

# A Logical Characterization of Agent-Based Models

James F. Lynch\*  
Department of Computer Science  
Box 5815  
Clarkson University  
Potsdam, NY 13699-5815  
Telephone: 315-268-2374  
Fax: 315-268-2371  
email: jlynch@clarkson.edu

## Abstract

Agent-based models are a relatively new approach to modeling dynamical systems of interacting entities, for example molecules in a biological cell. Although they are computationally expensive, they have the capability of modeling systems more realistically than traditional state-variable models. We give a formal definition of agent-based models, which includes state-variable models as a special case. We examine the questions of when state-variable models are sufficient for accurate modeling of a system, and when agent-based models are necessary. We define notions of abstraction and approximation, and give sufficient conditions that imply that an agent-based model can be approximated by a deterministic state-variable model. We also give negative results: examples of agent-based models that cannot be approximated by any state-variable model.

## 1 Introduction

Many of the systems studied in biology, chemistry, and physics consist of populations of interacting entities. These are often referred to as “complex systems.” For example, the metabolism within a biological cell is modeled by a system of molecules that interact through various types of chemical reactions. These systems can be described at two levels. The local, fine-grained, or microscopic level provides data on individual entities and their relations. In a model of cellular metabolism, this might include properties of individual molecules such as conformational state, phosphorylation, methylation, and other attributes, and

---

\*The author thanks the Isaac Newton Institute for Mathematical Sciences, Cambridge, UK, for hospitality and support of this research during the Programme in Logic and Algorithms.

relations between molecules such as the presence or absence of various types of molecular bonds. The global, coarse-grained, or macroscopic level of description consists of aggregate properties such as free energy, temperature, concentrations of molecular species, or even just the presence or absence of a particular type of molecule.

Although the global properties are defined in terms of the local properties and are usually the features of interest, it is difficult to bridge the gap between the two levels. In many cases, the local behavior of complex systems is well-understood. But in most cases, even when the local behavior is comparatively simple, the global behavior cannot be explained, much less predicted, from it.

The dynamics of these systems has traditionally been modeled by ignoring the fundamental stochastic interactions between the entities and using a fixed number of real-valued variables to denote population level properties. Evolution equations (systems of differential or difference equations) involving the variables describe the state transitions. These are often called state-variable models. It is assumed that as the population sizes increase, the behavior of the system is asymptotic to that of the state-variable model. An additional simplification that is often applied is to approximate these variables by their averages (mean-field approximation), thus getting a deterministic state-variable model. A well-known example is the logistic equation [45]

$$\frac{dx_i}{dt} = x_i \left( r_i - \frac{x_i}{K} + \frac{1}{K} \sum_{j=1}^k a_{i,j} x_j \right)$$

where  $x_1, \dots, x_k$  are population sizes of  $k$  different species. State-variable models, especially deterministic models, have been the standard methodology for modeling populations of interacting entities, even though it was clear from the outset that they had severe limitations. One reason for this dominance is the simplicity and mathematical elegance of differential equations as a way of describing the interactions among large populations. Another reason is that no viable alternative was evident.

Differential equations are powerful and essential tools in many areas of science, but they have not yet enabled comparable advances in the study of complex systems. One of the limitations of state variable models is that in most cases analytic solutions of differential equations are unknown, and numeric simulation is the main tool. More fundamentally, population level behavior is an emergent property of the individual interactions, and even when the local behavior is characterized in great detail, and evolution equations conceptually exist, they may not be known.

Complex systems can be modeled at the local level by using variants of von Neumann's cellular automaton [9]. But mathematical analysis of these systems appears to be at least as difficult as solving differential equations, and until recently, simulating them was not practical. This has changed with the increasing availability of inexpensive computing power. This approach appears to have originated independently in various fields, including chemistry, ecology,

and physics. There are many variations of this modeling methodology, for example hybrid continuous/discrete versions, and they go by several names, such as configuration models and object-oriented models. In ecology, they are called individual-based models, and the term has spread to other areas of biology, but the most generic term is agent-based models.

In spite of the increasing reliance on agent-based modelling, progress is hindered by some of the same issues that arise in more traditional modeling. Simulation is still the main tool for reaching conclusions. In fact, it can be more difficult not only because of the high computational cost, but because there is no widely accepted language for describing agent-based models. The basic question of which aspects of the model are necessary and which can be abstracted away becomes even more difficult with agent-based models.

In this article, we propose a formal syntax and semantics for agent-based models. It is based on a term logic that defines both the state of an agent-based model at any time and the dynamics of the system. There are variants of agent-based models depending on whether the state space and time are discrete or continuous. Using our formal framework, we define notions of abstraction and pose questions about the accuracy of abstractions. In some cases, we provide partial answers. We give sufficient conditions for agent-based models to be abstracted to state-variable models. We also give some examples of agent-based models that do not satisfy these conditions and cannot be abstracted to state-variable models. We conclude with an outline of future research. We begin with some examples of agent-based models and state-variable models from disparate areas of science.

## 2 Examples

This section is independent of the technical sections of the article. We will describe some typical agent-based models that illustrate the basic concepts that will be formalized. We will not attempt a thorough survey, except to note that agent-based models are used in a wide variety of scientific fields. In physics they model phenomena over an enormous range of scales—from statistical mechanics [39] and plasma physics [7] to the formation of galaxies [37]. They are also used in economics [56]. Our examples are from chemistry and biology.

### 2.1 The Lotka-Volterra Equations

The Lotka-Volterra equations

$$\begin{aligned}\frac{dx_1}{dt} &= k_1x_1 - k_2x_1x_2 \\ \frac{dx_2}{dt} &= -k_3x_2 + k_2x_1x_2\end{aligned}\tag{1}$$

are a classic example of a state-variable model derived from underlying assumptions about an implicit agent-based model involving two species of interacting

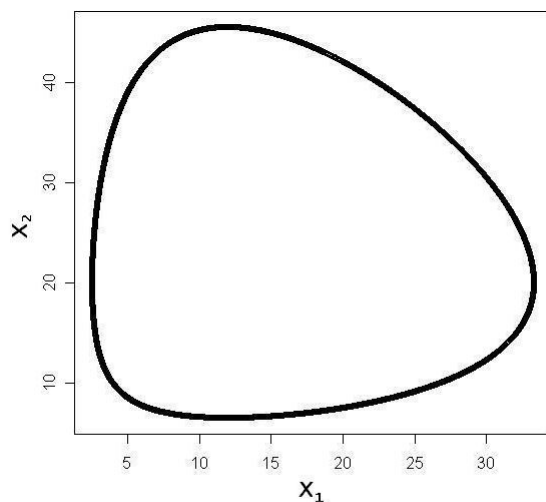


Figure 1: Trajectory of the Lotka-Volterra System

populations. Volterra [59] proposed them as a model of the interactions between a prey species whose population size is  $x_1$  and a predator species whose population size is  $x_2$ . In the first equation, the positive  $k_1$  term indicates that in the absence of predators, the prey will increase exponentially. The negative  $k_2$  term indicates that interactions with the predator cause a decline in the prey. The second equation has an analogous interpretation. Although the equations (1) are usually cited as a model of predator/prey dynamics, they were proposed independently and earlier by Lotka [43] as a model of an oscillating autocatalytic chemical reaction, where  $x_1$  and  $x_2$  are the amounts or concentrations of the two reactants. In both interpretations, the populations are spatially homogeneous and large enough so that the rate of interactions between two individuals from different populations is constant.

Analysis of these equations shows that, except for fixed points at  $(0, 0)$  and  $(k_3/k_2, k_1/k_2)$ , the system is periodic where  $x_1$  and  $x_2$  oscillate out of phase with each other (see Figure 1). This provides some insight into the many examples of oscillating chemical reactions and fluctuating populations of organisms. However, all of the actual examples discovered so far in nature involve different mechanisms. The chemical interpretation is perhaps more plausible because there are many systems of chemical reactions consisting of only two reactants and products. The interpretation in population biology is less so; an ecosystem comprised of exactly one prey and one predator species would be very rare indeed. Further, the model does not take into account the inevitable fluctuations in the prey's food supply. Thus the Lotka-Volterra model has limited

explanatory power and no predictive power. Nevertheless, it continues to be very influential as a paradigm for modeling complex systems. We list here some of the characteristics that it shares with many of the state-variable models and agent-based models that have followed it.

1. Populations of individuals are summarized by numeric variables such as population sizes, density, minimum and maximum values of certain attributes like age or weight, averages of these quantities, and so on. If this level of detail is inadequate, the populations can be partitioned into classes determined by certain attributes, for example many fish reproduce only once each year, and they can be categorized by their year-class. This may be significant in building an accurate model. But it still does not account for the individuals.
2. The states of the system are discrete, but the system is described by variables that vary continuously.
3. The atomic actions in the model (e.g., birth, death, predation) may be stochastic, but the rules governing the evolution of the aggregate properties of the system are deterministic.
4. Although the use of continuous, deterministic approximations may be appropriate in some regions of the state space, in other regions these approximations do not exhibit important behavioral features, and discrete, stochastic models must be used. We will illustrate this point with the Lotka-Volterra system shortly.
5. Some state-variable models or agent-based models are used to model phenomena on widely differing scales of space and time.
6. The Lotka-Volterra system has no known closed form analytic solution using elementary functions.

