiring that the state-change vector of a transition τ equals $\mathbf{d}' - \mathbf{d}$. When the summation set is empty, we assume the corresponding probability to be zero. As the transition functions form a sub-probability distribution in each state, we must concentrate the remaining probability mass[4] in the identity transition:

$$\mathbf{P}(\mathbf{d}, \mathbf{d}) = 1 - \sum_{\tau \in \mathcal{T}^{(N)} \mid \mathbf{v}_\tau \neq \mathbf{0}} p_\tau^{(N)}(\mathbf{d}).$$

CTMC semantics. The definition of a CTMC from a given CTMC model follows a similar pattern. In order to define a CTMC, we need to specify the state space and the infinitesimal generator matrix $\mathbf{Q}$. The former is simply the set $\mathcal{D}^{(N)}$, while the latter is defined by adding up all the rates inducing the same state change. Formally:

$$\mathbf{Q}(\mathbf{d}, \mathbf{d}') = \sum_{\tau \in \mathcal{T}^{(N)} \mid \mathbf{v}_\tau = \mathbf{d}' - \mathbf{d}} r_\tau^{(N)}(\mathbf{d}), \quad \mathbf{d} \neq \mathbf{d}'.$$

The diagonal elements of the matrix are defined as customary: $\mathbf{Q}(\mathbf{d}, \mathbf{d}) = -\sum_{\mathbf{d}' \neq \mathbf{d}} \mathbf{Q}(\mathbf{d}, \mathbf{d}')$.

3.3 Example continued

We turn back to the example of Section 3.1, specifying the transition functions in order to define both a DTMC model and a CTMC model. Hence, we will obtain two models, differing only in their transition functions.

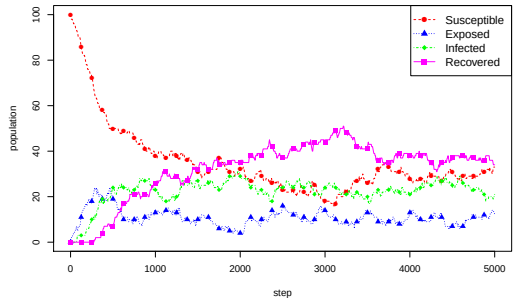
DTMC model of the network epidemic. We remind the reader that, for the DTMC model of our example, we assume that each step has a given, constant duration, say ϵ_N. In order to fully specify the DTMC model $\mathcal{E}_D^{(N)}$ of the network epidemic we have to define the transition probability functions. We do this below, where we also briefly explain them:

- Infection from an external source. We assume that each susceptible node is infected in any step with a constant probability α_e. Hence, the probability that we observe an external infection is α_e times the probability that a susceptible node is involved in the next transition, given that each node can be chosen with the same probability:
  $p_e^{(N)}(S, E, I, R) =_{\text{def}} \alpha_e \cdot \frac{S}{N}$.
- Infection from a malicious contact. A message sent from an infected node will arrive at a susceptible node with probability $\frac{S}{N}$. In this case, the infection will happen with probability α_i. Hence, the probability of observing an infection due to a malicious contact is $p_i^{(N)}(S, E, I, R) =_{\text{def}} \alpha_i \cdot \frac{S}{N} \cdot \frac{I}{N}$.
- Activation of the infection. A worm in an exposed node will activate with probability α_a, so that $p_a^{(N)}(S, E, I, R) =_{\text{def}} \alpha_a \cdot \frac{E}{N}$.

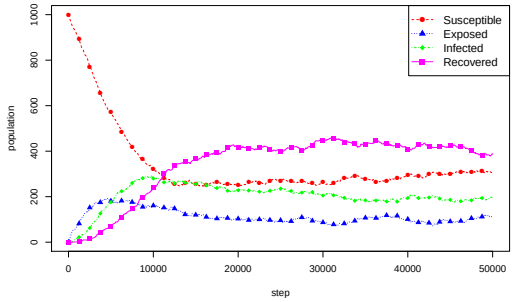
[4] A different possibility here would be that of allowing transition functions to take non-negative values, interpreting them as weights. In this case, the probability transition matrix is obtained by normalizing the weights state by state. Formally, each transition function $f_\tau^{(N)}$ is instantiated with transition *weight* function $w_\tau^{(N)}$; letting $W : \mathcal{D}^{(N)} \to \mathbb{R}_{\geq 0}$ with $W(\mathbf{d}) =_{\text{def}} \sum_{\tau \in \mathcal{T}^{(N)}} w_\tau^{(N)}(\mathbf{d})$, we have:

$$\mathbf{P}(\mathbf{d}, \mathbf{d}') = \sum_{\tau \in \mathcal{T}^{(N)} \mid \mathbf{v}_\tau = \mathbf{d}' - \mathbf{d}} \frac{w_\tau^{(N)}(\mathbf{d})}{W(\mathbf{d})}.$$

We do not need to treat $\mathbf{P}(\mathbf{d}, \mathbf{d})$ differently from other entries in the matrix, because the normalization step already guarantees that $\mathbf{P}$ is a stochastic matrix. However, we chose the other formulation because it is more general, allowing for sub-probabilities. In addition, weights can be incorporated easily by explicitly dividing each p_τ by W directly in the model.



(a) $N = 100$



(b) $N = 1000$

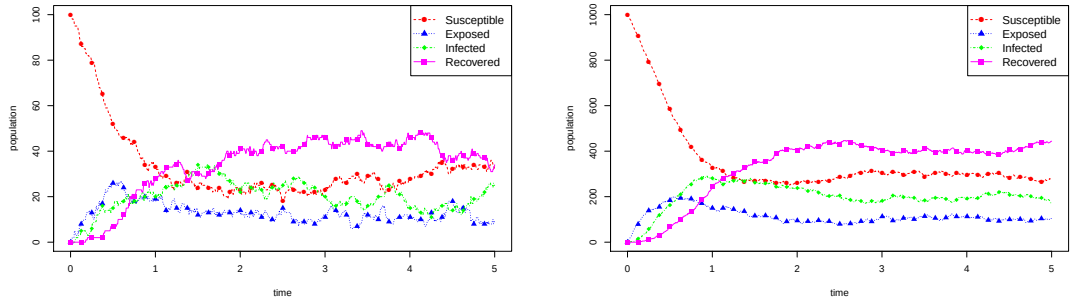
Figure 1: Trajectories of the DTMC model of the network infection example of Section 3.1, for different population levels. The parameters of the model are $\alpha_e = 0.1$, $\alpha_i = 0.2$, $\alpha_a = 0.4$, $\alpha_r = 0.2$, $\alpha_p = 0$, $\alpha_q = 0$, and $\alpha_s = 0.1$, with initial conditions $S_0 = N$, $E_0 = 0$, $I_0 = 0$, and $R_0 = 0$. All the simulations were performed with the following software tools: a dedicated Java implementation for asynchronous DTMC, Dizzy for CTMC [49], a dedicated R implementation for synchronous DTMC. Charts were generated with R.

- Patching of an infected node. Each infected node will be patched with probability α_r, so that $p_r^{(N)}(S,E,I,R) =_{\text{def}} \alpha_r \cdot \frac{I}{N}$.
- Patching of a non-infected node. Non-infected nodes can be patched with probability $\alpha_p, \alpha_q < \alpha_r$, so that $p_p^{(N)}(S,E,I,R) =_{\text{def}} \alpha_p \cdot \frac{S}{N}$ and $p_q^{(N)}(S,E,I,R) =_{\text{def}} \alpha_q \cdot \frac{E}{N}$.
- Loss of immunity. Each recovered node will lose immunity with probability α_s, so that $p_s^{(N)}(S,E,I,R) =_{\text{def}} \alpha_s \cdot \frac{R}{N}$.

The specific values of α constants must be defined in order to study the evolution model. They must satisfy some constraints to make the model compatible with the DTMC semantics based on probabilities. In particular, each α must be in $[0,1]$. Furthermore, $\alpha_e + \alpha_i + \alpha_p \leq 1$, $\alpha_r + \alpha_i \leq 1$, and $\alpha_a + \alpha_q \leq 1$. In Figure 1 we show a simulated trajectory for the DTMC model for two different population levels.

CTMC model of the network epidemic. In order to fully specify the CTMC model $\mathcal{E}_C^{(N)}$ of the network epidemic we have to define the transition rate functions. We obtain expressions which are similar to the DTMC case:

- Infection from an external source. $r_e^{(N)}(S,E,I,R) =_{\text{def}} \lambda_e \cdot S$. Notice that S is *not* divided by N, as for DTMC, as the global rate of an infection from an external source depends on the number of susceptibles, not on their relative frequency (we have to add up the infection rate for each individual).
- Infection from a malicious contact. $r_i^{(N)}(S,E,I,R) =_{\text{def}} \lambda_i \cdot S \cdot \frac{I}{N}$. Here λ_i is the basic rate of infection, multiplied by the total number of ways in which an I-node can contact an S-node. Notice that we do not want the rate of infection to depend on the number of pairs of infected-susceptible nodes, as we cannot reasonably assume that a susceptible node can be contacted by an unbounded number of infected nodes, i.e. the number of messages that can reach a given node tends to remain independent of the network size: each node can know the addresses of a constant number of nodes with respect to N, so that the rate of infection will depend on the *density* of infected nodes $\frac{I}{N}$, rather than on their *total* number.



(a) $N = 100$ (b) $N = 1000$

Figure 2: Trajectories of the CTMC model of the network infection example of Section 3.1, for different population levels. The parameters of the model are $\lambda_e = 1$, $\lambda_i = 2$, $\lambda_a = 4$, $\lambda_r = 2$, $\lambda_p = 0$, $\lambda_q = 0$, and $\lambda_s = 1$, with initial conditions $S_0 = N$, $E_0 = 0$, $I_0 = 0$, and $R_0 = 0$.

- Activation of the infection. $r_a^{(N)}(S, E, I, R) =_{\text{def}} \lambda_a \cdot E$.
- Patching of an infected node. $r_r^{(N)}(S, E, I, R) =_{\text{def}} \lambda_r \cdot I$.
- Patching of a non-infected node. $r_p^{(N)}(S, E, I, R) =_{\text{def}} \lambda_p \cdot S$ and $r_q^{(N)}(S, E, I, R) =_{\text{def}} \lambda_q \cdot E$ (here $\lambda_p, \lambda_q < \lambda_r$).
- Loss of immunity. $r_s^{(N)}(S, E, I, R) = \lambda_s R$.

Two simulation trajectories for the CTMC model are shown in Figure 2, for two different population levels.

Remark 3.1. Notice that $\mathcal{E}_D^{(N)}$ is *not* intended to be the embedded DTMC of $\mathcal{E}_C^{(N)}$. As we will see later on in the paper, for the specific choice of the α and λ parameters of $\mathcal{E}_D^{(N)}$ and $\mathcal{E}_C^{(N)}$, the relationship is intimately connected to *uniformization*. This will also explain the close correspondence of the simulation trajectories shown in Figures 1 and 2.

4 Basics of Approximation

The basic idea of the deterministic approximation of a stochastic model is that the effect of noise becomes more and more irrelevant as the "population size" of the system grows larger: different random individual choices will tend to average out when many individuals are interacting. For instance, in the computer epidemic example, if the size of the network (the number of connected computers) is large, the stochastic fluctuations of the model become irrelevant, and a deterministic description will capture all the relevant features of the dynamics (see Figures 4 and 10). Thus it becomes acceptable to describe the behaviour of the system in terms of continuously evolving variables and replace stochasticity with determinism. Whilst our motivation is generally the compact, *continuous* description of the system in terms of ordinary differential equations, the mathematical underpinning is based on the *deterministic* approximation of the stochastic behaviour. Thus, in this and following sections, we will primarily focus on deterministic approximation, on the understanding that for CTMCs and some DTMCs this leads to a continuous approximation.

In order to give a formal counterpart of the assertion that as the size of the system grows a deterministic description will capture the relevant dynamics, we need two things: a notion of the "size" of the system, and a limit theorem stating that the stochastic model converges to its deterministic counterpart (to be defined, too), as the size of the system goes to infinity. All the

approximation results in literature that we will consider follow this scheme [13, 26, 50, 51, 29, 52, 39].
Practical uses of deterministic approximation simply reason about the deterministic model, as an approximation of the stochastic one, in the hope that the error introduced is small. Estimates of the error may allow a better assessment, yet this is a difficult issue, see Remark 5.2.

Another important common feature of deterministic approximation of stochastic processes is that, in order to properly compare models at different scales, we need to normalize them to a *common scale*. This is intimately related to the notion of *size* of the system. Given a model $\mathcal{X}^{(N)} =_{\text{def}} (\mathbf{X}^{(N)}, \mathcal{D}^{(N)}, \mathcal{T}^{(N)}, \mathbf{d_0}^{(N)})$, indexed by N, we associate with it the size γ_N, which is a positive real number, for each N. As an example, consider again the computer network epidemic model. Here, a meaningful notion of size of the system is the number of nodes in the network, which is a quantity that remains constant throughout the evolution of the epidemic. Therefore, $\gamma_N = N$ for the epidemic model $\mathcal{E}^{(N)}$. In order to compare the evolution for different population levels, we divide each variable by γ_N. When we do this in the epidemic example, we are effectively looking at the fraction of nodes that are in each different state (susceptible S, exposed E, infected I, recovered R), as we divide the number of nodes in a given state by N (obtaining $\bar{S} = \frac{S}{N}$, and so on). Since in this model $S + E + I + R = N$, it holds that $\bar{S} + \bar{E} + \bar{I} + \bar{R} = 1$; thus the normalized variables can be interpreted as a probability distribution. Models in which the notion of size coincides with the total population, which remains constant in time, are quite common in the area of deterministic approximation, and in this case the normalized counting variables are usually referred to as the *occupancy measure* of the system.

If the total population of the model can vary, because there are birth and death events (e.g. the connection or disconnection of a computer in a network), the size γ_N can no longer be equal to the total population, as γ_N is required to be a constant. In such cases, the size of the system is often chosen to be the total initial population. Consequently, the normalized variables cannot be interpreted as a probability distribution, but rather as a count of the number of agents relative to the initial population. Deterministic approximation can be applied in both cases, as it does not depend on the conservation of the total population size.

For simplicity, in this paper we will always consider a constant population of size $\gamma_N = N$. In this setting, by normalizing counting variables with respect to N, the unit increment in the normalized variables is of the order of $\frac{1}{N}$; hence we can think of our process as being defined in a grid of width $\frac{1}{N}$ in $[0, 1]$. As N increases, the step size of this grid becomes smaller and smaller, and the limit process will live in the continuous world. Indeed, in many cases such a limit process is a continuous function, solution of an ordinary differential equation.

The notion of system size is by no means limited to a population level, as variables are not required to take integer values. For instance, in models of biochemical reactions, the system size is usually the volume of the container in which the reactions happen (multiplied by the Avogadro constant). In dividing the variables representing molecular counts by this volume, we obtain the molar concentration of reactants.

We turn now to a more precise description of the formal setting in which we will discuss deterministic limits. Suppose we have a sequence of models $(\mathcal{X}^{(N)})_{N \geq N_0}$, where $N \in \mathbb{N}$, $N_0 > 1$ is a problem-specific value[5], and $\mathcal{X}^{(N)} = (\mathbf{X}^{(N)}, \mathcal{D}^{(N)}, \mathcal{T}^{(N)}, \mathbf{d_0}^{(N)})$. The index N is closely related to the size γ_N of $\mathcal{X}^{(N)}$: we require $\lim_{N\to\infty} \gamma_N = \infty$. The basic step increment at level N is $\delta_N =_{\text{def}} \frac{1}{\gamma_N}$ with $\lim_{N\to\infty} \delta_N = 0$. For instance, in the computer network epidemic model, the range of each variable is $\{0, \ldots, N\}$ and $\delta_N = \frac{1}{N}$.

When we rescale a model according to this recipe, we need to rescale appropriately all the quantities involved; we do this as described below. For all (qualitative) models $\mathcal{X}^{(N)} = (\mathbf{X}^{(N)}, \mathcal{D}^{(N)}, \mathcal{T}^{(N)}, \mathbf{d_0}^{(N)})$, we define a *normalizing* operator $(\overline{\cdot})$ as follows:

- for all $\mathbf{d} \in \mathcal{D}^{(N)}$, $\overline{\mathbf{d}} =_{\text{def}} \delta_N \cdot \mathbf{d}$;
- $\overline{\mathcal{D}}^{(N)} =_{\text{def}} \{\overline{\mathbf{d}} \mid \mathbf{d} \in \mathcal{D}^{(N)}\}$;

[5]For instance, in the epidemic example we would have $N_0 = 2$.

- for all $\tau \in \mathcal{T}^{(N)}$, with $\tau = (a, \mathbf{s}^{(N)}, \mathbf{t}^{(N)}, f^{(N)})$, $\overline{\tau} =_{\text{def}} (a, \overline{\mathbf{s}}^{(N)}, \overline{\mathbf{t}}^{(N)}, \overline{f}^{(N)})$, where:
  - $\overline{\mathbf{s}}^{(N)} =_{\text{def}} \delta_N \cdot \mathbf{s}^{(N)}$;
  - $\overline{\mathbf{t}}^{(N)} =_{\text{def}} \delta_N \cdot \mathbf{t}^{(N)}$;
  - for all $\overline{\mathbf{d}} \in \overline{\mathcal{D}}$, $\overline{f}^{(N)}(\overline{\mathbf{d}}) =_{\text{def}} f^{(N)}(\gamma_N \cdot \overline{\mathbf{d}})$;
- $\overline{\mathcal{T}}^{(N)} =_{\text{def}} \{\overline{\tau} \mid \tau \in \mathcal{T}^{(N)}\}$

Given a model $\mathcal{X}^{(N)} = (\mathbf{X}^{(N)}, \mathcal{D}^{(N)}, \mathcal{T}^{(N)}, \mathbf{d_0}^{(N)})$, the corresponding *normalized* model $\overline{\mathcal{X}}^{(N)}$ is then $\overline{\mathcal{X}}^{(N)} =_{\text{def}} (\overline{\mathbf{X}}^{(N)}, \overline{\mathcal{D}}^{(N)}, \overline{\mathcal{T}}^{(N)}, \overline{\mathbf{d}_\mathbf{0}}^{(N)})$, where $\overline{\mathbf{X}}^{(N)}$ is a fresh new variable tuple of the same size as $\mathbf{X}^{(N)}$. Notice that the following relationship holds between the variables of the normalized model and those of the non-normalized one: $\overline{\mathbf{X}}^{(N)} = \delta_N \cdot \mathbf{X}^{(N)} = \frac{1}{\gamma_N} \cdot \mathbf{X}^{(N)}$. We stress moreover the fact that in the definition of the transition function for the normalized model we are simply changing variables, so that the (non-normalized) function has the same value for the corresponding tuples $\mathbf{d}$ and $\overline{\mathbf{d}} = \delta_N \cdot \mathbf{d}$, i.e. $\overline{f}^{(N)}(\overline{\mathbf{d}}) = f^{(N)}(\gamma_N \cdot \overline{\mathbf{d}}) = f^{(N)}(\mathbf{d})$.

Using the above procedure, we can build a sequence of normalized qualitative models $(\overline{\mathcal{X}}^{(N)})_{N \geq N_0}$ from a sequence of non-normalized ones $(\mathcal{X}^{(N)})_{N \geq N_0}$. With the same procedure, we get a sequence of normalized DTMC or CTMC models, $(\overline{\mathcal{X}}_D^{(N)})_{N \geq N_0}$ or $\{\overline{\mathcal{X}}_C^{(N)}\}_{N \geq N_0}$, from a sequence of DTMC or CTMC models, $(\mathcal{X}_D^{(N)})_{N \geq N_0}$ or $(\mathcal{X}_C^{(N)})_{N \geq N_0}$.

We will now present in detail the deterministic approximation results. We will focus mainly on situations where the deterministic limit is the solution of an ODE. This kind of result is more easily framed for CTMCs, as they already evolve in continuous time. For DTMC models, instead, we have to embed the discrete steps in continuous time, assuming a constant duration for each step and shrinking such a duration as N grows.

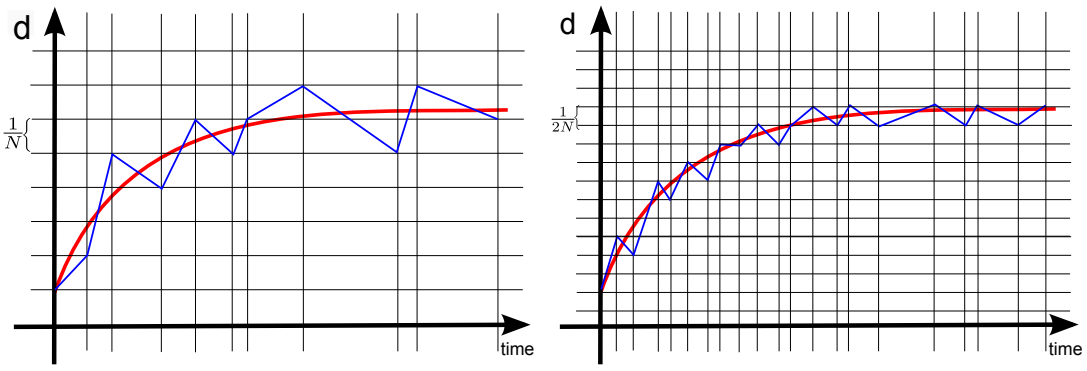
In the following, we will first state the main theorems for the CTMC interpretation. This will be done in the next section, where we will also discuss several examples. Then, we will turn our attention to results for DTMCs. We will interpret DTMCs as *timed* models, where the duration of each step of a DTMC of level N is ϵ_N, again presenting both theorems and examples. In particular, in Section 7 we will show that the asymptotic behaviour of such DTMCs and CTMCs, whenever certain conditions of probabilities and rates are fulfilled, is essentially the same: in the limit of large N, the difference between these two classes of processes is negligible.

After presenting the limit theorems in terms of ODE, we will discuss other aspects of deterministic approximation. First of all, we will present deterministic approximation results for DTMCs with synchronous evolution (Section 8). Afterwards, we will discuss more advanced topics, like the decoupling assumption and limit results for the stationary regime (Section 9).