The first three of the above points pertain to assumptions that are often implicit in agent-based modelling. This is in spite of the fact that they do not always hold. We will formalize these assumptions within our logical framework, and then give sufficient conditions for their validity. We will also give examples of agent-based models that do not satisfy these conditions and violate the assumptions.

## 2.2 Chemical Kinetics

Lotka's interpretation of the Equations (1) is an example of chemical kinetics. This branch of physical chemistry pertains to populations of molecules that move randomly, for example a mixture of gases or a solution. The spatial distribution of the molecules is often ignored by assuming that, in any small time interval, their trajectories reach every part of the volume containing them. It can be proven rigorously [29] that this strong mixing assumption implies that

the probability of a given reaction occurring between specified molecules in the interval  $dt$  approaches  $\lambda dt$  as  $dt \rightarrow 0$ , for some constant  $\lambda$ . That is, the time until the particular reaction occurs for those molecules is an exponentially distributed random variable with rate  $\lambda$ . We also assume that at most one reaction can occur in a small time interval. Thus the system is a discrete space, continuous time Markov process, where the state of the system at any time is a vector of the numbers of molecules of each species in the system. For a given state the time until a reaction occurs is exponential with rate depending on the numbers of molecules of the species involved in the reaction and  $\lambda$ .

The Lotka-Volterra system can be modeled in this way by three state transitions that correspond to three reactions. In the following,  $y_1$  and  $y_2$  are the numbers of prey and predators respectively at the present time, and  $\lambda_1$ ,  $\lambda_2$ , and  $\lambda_3$  are the rate constants of the reactions.

Transition	Rate
$(y_1, y_2) \mapsto (y_1 + 1, y_2)$	$\lambda_1 y_1$
$(y_1, y_2) \mapsto (y_1 - 1, y_2 + 1)$	$\lambda_2 y_1 y_2$
$(y_1, y_2) \mapsto (y_1, y_2 - 1)$	$\lambda_3 y_2$

Table 1: Discrete Stochastic Version of the Lotka-Volterra System

If a chemical reaction system is enclosed by, say a cell membrane, then it can be described by a corresponding vector of chemical concentrations. If the number of molecules is large enough with respect to the volume of the container, i.e., the concentration is high enough, then in many cases the evolution of the system can be approximated by a ordinary differential equations as in Equations (1). The variables and parameters in (1) are related to the variables and parameters of the discrete stochastic version of the Lotka-Volterra system as follows.

$$x_i = \frac{y_i}{n_A V} \text{ for } i = 1, 2, 3,$$

where  $n_A$  is Avogadro's number, and  $V$  is the volume of the container.

$$\begin{aligned} k_1 &= \lambda_1 \\ k_2 &= n_A V \lambda_2 \\ k_3 &= \lambda_3. \end{aligned}$$

The choice of which model to use, continuous deterministic or discrete stochastic, is a matter of considerable importance, especially when the population sizes are small. In such cases, certain modes of behavior may be observed in discrete stochastic models that cannot occur in more traditional models based on deterministic differential equations. For example, although both versions of the Lotka-Volterra system show good agreement over a limited time span when the populations are of even modest size, say several hundred (see Figure 2), the two

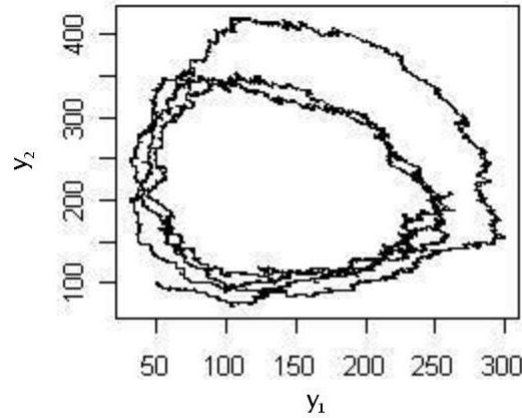


Figure 2: Discrete Stochastic Version of the Lotka-Volterra System

systems will eventually have very divergent behavior with probability 1. The continuous deterministic system will cycle through its states forever, but in the discrete stochastic system, extinction of one of the species is inevitable. If the prey vanish, then so will the predators; if the predators vanish, then the prey will experience a population explosion. In either case, the number of predators is ultimately 0 with probability 1. If either of the initial populations of prey or predator is small, this qualitative difference in behavior is likely to occur quickly.

There are many ways of implementing deterministic and stochastic kinetic models. One of the most widely used stochastic implementations is the so-called Gillespie algorithm [28], which is actually a method of discrete event simulation. It is not strictly an agent-based method because it works directly with population numbers and not individuals. Assuming there are  $m$  molecular species participating in  $r$  coupled reactions  $R_1, \dots, R_r$ , let the state of the system be  $\bar{x} = (x_1, \dots, x_m)$ , where  $x_i$  is the number of molecules of species  $i$ . The effect of applying reaction  $R_i$  to  $\bar{x}$  is denoted by  $R_i(\bar{x})$ . As with the stochastic version of the Lotka-Volterra system, each  $R_i$  has a rate constant  $\lambda_i$  for specified molecules, and there is a formula  $\rho_i(\bar{x}, \lambda_i)$  such that for a given state  $\bar{x}$ , the time until  $R_i$  occurs is an exponentially distributed random variable with rate  $\rho_i(\bar{x}, \lambda_i)$ .

Therefore the time until some reaction occurs will be exponential with rate

$$\rho_0(\bar{x}) = \sum_{i=1}^r \rho_i(\bar{x}, \lambda_i).$$

The Gillespie algorithm is an iteration of two random choices. At the beginning of each iteration, let  $t$  be the time and  $\bar{x}$  be the state. The following steps are executed:

1. Choose the time  $\Delta t$  until the next reaction using the exponential distribution with rate  $\rho_0(\bar{x})$ .
2. Choose  $i \in \{1, \dots, r\}$  with probability  $\rho_i(\bar{x}, \lambda_i)/\rho_0(\bar{x})$ .
3. Update  $t \mapsto t + \Delta t$  and  $\bar{x} \mapsto R_i(\bar{x})$ .

The Gillespie algorithm simulates the state transition probabilities exactly, subject to the limits of the computer’s accuracy. It is also very efficient provided the number of molecular species and reactions is not large. But many biological molecules can exist in different states, which affects their ability to react with other molecules. The Gillespie algorithm treats each molecular state as a separate species of molecule. Protein molecules can have more than one million states, and they can react with many other molecules, so the number of possible species and reactions can be much larger than the actual number of molecules being simulated. In such cases, it is more efficient to use a true agent-based simulator that represents each molecule as a software object whose attributes describe the state of the molecule.

StochSim (Stochastic Simulator) [24] is a well-known agent-based molecular simulator. It is a discrete time simulator. At each time interval  $dt$ , two molecules are randomly selected as candidates for a reaction. Unimolecular reactions are handled by including “dummy” molecules in the population, and allowing one of the selected molecules to be a dummy. Whether or not the selected molecules actually react is another random choice whose probability is determined by the attributes of the molecules. StochSim does not simulate the transition probabilities exactly since it is a discrete time simulator, but as  $dt \rightarrow 0$ , it approaches perfect accuracy.

The agent-based approach to molecular kinetics has other advantages. As we will discuss in the next subsection, the spatial arrangement of molecules can be significant, even in volumes as small as a bacterium. The location of a molecule can be one of its attributes. Also, the fate of an individual molecule can be followed.

## 2.3 Spatial Relationships

The choice of state transition (deterministic or stochastic, discrete or continuous) is not the only important decision facing modelers. The strong mixing assumption that is the basis of most kinetic modeling methodologies, including

the Gillespie algorithm, is not always valid for large molecules that move slowly or for larger volumes. Similar considerations apply at larger scales for certain models of biological dispersal and group behavior that we will describe later.

If the motion of the individuals obeys probabilistic laws, then another form of randomness must be accounted for. As before, there are choices between deterministic state-variable models and stochastic agent-based models. If deterministic differential equations are used, they are now partial differential equations because the transition rates are defined in terms of both space and time. If agent-based models are used and spatial values are represented exactly, then the state space is no longer discrete, and the system is a continuous space, continuous time Markov process.

Including an individual's spatial coordinates (and perhaps its velocity) among its attributes is easily done. But this approach is not widely used in modeling chemical reactions, probably due to the high computational cost. It is, however, used at larger biological scales by population biologists.

At the molecular level, a more common approach is to use some form of cellular automaton, where each "cell" represents a small region of the volume in which the reactions take place. Well-known examples of this type are reaction-diffusion models. They are also known as Turing models, after his 2-dimensional, two species model [58] of pattern formation during morphogenesis.