5 Deterministic approximation for CTMCs

The main results on deterministic approximation for continuous time Markov Chains date back at least to the work of Kurtz [13, 53]. The key concept is a sequence of Markov processes for which both the magnitude of jumps, i.e. $||\mathbf{v}_\tau^{(N)}||$, for all transitions τ, and the average time between consecutive jumps, go to zero. In this situation, fluctuations become negligible, and as time step and magnitude become infinitesimal, we can approximate a discrete jump with a continuous derivative, thus obtaining a deterministic model in terms of ordinary differential equations. This situation is graphically illustrated in Figure 3.

In order for these results to hold, the time step and the jump magnitude must go to zero in a consistent way, i.e. they must scale in the same way with respect to the system size parameter γ_N. One classical scaling condition requirement is *density dependence*, discussed below, which is satisfied by many models, such as models of chemical reactions and epidemic models. However, the conditions under which the limit theorem holds are far more general. Here we choose to follow [51] and [52]. These papers, in fact, present a formulation which is a good compromise



(a) (b)

Figure 3: Intuitive graphical visualization of deterministic approximation theorems. As the step size and the magnitude of jumps go to zero, the impact of fluctuations becomes negligible, hence stochastic trajectories start to look like the smooth deterministic one. Here we represented two situations, halving the step size and magnitude in passing from Figure 3(a) to Figure 3(b).

between generality and simplicity in verifying conditions that must hold for the theorem to work. In particular, [52] extends the validity of the results in [51], also providing an explicit formula for error bounds, which are shown to decay exponentially with N.

We decided to begin the presentation from CTMCs instead of DTMCs for two reasons. Firstly, continuous time must enter the picture in this class of deterministic approximation results, and in CTMCs continuous time is naturally considered, whilst introducing it in DTMCs is slightly less intuitive. Secondly, it is quite easy to deduce the limit results for DTMCs from those for CTMCs [54], and we will pursue this line of reasoning here.

In the following, we first introduce some notation. Then, we will list the scaling conditions that must be verified in order to apply the theorem, which is presented immediately after. No proof will be given (the reader interested in mathematical details can find them in the referenced papers), but we will comment on the proof techniques in Remark 5.1. Finally, we will show how to practically apply the theorem using the main example of the paper and working out the details. Other examples, taken from literature, are also discussed, in order to display some specific issues.

5.1 Notation

Let $\mathcal{X}_C^{(N)} = (\mathbf{X}^{(N)}, \mathcal{D}^{(N)}, \mathcal{T}^{(N)}, \mathbf{d_0}^{(N)})$ be a CTMC model and $\mathbf{X}^{(N)}$ a tuple of n variables. We define some derived quantities of interest, that will be important in defining the deterministic approximation. The point here is that the simple description of the model in terms of transitions allows the computation of these quantities in a straightforward way.

- The *exit rate* function $R : \mathcal{D}^{(N)} \to \mathbb{R}$ associates each state $\mathbf{d}$ to its exit rate, which is the speed at which we see some transition happening when the system is in state $\mathbf{d}$:

$$R(\mathbf{d}) =_{\text{def}} \sum_{\tau \in \mathcal{T}^{(N)}} r_\tau(\mathbf{d}). \tag{2}$$

- The *mean increment* function $\mu : \mathcal{D}^{(N)} \to \mathbb{R}^n$ associates each state $\mathbf{d}$ to the mean increment in $\mathbf{d}$, i.e. the average variation of each variable in a single discrete step of the CTMC. Since in each state $\mathbf{d}$, the probability that the next transition is τ equals $\frac{r_\tau^{(N)}(\mathbf{d})}{R(\mathbf{d})}$, and $\mathbf{v}_\tau^{(N)}$ is the variation of system variables due to transition τ, we get:

$$\mu(\mathbf{d}) = \sum_{\tau \in \mathcal{T}^{(N)}} \mathbf{v}_\tau \frac{r_\tau(\mathbf{d})}{R(\mathbf{d})}. \tag{3}$$

Notice that the $\mu(\mathbf{d})$ may have negative components.

- The *covariance matrix*[6] $\Sigma(\mathbf{d})$ of the increments in state $\mathbf{d}$, i.e. the matrix storing the covariances of the increments between each pair of variables, can be easily shown to be

$$\Sigma_{ij}(\mathbf{d}) = \sum_{\tau \in \mathcal{T}^{(N)}} v_{\tau,i} v_{\tau,j} \frac{r_\tau(\mathbf{d})}{R(\mathbf{d})} - \mu_i(\mathbf{d})\mu_j(\mathbf{d}). \tag{4}$$

  In particular, the *variance of the increments of variable* $X_i^{(N)}$ is

$$\Sigma_{ii}(\mathbf{d}) = \sum_{\tau \in \mathcal{T}^{(N)}} v_{\tau,i}^2 \frac{r_\tau(\mathbf{d})}{R(\mathbf{d})} - \mu_i(\mathbf{d})^2. \tag{5}$$

  Mean increment and the covariance matrix are concepts needed to describe the local average dynamics of the CTMC and the local structure of noise. By imposing suitable conditions on their dependence on the parameter N, which will be discussed in the next section, we can guarantee that noise goes to zero in the limit.

- The *mean dynamics* or *drift* $F : \mathcal{D}^{(N)} \to \mathbb{R}^n$ of the model in state $\mathbf{d}$ is

$$F(\mathbf{d}) =_{\text{def}} R(\mathbf{d})\mu(\mathbf{d}) = \sum_{\tau \in \mathcal{T}^{(N)}} \mathbf{v}_\tau r_\tau(\mathbf{d}). \tag{6}$$

  The concept of mean dynamics, by multiplying the mean increment by the exit rate, essentially captures the (average) local variation of the CTMC with respect to the passing of time.

All the above definitions apply also to the *normalized* model $\overline{\mathcal{X}}_C^{(N)}$ of $\mathcal{X}_C^{(N)}$. Moreover, we will use the notation $R_{\mathcal{X}_C^{(N)}}, \mu_{\mathcal{X}_C^{(N)}}, \Sigma_{\mathcal{X}_C^{(N)}}$, and $F_{\mathcal{X}_C^{(N)}}$ when the specific model $\mathcal{X}_C^{(N)}$ we refer to might be unclear from the context.

5.2 Scaling Assumptions

Let $\overline{\mathcal{X}}_C^{(N)} = (\overline{\mathbf{X}}^{(N)}, \overline{\mathcal{D}}^{(N)}, \overline{\mathcal{T}}^{(N)}, \overline{\mathbf{d}_\mathbf{0}}^{(N)})$ be a normalized model and let us consider the sequence $(\overline{\mathcal{X}}_C^{(N)})_{N \geq N_0}$ with respect to an increasing system size γ_N.

State Space. Let E be a (suitably chosen, see below) *closed* set in $\mathbb{R}^n$ such that $\bigcup_N \overline{\mathcal{D}}^{(N)} \subseteq E$. This is the space in which all processes of the sequence and their deterministic approximation live. We will, however, state the convergence result for a (appropriate) relatively open subset of $S \subseteq E$.[7] This is convenient because we can localize all the scaling assumptions to S. We will further denote $S \cap \overline{\mathcal{D}}^{(N)}$ by $S^{(N)}$.

Convergence of Initial Conditions. We assume that there is some point $\overline{\mathbf{d}_\mathbf{0}} \in S$ such that

$$\lim_{N \to \infty} \overline{\mathbf{d}_\mathbf{0}}^{(N)} = \overline{\mathbf{d}_\mathbf{0}} \tag{7}$$

[6] Formally, the covariance matrix $\Sigma(\mathbf{x})$ is defined by $\Sigma_{ij}(\mathbf{x}) = \mathbb{E}[(\tilde{X}_i(k+1) - \tilde{X}_i(k))(\tilde{X}_j(k+1) - \tilde{X}_j(k)) \mid \tilde{X}_i(k) = x_i, \tilde{X}_j(k) = x_j] - \mathbb{E}[(\tilde{X}_i(k+1) - \tilde{X}_i(k)) \mid \tilde{X}_i(k) = x_i]\mathbb{E}[(\tilde{X}_j(k+1) - \tilde{X}_j(k)) \mid \tilde{X}_j(k) = x_j]$, where $\tilde{\mathbf{X}}(\mathbf{k})$ is the embedded DTMC [1] associated to the CTMC $\mathbf{X}(t)$.

[7] A relatively open subset S is an open set in the subspace topology on E. The subspace topology is defined in the following way: the open sets of E are obtained by intersecting E with open sets $U \subseteq \mathbb{R}^N$. Therefore, S is relatively open in E if and only if there exists an open set $U \subseteq \mathbb{R}^N$ such that $S = E \cap U$. Notice that $E = E \cap \mathbb{R}^n$ is relatively open in E.

Convergence of Drift. We assume that the drift vectors behave coherently in the limit, i.e. we assume that there is a Lipschitz vector field $F : E \to \mathbb{R}^n$ such that the drift $F_{\overline{\mathcal{X}}_C^{(N)}}$ converges uniformly to F. This means that

$$\lim_{N\to\infty} \sup_{\overline{\mathbf{d}}\in S^{(N)}} ||F_{\overline{\mathcal{X}}_C^{(N)}}(\overline{\mathbf{d}}) - F(\overline{\mathbf{d}})|| = 0. \tag{8}$$

We further assume that all the trajectories $\mathbf{x}(t)$ which are solutions of the initial value problem $\frac{d\mathbf{x}(t)}{dt} = F(\mathbf{x}(t))$, when $\overline{\mathbf{d}}_\mathbf{0} \in E$, remain in E (for all time instants in which $\mathbf{x}$ is defined). This can be accomplished by choosing E appropriately. Note that the supremum is taken in $S^{(N)}$, i.e. we are just requiring convergence in $S^{(N)}$, not in E. This means that only what happens in S is relevant for the validity of the theorem. In particular, F is required to be Lipschitz only in S: its behaviour outside S is not relevant[8]. In our setting, it is usually easier to focus attention on single transitions, proving for each transition τ the existence of a Lipschitz function (in S) $f_\tau : E \to \mathbb{R}$ such that $\overline{r}_\tau^{(N)}$ converges uniformly to f_τ on S.

Convergence to Zero of Noise. The other hypotheses consider the dependence of the exit rate and of the jump size on N, plus a condition on their cross-relation (which is essentially a condition on the variance). Usually, the scaling conditions are such that the exit rate goes to infinity with N, while the step size goes to zero, both at the same speed (typically linearly with respect to N). Here, however, we consider more general scaling laws, that subsume the standard ones and allow the application of mean field results to a larger class of systems. We require three things:

1. The exit rate is bounded for each N, i.e. there exists $\Lambda_N \in \mathbb{R}_{\geq 0}, \Lambda_N < \infty$, such that

$$\sup_{\overline{\mathbf{d}}\in S^{(N)}} R_{\overline{\mathcal{X}}_C^{(N)}}(\overline{\mathbf{d}}) = \Lambda_N \tag{9}$$

   usually, $\lim_{N\to\infty} \Lambda_N = \infty$, and the frequency of jumps increases with N.

2. The magnitude of jumps goes to zero. More precisely, there exists $J_N \in \mathbb{R}_{\geq 0}$, such that

$$\max_{\tau\in\mathcal{T}^{(N)}} ||\mathbf{v}_\tau^{(N)}|| = J_N. \tag{10}$$

   Moreover $\lim_{N\to\infty} J_N = 0$ and, in particular, J_N is $O(N^{-1})$; this means that the magnitude of jumps goes to zero at least as quickly as N^{-1}.

3. The scaling of jump magnitude and exit rate must be compatible, according to the following condition:

$$J_N^2 \Lambda_N \text{ is } O(N^{-1}). \tag{11}$$

   This is essentially a condition on noise, and it enforces that the variance of the system goes to zero.

In our setting, the previous conditions can be simplified whenever the non-normalized increments are independent of N, i.e. whenever, for each transition τ, there is a vector $\mathbf{v}_\tau$ such that $\mathbf{v}_\tau^{(N)} = \mathbf{v}_\tau$ for all N. In particular, the second condition can be restated in terms of system size γ_N. We recall, in fact, that, by definition, $\delta_N = \frac{1}{\gamma_N}$ and $\overline{\mathbf{v}}_\tau^{(N)} = \delta_N \cdot \mathbf{v}_\tau^{(N)} = \delta_N \cdot \mathbf{v}_\tau$; consequently, we have that $||\overline{\mathbf{v}}_\tau^{(N)}|| = \delta_N \cdot ||\mathbf{v}_\tau||$, where $||\mathbf{v}_\tau||$ is $O(1)$. Hence $J_N = \delta_N \cdot J$, where $J = \max_{\tau\in\mathcal{T}^{(N)}} ||\mathbf{v}_\tau||$. Therefore, the second condition is satisfied as soon as δ_N converges to 0, with order $O(N^{-1})$, while the third condition can be restated as $\delta_N^2 \Lambda_N = O(N^{-1})$. In practical applications, one usually has that $\Lambda_N = \Theta(N)$ and $\gamma_N = \Theta(N)$, so that $\delta_N = \Theta(N^{-1})$, hence condition (11) holds straightforwardly.

[8] It is always possible to redefine F outside $S^{(N)}$ to make it Lipschitz in E.

Remark 5.1. The scaling assumptions we are taking into account are fairly general, and encompass many scaling laws found in the literature. For instance, in [51], the author assumes that $\Lambda_N = O(N)$ and he requires the following scaling condition on the variance of all variables:

$$\sup_{\overline{\mathbf{d}} \in S^{(N)}} \sum_{i=1}^{n} (\Sigma_{\overline{\mathcal{X}}_C^{(N)}})_{ii}(\mathbf{d}) + ||\mu_{\overline{\mathcal{X}}_C^{(N)}}(\overline{\mathbf{d}})||^2 = O(N^{-2}).$$

Provided $\mathbf{v}_\tau^{(N)}$ does not depend on N, this can be rewritten in our setting as

$$\sup_{\overline{\mathbf{d}} \in S^{(N)}} \sum_{\tau \in \mathcal{T}^{(N)}} \delta_N ||\mathbf{v}_\tau||^2 \frac{\overline{r}_\tau^{(N)}(\overline{\mathbf{d}})}{R_{\overline{\mathcal{X}}_C^{(N)}}(\overline{\mathbf{d}})} = O(N^{-2})$$

which essentially amounts to requiring that $\delta_N = O(N^{-1})$. Notice that, in this case, condition (11) is satisfied. In particular, this argument justifies the fact that we refer to (11) as a condition on noise.

Remark 5.2 (Density Dependence). A typical scaling law of rate functions, manifested in chemical reaction models and most epidemic models, is *density dependence*.
Given a (non-normalized) CTMC model $\mathcal{X}_C^{(N)} = (\mathbf{X}^{(N)}, \mathcal{D}^{(N)}, \mathcal{T}^{(N)}, \mathbf{d_0}^{(N)})$, this condition requires the following:

- the system size grows linearly with N, i.e. $\gamma_N = \Theta(N)$, and
- for each N and $\tau \in \mathcal{T}^{(N)}$:
  - there is a vector $\mathbf{v}_\tau$ such that $\mathbf{v}_\tau^{(N)} = \mathbf{v}_\tau$ (so increments are independent of N).
  - there is a function $g_\tau : E \to \mathbb{R}$ such that the rate function $r_\tau^{(N)} : \mathcal{D}^{(N)} \to \mathbb{R}$ scales with system size as $r_\tau^{(N)}(\mathbf{d}) = \gamma_N \cdot g_\tau(\frac{1}{\gamma_N} \cdot \mathbf{d})$, for all $\mathbf{d} \in \mathcal{D}^{(N)}$.

Notice that neither $\mathbf{v}_\tau$ nor g_τ depend on N, that g_τ takes values on the state space of the *normalized* model $\overline{\mathcal{X}}_C^{(N)}$, and that $r_\tau^{(N)}$ is the rate function on τ in the *non-normalized* model. Under these hypotheses, if we consider the normalized model $\overline{\mathcal{X}}_C^{(N)} =_{\text{def}} (\overline{\mathbf{X}}^{(N)}, \overline{\mathcal{D}}^{(N)}, \overline{\mathcal{T}}^{(N)}, \overline{\mathbf{d}}_{\mathbf{0}}{}^{(N)})$ of $\mathcal{X}_C^{(N)}$, we see that the following facts hold:

- The rate functions for the normalized system are $\overline{r}_\tau^{(N)}(\overline{\mathbf{d}}) = \gamma_N \cdot g_\tau(\overline{\mathbf{d}})$, and hence the exit rate is $R_{\overline{\mathcal{X}}_C^{(N)}}(\overline{\mathbf{d}}) = \Theta(\gamma_N) = \Theta(N)$ and condition (9) is satisfied, whenever functions g_τ are bounded in E (or in S).
- $F_{\overline{\mathcal{X}}_C^{(N)}}(\overline{\mathbf{d}}) = \sum_{\overline{\tau} \in \overline{\mathcal{T}}^{(N)}} \overline{\mathbf{v}}_{\overline{\tau}}^{(N)} \cdot \overline{r}_{\overline{\tau}}^{(N)}(\overline{\mathbf{d}}) = \sum_{\tau \in \mathcal{T}^{(N)}} \delta_N \cdot \mathbf{v}_\tau \cdot \gamma_N \cdot g_\tau(\overline{\mathbf{d}}) = \sum_{\tau \in \mathcal{T}^{(N)}} \mathbf{v}_\tau \cdot g_\tau(\overline{\mathbf{d}})$. Therefore, the mean dynamics are the same for each N, and, by letting vector field $F : E \to \mathbb{R}^n$ be defined as follows
$$F(\overline{\mathbf{d}}) =_{\text{def}} \sum_{\tau \in \mathcal{T}^{(N)}} \mathbf{v}_\tau \cdot g_\tau(\overline{\mathbf{d}}),$$
  Condition (8) is satisfied provided F is Lipschitz in E (or in S).
- As $\gamma_N = \Theta(N)$, Conditions (10) and (11) are satisfied.

Therefore, if the density dependence condition holds, one has simply to check the convergence of the initial conditions and the Lipschitzness of F to ensure Conditions (7-11).

5.3 Deterministic Approximation Theorem

We now state the main approximation theorems for CTMCs [13, 26, 50, 51, 54]. Consider a sequence of normalized CTMC models $(\overline{\mathcal{X}}_C^{(N)})_{N\geq N_0}$ with $\overline{\mathcal{X}}_C^{(N)} = (\overline{\mathbf{X}}^{(N)}, \overline{\mathcal{D}}^{(N)}, \overline{\mathcal{T}}^{(N)}, \overline{\mathbf{d}_\mathbf{0}}^{(N)})$, and denote by $\overline{\mathbf{X}}^{(N)}(t)$ the continuous-time Markov process associated with $\overline{\mathcal{X}}_C^{(N)}$. Furthermore, denote by $\overline{\mathbf{x}}(t)$ the solution of the initial value problem $\frac{d\overline{\mathbf{x}}(t)}{dt} = F(\overline{\mathbf{x}}(t))$, $\overline{\mathbf{x}}(0) = \overline{\mathbf{d}_\mathbf{0}}$, where F is as in (8). We will also need the following definitions: the *exit time* from S of the ODE solution $\overline{\mathbf{x}}(t)$ is $\zeta(S) =_{\text{def}} \inf\{t \geq 0 \mid \overline{\mathbf{x}}(t) \notin S\}$, and the *exit time* from S of the Markov processes $\overline{\mathbf{X}}^{(N)}(t)$ is $\zeta^{(N)}(S) =_{\text{def}} \inf\{t \geq 0 \mid \overline{\mathbf{X}}^{(N)}(t) \notin S\}$.

Theorem 5.1 (Deterministic approximation of CTMCs)**.** *Let the sequence* $(\overline{\mathbf{X}}^{(N)}(t))_{N\geq N_0}$ *of Markov processes and* $\overline{\mathbf{x}}(t)$ *be defined as above, and assume Conditions (7-11) apply. Then, for any* finite *time horizon* $T < \zeta(S)$*, it holds that:*

1. $\lim_{N\to\infty} \mathbb{P}\{\zeta^{(N)}(S) < T\} = 0$*;*
2. *for all* $\varepsilon \in \mathbb{R}_{>0}$*,* $\lim_{N\to\infty} \mathbb{P}\{\sup_{0\leq t\leq T} ||\overline{\mathbf{X}}^{(N)}(t) - \overline{\mathbf{x}}(t)|| > \varepsilon\} = 0$*.*

Furthermore, both $\mathbb{P}\{\zeta^{(N)}(S) < T\}$ *and* $\mathbb{P}\{\sup_{0\leq t\leq T} ||\overline{\mathbf{X}}^{(N)}(t) - \overline{\mathbf{x}}(t)|| > \varepsilon\}$ *are* $O(e^{-N})$ ■

This theorem states that, whenever the appropriate assumptions are in place, the probability of observing a significant difference between *any* single trajectory of the Markov process and the solution of the ODE goes to zero as N grows. Essentially, for large N *we cannot distinguish the individual trajectories of the CTMCs from the trajectory of the ODE.* The validity of this theorem is for any finite time horizon T, chosen so that the solution of the ODE remains in S in $[0, T]$. Nothing is said about asymptotic behavior. We will postpone the discussion of the behavior as T grows to ∞ to Section 9. We now give a slightly stronger result than the previous theorem, for the case in which the ODE solution leaves S in finite time, i.e. $\zeta(S) < \infty$.

Theorem 5.2 (Deterministic approximation for CTMCs with finite exit times)**.** *Let the sequence* $(\overline{\mathbf{X}}^{(N)}(t))_{N\geq N_0}$ *of Markov processes and* $\overline{\mathbf{x}}(t)$ *be defined as before, and assume Conditions (7-11) are in force. Then, if* $\zeta(S) < \infty$*, it holds that, for all* $\varepsilon \in \mathbb{R}_{>0}$*:*

1. $\lim_{N\to\infty} \mathbb{P}\{\sup_{0\leq t\leq \zeta(S)} ||\overline{\mathbf{X}}^{(N)}(\min\{t, \zeta^{(N)}(S)\}) - \overline{\mathbf{x}}(t)|| > \varepsilon\} = 0$*;*
2. $\lim_{N\to\infty} \mathbb{P}\{||\zeta^{(N)}(S) - \zeta(S)|| > \varepsilon\} = 0$*.*

Furthermore, both probabilities $\mathbb{P}\{\sup_{0\leq t\leq \zeta(S)} ||\overline{\mathbf{X}}^{(N)}(\min\{t, \zeta^{(N)}(S)\}) - \overline{\mathbf{x}}(t)|| > \varepsilon\}$ *and* $\mathbb{P}\{||\zeta^{(N)}(S) - \zeta(S)|| > \varepsilon\}$ *are* $O(e^{-N})$*.* ■

This second theorem states that the CTMC trajectories are indistinguishable from the ODE solution provided that we look at what happens while the ODE trajectory is in S. Furthermore, it states that the exit time from S of the CTMCs converges to that of the ODE, meaning that for large N we can estimate $\zeta^{(N)}(S)$ by $\zeta(S)$.

*Remark** 5.1. There are several proof techniques that have been used to prove these two theorems. Probably the most used approach is based on martingale theory [54]. Martingales are particular stochastic processes, whose conditional expected value at future times, given the present, is equal to the observed value at the present. Specifically, in this context [51, 52] one can construct a martingale as the difference between the CTMC and its accumulator (i.e. the process "accumulating" the mean increments, computed according to the mean dynamics). After showing that this is indeed a martingale with zero mean, the theorem is proved by applying standard martingale inequalities, like Doob's inequality or an exponential martingale inequality [52]. In this way, one

shows that the CTMC, in the limit, behaves like a deterministic process, which is the solution of an integral equation. Then, one proves that this is indeed the solution of the fluid ODE, by applying the Gronwall inequality [54], a functional inequality used to prove properties of ODE solutions in worst case scenarios (see also Remark 5.2).

Another approach has been presented in [50] which works for density dependent rates, using a representation of CTMCs in terms of Poisson processes, counting how many times each transition fired up to time t, and exploiting properties of Poisson processes to prove the result.

A different proof technique, having a broader spectrum of applications yet using more advanced mathematical tools, is based on infinitesimal generators [13, 50, 54], which are operators on functional spaces (usually the space of bounded measurable functions or the space of continuous functions vanishing at infinity) encoding information about the expected value of the CTMC for any function of the space. There is a large body of theory showing connections between properties of the stochastic processes and properties of their generators. In particular, convergence of stochastic processes is equivalent to the convergence of their infinitesimal generators, which may be easier to prove [13, 50, 54].

*Remark** 5.2. A useful companion of limit theorems are bounds on the error introduced in replacing a stochastic process by its deterministic limit. In Theorems 5.1 and 5.2, error bounds are provably exponentially decreasing in N. Despite this, they are still quite loose [52, 55]. In fact, the error bounds depend (doubly) exponentially on the time horizon T, so that, despite going exponentially to zero as N grows, they tend to be too large for any practical application. More precisely, error bounds have the following form:

$$\mathbb{P}\{\sup_{0\leq t\leq T} ||\overline{\mathbf{X}}^{(N)}(t) - \overline{\mathbf{x}}(t)|| > \varepsilon\} \leq 2 \cdot n \cdot e^{W},$$

where W stands for $-\frac{\varepsilon^2}{18} \cdot \frac{1}{T \cdot e^{2 \cdot K \cdot T}} \frac{1}{A_N}$, K is the Lipschitz constant of F and $A_N = \max\{e, (\delta_N \cdot \Lambda_N)^{-1}\} \cdot \delta_N^2 \cdot \Lambda_N$, so that $\frac{1}{A_N} = \Omega(N)$, if conditions (7) to (11) are in force. This is related to the fact that Theorem 5.1 holds for any arbitrary trajectory of the limit ODE. In particular, it holds for trajectories that are unstable, meaning that even small fluctuations can lead exponentially far away from them. Indeed, the appearance of T as an exponent is typical of the use of Gronwall's inequality, which is used to prove global properties of ODE solutions, including unstable trajectories.

In practical applications, however, we usually have to deal with trajectories with much nicer stability properties, hence the behaviour of the deterministic approximation in terms of error bounds is much better than the one predicted by the worst-case error bounds. Proving bounds for such situations is an active and challenging research area [55].

5.4 Back to the Main Example

We return to the main example of the computer network epidemic from Section 3.1. Recall that we have a fixed number of nodes, N, in the network, each being in one of four different states: susceptible S, exposed E, infected I, and recovered R. The interactions and their rates are described in Section 3.3.

In order to apply the deterministic approximation theorems, we need to check Conditions (7-11). First of all, we need to define a notion of system size γ_N and we need to construct a sequence of models for divergent γ_N. In this case, we can simply set $\gamma_N = N$, so that we are studying the behavior of the system for large populations. Consequently, the sequence of non-normalized models we are interested in is $(\mathcal{E}_C^{(N)})_{N \geq N_0}$, for $\mathcal{E}_C^{(N)}$ as defined in Section 3.3.

The sequence of normalized models is $(\overline{\mathcal{E}}_C^{(N)})_{N \geq N_0}$, where $\overline{\mathcal{E}}_C^{(N)}$ is derived according to the recipe described in Section 4. Notice that $\delta_N = \frac{1}{N}$ and $\overline{\mathcal{D}}^{(N)} \subseteq \mathcal{S}^4([0,1],1)$ for all N, i.e. $\overline{S}+\overline{E}+\overline{I}+\overline{R} = 1$. Consequently, we choose the state space E by $E =_{\text{def}} \mathcal{S}^4([0,1],1)$. For convenience, we explicitly report below the definition of the normalized rate functions as derived from the definition of

normalization:

$$\begin{array}{lll}
\overline{r}_e^{(N)}(\overline{S},\overline{E},\overline{I},\overline{R}) & =_{\text{def}} & N\cdot\lambda_e\cdot\overline{S}\\
\overline{r}_i^{(N)}(\overline{S},\overline{E},\overline{I},\overline{R}) & =_{\text{def}} & N\cdot\lambda_i\cdot\overline{S}\cdot\overline{I}\\
\overline{r}_a^{(N)}(\overline{S},\overline{E},\overline{I},\overline{R}) & =_{\text{def}} & N\cdot\lambda_a\cdot\overline{E}\\
\overline{r}_r^{(N)}(\overline{S},\overline{E},\overline{I},\overline{R}) & =_{\text{def}} & N\cdot\lambda_r\cdot\overline{I}\\
\overline{r}_p^{(N)}(\overline{S},\overline{E},\overline{I},\overline{R}) & =_{\text{def}} & N\cdot\lambda_p\cdot\overline{S}\\
\overline{r}_q^{(N)}(\overline{S},\overline{E},\overline{I},\overline{R}) & =_{\text{def}} & N\cdot\lambda_q\cdot\overline{E}\\
\overline{r}_s^{(N)}(\overline{S},\overline{E},\overline{I},\overline{R}) & =_{\text{def}} & N\cdot\lambda_s\cdot\overline{R}
\end{array}$$

We have now to check whether the conditions for the application of the deterministic approximation theorems hold for the example, given that $\gamma_N = \Theta(N)$. We first show that the rate functions of the non-normalized model $\mathcal{E}_C^{(N)}$ are density dependent, then we show convergence of the initial conditions, and finally Lipschitzness of the mean dynamics.

For density dependence, we define, for $\sigma \in \{e,i,a,r,p,q,s\}$, function $g_\sigma : E \to \mathbb{R}$ as follows:

$$\begin{array}{lll}
g_e(d_1,d_2,d_3,d_4) & =_{\text{def}} & \lambda_e\cdot d_1\\
g_i(d_1,d_2,d_3,d_4) & =_{\text{def}} & \lambda_i\cdot d_1\cdot d_3\\
g_a(d_1,d_2,d_3,d_4) & =_{\text{def}} & \lambda_a\cdot d_2\\
g_r(d_1,d_2,d_3,d_4) & =_{\text{def}} & \lambda_r\cdot d_3\\
g_p(d_1,d_2,d_3,d_4) & =_{\text{def}} & \lambda_p\cdot d_1\\
g_q(d_1,d_2,d_3,d_4) & =_{\text{def}} & \lambda_q\cdot d_2\\
g_s(d_1,d_2,d_3,d_4) & =_{\text{def}} & \lambda_s\cdot d_4
\end{array}$$

It is easy to check that, for all $0 < N \in \mathbb{N}$, $(S,E,I,R) \in \mathcal{S}^4(\{0,\ldots,,N\},N)$, and $\sigma \in \{e,i,a,r,p,q,s\}$ it holds that: $r_\sigma^{(N)}(S,E,I,R) = N\cdot g_\sigma(\frac{1}{N}\cdot(S,E,I,R))$, i.e. rate functions $r_\sigma^{(N)}$ are density dependent.

As far as convergence of initial conditions is concerned, let $\overline{\mathbf{d_0}}$ be an initial point in the state space E. We define $\overline{\mathbf{d_0}}^{(N)}$ as follows: $\overline{\mathbf{d_0}}^{(N)} =_{\text{def}} \frac{1}{N}\cdot\lfloor N\cdot\overline{\mathbf{d_0}}\rfloor$. Clearly, $\lim_{N\to\infty}\overline{\mathbf{d_0}}^{(N)} = \overline{\mathbf{d_0}}$.

As regards Lipschitzness, let us first compute the drift $F_{\overline{\mathcal{E}}_C^{(N)}}$ for our model. By definition, we get:

$$F_{\overline{\mathcal{E}}_C^{(N)}}\begin{pmatrix}\overline{S}\\ \overline{E}\\ \overline{I}\\ \overline{R}\end{pmatrix} =_{\text{def}} \sum_{\tau\in\{\tau_e,\tau_i,\tau_a,\tau_r,\tau_p,\tau_q,\tau_s\}} \overline{\mathbf{v}}_\tau^{(N)}\cdot\overline{r}_\tau^{(N)}(\overline{S},\overline{E},\overline{I},\overline{R}).$$

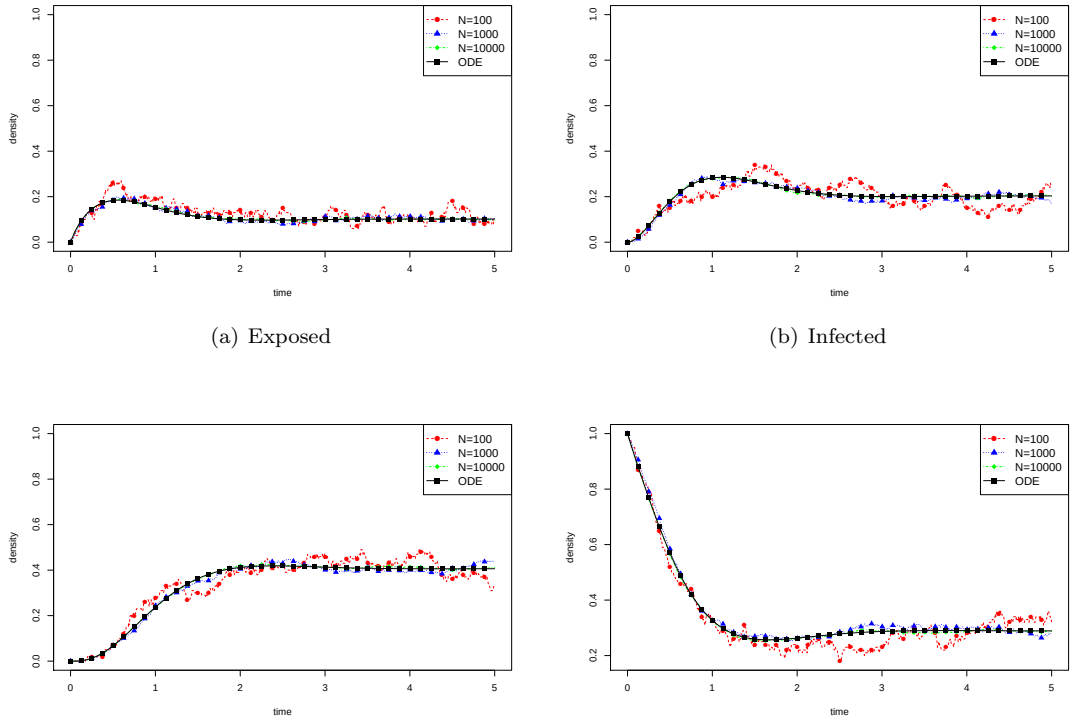
Using the definitions for $\overline{r}_\sigma^{(N)}$ and recalling that $\overline{\mathbf{v}}_\tau^{(N)} = \frac{1}{N}\cdot\mathbf{v}_\tau^{(N)} = \frac{1}{N}\cdot\mathbf{v}_\tau^{(N_0)}$ we get

$$\begin{aligned}
F_{\overline{\mathcal{E}}_C^{(N)}}\begin{pmatrix}\overline{S}\\ \overline{E}\\ \overline{I}\\ \overline{R}\end{pmatrix} &= \begin{pmatrix}-1\\1\\0\\0\end{pmatrix}\lambda_e\bar{S} + \begin{pmatrix}-1\\1\\0\\0\end{pmatrix}\lambda_i\bar{S}\bar{I} + \begin{pmatrix}0\\-1\\1\\0\end{pmatrix}\lambda_a\bar{E} + \begin{pmatrix}0\\0\\-1\\1\end{pmatrix}\lambda_r\bar{I} +\\
&+ \begin{pmatrix}-1\\0\\0\\1\end{pmatrix}\lambda_p\bar{S} + \begin{pmatrix}0\\-1\\0\\1\end{pmatrix}\lambda_q\bar{E} + \begin{pmatrix}1\\0\\0\\-1\end{pmatrix}\lambda_s\bar{R}
\end{aligned}$$

that is,

$$F_{\overline{\mathcal{E}}_C^{(N)}}\begin{pmatrix}\overline{S}\\ \overline{E}\\ \overline{I}\\ \overline{R}\end{pmatrix} = \begin{pmatrix}-\lambda_e\cdot\overline{S} - \lambda_i\cdot\overline{S}\cdot\overline{I} - \lambda_p\cdot\overline{S} + \lambda_s\cdot\overline{R}\\ \lambda_e\cdot\overline{S} + \lambda_i\cdot\overline{S}\cdot\overline{I} - \lambda_a\cdot\overline{E} - \lambda_q\cdot\overline{E}\\ \lambda_a\cdot\overline{E} - \lambda_r\cdot\overline{I}\\ \lambda_r\cdot\overline{I} + \lambda_p\cdot\overline{S} + \lambda_q\cdot\overline{E} - \lambda_s\cdot\overline{R}\end{pmatrix}$$

We note that the drift does *not* depend on N; moreover it is Lipschitz (since it is a continuously differentiable function in a compact set). Thus we can take the drift itself as the vector field F for the convergence condition (8).



(a) Exposed

(b) Infected

(c) Recovered

(d) Susceptible

Figure 4: Comparison between the solution of the ODE and stochastic trajectories for increasing population sizes. Parameters of the model are $\lambda_e = 0.2$, $\lambda_i = 0.2$, $\lambda_a = 0.4$, $\lambda_r = 0.2$, $\lambda_p = 0$, $\lambda_q = 0$, and $\lambda_s = 0.2$, while initial conditions of ODE are $\overline{s}_0 = 1$, $\overline{e}_0 = 0$, $\overline{i}_0 = 0$, and $\overline{r}_0 = 0$.

The limit set of ODEs of the sequence is easily constructed accordingly:

$$\begin{aligned}
\frac{d\overline{s}(t)}{dt} &= -\lambda_e \cdot \overline{s}(t) - \lambda_i \cdot \overline{s}(t) \cdot \overline{i}(t) - \lambda_p \cdot \overline{s}(t) + \lambda_s \cdot \overline{r}(t) \\
\frac{d\overline{e}(t)}{dt} &= \lambda_e \cdot \overline{s}(t) + \lambda_i \cdot \overline{s}(t) \cdot \overline{i}(t) - \lambda_a \cdot \overline{e}(t) - \lambda_q \cdot \overline{e}(t) \\
\frac{d\overline{i}(t)}{dt} &= \lambda_a \cdot \overline{e}(t) - \lambda_r \cdot \overline{i}(t) \\
\frac{d\overline{r}(t)}{dt} &= \lambda_r \cdot \overline{i}(t) + \lambda_p \cdot \overline{s}(t) + \lambda_q \cdot \overline{e}(t) - \lambda_s \cdot \overline{r}(t)
\end{aligned} \tag{12}$$

Note that this set of ODEs is defined in the state space E, and its trajectories can never leave it, as $\overline{s}(t)+\overline{e}(t)+\overline{i}(t)+\overline{r}(t)$ is a conserved quantity. Therefore, the exit time from set E is $\zeta(E) = \infty$.

In Figure 4 we can see a comparison between the solution of the ODEs and some trajectories of the stochastic process for different population levels. As we can see, as N increases, the trajectories can no longer be distinguished from the limit ODEs.

Theorems 5.1 and 5.2 can also be used to estimate other properties of the stochastic sequence of models, such as exit times. Consider, for instance, the following question: "when will one third of the nodes be contaminated by the virus?" In this case, we need to compute the exit time from the set $S = \{(d_1, d_2, d_3, d_4) \in E \mid d_2 + d_3 < \frac{1}{3}\}$. If we consider the set of ODEs, we obtain that $\zeta(S) = 0.43043$ (for parameters as in Figure 4). Let $\zeta^{(N)}(S)$ be the exit time for the sequence of CTMC models. Theorem 5.2 states that $\zeta^{(N)}(S)$ converges to $\zeta(S)$ in probability, so that, if N is large, $\zeta(S)$ is a good estimate. We can see this fact visually by looking at the distributions of exit

times for different values of N, as shown in Figure 5, obtained from a batch of simulations of the CTMC. Notice that for $N = 10^6$, $\zeta^{(N)}(S) = 0.43043$ in all the simulation runs.

5.5 Example: Crowd Dynamics

In this section, we consider a CTMC model of crowd dynamics, presented in [20]. The model tries to capture an emergent phenomenon happening in certain cities in southern Spain, where people wandering around city's squares during the evening suddenly start gathering in a single square, giving rise to a big (and noisy) party.

In this example, we assume we have 4 squares, connected in a ring topology (but more squares and more general topologies can be considered). The main idea is that each person is willing to remain in a square only if she finds someone to talk. Hence, she will encounter a certain number of other people in the square, and may talk with them, depending on several factors (she knows these people, she likes them, they want to talk to her or not, etc.). This event is modeled in a simple way, namely as a Bernoulli random variable with probability c, called the *chat probability*. Hence, a person will leave the square if she finds nobody to chat with, i.e. with a probability $(1-c)^k$, where k is the number of persons that she meets in the square. In the context of CTMCs, we will interpret this probability as a *rate*. We will describe the number of people in each square by the variables X_i, $i = 1, \ldots, 4$. The model assumes that a person will meet everybody currently in the square. Hence, the rate at which an individual in square i will leave that square in the next step is $(1-c)^{X_i - 1}$. Therefore, the rate at which someone will leave square i is $X_i \cdot (1-c)^{X_i - 1}$. When someone leaves, she may go to the preceding or the following square with the same probability, which in this case is $\frac{1}{2}$.

In this way, we can construct a sequence of models $\mathcal{B}_C^{(N)}$, for system size $\gamma_N = N$, with variables (X_1, X_2, X_3, X_4), state spaces $\mathcal{S}^4(\{0, \ldots, N\}, N)$, and with 8 transitions, two per square. For instance, the transition modeling the event of moving from square 1 to square 2 is $(a_{1,2}, \mathbf{e_1}, \mathbf{e_2}, r_{1,2}^{(N)})$ where $r_{1,2}^{(N)}(X_1, X_2, X_3, X_4) =_{\text{def}} \frac{1}{2} X_1 \cdot (1-c)^{X_1 - 1}$. The other transitions are defined similarly.

This model shows the following behaviour: if c is sufficiently high, larger than $\frac{4}{N}$, i.e. the number of squares divided by the population size, all the people tend to converge to the same square (a party emerges!). This phenomenon is observed both in the CTMC and in the fluid model, as can be seen in Figure 6, where we show the behavior of two models for different population levels, keeping $c = 0.01$ fixed. One model ($N = 80$) is below the threshold (no convergence, but random distribution of people among squares) and the other one ($N = 800$) is above it (convergence to a single square).