The cells in a  $d$ -dimensional reaction-diffusion model are  $d$ -dimensional cubes in some region of Euclidean  $d$ -space. Let the cells be indexed  $1, \dots, n$ . For each  $i = 1, \dots, n$ , let  $N_i$  be the set of indices of its nearest neighbors. For example, if  $d = 2$ ,  $N_i$  is the set of indices of the four cells adjacent to  $i$ . The state of cell  $i$  at time  $t$  is a vector  $\bar{c}(i) = (c_1(i), \dots, c_k(i))$  where  $c_j(i)$  is the concentration of species  $j$  in cell  $i$ . The transitions of the model are defined by a system of ordinary differential equations

$$\frac{dc_j(i)}{dt} = f_j(\bar{c}(i)) + D_j(\Delta c_j)(i) \quad (2)$$

for  $j = 1, \dots, k$ , where  $f_j(\bar{c}(i))$  is the change in  $c_j(i)$  due to reactions within cell  $i$ ,  $D_j$  is the diffusion coefficient of species  $j$ , and  $(\Delta c_j)$  is the discrete Laplacian

$$(\Delta c_j)(i) = \sum_{h \in N_i} (c_j(h) - c_j(i)).$$

That is,  $(\Delta c_j)(i)$  is the total difference in concentration levels of  $j$  between each neighbor of  $i$  and  $i$ , and  $D_j$  is a constant reflecting the rate at which the concentration of  $j$  will change due to diffusion. If the size of the cells gets vanishingly small, then using  $C_j(\bar{x})$  to represent the value of  $c_j(i)$  for any point  $\bar{x}$  in cell  $i$ , Each equation (2) approaches the partial differential equation

$$\frac{\partial C_j(\bar{x})}{\partial t} = f_j(\bar{C}(\bar{x})) + D_j(\nabla^2 C_j)(\bar{x}),$$

where  $\nabla^2$  is the differential Laplacian. More information on reaction-diffusion models and many other related classes of models is found in Kauffman's book [36].

Stochastic models of molecular kinetics can be extended to include spatial information. Gillespie [28] suggested a form of probabilistic cellular automaton, where the state of each cell is the vector of population numbers of the species. Within each cell, reactions are modeled by the Gillespie algorithm, and diffusion is modeled by motion between adjacent cells. His idea has been implemented in a variety of ways. See for example Ander et al. [2], Bernstein [6], and Burrage et al. [10].

Newer versions of StochSim implement a form of probabilistic 2-dimensional cellular automaton. Each cell represents exactly one molecule or molecular complex, and it can react only with its nearest neighbors. Shimizu et al. [53] have used it to model chemotaxis receptors embedded in a membrane. However, these complexes have fixed locations, and the system cannot model moving molecules.

## 2.4 Genomic and Biochemical Networks

The dynamical systems from chemistry, physics, and traditional biology pose substantial mathematical and computational problems. The essential difficulty is that their behavior is described by local, short-term interactions, while the properties that interest scientists are global and long-term. Physicists often refer to the two levels of behavior as microscopic and macroscopic. These systems do not provide much motivation for developing formal languages for agent-based models. The individuals and their interactions have concise mathematical descriptions. Their long-term temporal behavior can be very complex, but scientists are usually interested in specific temporal properties such as stability, lengths of transients, sizes of attractors, and chaos, rather than a large ensemble of properties that can be expressed in a formal language.

Systems biology has an additional level of complexity beyond physics and chemistry, which it shares with computer science. Genomic and biochemical networks are analogous to switching circuits, where genes, individual molecules or populations of molecules play the role of the switching elements. Software tools for specifying, testing, and verifying biological networks are becoming an integral part of systems biology research, for the same reasons that analogous tools are used in software and hardware development. Biologists are now realizing the limitations of informal, intuitive analysis of complex systems. For example, McAdams and Shapiro, writing about genomic regulatory systems [46], stated “As network size increases, intuitive analysis of feedback effects is increasingly difficult and error prone.” Von Dassow et al. [60] expressed similar thoughts: “Biologists’ maps of gene networks are rapidly outgrowing our ability to comprehend genetic mechanisms using human intuition alone.” This has led researchers to advocate a more “formal” approach [54, 55]. Quite often, these proposals merely mean using mathematical notation to describe a system quantitatively rather than qualitatively. But some biologists and computer scientists are now extending formal methods from software and hardware verification to biology [50]. Compared to computer science, formal methods for systems biology are still in a rather preliminary stage. This may be due partly to its relative

youth, and also to the greater complexity and range of microscopic behaviors. The review article of de Jong [19] covers a large repertoire of modeling and simulation methodologies that have been applied to genomic networks. Ultimately, analysis of biological systems will likely require a wide range of models for different aspects of each system. As Gibson and Mjolsness [27] have pointed out, progress in this area will require a deeper understanding of some fundamental problems about computational aspects of modeling, such as finding efficient algorithms for analyzing properties of models, and formulating rigorous justifications for the use of certain classes of models. These problems are somewhat metatheoretic because they pertain to classes of agent-based models from all areas of science. Indeed, ecologists have long recognized their importance.

## 2.5 Ecology

The growth of agent-based modelling in ecology, where it is referred to as individual-based modeling, has paralleled that in molecular biology; in fact it may have anticipated it. Many of the same issues that arise in molecular modeling are present in ecological modeling, except on much larger spatial and temporal scales. This is exemplified by the dual interpretation of the Lotka-Volterra system. Two areas of ecology where individual-based modeling has become widely accepted are trophic interactions and group behavior.

Trophic interactions include predation of one individual on another and competition between individuals (perhaps of the same species) for resources. An article by Huston et al. [35], which reviewed models of forest succession and fish population structure, was very influential in convincing ecologists of the value of individual-based modeling. Some of the advantages that Huston et al. described, and which we have already mentioned, depend on the ability of individual-based models to display variations in growth patterns that cannot be revealed if the individuals are aggregated. For example, if the initial height distribution of the trees in a forest has low variance, then the mature forest will consist of a large population of stunted trees, but a higher initial variance leads to a bimodal distribution in sizes, with a small proportion of very large trees and many small trees. Other examples of more realistic modeling resulting from individual-based modeling are given, but, according to Huston et al., its main contribution to ecology is conceptual. Ecosystems are very complex, and the traditional way of managing the complexity is to consider different levels of behavior in isolation, from physiological ecology, behavioral ecology, and so on up to community ecology and ecosystem ecology. These somewhat arbitrary and artificial distinctions are avoided by individual-based modeling which deals directly with the fundamental units of the ecosystem, and integrates the various levels in one model. This encourages more accurate modeling and can lead to a better understanding of the system. Huston et al. conclude with a prediction that individual-based modeling will develop rapidly and lead to important new insights into ecology.

It is interesting to note the tempered enthusiasm of the more recent retrospective article by Grimm [31]. While he does not deny the growing reliance

on individual-based modeling in ecology, he believes that some fundamental issues must be addressed before the full potential of the method can be realized. Specifically, a better understanding of the relationship between the individual-based and state-variable approaches is needed, and methods for determining the appropriate level of abstraction of a model must be developed. The book by Grimm and Railsback [32] goes even further in emphasizing the necessity of basing individual-based modeling on both theory and empirical evidence, and the authors also stress the importance of effective software development methodologies: “software engineering, not differential calculus, is the primary skill needed to implement and ‘solve’ models.”

Another kind of interaction is the cooperation of the members in a group. Behavioral ecologists are interested in the emergence of coordinated behavior of a group without global control, such as that found in fish schools [14, 51, 52]. In such groups, individuals may have different roles, but there are no designated leaders. The coordinated behavior is manifested by global features such as group extent, average velocity, and other properties defined by aggregate functions of the individual properties. Also, the transition rules often involve aggregate functions. For example, the state of each group member is typically determined by its location, velocity, and possibly other attributes. The rule for updating its state is an aggregate function of the relations of the individual to all other members of the group. In the cited articles, each fish  $x$  has a location specified by three coordinates  $X_1(x)$ ,  $X_2(x)$ , and  $X_3(x)$ . The terms

$$\nu_i(x, y) = \frac{X_i(y) - X_i(x)}{\sqrt{\sum_{j=1}^3 (X_j(y) - X_j(x))^2}}$$

for  $i = 1, 2, 3$  are the coordinates of the unit vector from  $x$  to  $y$ , which are used to compute the attraction or repulsion of  $x$  to or from  $y$ . For a given  $x$ , weighted averages of attraction/repulsion functions of the vectors  $(\nu_1(x, y), \nu_2(x, y), \nu_3(x, y))$  are used to update the location and velocity of  $x$ .

## 2.6 Graph Growth Models

The classic random graph model  $\mathfrak{G}(n, M)$  of Erdős and Rényi [23] can be regarded as a random process that starts with a set of  $n$  vertices and adds a random edge at each step. As is well-known, certain graph properties change rather rapidly as  $M$ , the number of edges increases. More recently, other random graph processes have been introduced as models of biological networks [13], the internet [38], and other types of networks that arise in computer science and engineering [1, 5, 11]. Like our previous examples, the update rules of these systems are local changes to the graph defined by aggregate functions.

A type of rule studied in [38] is known as preferential attachment. It models the tendency of popular web sites, i.e., ones that many sites link to, to become even more popular as the web grows. At each step, a new vertex is added to the graph, and a pre-existing vertex is selected, with probability proportional to its indegree. Then an edge from the new vertex to this vertex is added.

A duplication model imitates the way biological networks such as systems of genes add new vertices that are copies of older ones. The update rule is defined in terms of a probability  $p$ . At each step, a new vertex  $x$  is added and an older vertex  $y$  is randomly selected, all choices being equally likely. Then for every vertex  $z$  such that  $(y, z)$  is an edge, the edge  $(x, z)$  is added with probability  $p$ .

## 2.7 Other Formal Approaches

As we have indicated, some researchers have realized the value of formal languages for specifying and analyzing complex systems. Much, but not all, of this motivation comes from molecular biology. Currently, there are many efforts aimed at extending existing formal methods to systems biology and other areas, and some of them are agent-based. The following summary is not exhaustive; it merely illustrates the variety of approaches to modeling agents and their collective behavior.