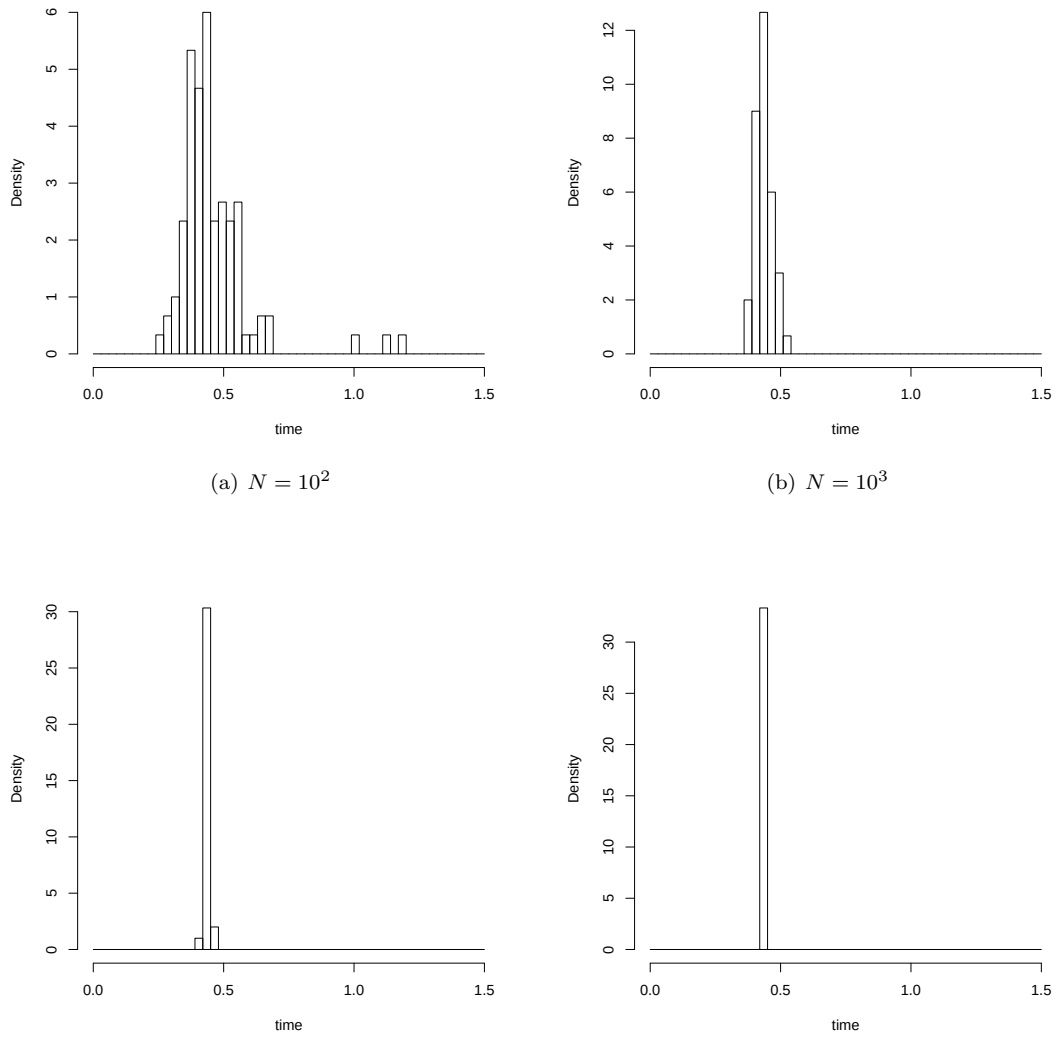
If we compute the drift for the normalized model with N people, we see that it depends on N (differently from other examples). For instance, the first component of $F_{\overline{\mathcal{B}}_C^{(N)}}$ is

$$-\overline{X}_1 \cdot (1-c)^{N \cdot \overline{X}_1 - 1} + 0.5 \cdot \overline{X}_2 \cdot (1-c)^{N \cdot \overline{X}_2 - 1} + 0.5 \cdot \overline{X}_4 \cdot (1-c)^{N \cdot \overline{X}_4 - 1}$$

For fixed c, letting the state space E be unit simplex in $[0,1]^4$, we have that

$$\lim_{N \to \infty} \sup_{\overline{X}_1, \overline{X}_2, \overline{X}_3, \overline{X}_4 \in E} ||F_{\overline{\mathcal{B}}_C^{(N)}}(\overline{X}_1, \overline{X}_2, \overline{X}_3, \overline{X}_4)|| = 0$$

so that the drift $F_{\overline{\mathcal{B}}_C^{(N)}}$ converges uniformly to the constant function yielding 0 for each point in E. Hence, the fluid limit of the sequence is the constant function $\overline{\mathbf{x}}(t) = \overline{\mathbf{x}}_\mathbf{0}$! The ODE dynamics shown Figure 6 is the solution of the ODE $\frac{d\overline{\mathbf{x}}^{(N)}(t)}{dt} = F_{\overline{\mathcal{B}}_C^{(N)}}(\overline{\mathbf{x}}^{(N)}(t))$, where we distinguish the different solutions for different N, hence it is not the fluid limit. However, increasing N, we observe that the ODE solution $\overline{\mathbf{x}}^{(N)}(t)$ converges to a constant function. In Figure 7, we see what happens for $N = 8000$. In this case, it seems that no-one moves either in the CTMC or in the ODE model (Figures 7(a) and 7(b)). However, running the simulation for a much longer time period (of the order of 10^{12}), we can still observe that people converge to the same square. This will happen for every finite N, even if the time of departure of $\overline{\mathbf{x}}^{(N)}(t)$ from the constant solution $\overline{\mathbf{x}}(t)$ will diverge.



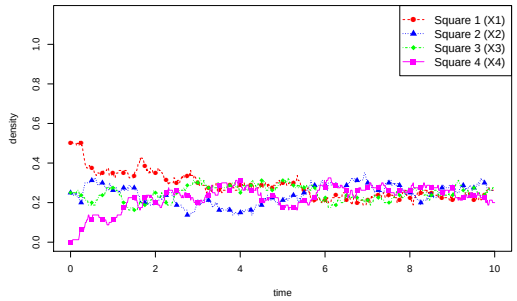
(a) $N = 10^2$

(b) $N = 10^3$

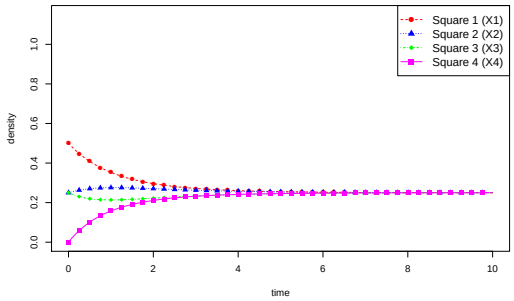
(c) $N = 10^4$

(d) $N = 10^6$

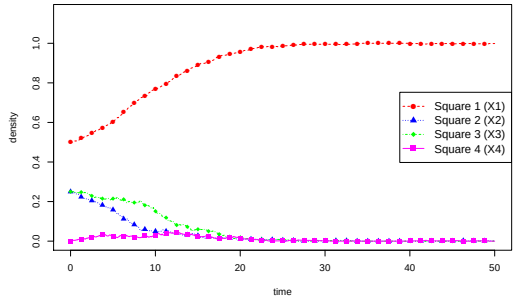
Figure 5: Distributions of exit times from the set $S = \{(d_1, d_2, d_3, d_4) \in E \mid d_2 + d_3 \leq \frac{1}{3}\}$, for different population levels, estimated from 1000 simulation runs. Parameters are as specified in Figure 4. The deterministic limit to which the exit time sequence converges is $\zeta(S) = 0.43043$.



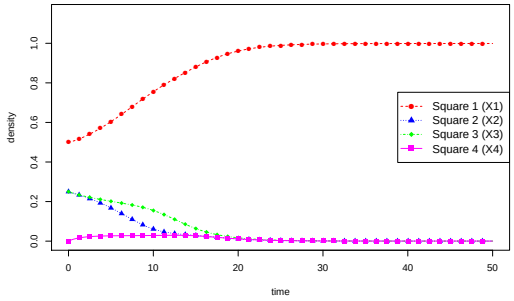
(a) CTMC $N = 80$



(b) ODE $N = 80$



(c) CTMC $N = 800$



(d) ODE $N = 800$

Figure 6: Comparison of deterministic and stochastic trajectories (for increasing population levels) of the crowd dynamics model for initial conditions $\overline{x}_1 = 0.5$, $\overline{x}_2 = 0.25$, $\overline{x}_3 = 0.25$, and $\overline{x}_4 = 0.0$, and chat probability $c = 0.01$. For $N = 80$ (Figures 6(a) and 6(b)), c is below the threshold, and no emergent crowding behaviour is observed. For $N = 800$, instead, this phenomenon holds.

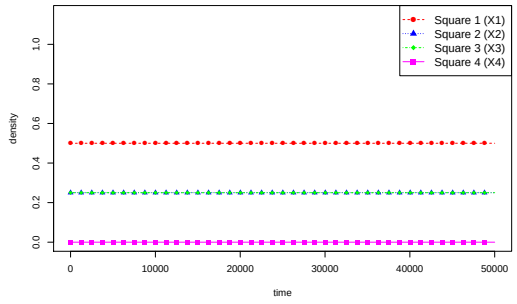
Although this behaviour seems to contradict the use of deterministic approximation theorems to study this model, there is a reasonable agreement between ODE and CTMCs for large N. Indeed, deterministic approximation results can still be used in this context for two complementary reasons:

1. $\overline{\mathbf{x}}^{(N)}(t)$ is, in any case, an approximation of the average of the stochastic process (cf. Section 9 and [20]);

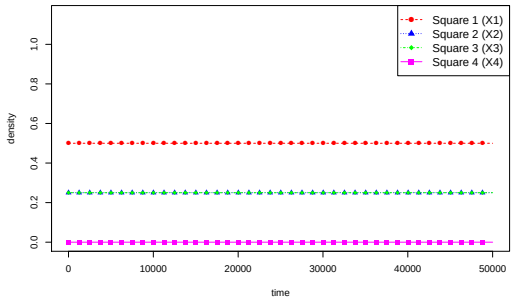
2. deterministic approximation theorems prove the convergence essentially by showing that $\lim_{N\to\infty} \overline{\mathbf{X}}^{(N)}(t) = \overline{\mathbf{x}}^{(N)}(t)$ and that $\lim_{N\to\infty} \overline{\mathbf{x}}^{(N)}(t) = \overline{\mathbf{x}}(t)$ (this is a consequence of hypothesis (9) on page 17); hence we can always approximate the CTMC with the solution of the ODE of level N, if N is sufficiently large.

5.6 Example: a Queue Model

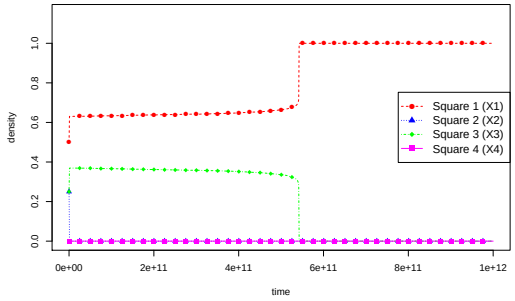
In this section we will briefly discuss a queue model studied in [40], in which the authors consider a CTMC model representing N servers, each with buffer size equal to 1. The policy for the incoming customers is to redirect them to the server with the shortest queue available, if there is one. Servers cannot be distinguished from one another, so we can describe the state of the queueing network by three variables, X_0, X_1, and X_2, describing the number of servers dealing with 0, 1, and 2 customers, respectively. We assume that the arrival rate scales with N, i.e. it is equal to $N \cdot \lambda$,



(a) CTMC $N = 8000$



(b) ODE $N = 8000$



(c) CTMC $N = 8000$, $T_f = 10^{12}$

Figure 7: Comparison of deterministic and stochastic trajectories of the crowd dynamics model with $N = 8000$ people, for initial conditions $\overline{x}_1 = 0.5$, $\overline{x}_2 = 0.25$, $\overline{x}_3 = 0.25$, and $\overline{x}_4 = 0.0$, and chat probability $c = 0.01$. In Figures 7(a) and 7(b), the system seems to remain constant. However, if we observe it for a very long time (up to time $T_f = 10^{12}$), we can still see the emergent behaviour typical of the model.

and that the service rate is ρ for each queue. We assume that arrivals are suspended when all queues are full. We can easily construct a model:

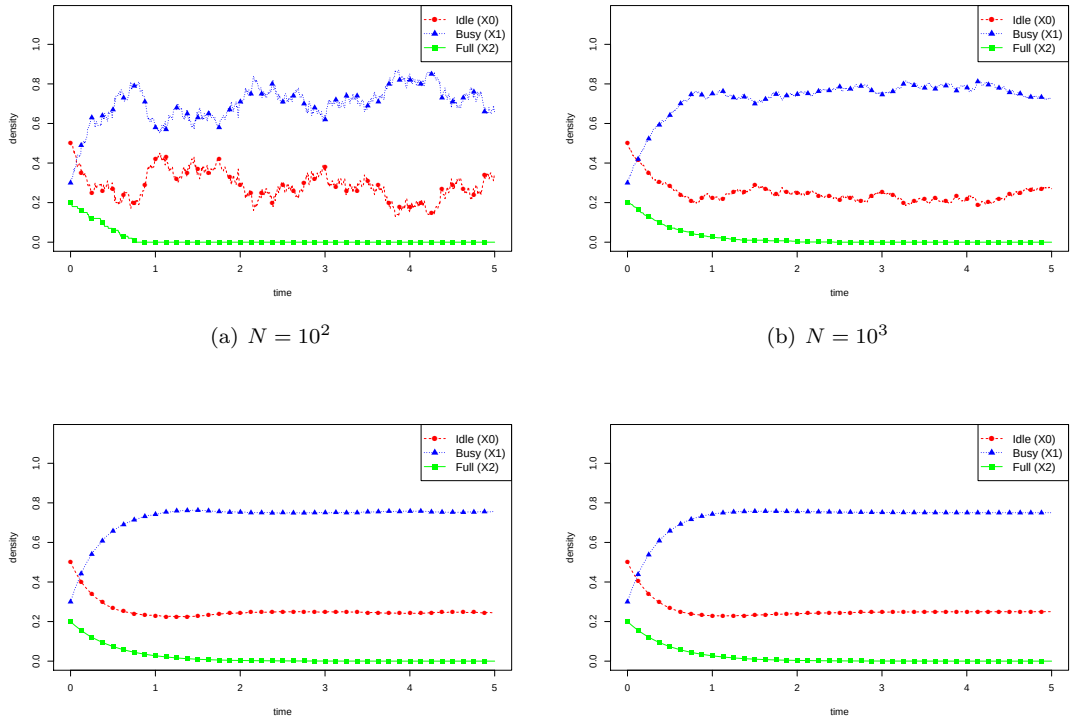
$$\mathcal{Q}_C^{(N)} =_{\text{def}} (\mathbf{X}^{(N)}, \mathcal{S}^3(\{0,\ldots,N\}, N), \mathcal{T}^{(N)}, {\mathbf{d_0}}^{(N)})$$

of such a system, where $\mathbf{X}^{(N)} = (X_0, X_1, X_2)$ and ${\mathbf{d_0}}^{(N)} \in \{0,\ldots,N\}^3$ such that we have $\lim_{N\to\infty} {\overline{\mathbf{d}}_\mathbf{0}}^{(N)} = \overline{\mathbf{d}}_\mathbf{0}$ for some $\overline{\mathbf{d}}_\mathbf{0} \in [0,1]^3$. For predicate $P : \{0,\ldots,N\}^3 \to \mathbb{B}$, we let function I_p be defined as follows:

$$I_P(\mathbf{d}) =_{\text{def}} \begin{cases} 1, \text{if } P(\mathbf{d}) \\ \\ 0, \text{otherwise} \end{cases}$$

The transition set $\mathcal{T}^{(N)}$ is composed of four transitions, defined below, where, in a similar way to before, $\mathbf{e_j}$ is the unit vector in $\mathbb{R}^3$ with 1 as j-component and 0 in the others:

- Incoming client directed to an idle server: state change vector $\mathbf{v}_{\tau_{\mathbf{ie}}} =_{\text{def}} \mathbf{e_1} - \mathbf{e_0}$ and rate function $r_{ie}^{(N)}(X_0, X_1, X_2) =_{\text{def}} N \cdot \lambda \cdot I_{X_0>0}$.
- Incoming client directed to a busy server with empty queue: state change vector $\mathbf{v}_{\tau_{\mathbf{io}}} =_{\text{def}} \mathbf{e_2} - \mathbf{e_1}$ and rate function $r_{io}^{(N)}(X_0, X_1, X_2) =_{\text{def}} N \cdot \lambda \cdot I_{X_0=0} \cdot I_{X_1>0}$.



(a) $N = 10^2$ (b) $N = 10^3$

(c) $N = 10^5$ (d) ODE

Figure 8: Comparison of deterministic and stochastic trajectories (for increasing population levels) of the queue model for initial conditions $\overline{x}_0 = 0.5$, $\overline{x}_1 = 0.3$, $\overline{x}_2 = 0.2$, and parameters $\lambda = 1.5$ and $\rho = 2$.

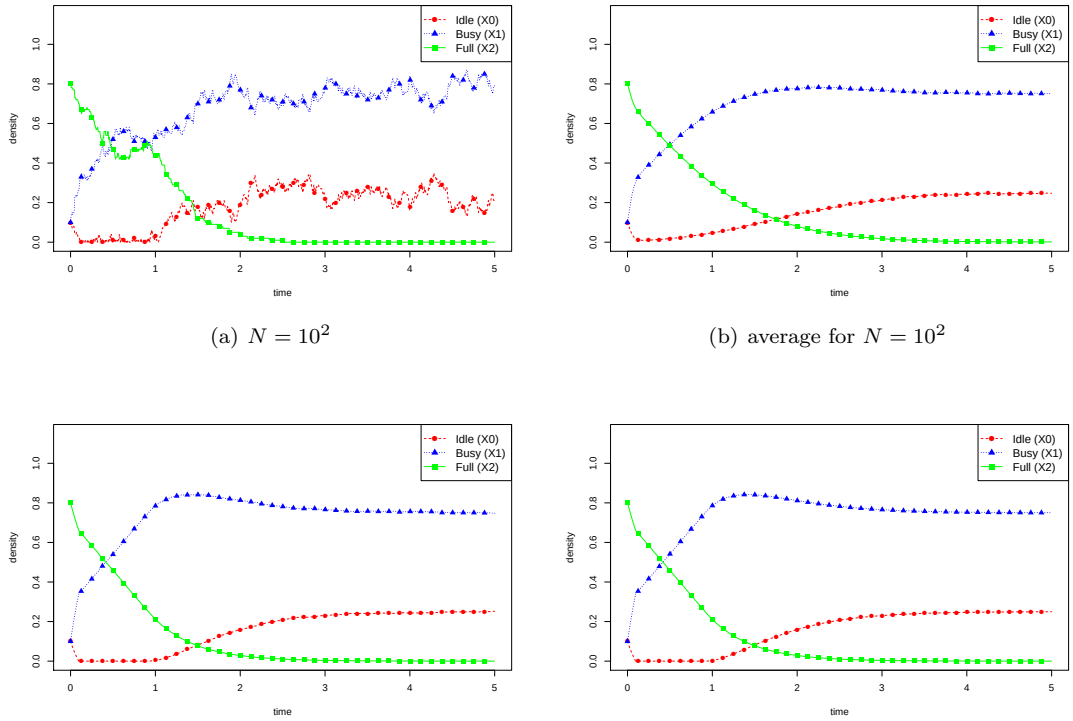
- Servicing of a client from a server with buffer empty: state change vector $\mathbf{v}_{\tau_{\mathbf{se}}} =_{\text{def}} \mathbf{e_0} - \mathbf{e_1}$ and rate function $r_{se}^{(N)}(X_0, X_1, X_2) =_{\text{def}} \rho \cdot X_1$.
- Servicing of a client from a server with buffer full: state change vector $\mathbf{v}_{\tau_{\mathbf{sf}}} =_{\text{def}} \mathbf{e_1} - \mathbf{e_2}$ and rate function $r_{sf}^{(N)}(X_0, X_1, X_2) =_{\text{def}} \rho \cdot X_2$.

Notice that the shortest queue policy requires the use of indicator functions (which are discontinuous). In fact, we must direct a customer to an idle server as long as there is one, while we direct the customer to a server already servicing a client (and with an empty queue) only if there are no idle servers.

We consider a sequence $(\overline{\mathcal{Q}}_C^{(N)})_{N \geq N_0}$ of models, normalized with respect to the system size $\gamma_N = N$, according to the recipe of Section 4, and check that scaling assumptions are satisfied. The state space of these models is the unit simplex E in $\mathbb{R}^3$, $E =_{\text{def}} \mathcal{S}^3([0,1],1)$. Convergence of initial conditions is immediate by definition. The exit rate is $\Theta(N)$ and the step size is $\delta_N = \frac{1}{N}$, hence the noise conditions hold. As for the drift, we have that

$$F_{\overline{\mathcal{Q}}_C^{(N)}} \begin{pmatrix} \overline{X}_0 \\ \overline{X}_1 \\ \overline{X}_2 \end{pmatrix} = \begin{pmatrix} \rho \cdot \overline{X}_1 - \lambda \cdot I_{\overline{X}_0 > 0} \\ \lambda \cdot I_{\overline{X}_0 > 0} + \rho \cdot \overline{X}_2 - \lambda \cdot I_{\overline{X}_0 = 0} \cdot I_{\overline{X}_1 > 0} - \rho \cdot \overline{X}_1 \\ \lambda \cdot I_{\overline{X}_0 = 0} \cdot I_{\overline{X}_1 > 0} - \rho \cdot \overline{X}_2 \end{pmatrix}$$

Also in this example, we have that function $F_{\overline{\mathcal{Q}}_C^{(N)}}$ does not depend on N; thus we denote it by F. However, note that there is a problem here: the function F is *not* Lipschitz in E, because of



(a) $N = 10^2$ (b) average for $N = 10^2$

(c) $N = 10^5$ (d) ODE

Figure 9: Comparison of deterministic and stochastic trajectories (for increasing population levels) of the queue model for initial conditions $\overline{x}_0 = 0.1$, $\overline{x}_1 = 0.1$, $\overline{x}_2 = 0.8$, and parameters $\lambda = 1.5$ and $\rho = 2$. Figure 9(b) shows the average over 1000 runs for the model with $N = 100$.

the discontinuities introduced by the guard functions. These functions introduce a discontinuity between (part of) the border of the simplex E and its interior — when a guard is activated F jumps from the boundary of E to its interior. This is the case when there are no idle servers and no busy servers with an empty buffer.

In order to circumvent this problem, we can focus our attention on a subset $S \subset E$, contained in the interior of E, in which F is Lipschitz. If we localize the theorem in this subset, we can still prove that the solution $\overline{\mathbf{x}}(t)$ of the ODE $\frac{d\overline{\mathbf{x}}(t)}{dt} = F(\overline{\mathbf{x}}(t))$, $\overline{\mathbf{x}}(0) = \overline{\mathbf{x}}_\mathbf{0}$, is the limit of the sequence of CTMCs $\overline{\mathbf{X}}^{(N)}(t)$ in S. A necessary condition for this to hold is that $\overline{\mathbf{x}}_\mathbf{0} \in S$. Hence we must choose an initial configuration with a non-null fraction of idle servers and busy servers with empty buffer. If the initial conditions are sufficiently far away from the boundary ∂E of E, and $\rho > \lambda$, then $\overline{\mathbf{x}}(t)$ will always remain inside E, hence the theorem is valid for all stop times T (see Figure 8). On the other hand, if the initial conditions are close to the point $(0, 0, 1)$, then the ODE solution will reach a state with no empty queues, remain in ∂E for a while (until $\overline{x}_1 > \frac{\lambda}{\rho}$, i.e. until the vector field just outside ∂E will point inside E, which will occur when the ODE for $\overline{x}_0$ in E is positive: $\rho \cdot \overline{x}_1 - \lambda > 0$), and then enter the interior of E again (see Figure 9(d)). During this period, we cannot invoke the theorems of Section 5 to obtain convergence. Nevertheless, the manifested behaviour of the system seems to obey the pattern of deterministic approximation (Figure 9). This switching behaviour of the ODE may be dealt with using hybrid techniques, cf. Section 9.

6 Deterministic Approximation for DTMCs

There are many approaches in the literature for the deterministic approximation of DTMCs. Some of them are focussed on specific examples [37, 36], while others present more general approaches [29, 39]. Usually, these approximation results are referred to under the name of *mean field approximation*. Broadly speaking, there are two main classes of mean field results for DTMC: those resulting in a set of ODEs characterising the limit model, and those resulting in a set of difference equations, characterising a discrete time deterministic system. With reference to the terminology introduced in Section 3, the basic difference between them is how the number of local transitions being executed in a single global transition scales with N: if this number is infinitesimal (with respect to N), we obtain a limit in continuous time, while if it is constant, we obtain a limit in discrete time.[9]

To clarify this last statement, consider a system in which components evolve asynchronously. We can construct a DTMC model of this system by observing its state every ϵ_N time units. In particular, if we sample it with a high frequency (i.e. if ϵ_N is sufficiently small), we can assume that *at most one* local transition can happen between two consecutive observations. Therefore, the number of local transitions for each global transition is zero or one. If we increase the number N of components, then the temporal density of local transitions also increases. Hence, if we want to preserve the property on the number of local transitions per global transition, we need to increase the observation frequency (i.e. decrease ϵ_N) as N increases. By increasing the sampling frequency, we also guarantee that the delay between the observations of two local transitions of a specific component is a constant independent of N. This means that we are requiring that the speed at which components evolve is independent on the number of components of the system.

Notice that, increasing N, the density of events in each time unit also increases, so that in the limit the evolution of the system becomes continuous. This argument must be made rigorous by stating clearly what the scaling assumptions are.

On the other hand, if in each step of the DTMC all nodes of the network perform a move *synchronously*, then the time associated with each step must remain constant with respect to N. In this case, as time spacing between events remains unaltered while increasing N, the limit of the sequence of DTMCs will also live in discrete time. Some details of this second class of models will be given in Section 8.

In the following, we will first introduce some notation and define the scaling assumptions. Then, we will state the main approximation theorems and discuss again the network epidemic example. In Section 7, instead, we will focus our attention on comparing the approximation results for DTMCs and CTMCs.

6.1 Notation and Scaling Assumptions

Consider a DTMC model $\mathcal{X}_D^{(N)} = (\mathbf{X}^{(N)}, \mathcal{D}^{(N)}, \mathcal{T}^{(N)}, \mathbf{d_0}^{(N)})$, where $\mathbf{X}^{(N)}$ is a tuple of n variables. The relevant notion in this context is that of *mean increment*, i.e. the function $\mu : \mathcal{D}^{(N)} \to \mathbb{R}^n$ defined as follows:

$$\mu(\mathbf{d}) = \sum_{\tau \in \mathcal{T}^{(N)}} \mathbf{v}_\tau p_\tau(\mathbf{d}). \tag{13}$$

Analogously to the CTMC case, the above definition applies also to the *normalized* model $\overline{\mathcal{X}}_D^{(N)}$ and we use subscripts, as in $\mu_{\mathcal{X}_D^{(N)}}(\mathbf{d})$ when necessary.

Let $\overline{\mathcal{X}}_D^{(N)} = (\overline{\mathbf{X}}^{(N)}, \overline{\mathcal{D}}^{(N)}, \overline{\mathcal{T}}^{(N)}, \overline{\mathbf{d}}_\mathbf{0}{}^{(N)})$ be a normalized model and let us consider the sequence $(\overline{\mathcal{X}}_D^{(N)})_{N \geq N_0}$ with respect to an increasing system size γ_N.

We highlight the fact that, in the context of mean field approximation of DTMCs, usually $\gamma_N = N$, and the normalized variables lie in $[0, 1]$ and define a probability distribution, called

[9] The reader can find a brief discussion on the relationship between these two approaches in Remark 8.2 at the end of Section 8.

the *occupancy measure*. In fact, typical models describe N objects that can only change their internal state (hence, no births or deaths are taken into account), and we are concerned with the probability distribution of internal states (cf. Section 4). However, this is not a strict requirement, and we can consider either different notions of γ_N or models with birth and death. In such cases, however, we can no longer interpret the normalized variables as a probability distribution.

We now list the scaling assumptions required for the theorem to hold.

State Space. The state space is defined as for the CTMC case, i.e. it is a closed set E in $\mathbb{R}^n$ such that $\bigcup_N \overline{\mathcal{D}}^{(N)} \subseteq E$.

Convergence of the Initial Conditions. Also the requirement on the initial condition is the same as for the CTMC case: there must exist some point $\overline{\mathbf{d}}_\mathbf{0} \in E$ such that $\lim_{N\to\infty} \overline{\mathbf{d}}_\mathbf{0}^{(N)} = \overline{\mathbf{d}}_\mathbf{0}$.

Intensity and scaling of time. We assume that the temporal duration equals ϵ_N, with $\lim_{N\to\infty} \epsilon_N = 0$. ϵ_N is usually referred to as the *intensity* [39], cf. also Remark 6.1.

Convergence of Drifts. Intensity is further involved in the scaling of the mean increments. In the DTMC case, the *drift* is defined as follows:

$$F_{\overline{\mathcal{X}}_D^{(N)}}(\overline{\mathbf{d}}) =_{\text{def}} \frac{\mu_{\overline{\mathcal{X}}_D^{(N)}}(\overline{\mathbf{d}})}{\epsilon_N} \tag{14}$$

As in the CTMC case, we assume that there exists a Lipschitz vector field $F : E \to \mathbb{R}^n$, such that $F_{\overline{\mathcal{X}}_D^{(N)}}$ converges uniformly to F:

$$\lim_{N\to\infty} \sup_{\overline{\mathbf{d}}\in E} ||F_{\overline{\mathcal{X}}_D^{(N)}}(\overline{\mathbf{d}}) - F(\overline{\mathbf{d}})|| = 0. \tag{15}$$

We furthermore assume that all the trajectories of the vector field F lie in E (this can be accomplished by choosing E appropriately). In our framework, the notion of intensity coincides with the step-size $\delta_N = \frac{1}{\gamma_N}$: $\epsilon_N = \delta_N$, and we usually have $\epsilon_N = \frac{1}{N}$. This is consistent with many applications, in which the number of local entities evolving in each step of the DTMC is constant, hence the fraction of entities evolving goes to zero as $\frac{1}{N}$. It is easy to see that, under the assumption $\epsilon_N = \delta_N$, and assuming that for each N and τ there is a vector $\mathbf{v}_\tau$ such that $\overline{\mathbf{v}}_\tau^{(N)} = \delta_N \cdot \mathbf{v}_\tau$, we furthermore have that:

$$F_{\overline{\mathcal{X}}_D^{(N)}}(\overline{\mathbf{d}}) = \sum_{\tau\in\overline{\mathcal{T}}^{(N)}} \mathbf{v}_\tau \cdot \overline{p}_\tau(\overline{\mathbf{d}}). \tag{16}$$

so that property (15) can be proved by showing that each $\overline{p}_\tau^{(N)}(\overline{\mathbf{d}})$ has a Lipschitz limit $p_\tau(\overline{\mathbf{d}})$. This condition will be automatically verified if $\overline{p}_\tau^{(N)}(\overline{\mathbf{d}})$ does not depend on N, as is the case in many applications.

*Remark** 6.1. The notion of intensity is often found in the mean field literature for DTMCs with a more general flavour. In general, it is introduced in the context of models of N interacting objects the evolution of which is described from a local perspective, namely from the point of view of a single object. In contrast, in our modeling approach we take a global perspective, describing the evolution at the system level.

According to the local description approach, if we have a system composed of N objects/entities, as is the case for the example of Section 3.1, the intensity ϵ_N is usually interpreted as the probability with which a specific object makes a transition in a step of the global DTMC (derived from the local description), or equivalently as the expected fraction of objects performing a transition in a single step[10].

[10]We are implicitly assuming here that all local components can always perform at least one transition in each step. This can be always enforced by appropriately adding *dummy* transitions.

According to this interpretation, in our context, by assigning temporal duration δ_N to each step of the (global) DTMC, we are essentially assuming that the expected number of global steps necessary in order to let each object execute a local transition is $\frac{1}{\delta_N}$, so that the expected time for an object to make two successive transitions (i.e. the expected delay between the same object being selected twice to execute a transition) is kept equal to 1 for each N. This justifies the use of the word intensity to denote the scaling factor for time.

Note that, in general, the intensity is not forced to be equal to $\frac{1}{\delta_N}$, but can encompass more general situations, like those in which the number of entities evolving per step is $\Theta(\log N)$. The results of the following section, however, remain valid also in this more general setting, as long as intensity goes to zero with N and drift scales appropriately.

6.2 Deterministic Approximation Theorems

We can now state the main approximation theorem for DTMCs [29, 39]. Consider a sequence of normalized DTMC models $(\overline{\mathcal{X}}_D^{(N)})_{N\geq N_0}$ with $\overline{\mathcal{X}}_D^{(N)} = (\overline{\mathbf{X}}^{(N)}, \overline{\mathcal{D}}^{(N)}, \overline{\mathcal{T}}^{(N)}, \overline{\mathbf{d_0}}^{(N)})$, and denote by $\overline{\mathbf{X}}^{(N)}(k)$ the discrete-time Markov process associated with $\overline{\mathcal{X}}_D^{(N)}$. In addition, let $\overline{\mathbf{X}}_c^{(N)}(t) =_{\text{def}} \overline{\mathbf{X}}^{(N)}(\lfloor \frac{t}{\epsilon_N} \rfloor)$ be the process in continuous time associated with $\overline{\mathbf{X}}^{(N)}(k)$ and let $\overline{\mathbf{x}}(t)$ be the solution of the initial value problem $\frac{d\overline{\mathbf{x}}(t)}{dt} = F(\overline{\mathbf{x}}(t))$, $\overline{\mathbf{x}}(0) = \overline{\mathbf{d}}_{\mathbf{0}}$, where F is as in (15).

Theorem 6.1 (Deterministic approximation of DTMCs)**.** *Let the sequence $(\overline{\mathbf{X}}_c^{(N)}(t))_{N\geq N_0}$ and $\overline{\mathbf{x}}(t)$ be defined as above, and assume Conditions (7) and (15) hold. Then, for any $T < \infty$, it holds that:*

$$\lim_{N\to\infty} \mathbb{P}\{ \sup_{0\leq t\leq T} ||\overline{\mathbf{X}}_c^{(N)}(t) - \overline{\mathbf{x}}(t)|| > \varepsilon\} = 0$$

■

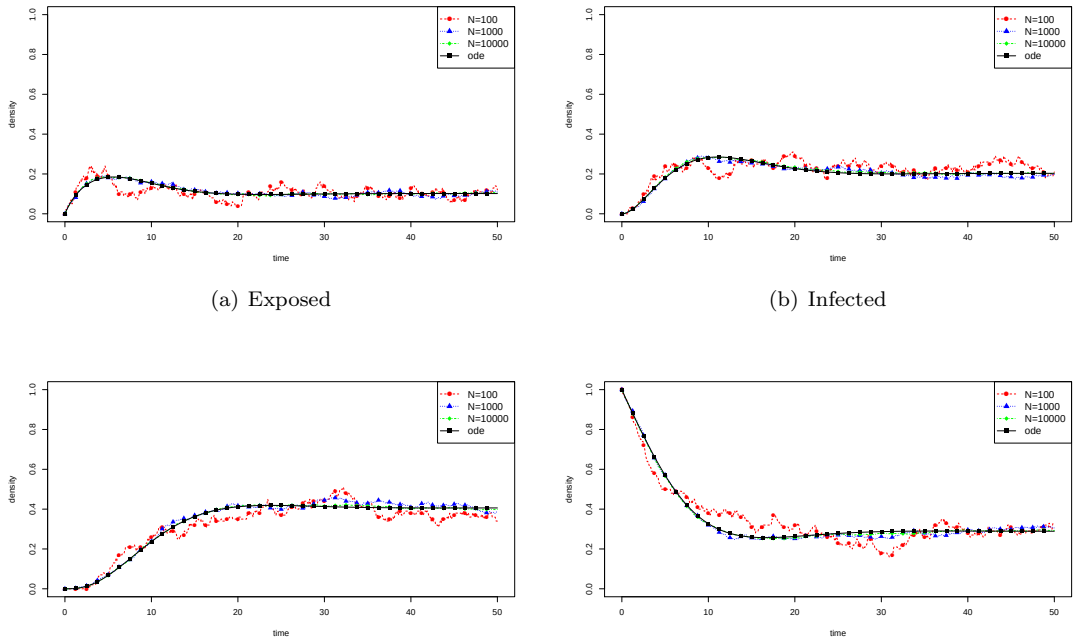
Note that this result is different from those for CTMCs, as we deliberately avoided any discussion about exit times and restrictions of the state space. We will present the justification for this in Section 7.

6.3 Example revisited

We consider again the DTMC version of the network epidemic model of Section 3.1. Proceeding as in Section 5.4, we need to define a notion of system size γ_N and we need to construct a sequence of models for divergent γ_N. Also in the DTMC case, we can simply set $\gamma_N = N$, so that we are studying the behaviour of the system for large populations. Consequently, the sequence of non-normalized models of interest to us is $(\mathcal{E}_D^{(N)})_{N\geq N_0}$, for $\mathcal{E}_D^{(N)}$ as defined in Section 3.3.

The sequence of normalized models is $(\overline{\mathcal{E}}_D^{(N)})_{N\geq N_0}$, where $\overline{\mathcal{E}}_D^{(N)}$ is derived as described in Section 4. Notice that $\delta_N = \frac{1}{N}$ and $\overline{\mathcal{D}}^{(N)} \subseteq \mathcal{S}^4([0,1],1)$ for all N, i.e. $\overline{S} + \overline{E} + \overline{I} + \overline{R} = 1$. Consequently, we choose the state space E by $E =_{\text{def}} \mathcal{S}^4([0,1],1)$. For convenience, we explicitly report below the definition of the normalized probability functions as they result from the definition of normalization:

$$\begin{array}{lll}
\overline{p}_e^{(N)}(\overline{S},\overline{E},\overline{I},\overline{R}) & =_{\text{def}} & \alpha_e \cdot \overline{S} \\
\overline{p}_i^{(N)}(\overline{S},\overline{E},\overline{I},\overline{R}) & =_{\text{def}} & \alpha_i \cdot \overline{S} \cdot \overline{I} \\
\overline{p}_a^{(N)}(\overline{S},\overline{E},\overline{I},\overline{R}) & =_{\text{def}} & \alpha_a \cdot \overline{E} \\
\overline{p}_r^{(N)}(\overline{S},\overline{E},\overline{I},\overline{R}) & =_{\text{def}} & \alpha_r \cdot \overline{I} \\
\overline{p}_p^{(N)}(\overline{S},\overline{E},\overline{I},\overline{R}) & =_{\text{def}} & \alpha_p \cdot \overline{S} \\
\overline{p}_q^{(N)}(\overline{S},\overline{E},\overline{I},\overline{R}) & =_{\text{def}} & \alpha_q \cdot \overline{E} \\
\overline{p}_s^{(N)}(\overline{S},\overline{E},\overline{I},\overline{R}) & =_{\text{def}} & \alpha_s \cdot \overline{R}
\end{array}$$



(a) Exposed

(b) Infected

(c) Recovered

(d) Susceptible

Figure 10: Comparison between the solution of the ODE and stochastic trajectories of DTMCs for increasing population sizes. Parameters of the model are $\alpha_e = 0.1$, $\alpha_i = 0.2$, $\alpha_a = 0.4$, $\alpha_r = 0.2$, $\alpha_p = 0$, $\alpha_q = 0$, and $\alpha_s = 0.1$, while initial conditions of ODE are $\overline{S}_0 = 1$, $\overline{E}_0 = 0$, $\overline{I}_0 = 0$, and $\overline{R}_0 = 0$.

Clearly the probability functions do not depend on N and they are Lipschitz; consequently, also the drift does not depend on N and is Lipschitz. Therefore, we take the drift itself as the vector field according to (15) in order to get the following system of ODEs:

$$\begin{aligned}
\frac{d\overline{s}(t)}{dt} &= -\alpha_e \cdot \overline{s}(t) - \alpha_i \cdot \overline{s}(t) \cdot \overline{i}(t) - \alpha_p \cdot \overline{s}(t) + \alpha_s \cdot \overline{r}(t) \\
\frac{d\overline{e}(t)}{dt} &= \alpha_e \cdot \overline{s}(t) + \alpha_i \cdot \overline{s}(t) \cdot \overline{i}(t) - \alpha_a \cdot \overline{e}(t) - \alpha_q \cdot \overline{e}(t) \\
\frac{d\overline{i}(t)}{dt} &= \alpha_a \cdot \overline{e}(t) - \alpha_r \cdot \overline{i}(t) \\
\frac{d\overline{r}(t)}{dt} &= \alpha_r \cdot \overline{i}(t) + \alpha_p \cdot \overline{s}(t) + \alpha_q \cdot \overline{e}(t) - \alpha_s \cdot \overline{r}(t)
\end{aligned} \tag{17}$$

A visual depiction of the theorem can be found in Figure 10, where the ODE trajectories are compared with trajectories of the DTMCs with different population levels.

7 Comparing CTMC and DTMC deterministic approximations*

In this section we will compare the deterministic approximation results for CTMCs and DTMCs. We will start by recalling the notion of *uniformization* of a CTMC, which plays a central role in this discussion, and explore its implications for the example of Section 3.1. We will observe a striking "coincidence": the deterministic limit processes coincide (cf. equations (12) on page 22 and equations (17) on page 32). Then, we will show that this is no "coincidence" at all, but rather another consequence of the law of large numbers (this time for Poisson processes), commenting on

the implications of this result. We begin by defining the uniformization construction of a DTMC from a CTMC.

Uniformization. Consider a $\mathcal{D}$-valued CTMC $\mathbf{X_C}(t)$ with infinitesimal generator matrix $\mathbf{Q}$. Assume that the exit rate of $\mathbf{X_C}(t)$ is bounded, i.e. there exists a constant Λ such that $\sup_{\mathbf{d}\in\mathcal{D}}(-\mathbf{Q}(\mathbf{d},\mathbf{d})) \leq \Lambda < \infty$. The Λ-*uniformization* [56] (or simply *uniformization*) of $\mathbf{X_C}(t)$, denoted by $\mathsf{UFZ}_\Lambda(\mathbf{X_C}(t))$ in the sequel, is obtained by decoupling the evolution of $\mathbf{X_C}(t)$ into two independent stochastic processes, a DTMC describing the jumps, and their probabilities, and a Poisson process describing how many jumps have been made up to time t.

Formally, we define $\mathsf{UFZ}_\Lambda(\mathbf{X_C}(t))$ as a pair $\mathsf{UFZ}_\Lambda(\mathbf{X_C}(t)) =_{\text{def}} (\mathbf{X_D}(k),\mathcal{N}(t))$ where $\mathbf{X_D}(k)$ is a $\mathcal{D}$-valued DTMC with probability transition matrix $\mathbf{P} = \mathbf{I} + \mathbf{Q}/\Lambda$, and $\mathcal{N}(t)$ the companion Poisson process, i.e. $\mathcal{N}_t \in POI[\Lambda \cdot t]$. $\mathbf{X_C}(t)$ is equivalent to the stochastic process $\mathbf{X_D}(\mathcal{N}(t))$, that is for all $t \in \mathbb{R}_{\geq 0}, \mathbf{d} \in \mathcal{D}$ we have

$$\mathbb{P}\{(\mathbf{X_C})_t = \mathbf{d}\} = \sum_{k=0}^{\infty} \mathbb{P}\{(\mathbf{X_D})_k = \mathbf{d}\} \cdot \mathbb{P}\{\mathcal{N}_t = k\}.$$

Unlike the embedded DTMC, the DTMC associated with a CTMC via uniformization is known to have the same steady state probability distribution as the CTMC.

The uniformization procedure can easily be lifted from CTMCs to CTMC models specified in the language defined in Section 3, as described in the sequel. Let $\mathcal{X}_C^{(N)} = (\mathbf{X_C}^{(N)}, \mathcal{D}^{(N)}, \mathcal{T}_C^{(N)}, \mathbf{x_0}^{(N)})$ be such a model, with transition rate functions $r_\tau^{(N)}(\mathbf{d})$ and exit rate function $R_{\mathcal{X}_C^{(N)}}$. Assuming $R_{\mathcal{X}_C^{(N)}}$ is bounded, fix $\Lambda_N \geq \sup_{\mathbf{d}\in\mathcal{D}} R_{\mathcal{X}_C^{(N)}}(\mathbf{d})$, and define the DTMC model as $\mathcal{X}_D^{(N)} = (\mathbf{X_D}^{(N)}, \mathcal{D}^{(N)}, \mathcal{T}_D^{(N)}, \mathbf{x_0}^{(N)})$ where

$$\mathcal{T}_D^{(N)} =_{\text{def}} \{(a, \mathbf{s}, \mathbf{t}, p_\tau^{(N)}) \mid (a, \mathbf{s}, \mathbf{t}, r_\tau^{(N)}) \in \mathcal{T}_C^{(N)}\}$$

with $p_\tau^{(N)}(\mathbf{d}) =_{\text{def}} \frac{r_\tau^{(N)}(\mathbf{d})}{\Lambda_N}$, for all $\mathbf{d} \in \mathcal{D}^{(N)}$. Furthermore, let $\mathcal{N}^{(N)}(t)$ be the Poisson process such that $\mathcal{N}_t^{(N)} \in POI[\Lambda_N \cdot t]$. We let the Λ_N-uniformization of CTMC model $\mathcal{X}_C^{(N)}$ be defined as follows: $\mathsf{UFZ}_{\Lambda_N}(\mathcal{X}_C^{(N)}) =_{\text{def}} (\mathcal{X}_D^{(N)}, \mathcal{N}^{(N)}(t))$. Of course, the uniformization procedure can be applied to normalized CTMC models and to sequences of CTMC models as well.