The  $\pi$ -calculus [47] and its extensions can model systems where changes occur to one or two agents at a time. A basic feature of the  $\pi$ -calculus is communication between two processes or agents. This is a natural way to model chemical reactions because they involve at most two reactants. The stochastic  $\pi$ -calculus [49] can specify probabilistic reaction rates and has been used to model biomolecular networks. It can also model binary interactions at other scales, for example social organisms [57].

MGS [26] is an implementation of L-systems [42] augmented with rules describing transformations on multisets of symbols. It has been used to model developmental pathways in multicellular organisms. Dynamical Grammars [48] are another class of models with multiset rewriting rules. The rules also include constructs for specifying probabilities of rule execution.

Some modeling systems tailored toward specific applications are the brane calculi [12] for biological membrane interactions, the stochastic simulator of [17] for chemical kinetics, and the graph rewrite system of [18] for protein interactions.

## 3 The State Space of Agent-Based Models.

An agent-based model is a set of finite structures that evolve according to locally defined probabilistic rules. From this description, it might seem that finite models with probabilistic first-order update rules could describe all agent-based models. However, in many cases the interactions are defined by real numbers or arbitrarily large integers, and the properties of interest involve aggregations of the populations. Thus we must augment finite models with functions that assign numeric values to the relations and have aggregate functions, and we must use logics stronger than first-order logic. Similar considerations led Grädel and Gurevich [30] to propose metafinite structures as models of states of dynamical systems in computer science. We will define a closely related version of metafinite structure and logic. The only significant distinction is that our logic

(described in the next section) has two types of variables: those whose values range over the finite universe and those whose values range over the real numbers, whereas the terms in the logic of Grädel and Gurevich have only variables of the first type.

There are other ways of formalizing agent-based models, such as process algebras augmented with probabilistic operators [49], Abstract State Machines [33] with probabilistic updates, CLU models with probabilistic next-state expressions [8], or relational databases with probabilistic updates. We have chosen a formalism that is simple and direct and includes all examples of agent-based models we have encountered in the scientific literature.

A multiset  $M$  over a set  $S$  is a collection of elements from  $S$  that distinguishes the multiplicity (number of occurrences) of elements but not their order. We require that the multiplicity of each element is finite; thus  $M$  can be identified with a function  $m: S \rightarrow \mathbb{N}$ , where  $m(s)$  is the multiplicity of  $s$  in  $M$ . Our multisets will be finite, i.e.,  $m(s) > 0$  for only finitely many  $s \in S$ . We use  $\{\!\!\{ \}$  to enclose the members of a multiset. Let  $\text{fm}(S)$  be the collection of finite multisets over  $S$ . A multiset operation on  $S$  is a function  $\Gamma: \text{fm}(S) \rightarrow S$ .

Metafinite models are functional structures with three kinds of functions: weight functions, numeric functions, and multiset operations. A vocabulary is a triple  $(\mathcal{W}, \mathcal{F}, \mathcal{G})$  where  $\mathcal{W}$  is a set of weight function symbols,  $\mathcal{F}$  is a set of numeric function symbols, and  $\mathcal{G}$  is a set of multiset operation symbols. Each symbol is associated with an arity.

**Definition 1.** *A metafinite model  $\mathfrak{A}$  over the vocabulary  $(\mathcal{W}, \mathcal{F}, \mathcal{G})$  is a structure*

$$\langle A, \mathcal{W}^{\mathfrak{A}}, \mathcal{F}^{\mathfrak{A}}, \mathcal{G}^{\mathfrak{A}} \rangle$$

where

*$A$  is a finite set (the universe).*

*$\mathcal{W}^{\mathfrak{A}} = \langle w^{\mathfrak{A}} : w \in \mathcal{W} \rangle$ , where each  $w^{\mathfrak{A}}$  is a partial function from  $A^k$  to  $\mathbb{R}$  and  $w$  is  $k$ -ary.*

*$\mathcal{F}^{\mathfrak{A}} = \langle f^{\mathfrak{A}} : f \in \mathcal{F} \rangle$ , where each  $f^{\mathfrak{A}}$  is a function from  $\mathbb{R}^k$  to  $\mathbb{R}$  and  $f$  is  $k$ -ary.*

*$\mathcal{G}^{\mathfrak{A}} = \langle \Gamma^{\mathfrak{A}} : \Gamma \in \mathcal{G} \rangle$ , where each  $\Gamma^{\mathfrak{A}}$  is a multiset operation on  $\mathbb{R}$ .*

We will use uppercase Fraktur letters to denote metafinite models and their corresponding uppercase Roman letters (primed or subscripted) to denote their universes. We sometimes put  $w^{\mathfrak{A}}(a_1, \dots, a_k) = \text{undef}$  for  $(a_1, \dots, a_k) \notin \text{dom}(w^{\mathfrak{A}})$ .

A simple example of a metafinite model is a weighted graph  $\mathfrak{G}$  with vertex set  $G$  and edge weight function  $w^{\mathfrak{G}}: G^2 \rightarrow \mathbb{R}$ .

The states of a classical state-variable model are special cases of metafinite model. Their universe is  $\emptyset$ , and the weight functions are 0-ary. Thus the state can be identified with the vector  $\mathcal{W}^{\mathfrak{A}}$ .

The states of an agent-based model evolve by changing their weight functions. But their numeric functions and multiset operations are fixed, hence we will omit their superscripts.

The definition of isomorphism between two models easily extends to metafinite models.

**Definition 2.** *Let*

$$\begin{aligned}\mathfrak{A} &= \langle A, \mathcal{W}^{\mathfrak{A}}, \mathcal{F}, \mathcal{G} \rangle \text{ and} \\ \mathfrak{B} &= \langle B, \mathcal{W}^{\mathfrak{B}}, \mathcal{F}, \mathcal{G} \rangle\end{aligned}$$

*be metafinite models over the same vocabulary  $(\mathcal{W}, \mathcal{F}, \mathcal{G})$ . Then  $\mathfrak{A}$  and  $\mathfrak{B}$  are isomorphic, written  $\mathfrak{A} \cong \mathfrak{B}$ , if there is a 1-1 onto function  $f: A \rightarrow B$  such that for all  $w \in \mathcal{W}$  and  $a_1, \dots, a_k \in A$  where  $w$  is  $k$ -ary,  $w^{\mathfrak{A}}(a_1, \dots, a_k) = w^{\mathfrak{B}}(f(a_1), \dots, f(a_k))$ .*

## 4 A Logic for Agent-Based Models

We will use a pure term calculus to express properties of the states of agent-based models. Such properties are often called observables. We will use a closely related logic to define the transition rules. These logics have two types of variables: those that take values in the finite universe of the metafinite model, and those that take values in  $\mathbb{R}$ . We distinguish the two types by referring to them as agent and numeric variables respectively. In any term, we assume that the type of each variable has been declared. As in first-order logic, agent variables can be free or bound, but numeric variables cannot be bound. Thus when we say “free variable” or “bound variable,” we are implying that it is an agent variable. Let  $\mathfrak{A}$  be a metafinite model over the vocabulary  $(\mathcal{W}, \mathcal{F}, \mathcal{G})$  and  $\tau$  be a term with free variables among  $x_1, \dots, x_i$  and numeric variables among  $y_1, \dots, y_j$ . We will define  $\tau^{\mathfrak{A}}$  to be a function on  $A^i \times \mathbb{R}^j$ . If  $a_1, \dots, a_i \in A$  and  $r_1, \dots, r_j \in \mathbb{R}$ , then  $\tau^{\mathfrak{A}}(a_1, \dots, a_i; r_1, \dots, r_j)$  will denote the value of  $\tau^{\mathfrak{A}}$  when each  $a_k$  is assigned to  $x_k$  and each  $r_k$  is assigned to  $y_k$ . If  $\tau$  has exactly  $i$  distinct free variables and  $j$  distinct numeric variables, then it is of arity  $(i, j)$ .

**Definition 3.** *If  $x$  is declared as a numeric variable, then  $x$  is a term with numeric variable  $x$ .*

*If  $w \in \mathcal{W}$  is  $k$ -ary and  $x_1, \dots, x_k$  are agent variables, then  $w(x_1, \dots, x_k)$  is a term with free variables  $x_1, \dots, x_k$ .*

*If  $\tau_1, \dots, \tau_k$  are terms,  $x_1, \dots, x_i$  and  $y_1, \dots, y_j$  are the free and numeric variables respectively that occur in  $\tau_1, \dots, \tau_k$ , and  $f \in \mathcal{F}$  is  $k$ -ary, then  $f(\tau_1, \dots, \tau_k)$  is a term with free variables  $x_1, \dots, x_i$  and numeric variables  $y_1, \dots, y_j$ . For  $a_1, \dots, a_i \in A$  and  $r_1, \dots, r_j \in \mathbb{R}$ ,*

$$\begin{aligned}f(\tau_1, \dots, \tau_k)^{\mathfrak{A}}(a_1, \dots, a_i; r_1, \dots, r_j) \\ = f(\tau_1^{\mathfrak{A}}(a_1, \dots, a_i; r_1, \dots, r_j), \dots, \tau_k^{\mathfrak{A}}(a_1, \dots, a_i; r_1, \dots, r_j)).\end{aligned}$$

If  $\tau$  is a term with free variables  $x_1, \dots, x_i, z$ , numeric variables  $y_1, \dots, y_j$ , and  $\Gamma \in \mathcal{G}$ , then  $(\Gamma z \tau)$  is a term with free variables  $x_1, \dots, x_i$  and numeric variables  $y_1, \dots, y_j$ . For  $a_1, \dots, a_i \in A$  and  $r_1, \dots, r_j \in \mathbb{R}$ ,

$$(\Gamma z \tau)^{\mathfrak{A}}(a_1, \dots, a_i; r_1, \dots, r_j) = \Gamma(\{\tau^{\mathfrak{A}}(a_1, \dots, a_i, b; r_1, \dots, r_j) \mid b \in A\}).$$

Term logics can define aggregate properties of agent-based models. For example, the average outdegree of a directed graph can be defined by the term  $(\Sigma x (\Sigma y E(x, y))) / (\Sigma x V(x))$  where  $E^{\mathfrak{G}}(u, v) = 1$  if  $(u, v)$  is an edge of the graph  $\mathfrak{G}$  and  $= 0$  otherwise,  $V^{\mathfrak{G}}(v) = 1$  for all vertices  $v$ , and  $\Sigma$  is the summation operation on multisets of real numbers.