Consider again the network epidemic example of Section 3.1, which we modeled both as a CTMC and as a DTMC. It is easily seen that the exit rate $R_{\mathcal{E}_C^{(N)}}$ function is bounded in $\mathcal{S}^4(\{0,\ldots,N\}, N)$ by $N \cdot \sum_{\sigma\in\{e,i,a,r,p,q,s\}} \lambda_\sigma$, since $S + E + I + R = N$ implies $r_\sigma^{(N)}(S,E,I,R) \leq N \cdot \lambda_\sigma$ for $\sigma \in \{e,i,a,r,p,q,s\}$. Letting $\Lambda =_{\text{def}} \sum_{\sigma\in\{e,i,a,r,p,q,s\}} \lambda_\sigma$ and $\Lambda_N =_{\text{def}} N \cdot \Lambda$, we get that the transition probabilities of the DTMC model associated with $R_{\mathcal{E}_C^{(N)}}$ by Λ_N-uniformization are $p_\sigma^{(N)}(S,E,I,R) =_{\text{def}} \frac{r_\sigma^{(N)}(S,E,I,R)}{\Lambda_N}$ for σ as above. For instance, for the rate of external infection, we have that $r_e^{(N)}(S,E,I,R) = \lambda_e \cdot S$, so that $p_e^{(N)}(S,E,I,R) =_{\text{def}} \frac{\lambda_e \cdot S}{N \cdot \Lambda}$. Under the assumption that $\alpha_e = \frac{\lambda_e}{\Lambda}$ we get that $p_e^{(N)}$ is exactly the same as the external infection probability in the DTMC model $\mathcal{E}_D^{(N)}$ we defined in Section 3.3. Therefore, the two network epidemic models $\mathcal{E}_D^{(N)}$ and $\mathcal{E}_C^{(N)}$ we have considered are strongly related, via the uniformization construction.

We now recall that the vector field for the limit ODE of the normalized DTMC model $\overline{\mathcal{E}}_D^{(N)}$ is its drift $F_{\overline{\mathcal{E}}_D^{(N)}}$ and, since it does not depend on N, we use the abbreviation F_D. Similarly, we use F_C for the drift $F_{\overline{\mathcal{E}}_C^{(N)}}$ of the normalized CTMC model $\overline{\mathcal{E}}_C^{(N)}$. Again, under the assumption that $\alpha_\sigma = \frac{\lambda_\sigma}{\Lambda}$, we get that $F_D(\overline{S},\overline{E},\overline{I},\overline{R}) = \frac{1}{\Lambda} \cdot F_C(\overline{S},\overline{E},\overline{I},\overline{R})$: the two limit dynamics are essentially the same (modulo a rescaling of time, see below), as can also be seen looking at Figures 4 and 10.

This is not a coincidence: the previous relation is always in force between a CTMC model and the DTMC model associated with it via uniformization. It is worth emphasising here that

what we are comparing is *not* a CTMC model $\mathcal{X}_C^{(N)}$ and its uniformization $\mathsf{UFZ}_{\Lambda_N}(\mathcal{X}_C^{(N)}) = (\mathcal{X}_D^{(N)}, \mathcal{N}^{(N)}(t))$, where, as we have seen, the time dimension is added to the DTMC $\mathcal{X}_D^{(N)}$ by means of the Poisson process $\mathcal{N}^{(N)}(t)$, i.e. in a way in which the duration of the DTMC steps is *intrinsically randomized*. We are comparing, instead, the CTMC model $\mathcal{X}_C^{(N)}$ with the DTMC model $\mathcal{X}_D^{(N)}$ where each step takes a *fixed* and *fully deterministic* amount of time ε_N.

So, for each sequence of CTMC models we can construct, using the probabilities used for uniformization, a sequence of DTMC models sharing the same limit behaviour. Furthermore, given a sequence of DTMC models, we can always construct a sequence of CTMC models having the same limit behaviour. Essentially, this means that there is a very close relationship between the limit theorems of Sections 5 and 6: in a certain sense, they are two facets of the same result. This relationship is not only of theoretical interest, but also pragmatic. It means that one can apply results about limit behaviour proved for DTMCs to CTMCs and vice versa. A formal argument for this will be given in Theorem 7.1 below. First we present the relationship between a sequence of CTMCs and the corresponding sequence of DTMCs and then the relationship between a sequence of DTMCs and the corresponding sequence of CTMCs before presenting the convergence results.

From CTMCs to DTMCs. Let $\overline{\mathcal{X}}_C^{(N)} = (\overline{\mathbf{X}}^{(N)}, \overline{\mathcal{D}}^{(N)}, \overline{\mathcal{T}}^{(N)}, \overline{\mathbf{d}}_{\mathbf{0}}{}^{(N)})$ be a normalized CTMC model and suppose the sequence $(\overline{\mathcal{X}}_C^{(N)})_{N \geq N_0}$ with respect to an increasing system size γ_N admits deterministic approximation. Let us construct the sequence of DTMC models $(\overline{\mathcal{X}}_D^{(N)})_{N \geq N_0}$, associated by uniformization, for constants $\Lambda_N \geq \sup_{\overline{\mathbf{d}} \in E} R_{\overline{\mathcal{X}}_C^{(N)}}(\overline{\mathbf{d}})$, such that $\Lambda_N = \Theta(\gamma_N)$. The previous hypothesis on Λ_N implies that $\lim_{N\to\infty} \Lambda_N \delta_N = \Lambda > 0$.

Furthermore let F_C and F_D be the vector fields according to equations (8) and (15) respectively and recall our assumption that $\epsilon_N = \delta_N$. We get the following:

$$
\begin{aligned}
& F_D(\overline{\mathbf{d}}) \\
= \;& \{\text{Equation (15)}\} \\
& \lim_{N\to\infty} F_{\overline{\mathcal{X}}_D^{(N)}}(\overline{\mathbf{d}}) \\
= \;& \{\text{Def. } F_{\overline{\mathcal{X}}_D^{(N)}}; \text{Def. } \mu_{\overline{\mathcal{X}}_D^{(N)}}; \text{Def. } \mathsf{UFZ}_{\Lambda_N}(\overline{\mathcal{X}}_C^{(N)}); \epsilon_N = \delta_N\} \\
& \lim_{N\to\infty} \textstyle\sum_{\overline{\tau} \in \overline{\mathcal{T}}^{(N)}} \overline{\mathbf{v}}_{\overline{\tau}}^{(N)} \cdot \frac{\overline{r}_{\overline{\tau}}^{(N)}(\overline{\mathbf{d}})}{\delta_N \cdot \Lambda_N} \\
= \;& \{\text{Algebra}\} \\
& \lim_{N\to\infty} \tfrac{1}{\delta_N \cdot \Lambda_N} \cdot \textstyle\sum_{\overline{\tau} \in \overline{\mathcal{T}}^{(N)}} \overline{\mathbf{v}}_{\overline{\tau}}^{(N)} \cdot \overline{r}_{\overline{\tau}}^{(N)}(\overline{\mathbf{d}}) \\
= \;& \{\text{Def. } F_{\overline{\mathcal{X}}_C^{(N)}}\} \\
& \lim_{N\to\infty} \tfrac{1}{\delta_N \cdot \Lambda_N} \cdot F_{\overline{\mathcal{X}}_C^{(N)}}(\overline{\mathbf{d}}) \\
= \;& \{\lim_{N\to\infty} \Lambda_N \cdot \delta_N = \Lambda \text{ by hypothesis; Equation (8)}\} \\
& \tfrac{1}{\Lambda} \cdot F_C(\overline{\mathbf{d}}).
\end{aligned}
$$

Hence, we have shown that the deterministic limit of the associated sequence of DTMCs satisfies the ODE $\frac{d\overline{\mathbf{x}}(t)}{dt} = \frac{1}{\Lambda} F_C(\overline{\mathbf{x}}(t))$, which is equivalent, modulo time rescaling by a factor $\frac{1}{\Lambda}$. This means that if we rescale time from t to $t' = \frac{t}{\Lambda}$, we obtain from the sequence of DTMCs obtained by uniformization the same set of ODE as the one derived from the sequence of CTMCs: $\frac{d\overline{\mathbf{x}}(t')}{dt'} = F_C(\overline{\mathbf{x}}(t'))$.

From DTMCs to CTMCs. Let $\overline{\mathcal{X}}_D^{(N)} = (\overline{\mathbf{X}}^{(N)}, \overline{\mathcal{D}}^{(N)}, \overline{\mathcal{T}}^{(N)}, \overline{\mathbf{d}}_{\mathbf{0}}{}^{(N)})$ be a normalized DTMC model and consider the sequence $(\overline{\mathcal{X}}_D^{(N)})_{N \geq N_0}$.

We can associate it to a sequence $(\overline{\mathcal{X}}_C^{(N)})_{N\geq N_0}$ of normalized CTMC models simply by interpreting probabilities, multiplied by the system size, as rates. Then we define the set of transitions $\overline{\mathcal{T}}^{d^{(N)}}$. These are formed from the transitions of $\overline{\mathcal{T}}^{(N)}$ by replacing the probability transition functions with rate functions defined as $\overline{r}_\tau^{(N)} = \gamma_N \cdot \overline{p}_\tau^{(N)}$. The associated CTMC model is therefore $\overline{\mathcal{X}}_C^{(N)} = (\overline{\mathbf{X}}^{(N)}, \overline{\mathcal{D}}^{(N)}, \overline{\mathcal{T}}^{d^{(N)}}, \overline{\mathbf{d}_\mathbf{0}}^{(N)})$ whose mean dynamics is

$$F_{\overline{\mathcal{X}}_C^{(N)}}(\overline{\mathbf{d}}) = \sum_{\tau\in\overline{\mathcal{T}}^{d^{(N)}}} \delta_N \cdot \mathbf{v}_\tau^{(N)} \cdot \overline{r}_\tau^{(N)}(\overline{\mathbf{d}}) = \sum_{\tau\in\overline{\mathcal{T}}^{d^{(N)}}} \mathbf{v}_\tau^{(N)} \overline{p}_\tau^{(N)}(\overline{\mathbf{d}}) = F_{\overline{\mathcal{X}}_D^{(N)}}(\overline{\mathbf{d}}).$$

Thus the mean dynamics of the constructed CTMC model is equal to the mean dynamics of the DTMC model. As a consequence, $(\overline{\mathcal{X}}_D^{(N)})_{N\geq N_0}$ and $(\overline{\mathcal{X}}_C^{(N)})_{N\geq N_0}$ will have the same deterministic limit.

Convergence. If we inspect a (normalized) CTMC model $\overline{\mathcal{X}}_C^{(N)}$ and its associated DTMC model, $\overline{\mathcal{X}}_D^{(N)}$, or vice versa, we can observe that the jump process underlying the CTMC model is in fact that of the DTMC model. The only difference between the two is that the CTMC model has a variable delay between two consecutive events, exponentially distributed with rate $R_{\overline{\mathcal{X}}_C^{(N)}}(\overline{\mathbf{d}})$, while the time delay between two steps is constant for the DTMC model. Nevertheless, in the limit, both delays go to zero at the same speed, namely as δ_N, which equals the intensity ϵ_N.

We now give an intuitive argument showing that in this setting these two processes must behave in the same way in the limit (i.e. for large N), before stating a theorem which gives a precise statement of this property.

Consider a sequence of (normalized) DTMC models $(\overline{\mathcal{X}}_D^{(N)})_{N\geq N_0}$;
let $(\overline{\mathbf{X}}_\mathbf{D}^{(N)}(k))_{N\geq N_0}$ be the corresponding sequence of DTMCs, and let ϵ_N be the intensity of the sequence, with $\lim_{N\to\infty}\epsilon_N = 0$. We construct two sequences of stochastic process on $t \in [0,\infty)$. One is $\overline{\mathbf{X}}_\mathbf{ds}^{(N)}(t) =_{\text{def}} \overline{\mathbf{X}}_\mathbf{D}^{(N)}(\lfloor \frac{t}{\epsilon_N} \rfloor)$, obtained by assuming that the length of each step in the DTMC is deterministically fixed and equal to ϵ_N. The other one is $\overline{\mathbf{X}}_\mathbf{rs}^{(N)}(t) =_{\text{def}} \overline{\mathbf{X}}_\mathbf{D}^{(N)}(\mathcal{N}(\frac{t}{\epsilon_N}))$, where $\mathcal{N}$ is a Poisson random variable, hence the number of steps at time t is random (and, consequently, $\overline{\mathbf{X}}_\mathbf{rs}^{(N)}(t)$ is a CTMC).

We now give an argument showing that $\overline{\mathbf{X}}_\mathbf{ds}^{(N)}(t)$ and $\overline{\mathbf{X}}_\mathbf{rs}^{(N)}(t)$ are basically the same. Fix a time t, and look at $\overline{\mathbf{X}}_\mathbf{ds}^{(N)}(t)$. Let $k(t,N)$ denote the number of steps which have occurred at time t; then $k(t,N) = \lfloor \frac{t}{\epsilon_N} \rfloor$ by definition. Note that the time of occurrence of the $k(t,N)$-th event, which is $\epsilon_N k(t,N)$, trivially converges to t as N goes to infinity. Note also that $k(t,N)$ depends on N; indeed $\lim_{N\to\infty} k(t,N) = \infty$. Now, consider the CTMC-process $\overline{\mathbf{X}}_\mathbf{rs}^{(N)}(t)$. The $k(t,N)$-th event for this process occurs at random time $T_{k(t,N)}^{(N)} = \sum_{i=1}^{k(t,N)} \cdot \bar{\xi}_i = \epsilon_N \cdot \sum_{i=1}^{k(t,N)} \xi_i$, where $\bar{\xi}_i \in EXP[\frac{1}{\epsilon_N}]$ and $\xi_i \in EXP[1]$[11].

Then the time of the $k(t,N)$-th event, $T_{k(t,N)}^{(N)} = \epsilon_N \cdot k(t,N) \cdot \frac{\sum_{i=1}^{k(t,N)} \cdot \xi_i}{k(t,N)}$, converges to t almost surely (a.s.), as N increases, i.e.

$$\mathbb{P}\{\lim_{N\to\infty} |T_{k(t,N)}^{(N)} - t| = 0\} = 1.$$

In fact, by the strong law of large numbers, $\frac{\sum_{i=1}^{k} \xi_i}{k}$ converges to 1 as k goes to ∞ a.s. since this is the average of independent identically distributed variables with mean 1.

This discussion intuitively shows that, at any time t, the two processes $\overline{\mathbf{X}}_\mathbf{ds}^{(N)}(t)$ and $\overline{\mathbf{X}}_\mathbf{rs}^{(N)}(t)$ will have the same distribution in the limit N to ∞: the same number of events will have fired. This intuition can be formalized in the following theorem, whose proof can be found, for instance, in [54] (cf. proof of Theorem 17.28 of [54]).

[11] The rate of each event is $\frac{1}{\epsilon_N}$, so that the time to see event k is the sum of k exponentially distributed times with rate ϵ_N^{-1}.

Theorem 7.1. *Let* $\overline{\mathbf{X}}_{\mathbf{ds}}^{(N)}(t)$ *and* $\overline{\mathbf{X}}_{\mathbf{rs}}^{(N)}(t)$ *be defined as above. Then* $\overline{\mathbf{X}}_{\mathbf{ds}}^{(N)}(t)$ *converges to* $\overline{\mathbf{X}}_{\mathbf{rs}}^{(N)}(t)$ *weakly*[12]*. Furthermore, if either of the processes* $\overline{\mathbf{X}}_{\mathbf{ds}}^{(N)}(t)$ *and* $\overline{\mathbf{X}}_{\mathbf{rs}}^{(N)}(t)$ *converges to a limit process* $\overline{\mathbf{X}}(t)$ *in distribution, then so does the other*[13]*.*

This result has an interesting corollary, namely that in the limit, considering steps of fixed (DTMCs) or variable (CTMCs) length is irrelevant, as the limit behaviour is the same in both cases. Essentially, given a sequence of DTMCs which have a deterministic limit, we can always find a sequence of CTMCs which have the same limit, and vice versa, given a sequence of CTMCs with deterministic limit, there is a corresponding sequence of DTMCs with the same limit, both behaving in the same way for large N, as stochastic processes. Consequently, every limit result holding for DTMCs holds also for CTMCs, and vice versa, providing that both models satisfy the same scaling assumptions.

An immediate application of this fact is the extension of Theorem 6.1 (deterministic approximation for DTMC) along the lines of Theorem 5.2 (deterministic approximation of CTMC): if we have a subset $S \subseteq E$, we can study the exit time distribution from S in the limit of $N \to \infty$ in the same way as for CTMCs. Moreover, we can also estimate the number of steps required to escape from S, for large N.

8 Deterministic Approximation for synchronous DTMCs

In this section we will deal with a different kind of deterministic approximation result for DTMCs, in which the limit process is a discrete-time (deterministic) dynamical system.[14] The main difference with the approach presented in the previous sections is in how the DTMC is defined. In fact, we will consider models in which the evolution is synchronous, meaning, in essence, that at each step of the DTMC we perform the maximum possible number of transitions.

To explain this concept, consider again the example of the computer epidemic of Section 3.1, with a network of N interacting nodes. In the asynchronous model considered so far, at most a single node changes its state at each step of the evolution. In a synchronous model, instead, at each step all N nodes undergo a transition (not necessarily changing their state). Such evolution is better modeled by describing the behaviour of each single node, and then deriving the behaviour of the entire network, under the assumption that each node moves independently from the others. For instance, consider the activation of the worm in a latent node (i.e. the changing of state from E to I). Looking at a single node in state E, we can assume that this event happens with constant probability α_a. Therefore, at the network level, the number B of agents in the latent state that will transit to the infected state in the next step follows a Binomial distribution: $B \sim \mathcal{B}(E, \alpha_a)$. If we look at the fraction $\overline{E} = \frac{E}{N}$ of agents in the E state which become infected, this quantity is clearly $\frac{1}{N}B$. It is easy to check that, by the strong law of large numbers, $\frac{1}{N}B$ will almost surely converge to $\alpha_a \overline{E}$. Hence, it seems reasonable, for large N, to replace the synchronous DTMC with a deterministic process in discrete time, computing the fraction of nodes that change state in one time step as the average of the DTMC process.

In order to work with the synchronous semantics, we face a problem with the modeling language of Section 3: the transition-based description we use works well for defining the asynchronous one-transition-per-step semantics but not for defining the synchronous one. In fact, in this latter case,

[12] A sequence of processes $\overline{\mathbf{X}}^{(N)}(t)$, with values in E, converges weakly (or in distribution) to a process $\overline{\mathbf{X}}(t)$ on E if and only if the sequence of measures of the trajectory space associated with $\overline{\mathbf{X}}^{(N)}(t)$ (i.e., the space $D = D(\mathbb{R}, E)$ of right-continuous functions with left limits from $\mathbb{R}$ to E) converges (in the weak sense) to the measure of the trajectory space associated with $\overline{\mathbf{X}}(t)$. This, in turn, holds if and only if for each bounded continuous functional $f : D \to \mathbb{R}$, it holds that $E^{(N)}[f] \to E[f]$, where $E^{(N)}$ (resp. E) is the expectation with respect to the measure on D defined by $\overline{\mathbf{X}}^{(N)}(t)$ (resp. $\overline{\mathbf{X}}(t)$). The interested reader is referred to [54]

[13] If the limit process $\overline{\mathbf{X}}(t)$ is deterministic, then convergence in distribution of $\overline{\mathbf{X}}_{\mathbf{ds}}^{(N)}(t)$ or $\overline{\mathbf{X}}_{\mathbf{rs}}^{(N)}(t)$ to $\overline{\mathbf{X}}(t)$ implies convergence in probability, uniformly in $[0, T]$, for any finite time horizon T, see [54] for further details.

[14] A discrete time dynamical system (DTDS) is defined in the following way: given a function $f : E \to E$, and an initial point $\mathbf{x_0} \in E$, the DTDS $\mathbf{x}(k)$ at step k satisfies the relation $\mathbf{x}(k) = f(\mathbf{x}(k-1))$, giving rise to what is usually referred to as a set of difference equations.

the number of transitions increases with N. Think again of the epidemic model: we have to combine the behaviour of each node to produce transitions describing the evolution at the global level. For instance, the transition activating a latent worm (i.e. one changing internal state from E to I), can apply to a number of nodes in the latent state ranging from 0 to E. Hence we need $E+1$ global transitions (times the number of possible evolutions of the other "basic" transitions in the model). This number is very large, and it goes to infinity with N.

Nevertheless, even if our description language is not suitable to directly represent DTMCs with synchronous evolution, we will consider a specific class of models from which a synchronous semantics can be constructed in a relatively simple way. This class of models is essentially the one considered in [38]: we have systems of N interacting objects, which can change internal state (without births or deaths), each according to a given set of possible transitions. These systems will be modeled in our language according to the one-step clock-asynchronous paradigm. Then, we will show how to construct a synchronous semantics from this model.

In the following, we will first define this class of models, which will be termed *System-of-Independent-Objects* models (*SIO-models*), then we will present the deterministic approximation result, and finally we will go back to the network epidemic example.

SIO-model. A SIO-model captures the idea of a system composed of N objects, each of which can be in many different states, evolving independently of one another, with probabilities depending on the global state of the system (i.e. on the number of objects in any state).

Definition 8.1. $\mathcal{X} = (\mathbf{X}, \mathcal{D}, \mathcal{T}, \mathbf{x_0})$ is a SIO-model if it satisfies the following constraints:

1. The domain of each variable X_i is $\mathcal{D}_i = \{0, 1, \ldots, N\}$, for $i = 1 \ldots n$.
2. Let $\tau_{i,j} \in \mathcal{T}$ be a transition from state i to state j, $\tau_{i,j} = (a, \mathbf{e_i}, \mathbf{e_j}, p_{\tau_{i,j}}(\mathbf{X}))$, and $q_{\tau_{i,j}}(\mathbf{X})$ be the probability that a given entity in state i will make a transition to state j. (For simplicity, we require that, for each i and j, there is at most one transition leaving i and entering j.) Then $p_{\tau_{i,j}}(\mathbf{X}) = X_i \cdot q_{\tau_{i,j}}(\mathbf{X})$ is the probability that an entity changes state from i to j.
3. For each i and $\mathbf{X}$, $\sum_{l=1}^{N} q_{\tau_{i,l}}(\mathbf{X}) \leq 1$.