Numeric variables are used to describe structural properties of a metafinite model. In a model of diffusion, the weight functions  $X(p)$ ,  $Y(p)$ , and  $Z(p)$  could indicate the coordinates of a particle  $p$ . If  $x, y, z$ , and  $r$  are numeric variables, then the concentration of particles in the neighborhood of radius  $r$  of  $(x, y, z)$  can be defined by the  $(0, 4)$ -ary term

$$\frac{\sum(\{(x - X(p))^2 + (y - Y(p))^2 + (z - Z(p))^2 < r^2 \mid p \in A\})}{4\pi r^3/3}.$$

(We are using an informal syntax, where the relation  $<$  is actually a binary 0-1 valued function.) Another use of numeric variables, as shown in the next section, is to denote time in terms that describe state transitions of agent-based models.

The first-order logic  $\mathcal{L}$  of any predicate vocabulary can be embedded in a term logic  $\mathcal{L}'$  in the following way. For every  $k$ -ary relation symbol  $R$  in  $\mathcal{L}$ , the vocabulary of  $\mathcal{L}'$  has the  $k$ -ary weight function symbol  $\chi_R$ . Every  $\mathcal{L}$  model  $\mathfrak{A}$  can be expanded to an  $\mathcal{L}'$  model where each  $\chi_R$  is interpreted as the characteristic function of  $R$ : for  $a_1, \dots, a_k \in A$ ,

$$\chi_R^{\mathfrak{A}}(a_1, \dots, a_k) = \begin{cases} 1 & \text{if } \mathfrak{A} \models R(a_1, \dots, a_k) \\ 0 & \text{if } \mathfrak{A} \not\models R(a_1, \dots, a_k). \end{cases}$$

Function symbols of arity  $k$  are treated as  $(k + 1)$ -ary relation symbols. By also expanding  $\mathfrak{A}$  with numeric functions for the Boolean operations  $\vee$ ,  $\wedge$ , and  $\neg$  and the multiset operator  $\max$ , for every formula in  $\mathcal{L}$ , we can construct a term in  $\mathcal{L}'$  that is interpreted as the characteristic function of the formula. Recursively, let  $\phi$  be a formula in  $\mathcal{L}$  with free variables  $x_1, \dots, x_k, y$  and  $\chi_\phi$  be the term expressing its characteristic function. Then  $(\max y \chi_\phi)$  expresses the characteristic function of  $(\exists y \phi)$ .

## 5 Transition Rules

The evolution of an agent-based model is a stochastic process whose states are metafinite models over some specified vocabulary  $(\mathcal{W}, \mathcal{F}, \mathcal{G})$ . We put  $\mathfrak{A}_t$  for the state at time  $t$ . We assume that, for any  $t$  and later time  $t'$ , the probability

distribution of  $\mathfrak{A}_{t'}$  is determined by  $\mathfrak{A}_t$ . That is, the evolution is a Markov process. We will use an expanded vocabulary  $(\{A, A'\} \cup \mathcal{W} \cup \{w' | w \in \mathcal{W}\}, \mathcal{F}, \mathcal{G})$  to define the transition probabilities. A pair of states  $(\mathfrak{A}, \mathfrak{A}')$  can be regarded as a metafinite model over this expanded vocabulary, whose universe is  $A \cup A'$ . For  $a \in A \cup A'$ ,

$$A^{(\mathfrak{A}, \mathfrak{A}')} (a) = \begin{cases} 1 & \text{if } a \in A \\ 0 & \text{if } a \notin A \end{cases}$$

and similarly for  $A'$ . For a  $k$ -ary weight function  $w \in \mathcal{W}$  and  $a_1, \dots, a_k \in A \cup A'$ ,

$$w^{(\mathfrak{A}, \mathfrak{A}')} (a_1, \dots, a_k) = \begin{cases} w^{\mathfrak{A}}(a_1, \dots, a_k) & \text{if } a_1, \dots, a_k \in A \\ \text{undef} & \text{otherwise.} \end{cases}$$

and similarly for  $w'$ .

In general, the transition probabilities are defined from states to measurable sets of states by terms in the expanded logic. The precise form of these terms depends on the type of process (discrete or continuous time or space, time-dependent or homogeneous) and whether we are using declarative or operational semantics. For example, suppose  $\mathcal{S}$  is a set of metafinite models over  $(\mathcal{W}, \mathcal{F}, \mathcal{G})$ , we are using declarative semantics, and the sequence of terms  $\delta_1(t, t'), \dots, \delta_k(t, t')$  describes the transition from state  $\mathfrak{A}_t$  to  $\mathfrak{A}_{t'}$ . Fixing  $\mathfrak{A} \in \mathcal{S}$  and  $t, t' \in \mathbb{R}$  such that  $0 \leq t \leq t'$ , and letting  $\mathfrak{A}'$  range over  $\mathcal{S}$ ,  $(\delta_1^{(\mathfrak{A}, \mathfrak{A}')} (t, t'), \dots, \delta_k^{(\mathfrak{A}, \mathfrak{A}')} (t, t'))$  is a multivariate random variable on  $\mathcal{S}$  that characterizes the transition from  $\mathfrak{A}$  at time  $t$  to  $\mathfrak{A}'$  at time  $t'$ . We use a term  $F(x_1, \dots, x_k, t, t')$  in the expanded logic of  $(\mathcal{W}, \mathcal{F}, \mathcal{G})$  to define the conditional cumulative distribution function of the transition. For  $\mathfrak{A} \in \mathcal{S}$ ,  $r_1, \dots, r_k \in \mathbb{R}$ , and  $0 \leq t \leq t'$ ,

$$F^{\mathfrak{A}}(r_1, \dots, r_k, t, t') = \Pr \left( \bigwedge_{i=1}^k \delta_i^{(\mathfrak{A}, \mathfrak{A}')} (t, t') \leq r_i \mid \mathfrak{A}_t = \mathfrak{A} \right).$$

There are numerous alternatives, depending on additional assumptions about the process. For example, if  $F$  is differentiable in  $x_1, \dots, x_k$ , we can use the conditional probability density function  $f(x_1, \dots, x_k, t, t')$ :

$$f^{\mathfrak{A}}(r_1, \dots, r_k, t, t') = \frac{\partial^k F^{\mathfrak{A}}(x_1, \dots, x_k, t, t')}{\partial x_1 \dots \partial x_k} \Big|_{x_1=r_1, \dots, x_k=r_k}.$$

If the process is homogeneous, then the transition probability does not depend on  $t$ , and we can write  $F(x_1, \dots, x_k, \Delta t)$ , where  $\Delta t = t' - t$ . If the process operates in discrete time steps, then we take  $t' = t + 1$ , and we can write  $F(x_1, \dots, x_k, t)$ . Note, however, that it is not necessary to make these distinctions because time can be built into the state as a 0-ary weight function, and we really need to consider only transitions described by a term of the form  $F(x_1, \dots, x_k)$ .

If  $\mathcal{S}$  is discrete, i.e., the process is a Markov chain, we can define a conditional probability function  $g(x_1, \dots, x_k, t, t')$ :

$$g^{\mathfrak{A}}(r_1, \dots, r_k, t, t') = \Pr \left( \bigwedge_{i=1}^k \delta_i^{(\mathfrak{A}, \mathfrak{A}')} (t, t') = r_i \mid \mathfrak{A}_t = \mathfrak{A} \right).$$

Alternatively, we can use the transition rate function  $h(x_1, \dots, x_k, t)$ :

$$h^{\mathfrak{A}}(r_1, \dots, r_k, t) = \lim_{\Delta t \rightarrow 0} \frac{g^{\mathfrak{A}}(r_1, \dots, r_k, t, t + \Delta t) - g^{\mathfrak{A}}(r_1, \dots, r_k, t, t)}{\Delta t}$$

if the limit exists. Note that, since  $\Pr(\delta_i^{(\mathfrak{A}, \mathfrak{A}')} (t, t) = \delta_i^{(\mathfrak{A}, \mathfrak{A})} (t, t) \mid \mathfrak{A}_t = \mathfrak{A}) = 1$ ,

$$g^{\mathfrak{A}}(r_1, \dots, r_k, t, t) = \begin{cases} 1 & \text{if } \delta_i^{(\mathfrak{A}, \mathfrak{A})} (t, t) = r_i \text{ for all } i = 1, \dots, k \\ 0 & \text{otherwise.} \end{cases}$$

Another possibility for discrete state spaces is to define the transition probabilities directly for pairs of states:

$$g^{(\mathfrak{A}, \mathfrak{A}')} (t, t') = \Pr(\mathfrak{A}_{t'} = \mathfrak{A}' \mid \mathfrak{A}_t = \mathfrak{A}), \quad (3)$$

and similarly for the transition rate function.

Further simplifications are possible if the chain is homogeneous or operates in discrete time.

**Definition 4.** *An agent-based model is a pair  $(\mathcal{S}, F)$  where  $\mathcal{S}$  is a set of metafinite models over some vocabulary, and  $F$  is a term that defines the state transitions.*

Update rules can also be expressed in an imperative style, as in Abstract State Machines [33]. By expanding the set of numeric functions with random functions, probabilistic update rules can be defined. In some cases, for example diffusion processes, this may be the more natural style. The state of a diffusion process is a real-valued vector  $(v_1, \dots, v_k)$ . That is, a diffusion process is a state-variable model. The update rule of a diffusion process has the form

$$(v_1, \dots, v_k) := (v_1, \dots, v_k) + \mu(v_1, \dots, v_k) \Delta t + (W(t + \Delta t) - W(t)) \Lambda(v_1, \dots, v_k),$$

where  $\mu(v_1, \dots, v_k)$  is a  $k$ -dimensional drift vector,  $\Lambda(v_1, \dots, v_k)$  is a  $k \times k$ -dimensional diffusion matrix, and  $W$  is a  $k$ -dimensional Wiener process.

## 6 Abstractions of Agent-Based Models

All attempts at modeling a system, whether it is software, hardware, or biological, must balance accuracy of the model with simplicity of its description. This tradeoff may be the most significant decision faced by modelers. Ignoring important features results in models that do not accurately portray the system,

but including superfluous features leads to other difficulties. It decreases the efficiency of simulation, masks the important features, and makes it harder to understand the model. The agent-based approach makes it very easy to construct highly detailed models, but as pointed out by Grimm [31], “it seems as if many details are in the models simply because they make the model look more ‘realistic’.” Thus choosing the appropriate features becomes even more difficult in agent-based modelling.

The process of simplifying a model is called abstraction. The two most common forms of abstraction in agent-based modelling are removing some of the weight functions and summarizing population data with observables, which are often aggregate functions of the populations. Ignoring spatial information in kinetic models is an example of the former, and replacing populations of individuals with their sizes or concentrations is an example of the latter. Classical state-variable models completely abstract the individuals, replacing them with observables, and defining the state transitions in terms of the observables.

In many cases, the exact value of an observable is unknown, and an approximation is the best possible result. This may be due to the stochastic nature of the underlying phenomena, or to the experimental margin of error. It is often assumed that the error becomes small relative to the population sizes as they get large. Another common assumption is that as population sizes increase, the dynamics of the agent-based model is approximated by deterministic transitions. The systems of deterministic differential equations in Section 2 are examples.

Thus the common practice of modeling an agent-based model with a system of deterministic differential equations is really based on two simplifications:

1. Abstracting the agent-based model to a finite number of observables.
2. Assuming that as the size of the agent-based model increases, with high probability the observables follow a deterministic transition rule.

We next formalize a notion of abstraction. It will include as special cases the kinds of abstractions described above. We then define a notion of accuracy of an abstraction, and use it to justify the practice of approximating an agent-based model with a deterministic state-variable model. That is, we give sufficient conditions for an agent-based model to be approximated in this way and show that some of the previous examples satisfy these conditions. We also describe a recent technique for equation-free modeling of dynamical systems that is based on this kind of approximation.

**Definition 5.** *Let  $\mathcal{V} = (\mathcal{W}, \mathcal{F}, \mathcal{G})$  and  $\mathcal{V}' = (\mathcal{W}', \mathcal{F}, \mathcal{G})$  be vocabularies such that, for every  $w \in \mathcal{W}'$ , there is a term  $\tau_w$  in the logic of  $\mathcal{V}$  with the same arity as  $w$ . Let  $\mathcal{I} = (\mathcal{S}, F)$  and  $\mathcal{I}' = (\mathcal{S}', F')$  be agent-based models over vocabularies  $\mathcal{V}$  and  $\mathcal{V}'$  respectively. Suppose there is a map  $\alpha: \mathcal{S} \rightarrow \mathcal{S}'$  such that for every  $\mathfrak{A} \in \mathcal{S}$ , if  $\mathfrak{A}' = \alpha(\mathfrak{A})$ , then  $A' \subseteq A$ , and if  $w \in \mathcal{W}'$  has arity  $(i, j)$ , then for all  $a_1, \dots, a_i$  in  $A'$  and all  $r_1, \dots, r_j \in \mathbb{R}$ ,  $w^{\mathfrak{A}'}(a_1, \dots, a_i; r_1, \dots, r_j) = \tau_w^{\mathfrak{A}}(a_1, \dots, a_i; r_1, \dots, r_j)$ . We say that  $\alpha$  is an abstraction function and  $\mathcal{I}'$  is an abstraction of  $\mathcal{I}$  via  $\alpha$ .*

We now give a characterization of the accuracy of an abstraction.

**Definition 6.** Using the notation of Definition 5, for  $\mathfrak{A} \in \mathcal{S}$ ,  $\mathfrak{B} \in \mathcal{S}'$ ,  $s \in \mathbb{R}$ , and  $0 \leq t \leq t'$ , let

$$G^{\mathfrak{A}}(a_1, \dots, a_i; r_1, \dots, r_j, s, t, t') = \Pr(\tau_w^{\mathfrak{A}t'}(a_1, \dots, a_i; r_1, \dots, r_j) \leq s \mid \mathfrak{A}_t = \mathfrak{A})$$

and

$$H^{\mathfrak{B}}(a_1, \dots, a_i; r_1, \dots, r_j, s, t, t') = \Pr(w^{\mathfrak{B}t'}(a_1, \dots, a_i; r_1, \dots, r_j) \leq s \mid \mathfrak{B}_t = \mathfrak{B})$$

be the conditional cumulative distribution functions of  $\tau_w^{\mathfrak{A}t'}(a_1, \dots, a_i; r_1, \dots, r_j)$  and  $w^{\mathfrak{B}t'}(a_1, \dots, a_i; r_1, \dots, r_j)$  respectively.

Let  $\gamma \geq 0$  and  $\epsilon \in [0, 1]$ . For a given state  $\mathfrak{A} \in \mathcal{S}$ ,  $w \in \mathcal{W}'$ , and times  $t \leq t'$ , we say that  $\mathcal{I}'$  approximates  $\mathcal{I}$  with respect to  $\tau_w$  with accuracy  $\gamma$  and confidence  $\epsilon$  if for  $\mathfrak{A}' = \alpha(\mathfrak{A})$ , all  $a_1, \dots, a_i \in A'$  and  $r_1, \dots, r_j, s \in \mathbb{R}$ , there exists  $s' \in \mathbb{R}$  such that  $|s - s'| \leq \gamma$  and  $|G^{\mathfrak{A}}(a_1, \dots, a_i; r_1, \dots, r_j, s, t, t') - H^{\mathfrak{A}'}(a_1, \dots, a_i; r_1, \dots, r_j, s', t, t')| \leq 1 - \epsilon$ .

Note that we do not require that  $\mathcal{I}$  and  $\mathcal{I}'$  have the same kind of transitions (continuous or discrete in space or time), only that  $t$  is an integer if either  $\mathcal{I}$  or  $\mathcal{I}'$  has discrete transitions.

If all of the  $w \in \mathcal{W}'$  satisfy Definition 6, then we say that  $\mathcal{I}'$  approximates  $\mathcal{I}$  with accuracy  $\gamma$  and confidence  $\epsilon$  for the specified  $\mathfrak{A}$ ,  $t$  and  $t'$ .

A special case of approximation implies a form of lumpability or probabilistic bisimulation [20] of  $\mathcal{I}$ . An equivalence relation  $\equiv$  on the states of a Markov process is a lumping if it respects the transition probabilities. That is, for any states  $u$  and  $v$ ,  $u \equiv v$  implies that for any measurable set  $S$  of states that is closed under  $\equiv$ , the probability of a transition from  $u$  to  $S$  equals the probability of a transition from  $v$  to  $S$ . Thus, if the probability space of  $\mathcal{I}$  is generated by the sets  $\{\mathfrak{A} \mid \tau_w^{\mathfrak{A}}(a_1, \dots, a_i; r_1, \dots, r_j) \leq s\}$ , and  $\mathcal{I}'$  approximates  $\mathcal{I}$  with accuracy 0 and confidence 1 for all  $\mathfrak{A}$ ,  $t$ , and  $t'$ , then the equivalence relation  $\mathfrak{A} \equiv \mathfrak{B}$  if and only if  $\alpha(\mathfrak{A}) = \alpha(\mathfrak{B})$  is a lumping.