Of course SIO-models are also parametric with respect to the size N and sequences of such models are considered. Given a SIO-model $\mathcal{X} = (\mathbf{X}, \mathcal{D}, \mathcal{T}, \mathbf{x_0})$, we can construct the probability transition matrix $\mathbf{P}(\mathbf{X})$ for a single object in the model as, for $i \neq j$, $\mathbf{P}_{i,j}(\mathbf{X}) = q_{\tau_{i,j}}(\mathbf{X})$, if $\tau = (a, \mathbf{e_i}, \mathbf{e_j}, p_{\tau_{i,j}}(\mathbf{X}))$, $\mathbf{P}_{i,j}(\mathbf{X}) = 0$ if there is no such a transition, and $\mathbf{P}_{i,i}(\mathbf{X}) = 1 - \sum_{j \neq i} \mathbf{P}_{i,j}(\mathbf{X})$.

The evolution of this model is given by assuming that each object evolves independently from other objects, according to the probabilities given by matrix $\mathbf{P}(\mathbf{X})$. More precisely, the evolution in terms of the state counting variables $\mathbf{X}$ can be defined introducing N random vectors $\mathbf{B_i}(\mathbf{X})$ (depending on the current state $\mathbf{X}$), each distributed according to a multinomial distribution on X_i objects with probabilities $\mathbf{P}_{i,1}(\mathbf{X}), \ldots, \mathbf{P}_{i,n}(\mathbf{X})$. Denoting by $B_{i,j}(\mathbf{X})$ the j-th component of the random vector $\mathbf{B_i}(\mathbf{X})$ and by $\mathbf{X}(k)$ the DTMC defined in this way, we have that

$$X_j(k+1) = \sum_{i=1}^{n} B_{i,j}(\mathbf{X}(k)). \tag{18}$$

Deterministic Approximation Theorem. Consider now a sequence of normalized SIO-models $\overline{\mathcal{X}}^{(N)} = (\overline{\mathbf{X}}^{(N)}, \overline{\mathcal{D}}^{(N)}, \overline{\mathcal{T}}^{(N)}, \overline{\mathbf{d}_\mathbf{0}}^{(N)})$, with respect to system size $\gamma_N = N$, and assume that $\overline{\mathbf{d}_\mathbf{0}}^{(N)} \to \overline{\mathbf{d}}_\mathbf{0}$ (initial conditions converge). Notice that $\overline{\mathcal{D}}^{(N)}$ is a subset of the unit simplex $E = \mathcal{S}^n([0,1], 1)$. We can easily see that the synchronous evolution of $\overline{\mathbf{X}}^{(N)}$ is given by normalizing (18):

$$\overline{X}_j(k+1) = \sum_{i=1}^{n} \frac{1}{N} B_{i,j}(N \cdot \overline{\mathbf{X}}(k)), \tag{19}$$

where $B_{i,j}(\overline{\mathbf{X}})$ is the j-th component of the random vector $\mathbf{B_i}(\overline{\mathbf{X}})$, distributed according to the multinomial distribution on $N\overline{X}_i$ objects with probabilities $\mathbf{P}_{i,1}(\overline{\mathbf{X}}), \ldots, \mathbf{P}_{i,n}(\overline{\mathbf{X}})$. Applying the law of large numbers, we can see that $\frac{1}{N} B_{i,j}(\overline{\mathbf{X}}(k)) \to \overline{X}_i \mathbf{P}_{i,j}(\overline{\mathbf{X}})$ as $N \to \infty$. Hence, we expect that in the limit of N going to infinity, the synchronous model will behave as the following discrete time deterministic system:

$$\begin{cases} \overline{\mathbf{x}}(k+1) = \overline{\mathbf{x}}(k) \cdot \mathbf{P}(\overline{\mathbf{x}}(k)) \\ \overline{\mathbf{x}}(0) = \overline{\mathbf{d}}_{\mathbf{0}} \end{cases} \tag{20}$$

This fact is confirmed by the following theorem.

Theorem 8.1 (Mean Field Limit in Discrete Time [38])**.** *Let* $\overline{\mathbf{X}}^{(N)}(k)$ *and* $\overline{\mathbf{x}}(k)$ *be defined as above. Then, for any* $k > 0$*, almost surely*

$$\lim_{N \to \infty} \overline{\mathbf{X}}^{(N)}(k) = \overline{\mathbf{x}}(k).$$

The theorem states that, for any finite time horizon k, the behavior of $\overline{\mathbf{X}}^{(N)}(k)$ can be approximated by $\overline{\mathbf{x}}(k)$.

Example revisited. We now turn back to the example of the computer network epidemic of Section 3.1. Inspection of the definition of the DTMC transitions in Section 3.3 confirms that the model satisfies all the constraints of Definition 8.1; hence this is a SIO-model. Consequently, we can construct the probability transition matrix $\mathbf{P}$ for a single object:

$$\mathbf{P}(\overline{S}, \overline{E}, \overline{I}, \overline{R}) = \begin{pmatrix} 1 - (\alpha_e + \alpha_i \overline{I} + \alpha_p) & \alpha_e + \alpha_i \overline{I} & 0 & \alpha_p \\ 0 & 1 - (\alpha_a + \alpha_q) & \alpha_a & \alpha_q \\ 0 & 0 & 1 - \alpha_r & \alpha_r \\ \alpha_s & 0 & 0 & 1 - \alpha_s \end{pmatrix}$$

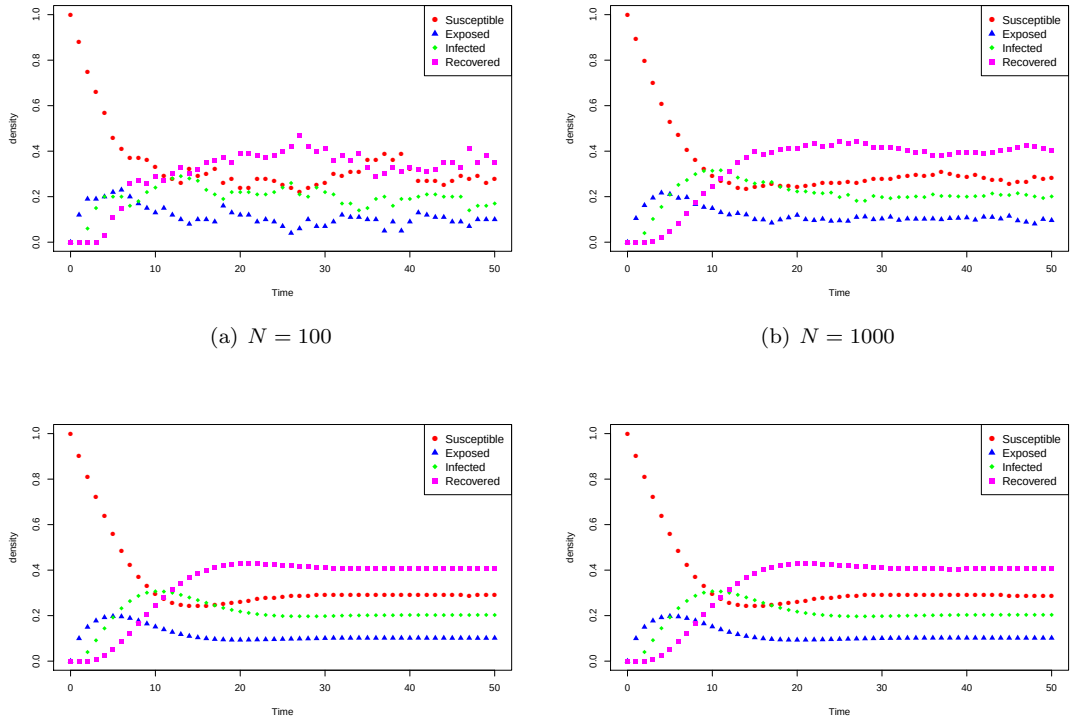
According to Theorem 8.1, the synchronous DTMC is asymptotically approximated by the following deterministic system:

$$\begin{cases} \overline{s}(k+1) = \overline{s}(k) - \alpha_e \cdot \overline{s}(k) - \alpha_i \cdot \overline{i}(k) \cdot \overline{s}(k) - \alpha_p \cdot \overline{s}(k) + \alpha_s \cdot \overline{r}(k) \\ \overline{e}(k+1) = \overline{e}(k) + \alpha_e \cdot \overline{s}(k) + \alpha_i \cdot \overline{i}(k) \cdot \overline{s}(k) - \alpha_a \cdot \overline{e}(k) - \alpha_q \cdot \overline{e}(k) \\ \overline{i}(k+1) = \overline{i}(k) + \alpha_a \cdot \overline{e}(k) - \alpha_r \cdot \overline{i}(k) \\ \overline{r}(k+1) = \overline{r}(k) + \alpha_p \cdot \overline{s}(k) + \alpha_q \cdot \overline{e}(k) + \alpha_r \cdot \overline{i}(k) - \alpha_s \cdot \overline{r}(k) \end{cases} \tag{21}$$

A visual representation of this result is given in Figure 11 , where we compare trajectories of the DTMC for increasing N with the trajectory of the deterministic limit.

*Remark** 8.1. In this section, we considered a restricted form of interaction between the objects constituting a given system. Essentially, the interaction between two objects is mediated by the environment (more precisely, by sensing the global distribution of object states), and direct cooperation is not allowed. This restriction has the advantage of simplifying the description of the functional form of interaction probabilities as well as simplifying the description of the one-step evolution of the system (cf. equation (18)). Considering direct forms of cooperation, like those encoded in transitions involving two or more objects, makes the matter much more complex. One approach in this direction is the one of [48], where the authors define a mean field semantics for a (clock-)synchronous process algebra (WSCCS — Weighted Synchronous Calculus of Communicating Systems) which encompasses a (restricted) form of direct interaction. Exporting their approach to our formalism, however, introduces some notational overhead, hence we omitted it.

Dealing with birth and death events in the context of SIO-models, instead, is less problematic. The only difference is that the normalized variables no longer represent a probability distribution, but rather they count the fraction of objects in any state, relatively to the initial population size.



(a) $N = 100$ (b) $N = 1000$

(c) $N = 10^6$ (d) Det. Limit

Figure 11: Comparison between the deterministic system (21) and stochastic trajectories of synchronous DTMC of the network epidemic example for increasing population sizes. Parameters of the model are $\alpha_e = 0.1$, $\alpha_i = 0.2$, $\alpha_a = 0.4$, $\alpha_r = 0.2$, $\alpha_p = 0$, $\alpha_q = 0$, and $\alpha_s = 0.1$, while initial conditions are $\overline{s}_0 = 1$, $\overline{e}_0 = 0$, $\overline{i}_0 = 0$, and $\overline{r}_0 = 0$.

*Remark** 8.2. The relationship between synchronous and asynchronous deterministic approximation limits of SIO-models is less clear and somehow less relevant than that between CTMC and DTMC. More specifically, the synchronous discrete time deterministic system associated with a SIO-model is easily shown to be the first-order Euler approximation [57] of the fluid ODE, with step-size equal to 1. Whether the discrete-time deterministic system shows a dynamics comparable to the one of the fluid ODE depends on the degree of stiffness and non-linearity of the latter, i.e. it depends on the accuracy of integrating the fluid ODE with the first-order Euler scheme with step-size equal to 1.

9 Discussion and Conclusions

In this section, we briefly discuss some more advanced topics in order to give the reader an impression of some of the consequences of the limit results. Detailed coverage of these research areas would be the basis of a further tutorial, and beyond the scope of this paper. Our intention here is just to offer a broader perspective on continuous approximation. Of course even this cannot be exhaustive and we do not consider mean field methods for stochastic models with spatial aspects, whose limits are generally expressed in terms of partial differential equations [58, 59], continuous approximation of Markov Decision Processes [60], in which limit results are used for the solution of optimization problems, or deterministic approximation algorithms [30], where these techniques, and many variations of them, have been widely applied. After this discussion we offer

our conclusions.

Fluid ODE and average behaviour of CTMC models Consider a sequence $\overline{\mathcal{X}}^{(N)}$ of CTMC models amenable of deterministic approximation. For any given N, one can compute the average value $\mathbb{E}[\overline{\mathbf{X}}^{(N)}(t)]$ of variables $\overline{\mathbf{X}}^{(N)}$, for any fixed time t. The deterministic approximation results of Section 5 guarantee that $\mathbb{E}[\overline{\mathbf{X}}^{(N)}(t)]$ converges in probability to $\overline{\mathbf{x}}(t)$, uniformly in any bounded time interval.

However, for any N, it is also possible to derive an ODE describing the exact evolution of $\mathbb{E}[\overline{\mathbf{X}}^{(N)}(t)]$ by standard arguments involving the Kolmogorov Forward Equation (or Master Equation) of the CTMC [61, 62] — the set of ODEs describing the time evolution of the probability mass in the state space of the CTMC. Using Taylor expansion or generating function arguments [61, 62, 8], it can easily be shown that the limit ODE is a *first order approximation* of the true average equation. It is also possible to show that higher order terms in the approximation vanish in the limit as N goes to infinity (hence, the larger N, the better the approximation works). Furthermore, the second order terms in the approximation depend on the variance and covariance of the process, third order terms depend on third order moments, and so on. Hence, it is also possible to derive coupled differential equations describing the variance and higher order moments, at the price of introducing further variables in the system of ODEs (precisely, $\Theta(n^k)$ new variables to include terms up to order k) [62, 8].

Decoupling of Joint Probability A perhaps surprising consequence of the deterministic results is that in a system comprised of N interacting objects, each of which may be in k distinct states, the joint distribution over the local states of the model displays asymptotic independence. Let $W_i^{(N)}(t) \in \{1, \ldots, k\}$ denote the state of the i-th object, in the N-th CTMC, at time t, and let $W_i(t)$ denote its state in the limit model. Then it follows from the deterministic nature of the limit and the (implicit) assumption that objects are indistinguishable, that

$$\begin{aligned}\mathbb{P}\{W_1(t) = k_1, \ldots, W_h(t) = k_h\} &= \mathbb{P}\{W_1(t) = k_1\} \cdot \ldots \cdot \mathbb{P}\{W_h(t) = k_h\} \\ &= \overline{x}_{k_1}(t) \cdot \ldots \cdot \overline{x}_{k_h}(t).\end{aligned}$$

Moreover, the deterministic approximation theorems presented in Sections 5, 6, and 8, guarantee that $\mathbb{P}\{W_1^{(N)}(t) = k_1, \ldots, W_h^{(N)}(t) = k_h\} \to \mathbb{P}\{W_1(t) = k_1, \ldots, W_h(t) = k_h\}$, and hence, for large N,

$$\mathbb{P}\{W_1^{(N)}(t) = k_1, \ldots, W_h^{(N)}(t) = k_h\} \approx \overline{x}_{k_1}(t) \cdot \ldots \cdot \overline{x}_{k_h}(t).$$

This asymptotic independence property is known as the *decoupling assumption* [38, 39]. Basically, it means that, in the limit, the evolution of each object becomes independent from the other objects. This happens even in models with explicit cooperation between objects. Consider, for instance, a model in which an object in state s_1 and an object in state s_2 cooperate and evolve to states s_3 and s_4, respectively, at rate kX_1X_2. In this case, there is a clear dependency in the model between objects in states s_1 and s_2, as their state can be changed simultaneously. Nevertheless, in the limit, this dependence is lost. In fact, we obtain the same limit process as for a model in which objects in state s_1 evolve to s_3 independently of objects in s_2 evolving to s_4, in both cases with rate kX_1X_2. In terms of transitions, the model containing the transition $(a_\tau, \mathbf{e_1} + \mathbf{e_2}, \mathbf{e_3} + \mathbf{e_4}, kX_1X_2)$ has the same *limit behaviour* of the model containing the transitions $(a_{\tau_1}, \mathbf{e_1}, \mathbf{e_3}, kX_1X_2)$ and $(a_{\tau_2}, \mathbf{e_2}, \mathbf{e_4}, kX_1X_2)$. Note that, by splitting the transition into two, the global rate R doubles, and so does the density of events.

It is also important to note that the asymptotic independence holds for any fixed time t, but not necessarily in the limit for $t \to \infty$. This is because the decoupling is based on theorems of Sections 5, 6, and 8, which hold in a finite time horizon.

Stationary Regime The deterministic approximation theorems for CTMCs and (asynchronous) DTMCs we presented in Sections 5 and 6 do not assume any special property of the limit system

of ODEs (apart from the existence and uniqueness of solutions, provided by the Lipschitz property of the vector field). A consequence of this very general setting is that these theorems guarantee convergence in any finite time horizon. However, they do not give any indication about the relationship to the stationary regime of the sequence of Markov Models, i.e. their behaviour as time goes to infinity. Indeed, as we argued in Remark 5.2, this is connected with the fact that the theorems have to work even for unstable trajectories, or in chaotic situations, i.e. for systems extremely sensitive to their initial conditions. It is therefore reasonable to expect that, assuming some stability properties of the system of ODEs, we may be able to obtain results also for the stationary regime. Indeed, this is the case: if the system of ODEs has a unique globally stable stationary point[15], we can show that the sequence of stochastic processes will end up precisely in that point [29, 39, 55].

Such results have been pursued for the (asynchronous) DTMC case with restriction on the possible transitions [29, 39]. Furthermore, they can be also applied to CTMC models, by virtue of the discussion in Section 7. Moreover, the results essentially depend on having exponential bounds for convergence in probability; hence they can be straightforwardly exported to the more general setting of this paper (see Remark 5.2).

For any vector field and initial starting point, the *Birkhoff centre*, $\mathcal{B}$, is the set of limit points within the trajectories generated by this field from this point. If we consider a sequence of Markov chains (in either deterministic or continuous time) that admit a deterministic limit, that limit will be defined by a vector field F on a compact space E. Each such Markov chain model has a finite state space (since E is compact) and so has one or more invariant probability distributions (invariant measure or steady state) [1], i.e. distributions that are invariant under the dynamics of the process.[16] If the chain is aperiodic and irreducible, then the invariant measure is unique [1]. Consider such an invariant measure μ_N for each N. Generally, the sequence μ_N does not have a limit, but it can be proved that from each subsequence one can extract another subsequence convergent (weakly) to a limit measure [63].

Theorem 9.1 ([39, 55])**.** *Every limit measure μ of a sequence of invariant measures μ_N has support contained in the Birkhoff centre $\mathcal{B}$.*

If the system of ODEs has a unique globally stable fixed point $\overline{\mathbf{d}}_{\mathbf{0}}$, meaning that the equation $F(\overline{\mathbf{d}}) = 0$ has a unique solution $\overline{\mathbf{d}}_{\mathbf{0}}$, then its Birkhoff centre is $\mathcal{B} = \{\overline{\mathbf{d}}_{\mathbf{0}}\}$. In this case, the following corollary can be derived from the previous theorem.

Corollary 9.1 ([39, 55])**.** *Every sequence of invariant measures μ_N converges weakly to the Dirac distribution on $\overline{\mathbf{d}}_{\mathbf{0}}$ (i.e. the distribution with mass concentrated on $\overline{\mathbf{d}}_{\mathbf{0}}$). Furthermore, if each $\overline{\mathbf{X}}^{(N)}(t)$ has a unique invariant distribution, then*

$$\lim_{N\to\infty}\lim_{t\to\infty}\overline{\mathbf{X}}^{(N)}(t) = \lim_{t\to\infty}\lim_{N\to\infty}\overline{\mathbf{X}}^{(N)}(t) = \overline{\mathbf{d}}_{\mathbf{0}}.$$

Hence, in this case we can exchange the limit in t and in N, an operation not possible in general.

When $\overline{\mathcal{X}}^{(N)}$ is a (normalized) model of a system of N objects, without births or deaths, then we can combine this corollary with the decoupling assumption, to obtain independence also for the stationary regime:

$$\lim_{t\to\infty}\mathbb{P}\{W_1(t) = k_1, \ldots, W_h(t) = k_h\} = \bar{x}_{k_1}\cdot\ldots\cdot\bar{x}_{k_h}.$$

This approach is at the basis of the so-called fixed point method [39], which approximates the stationary distribution of $\overline{\mathbf{X}}^{(N)}(t)$ with $\overline{\mathbf{d}}_{\mathbf{0}}$, when the vector field F has a unique solution. As

[15] A system of ODEs has a unique globally stable stationary point if and only if there is a point $\mathbf{y} \in E$ such that all trajectories converge to it in the limit, i.e. if for all $\mathbf{x_0} \in E$, denoting $\mathbf{x}(t)$ be the solution of ODE with initial point $\mathbf{x_0}$, it holds that $\lim_{t\to\infty}\mathbf{x}(t) = \mathbf{y}$.

[16] If the process $\overline{\mathbf{X}}^{(N)}(t)$ is initially distributed according to the invariant distribution μ, then its probability distribution at each time t will always be μ.

pointed out in [39], in order to apply this method one has further to check the global stability of this fixed point.

Fast Simulation The asymptotic independence property shared by models of interacting objects discussed above has an interesting consequence: it allows us to track the evolution of a single object ignoring all other objects, if N is sufficiently large. More specifically, we can consider the process $(W^{(N)}(t), \overline{\mathbf{X}}^{(N)}(t))$, where W tracks the evolution of the selected object, which is a Markov process. A consequence of the deterministic approximation theorems shows that $(W^{(N)}(t), \overline{\mathbf{X}}^{(N)}(t))$ converges to a jump process $(W(t), \overline{\mathbf{x}}(t))$, where $\overline{\mathbf{x}}(t)$ follows the solution of the fluid differential equation (and it is not influenced by the discrete state of the tracked object), while $W(t)$ evolves according to a time-inhomogeneous process depending on the state of the other objects only through $\overline{\mathbf{x}}(t)$. This kind of fast simulation scheme can be defined both for a DTMC [64] or a CTMC [52] model.