Generally, lumpability is too strict to be useful in practice, and it usually suffices to have an abstraction that approximates the agent-based model. The form of the approximation depends on the dynamical properties under investigation. For example, if the behavior of the discrete predator-prey system (Table 1 and Figure 2) during a bounded time interval is being studied, then the Lotka-Volterra approximation (Equations (1) and Figure 1) may be adequate. But if the long-term behavior of the system is important, then the Lotka-Volterra model is inappropriate. Our notion of approximation pertains the first situation, where an abstraction must approximate the agent-based model with a known degree of accuracy over a specified length of time. A common assumption is that the accuracy of an abstraction increases relative to the time interval as the state size increases. The notion of size is usually more complex than the number of individuals. For example, it could be the minimum number of individuals of certain types. Further, the approximation may hold only in certain regions of state space. For  $n \in \mathbb{N}$ ,  $\mathcal{R}_n$  will denote a region in  $\mathcal{S}$  containing states of size  $n$ .

**Definition 7.** Let  $w \in \mathcal{W}'$  and  $t \leq t'$ . Suppose, for every  $\gamma > 0$  and  $\epsilon \in (0, 1)$ , there is  $N$  such that if  $n \geq N$ , then  $\mathcal{I}'$  approximates  $\mathcal{I}$  with accuracy  $\gamma$  and confidence  $\epsilon$  with respect to  $\tau_w$  for all  $\mathfrak{A} \in \mathcal{R}_n$ . Then we say that  $\mathcal{I}$  converges to  $\mathcal{I}'$  with respect to  $\tau_w$  over  $[t, t']$ .

If Definition 7 holds for all  $w \in \mathcal{W}'$ , then we say that  $\mathcal{I}$  converges to  $\mathcal{I}'$ .

A further assumption is that the dynamics of an agent-based model converges to a deterministic state-variable model. In this case a simpler form of approximation holds. If  $\mathcal{I}'$  is a deterministic state-variable model with state set  $\mathcal{S}'$  and weight function symbol set  $\mathcal{W}'$ , then for every  $\mathfrak{B} \in \mathcal{S}'$ ,  $w \in \mathcal{W}'$ , and  $t \leq t'$ , there is  $s \in \mathbb{R}$  such that  $\Pr(w^{\mathfrak{B}_{t'}} = s \mid \mathfrak{B}_t = \mathfrak{B}) = 1$ . Therefore if  $\mathcal{I}'$  approximates  $\mathcal{I}$  with accuracy  $\gamma$  and confidence  $\epsilon$  for a given  $\mathfrak{A} \in \mathcal{S}$ ,

$$\Pr\left(\left|\tau_w^{\mathfrak{A}_{t'}} - w^{\mathfrak{B}_{t'}}\right| < \gamma \mid \mathfrak{A}_t = \mathfrak{A}\right) > 2\epsilon - 1.$$

If  $\mathcal{I}$  converges to  $\mathcal{I}'$  then for every  $\gamma > 0$  and  $\epsilon \in (0, 1)$ , for sufficiently large  $\mathfrak{A}$ ,

$$\Pr\left(\left|\tau_w^{\mathfrak{A}_{t'}} - w^{\mathfrak{B}_{t'}}\right| < \gamma \mid \mathfrak{A}_t = \mathfrak{A}\right) > \epsilon.$$

The following diagram illustrates the idea behind this definition.

$$\begin{array}{ccc} \mathfrak{A}_t & \xrightarrow{\Delta t} & \mathfrak{A}_{t+\Delta t} \\ \alpha \downarrow & & \downarrow \alpha \\ \alpha(\mathfrak{A}_t) = \mathfrak{B}_t & \xrightarrow{\Delta t} & \mathfrak{B}_{t+\Delta t} \approx \alpha(\mathfrak{A}_{t+\Delta t}) \end{array}$$

If  $\mathcal{I}'$  is a mean-field approximation, then  $w^{\mathfrak{B}_{t'}} = \mathbf{E}\left(\tau_w^{\mathfrak{A}_{t'}} \mid \mathfrak{A}_t = \mathfrak{A}\right)$ .

Although a state-variable model may be a continuous approximation of an underlying discrete system, it may reveal significant modes of behavior that are not evident in the discrete model. For example, questions of stability and existence of fixed points in the state space of a state-variable model are central to many of the modeling efforts described in Section 2. Of course, continuous state spaces must be approximated by discrete spaces in order to simulate them or apply formal methods. Thus there are two levels of approximation: from populations of discrete individuals to continuous state variables, and then to discrete approximations of the state variables. In their article on the second type of approximation, Desharnais, Edalat, and Panangaden [20] ask: “how does one argue that the discretized system is a faithful model of the underlying continuous system?” We ask a form of converse of this question, with regard to the first type of approximation.

Erban et al. [22] have used an equation-free method to model chemotaxis and other mechanisms of biological dispersal. Their method is applicable in regions of the state space of an agent-based model where the behavior is approximated by a deterministic state-variable model. The weight functions of the state-variable model are typically the lower moments of certain aggregate functions, such as concentration and distribution of momenta. Since these functions depend on space and time, the evolution equations are partial differential

equations. It is not necessary to use or even to know these equations. Instead, a Monte-Carlo simulation is run for a brief time, and the results are extrapolated to a much larger time. This can enable very significant speedups in total simulation time - a factor of one thousand according to [22]. The method consists of four steps, as illustrated below.

$$\begin{array}{ccc}
 \mathfrak{A}_t & \xrightarrow{\Delta t} & \mathfrak{A}_{t+\Delta t} \\
 \alpha^{-1} \uparrow & & \downarrow \alpha \\
 \mathfrak{B}_t & & \mathfrak{B}_{t+\Delta t} \xrightarrow{T} \mathfrak{B}_{t+\Delta t+T}
 \end{array}$$

The flow of the arrows shows the four steps:

1. Given  $\mathfrak{B}_t \in \alpha(\mathcal{S})$ , choose some  $\mathfrak{A}_t \in \alpha^{-1}(\mathfrak{B}_t)$ .
2. Use a Monte-Carlo simulator to evolve  $\mathfrak{A}_t$  for time  $\Delta t$ , getting  $\mathfrak{A}_{t+\Delta t}$ .
3. Estimate  $w^{\mathfrak{B}_{t+\Delta t}} \approx \tau_w^{\mathfrak{A}_{t+\Delta t}}$  and

$$\frac{\partial w}{\partial t}(t + \Delta t) \approx \frac{\tau_w^{\mathfrak{A}_{t+\Delta t}} - \tau_w^{\mathfrak{A}_t}}{\Delta t}.$$

4. Project

$$w^{\mathfrak{B}_{t+\Delta t+T}} \approx w^{\mathfrak{B}_{t+\Delta t}} + T \left( \frac{\tau_w^{\mathfrak{A}_{t+\Delta t}} - \tau_w^{\mathfrak{A}_t}}{\Delta t} \right).$$

## 7 Concentration Bounds

Here, we examine the rate of convergence of an observable to its mean-field approximation. We give properties of observables and update probabilities that guarantee a specified accuracy and confidence of the mean-field approximation for a specified starting state and time interval. These are properties of the observables and update probabilities as functions on the state space, and do not involve their logical definitions. In the following sections, we regard the observables and update probabilities as terms in our logic, and formulate logical conditions that imply the properties described in this section.

An estimate of the closeness of a random variable to its expectation is called a concentration bound. There are at least two reasons for obtaining tight concentration bounds in agent-based modeling. In general, they help in estimating the size of a model required for a given accuracy and confidence. More specifically, when an agent-based model converges to a deterministic state-variable model, concentration bounds can indicate the scale at which approximation by state-variable models is satisfactory. Since this size can be modest (on the order of several hundred individuals according to Gillespie [28]), lowering the bounds on population sizes needed for accurate simulation is more than just an academic exercise.

There are a variety of methods for obtaining concentration bounds. Classical methods such as Chebyshev's inequality and the Central Limit Theorem can be used, but they do not give very good bounds on the size needed to guarantee desired accuracy and confidence. Much smaller bounds can be derived from exponential tail bounds. We will use one that is based on Azuma's inequality for martingales [4]. We assume that the state space is discrete and the Markov chain is homogeneous and begin with the case when time is discrete.

## 7.1 Discrete Time

For simplicity, we will consider only one term  $\tau$ , but we could generalize our results to a finite sequence of terms. Also, since the free and numeric variables of  $\tau$  are fixed, we omit them. For  $t = 0, 1, \dots, T$ , let

$$\begin{aligned} Z_t &= \tau^{\mathfrak{A}_t} \text{ and} \\ Y_t &= \mathbf{E}(Z_T | \mathfrak{A}_0, \dots, \mathfrak{A}_t), \end{aligned}$$

the expectation of  $Z_T$ , conditioned on the first  $t + 1$  states  $\mathfrak{A}_0, \dots, \mathfrak{A}_t$ . Then

$$\begin{aligned} Y_0 &= \mathbf{E}(Z_T) \text{ and} \\ Y_T &= Z_T. \end{aligned}$$

The sequence of random variables  $Y_0, \dots, Y_T$  is a martingale with respect to the sequence  $\mathfrak{A}_0, \dots, \mathfrak{A}_T$ . That is, for every  $t = 0, \dots, T - 1$ ,

$$\mathbf{E}(Y_{t+1} | \mathfrak{A}_0, \dots, \mathfrak{A}_t) = Y_t.$$

(Of course, since  $\mathfrak{A}_0, \dots, \mathfrak{A}_t$  is a Markov chain,  $Y_t$  depends only on  $\mathfrak{A}_t$ , but the proof of the martingale property is just as easy for arbitrary stochastic processes.) To prove this fact,

$$\begin{aligned} \mathbf{E}(Y_{t+1} | \mathfrak{A}_0, \dots, \mathfrak{A}_t) &= \mathbf{E}(\mathbf{E}(Z_T | \mathfrak{A}_0, \dots, \mathfrak{A}_{t+1}) | \mathfrak{A}_0, \dots, \mathfrak{A}_t) \\ &= \mathbf{E}(Z_T | \mathfrak{A}_0, \dots, \mathfrak{A}_t) \end{aligned}$$

by the law of total probability for expectation

$$= Y_t.$$

In order to apply Azuma's inequality, we need to prove that the martingale  $Y_0, \dots, Y_T$  satisfies a Lipschitz or bounded differences condition. That is, we need to show that  $|Y_{t+1} - Y_t|$  is much smaller than  $T$  for  $t = 0, \dots, T - 1$ . Indeed, any method for proving a concentration bound requires some bound on the volatility of the sequence. We will state some conditions that are satisfied by many of the best-known examples of agent-based models, which imply a Lipschitz condition. Our conditions are smoothness properties on the term  $\tau$  that we wish to approximate and the transition probability function  $g^{(\mathfrak{A}, \mathfrak{A}' )}$ . Essentially, they say that transitions change  $\tau$  and its transition probabilities by a small amount.