As an example, consider again the CTMC network epidemic model of Section 3.1. If we look at a single node in the network, we can define the time dependent infinitesimal generator matrix $\mathbf{Q}^{(N)}(t)$ at level N for $W^{(N)}(t)$ as

$$\mathbf{Q}^{(N)}(t) = \mathbf{Q}(t) = \begin{pmatrix} -\lambda_e - \lambda_i \overline{I}(t) - \lambda_p & \lambda_e + \lambda_i \overline{I}(t) & 0 & \lambda_p \\ 0 & -\lambda_a - \lambda_q & \lambda_a & \lambda_q \\ 0 & 0 & -\lambda_r & \lambda_r \\ \lambda_s & 0 & 0 & -\lambda_s \end{pmatrix},$$

which depends on N only via the fraction of infected agents present in the network at time t. Hence, in the limit of N going to infinity, $(W^{(N)}(t), \overline{\mathbf{X}}^{(N)}(t))$ converges to the process $(W(t), \overline{\mathbf{x}}(t))$, in which $\overline{\mathbf{x}}(t)$ is the solution of the system of ODE shown in Section 5, and $W(t)$ is a continuous-time jump process on the state state space $\{\overline{s}, \overline{e}, \overline{i}, \overline{r}\}$, with time-dependent generator matrix $\mathbf{Q}(t)$.

Hybrid Approximations The deterministic approximation results we considered in this paper rely on crucial assumptions on how the system scales as its size increases. In particular, they require that all variables of the model increase proportionally to the system size. This condition holds in many cases, but it may happen that some parts of our system cannot be assumed to behave in this way. Typical examples include network models where we have a constant number of servers and an increasing number of clients, or models of biological cellular systems explicitly representing genes, which are present in one or few copies within a cell.

In these cases, one may still want to consider deterministic approximation results only for those parts of the model which can be assumed to scale correctly with system size. The rest of the system, instead, has to be kept discrete. This leads to different deterministic approximation schemes, in which a sequence of stochastic models converges to a limit process whose nature depends on the structure of the model. In particular, we will briefly consider here sequences of CTMCs converging to hybrid limit processes, i.e. processes presenting an evolution in terms of differential equations and stochastic jumps [65, 66, 67], and sequences of DTMCs in which the intrinsically discrete component evolves so fast that it immediately reaches the equilibrium [39].

The former class of systems emerges naturally when we consider systems with entities whose number does not grow with N, such as genes in a cellular model. The evolution of the rest of the system, however, can depend on the internal state of these discrete entities. In this case, as we increase the size of the system, the sequence of CTMC models converges [66] to a hybrid system, in which the evolution is given in terms of ODEs (for those components scaling with N) and of a time-inhomogeneous jump process (for the components independent of N). These models belong to the class of Piecewise Deterministic Markov Processes [65].

The second class has been studied in [39], and includes models of objects interacting with a rapidly changing environment. In these models, we have N objects interacting with a resource R, which changes its state at each step of the DTMC. When we consider the sequence of DTMC models for increasing population levels and scale the time accordingly, the resource performs an

increasing number of transitions per unit time. When R has an equilibrium distribution, in the limit this equilibrium will provably have been reached, and the ODEs are consequently modified to include these effects, by averaging transition probabilities that depend on the state of the resource with respect to the equilibrium distribution of R. For example, if the limit probability p_τ of a transition τ depends on the state r of the resource, $p_\tau = p_\tau(\overline{\mathbf{x}}, r)$, then in the ODE we use $p_\tau = \sum_r \pi_R(r) \cdot p_\tau(\overline{\mathbf{x}}, r)$. The convergence to such a limit ODE has been proved in [39].

Finally, a different class of hybrid limit processes is obtained when the rate functions of a CTMC are discontinuous [68]. In these situations, assuming a certain regularity in the nature of discontinuity of the rate functions, the sequence of CTMC converges to the solution of a Piecewise Smooth Dynamical System [69]. In particular, the queue model of Section 5.6 falls into this last class of models.

Conclusions In this paper, we have presented continuous approximation results for stochastic Markov models, both in discrete time and in continuous time. These results have been well known in literature of stochastic approximations for 40 years. However they have only relatively recently attracted the attention of the performance community. Nevertheless they are rapidly becoming important tools for studying models of systems composed of many interacting objects, thus circumventing the state space explosion problem.

Our objective has been to give a tutorial that presents the main ideas of the field at a level which is accessible to those who do not necessarily have a detailed background in the relevant theory of stochastic processes and deterministic approximation. We have included a bibliography which references the relevant papers for the theory as well as some of those presenting applications and more advanced approximation techniques.

Our presentation has discussed the main approximation theorems, and several applications and extensions, in a uniform way. We have refrained from showing the mathematical details of the proofs, which can be found in excellent reference papers and books, such as [52, 50, 54], in order to emphasize a more intuitive account of the results. In particular, we have focused on the conditions that must be satisfied in order for the theorems to be applied, showing how to check them in practical examples.

References

[1] J. R. Norris, Markov Chains, Cambridge University Press, 1997.

[2] G. Bolch, S. Greiner, H. de Meer, K. Trivedi, Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications, Wiley, 2006.

[3] F. Ciocchetta, J. Hillston, Formal methods for computational systems biology, Springer-Verlag, 2008, Ch. Process algebras in systems biology, pp. 265–312.

[4] C. Baier, B. Haverkort, H. Hermanns, J.-P. Katoen, Model-Checking Algorithms for Continuous-Time Markov Chains, IEEE Trans. on Soft. Eng. 29 (6) (2003) 524–541.

[5] J. Hillston, Fluid flow approximation of PEPA models, in: Proceedings of the Second International Conference on the Quantitative Evaluaiton of Systems (QEST 2005), 2005, pp. 33–43.

[6] N. Geisweiller, J. Hillston, M. Stenico, Relating continuous and discrete PEPA models of signalling pathways, Theoretical Computer Science 404 (1-2) (2008) 97–111.

[7] L. Bortolussi, A. Policriti, Dynamical systems and stochastic programming: To ordinary differential equations and back, T. Comp. Sys. Biology 11 (2009) 216–267.

[8] R. Hayden, J. Bradley, A fluid analysis framework for a Markovian process algebra, Theor. Comput. Sci. 411 (22-24) (2010) 2260–2297.

[9] L. Cardelli, On process rate semantics, Theoretical Computer Science 391 (3) (2008) 190–215.

[10] J. Hillston, A Compositional Approach to Performance Modelling, Cambridge University Press, 1996.

[11] U. Herzog, EXL: Syntax, semantics and examples, Tech. Rep. IMMD7-16-90, Universität Erlangen-Nürnberg (1990).

[12] H. Hermanns, M. Rettelbach, Syntax, semantics, equivalences, and axioms for MTIPP, in: in Proc. of the 2nd Workshop on Process Algebras and Performance Modelling (PAPM '94), 1994, pp. 71–87.

[13] T. Kurtz, Solutions of ordinary differential equations as limits of pure jump Markov processes, Journal of Applied Probability 7 (1970) 49–58.

[14] J. Bradley, S. Gilmore, J. Hillston, Analysing distributed Internet worm attacks using continuous state-space approximation of process algebra models, Journal of Computer and System Sciences 74 (6) (2008) 1013–1032.

[15] M. Calder, S. Gilmore, J. Hillston, Automatically deriving ODEs from process algebra models of signalling pathways, in: Proceedings of Computational Methods in Systems Biology (CMSB), 2005.

[16] F. Ciocchetta, J. Hillston, Bio-PEPA: A framework for the modelling and analysis of biological systems, Theoretical Computer Science 410 (33-34) (2009) 3065 – 3084.

[17] M. Tribastone, S. Gilmore, Rigorous Software Engineering for Service-Oriented Systems, Vol. 6582 of LNCS, Springer-Verlag, 2011, Ch. Scaling Performance Analysis using Fluid-Flow Approximation.

[18] M. Tribastone, Relating layered queueing networks and process algebra models, in: Proceedings of the first joint WOSP/SIPEW International Conference on Performance Engineering, 2010, pp. 183–194.

[19] M. Massink, D. Latella, A. Bracciali, M. Harrison, A Scalable Fluid Flow Process Algebraic Approach to Emergency Egress Analysis, in: Proceedings of 8th IEEE International Conference on Software Engineering and Formal Methods (SEFM 2010), 2010, pp. 169–180.

[20] M. Massink, D. Latella, A. Bracciali, J. Hillston, Modelling non-linear crowd dynamics in Bio-PEPA, in: Proceeding of 14th International Conference on Fundamental Approaches to Software Engineering (FASE 2011), Vol. 6603 of LNCS, Springer, 2011, pp. 96–110.

[21] R. David, H. Alla, Discrete, Continuous, and Hybrid Petri Nets, 2nd ed., Springer-Verlag, 2010. doi:10.1007/978-3-642-10669-9_5.

[22] K. Trivedi, V. Kulkarni, FSPNs: Fluid stochastic Petri net, in: M. A. M. et al. (Ed.), Application and Theory of Petri Nets 1993, Vol. 691 of LNCS, Springer, 1993, pp. 24–31.

[23] G. Horton, V. Kulkarni, D. Nicol, K. Trivedi, Fluid stochastic Petri nets: Theory, applications and solution techniques, European Journal of Operational Research 105 (1998) 184–201.

[24] M. Silva, L. Recalde, On fluidification of petri net models: from discrete to hybrid and continuous models, Annual Reviews in Control 28 (2004) 253–266.

[25] D. Sumpter, D. Broomhead, Relating individual behaviour to population dynamics, Proceedings of the Royal Society B 268 (1470) (2001) 925–932. doi:10.1098/rspb.2001.1604.

[26] T. Kurtz, Approximation of population processes, SIAM, 1981.

[27] H. Andersson, T. Britton, Stochastic Epidemic Models and Their Statistical Analysis, Springer-Verlag, 2000.

[28] M. Benaïm, Recursive algorithms, urn processes and chaining number of chain recurrent sets, Ergodic Theory and Dynamical Systems 18 (01) (1998) 53–87.

[29] M. Benaïm, J. Weibull, Deterministic approximation of stochastic evolution in games, Econometrica 71 (3) (2003) 873–903.

[30] H. Kushner, G. Yin, Stochastic approximation and recursive algorithms and applications, Springer-Verlag, 2003.

[31] F. Baccelli, A. Chaintreau, D. D. Vleeschauwer, D. McDonald, HTTP turbulence, `hal.archives-ouvertes.fr`.

[32] F. Baccelli, M. Lelarge, D. McDonald, Metastable regimes for multiplexed TCP flows, in: Proceedings of 42nd Annual Allerton Conference on Communication, Control and Computing, 2004.

[33] A. Martinoli, K. Easton, W. Agassounon, Modeling swarm robotic systems: a case study in collaborative distributed manipulation, The International Journal of Robotics Research 23 (4-5) (2004) 415–436.

[34] S. Kumar, L. Massouli, Integrating streaming and file-transfer internet traffic: fluid and diffusion approximations, Queueing Systems 55 (4) (2007) 195–205.

[35] C. Bordenave, D. McDonald, A. Proutiere, A particle system in interaction with a rapidly varying environment: Mean field limits and applications, ArXiv Mathematics e-printsarXiv:arXiv:math/0701363v3.

[36] C. Bordenave, D. McDonald, A. Proutiere, Performance of random medium access control, an asymptotic approach, in: Proceedings of the 2008 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, 2008, pp. 1–12.

[37] G. Sharma, A. Ganesh, P. Key, Performance analysis of contention based medium access control protocols, IEEE Transactions on Information Theory 55 (4) (2009) 1665–1682.

[38] J. L. Boudec, D. McDonald, J. Mundiger, A generic mean field convergence result for systems of interacting objects, in: Proceedings of Fourth International Conference on the Quantitative Evaluaiton of Systems (QEST 2007), 2007, pp. 3–18.

[39] M. Benaïm, J. L. Boudec, A class of mean field interaction models for computer and communication systems, Performance Evaluation 65 (11-12) (2008) 823–838.

[40] A. Bobbio, M. Gribaudo, M. Telek, Analysis of large scale interacting systems by mean field method, in: Proceedings of Fifth International Conference on the Quantitative Evaluaiton of Systems (QEST 2008), 2008, pp. 215–224.

[41] R. Bakhshi, L. Cloth, W. Fokkink, B. Haverkort, Mean-field analysis for the evaluation of gossip protocols, in: Proceedings of Sixth International Conference on the Quantitative Evaluation of Systems (QEST 2009), 2009, pp. 247–256.

[42] W. Rudin, Principles of Mathematical Analysis, McGraw-Hill, 1976.

[43] R. De Nicola, D. Latella, M. Loreti, M. Massink, On a uniform framework for the definition of stochastic process languages, in: Proceedings of 14th International Workshop on Formal Methods for Industrial Critical Systems (FMICS 2009), 2009, pp. 9–25.

[44] M. Kwiatkowska, G. Norman, D. Parker, PRISM: Probabilistic symbolic model checker, in: T. Field, P. Harrison, J. Bradley, U. Harder (Eds.), Proc. 12th International Conference on Modelling Techniques and Tools for Computer Performance Evaluation (TOOLS'02), Vol. 2324 of LNCS, Springer, 2002, pp. 200–204.

[45] T. Dayar, H. Hermanns, D. Spieler, V. Wolf, Bounding the equilibrium distribution of Markov population models, Numerical Linear Algebra with Applications.

[46] M. Tribastone, S. Gilmore, J. Hillston, Scalable Differential Analysis of a Process Algebra Model, Transactions on Software Engeneering.

[47] L. Bortolussi, A. Policriti, Modeling biological systems in concurrent constraint programming, Constraints 13 (1) (2008) 66–90.

[48] C. McCaig, R. Norman, C. Shankland, From individuals to populations: A mean field semantics for process algebra, Theor. Comput. Sci. 412 (17) (2011) 1557–1580.

[49] S. CompBio Group, Institute for Systems Biology, Dizzy home page.

[50] T. Kurtz, S. Ethier, Markov Processes - Characterisation and Convergence, Wiley, 1986.

[51] R. Darling, Fluid limits of pure jump Markov processes: A practical guide, ArXiv Mathematics e-printsarXiv:arXiv:math/0210109.

[52] R. Darling, J. Norris, Differential equation approximations for markov chains, Probability Surveys 5 (2008) 37–79. doi:10.1214/07-PS121.

[53] T. Kurtz, Limit theorems for sequences of jump Markov processes approximating ordinary differential processes, Journal of Applied Probability 8 (1971) 244–356.

[54] O. Kallenberg, Foundations of Modern Probability, Springer-Verlag, 2002.

[55] R. Hayden, Convergence of ODE approximations and bounds on performance models in the steady-state, in: Proceedings of PASTA 2010, 2010.

[56] A. Jensen, Markov chains as an aid in the study of Markov processes, Skandinavisk Aktuarietidskriff 36 (1953) 8791.

[57] R. Burden, J. D. Faires, Numerical analysis, Thomson Brooks/Cole, 2005.

[58] C. Kipnis, C. Landim, Scaling limits of interacting particle systems, Springer-Verlag, 1999.

[59] A. Chaintreau, J. L. Boudec, N. Ristanovic, The age of gossip: spatial mean field regime, in: Proceedings of the Eleventh International ACM SIGMETRICS Joint Conference on Measurement and Modeling of Computer Systems, 2009, pp. 109–120.

[60] N. Gast, B. Gaujal, J. L. Boudec, Mean field for Markov decision processes: from discrete to continuous optimization, CoRR abs/1004.2342.

[61] N. V. Kampen, Stochastic Processes in Physics and Chemistry, Elsevier, 1992.

[62] L. Bortolussi, A master equation approach to differential approximations of stochastic concurrent constraint programming, in: Proceedings of the Sixth Workshop on Quantitative Aspects of Programming Languages (QAPL 2008), Vol. 220 of Electr. Notes Theor. Comput. Sci., 2008, pp. 163–180.

[63] P. Billingsley, Probability and Measure, John Wiley and Sons, 1979.

[64] H. Tembine, J. L. Boudec, R. El-Azouzi, E. Altman, Mean field asymptotic of Markov decision evolutionary games and teams, in: Proceedings of GameNets 2009, 2010, pp. 140–150. doi:10.1109/GAMENETS.2009.5137395.

[65] M. Davis, Markov Models and Optimization, Chapman & Hall, 1993.

[66] L. Bortolussi, Limit behavior of the hybrid approximation of stochastic process algebras, in: Proceedings of the 17th International Conference on Analytical and Stochastic Modeling Techniques and Applications (ASMTA), 2010, pp. 367–381.

[67] L. Bortolussi, V. Galpin, J. Hillston, M. Tribastone, Hybrid semantics for PEPA, in: Proceedings of the Seventh International Conference on the Quantitative Evaluation of Systems (QEST 2010), 2010, pp. 181–190.

[68] L. Bortolussi, Hybrid limits of continuous time Markov chains, in: Proceedings of the 8th International Conference on Quantitative Evaluation of SysTems (QEST), 2011.

[69] J. Cortez, Discontinuous dynamical systems: A tutorial on solutions, nonsmooth analysis, and stability, IEEE Control Systems Magazine 28 (3) (2008) 36–73.

A From SPAs to the Low Level Modeling Langage

In this section we show, by means of an example, how a Stochastic Process Algebra model can be easily translated into the low level language introduced in Section 3. We consider a simple example, taken from [5], of a system consisting of interacting processors and resources described in PEPA:

$$\begin{aligned} \mathit{Processor}_0 &= (\text{task}_1, r_1).\mathit{Processor}_1 \\ \mathit{Processor}_1 &= (\text{task}_2, r_2).\mathit{Processor}_0 \\ \mathit{Resource}_0 &= (\text{task}_1, r_1).\mathit{Resource}_1 \\ \mathit{Resource}_1 &= (\text{reset}, s).\mathit{Resource}_0 \end{aligned}$$

$$(\mathit{Resource}_0 || \mathit{Resource}_0) \underset{\{\mathit{task}_1\}}{\bowtie} (\mathit{Processor}_0 || \mathit{Processor}_0)$$

where:

- *rated action prefix* $(a, \lambda).P$ specifies the behaviour where an action of type a is executed with duration characterized by an exponentially distributed random variable the rate of which is λ;
- the *cooperation* composition $P_1 \underset{L}{\bowtie} P_2$ specifies the behaviour where any action of P_1 (P_2, respectively) which is not of any type in set L can be executed independently from P_2, while actions with type in L must be executed *together* by P_1 and P_2 and the rate of the resulting cooperation action is that of the *slowest* cooperating action.

The initial global state of the system above is composed of two (instances of) *Resources*, both in local state $\mathit{Resource}_0$, and two (instances of) *Processors*, both in local state $\mathit{Processor}_0$. Either of the Processors can cooperate with either of the Resources via the execution of a task_1, with rate $2 \cdot r_1$, thus bringing the system to a global state where the local state of one Processor is $\mathit{Processor}_1$, while the other one remains in local state $\mathit{Processor}_0$, and the local state of one Resource is $\mathit{Resource}_1$, while the other one remains in local state $\mathit{Resource}_0$. From this new state, three different transitions can take place:

- a task_1-transition, with rate r_1, bringing the system to a global state where both resources are used and both processors assigned to tasks of type task_1;
- a *reset*-transition, with rate s, bringing the system to a global state where both resources are free and one processor can be assigned to task_1 and the other to task_2;
- a task_2-transition, with rate r_2, bringing the system to a global state where both processors are free and one resource is still used;

The behaviour of the systems then continues in a similar way. We can, of course, imagine a generalization of the above specification, where instead of having just two resources and two processors we have to deal with P processors and R resources. The initial global state of such a system would have a total of $P + R$ entities. In particular P in local state $\mathit{Processor}_0$ and R in local state $\mathit{Resource}_0$ while no entity would initially be in state $\mathit{Processor}_1$ or $\mathit{Resource}_1$. In such a system, an initial cooperation transition on task_1 would occur with rate $2 \cdot \min\{P, R\}$.

For given values for r_1, r_2 and s, we can model the above system using our low level language as follows, letting $N =_{\text{def}} P + R$:

$$\mathcal{PR}^{(N)} =_{\text{def}} ((\mathit{Proc}_0, \mathit{Proc}_1, \mathit{Res}_0, \mathit{Res}_1), \mathcal{S}^4(\{0, \ldots, N\}, N), \mathcal{T}^{(N)}, (P, 0, R, 0))$$

where set $\mathcal{T}^{(N)}$ contains three global transitions, $\tau_{t1}, \tau_{t2}, \tau_{rs}$ with:

- $\tau_{t1} =_{\text{def}} (\mathit{task}_1, (1, 0, 1, 0), (0, 1, 0, 1), r_{t1}^{(N)})$ and $r_{t1}^{(N)}(\mathit{Proc}_0, \mathit{Proc}_1, \mathit{Res}_0, \mathit{Res}_1) =_{\text{def}} \min\{\mathit{Proc}_0, \mathit{Res}_0\} \cdot r_1$;
- $\tau_{t2} =_{\text{def}} (\mathit{task}_2, (0, 1, 0, 0), (1, 0, 0, 0), r_{t2}^{(N)})$ and $r_{t2}^{(N)}(\mathit{Proc}_0, \mathit{Proc}_1, \mathit{Res}_0, \mathit{Res}_1) =_{\text{def}} \mathit{Proc}_1 \cdot r_2$;

- $\tau_{rs} =_{\text{def}} (\mathit{reset}, (0,0,0,1), (0,0,1,0), r_{rs}^{(N)})$ and
  $r_{rs}^{(N)}(\mathit{Proc}_0, \mathit{Proc}_1, \mathit{Res}_0, \mathit{Res}_1) =_{\text{def}} \mathit{Res}_1 \cdot s.$

We want to stress that the translation from Stochastic Process Algebras to the low level language can be defined on the syntax of the algebras and fully automatized. The interested reader is referred to [5] for details. It should also be clear that similar translations can be defined for Probabilistic Process Algebras as well as just non-deterministic (i.e. qualitative) Process Algebras.