**Definition 8.** Let  $\mathcal{I} = (\mathcal{S}, g)$  be an agent-based model where  $g$  is its transition probability function as in Equation (3). For any states  $\mathfrak{A}, \mathfrak{B} \in \mathcal{S}$ , we put  $\mathfrak{A} \rightarrow \mathfrak{B}$  if  $g^{(\mathfrak{A}, \mathfrak{B})} > 0$ .

We say  $g$  has bounded support if  $|\tau^{\mathfrak{B}} - \tau^{\mathfrak{A}}| \leq c$  for some  $c \in \mathbb{R}$  and all  $\mathfrak{A}, \mathfrak{B} \in \mathcal{S}$  such that  $\mathfrak{A} \rightarrow \mathfrak{B}$ .

For  $r \in \mathbb{R}$ , let

$$q^{\mathfrak{A}}(r) = \sum_{\substack{\mathfrak{B} \in \mathcal{S} \\ \tau^{\mathfrak{B}} - \tau^{\mathfrak{A}} = r}} g^{(\mathfrak{A}, \mathfrak{B})}$$

We say that  $g$  is smooth if, as  $n \rightarrow \infty$ , for all  $\mathfrak{A} \in R_n$  and  $\mathfrak{B}$  such that  $\mathfrak{A} \rightarrow \mathfrak{B}$ ,

$$q^{\mathfrak{B}}(r) \sim q^{\mathfrak{A}}(r).$$

For the remainder of this section, we assume the boundedness and smoothness conditions are satisfied.

**Lemma 1.** For any  $d, t \in \mathbb{N}$ , as  $n \rightarrow \infty$ , for  $\mathfrak{A} \in R_n$  and  $\mathfrak{B}$  such that  $\mathfrak{A} \xrightarrow{d} \mathfrak{B}$ ,

$$\mathbf{E}(Z_t - Z_0 | \mathfrak{A}_0 = \mathfrak{B}) \sim \mathbf{E}(Z_t - Z_0 | \mathfrak{A}_0 = \mathfrak{A}).$$

*Proof.* We will prove the lemma by induction on  $t$ . The lemma is trivial when  $t = 0$ .

Now assume the lemma for  $t$ , and prove it for  $t + 1$ .

$$\begin{aligned} \mathbf{E}(Z_{t+1} - Z_0 | \mathfrak{A}_0 = \mathfrak{B}) &= \mathbf{E}(Z_{t+1} - Z_1 | \mathfrak{A}_0 = \mathfrak{B}) + \mathbf{E}(Z_1 - Z_0 | \mathfrak{A}_0 = \mathfrak{B}) \\ &= \sum_{\mathfrak{B}'} g^{(\mathfrak{B}, \mathfrak{B}')} \mathbf{E}(Z_{t+1} - Z_1 | \mathfrak{A}_1 = \mathfrak{B}') + \sum_r q^{\mathfrak{B}}(r)r \\ &= \sum_{\mathfrak{B}'} g^{(\mathfrak{B}, \mathfrak{B}')} \mathbf{E}(Z_{t+1} - Z_1 | \mathfrak{A}_1 = \mathfrak{A}) (1 + o(1)) + \sum_r q^{\mathfrak{B}}(r)r \end{aligned}$$

by the induction assumption for  $t$

$$\sim \mathbf{E}(Z_{t+1} - Z_1 | \mathfrak{A}_1 = \mathfrak{A}) + \sum_r q^{\mathfrak{A}}(r)r$$

by the smoothness condition. Then reversing the above steps with  $\mathfrak{A}$  instead of  $\mathfrak{B}$ ,

$$\sim \mathbf{E}(Z_{t+1} - Z_0 | \mathfrak{A}_0 = \mathfrak{A}).$$

□

**Lemma 2.** For any  $t \in \mathbb{N}$ , as  $n \rightarrow \infty$ , for  $\mathfrak{A} \in R_n$  and  $\mathfrak{B}$  such that  $\mathfrak{A} \xrightarrow{1} \mathfrak{B}$ ,

$$|\mathbf{E}(Z_t - Z_1 | \mathfrak{A}_1 = \mathfrak{B}) - \mathbf{E}(Z_t - Z_0 | \mathfrak{A}_0 = \mathfrak{A})| < 2c.$$

*Proof.* We have

$$\begin{aligned}
& |\mathbf{E}(Z_t - Z_1 | \mathfrak{A}_1 = \mathfrak{B}) - \mathbf{E}(Z_t - Z_0 | \mathfrak{A}_0 = \mathfrak{A})| \\
&= |\mathbf{E}(Z_{t-1} - Z_0 | \mathfrak{A}_0 = \mathfrak{B}) - \mathbf{E}(Z_t - Z_{t-1} | \mathfrak{A}_0 = \mathfrak{A}) - \mathbf{E}(Z_{t-1} - Z_0 | \mathfrak{A}_0 = \mathfrak{A})| \\
&\leq |\mathbf{E}(Z_{t-1} - Z_0 | \mathfrak{A}_0 = \mathfrak{B}) - \mathbf{E}(Z_{t-1} - Z_0 | \mathfrak{A}_0 = \mathfrak{A})| + |\mathbf{E}(Z_t - Z_{t-1} | \mathfrak{A}_0 = \mathfrak{A})| \\
&\leq |\mathbf{E}(Z_{t-1} - Z_0 | \mathfrak{A}_0 = \mathfrak{A})| o(1) + c,
\end{aligned}$$

by Lemma 1 and the boundedness condition. By taking  $n$  sufficiently large, the lemma follows.  $\square$

**Corollary 1.** *For any  $T$ , for sufficiently large  $n$ , for all  $t = 1, \dots, T$ ,  $|Y_t - Y_{t-1}| \leq 3c$ .*

*Proof.* Restating Lemma 2 in terms of  $\mathfrak{A}_{t-1}$  and  $\mathfrak{A}_t$ ,

$$|Y_t - Z_t - (Y_{t-1} - Z_{t-1})| \leq 2c,$$

and since

$$|Y_t - Y_{t-1}| \leq |Y_t - Z_t - (Y_{t-1} - Z_{t-1})| + |Z_t - Z_{t-1}|,$$

the result follows.  $\square$

**Theorem 1.** *For any  $\gamma \in (0, \infty)$  and  $\epsilon \in [0, 1)$ , for sufficiently large  $T$  and  $n$ ,*

$$\Pr(|Z_T - \mathbf{E}(Z_T)| < \gamma T | \mathfrak{A}_0 = \mathfrak{A}) > \epsilon.$$

*Proof.* When  $n$  is sufficiently large with respect to  $T$ , the Lipschitz condition stated in Corollary 1 holds, and by Azuma's inequality,

$$\Pr(|Z_T - \mathbf{E}(Z_T)| > \gamma T) < 2 \exp\left(-\frac{\gamma^2 T^2}{18Tc^2}\right).$$

So by taking

$$\frac{18c^2 \ln(2/(1-\epsilon))}{\gamma^2} < T \tag{4}$$

the theorem follows.  $\square$

By strengthening the smoothness condition, we can show convergence to a deterministic state-variable model.

**Definition 9.** *We say that  $g$  is closed if, as  $n \rightarrow \infty$ , for all  $\mathfrak{A} \in R_n$  and  $\mathfrak{B}$  such that  $|\tau^{\mathfrak{A}} - \tau^{\mathfrak{B}}| \leq c$ ,*

$$q^{\mathfrak{B}}(r) \sim q^{\mathfrak{A}}(r).$$

**Theorem 2.** *If  $g$  is closed, there is a deterministic state-variable model  $\mathcal{I}'$  such that for any  $\gamma \in (0, \infty)$  and  $\epsilon \in [0, 1)$ , for sufficiently large  $T$  and  $n$ ,  $\mathcal{I}$  can be approximated by  $\mathcal{I}'$  with accuracy  $\gamma T$  and confidence  $\epsilon$  with respect to  $\tau$ .*

*Proof.* We will construct a deterministic state-variable model  $\mathcal{I}' = (\mathcal{S}', h)$  where  $\mathcal{S}' \subseteq \mathbb{R}$  and  $h: \mathcal{S}' \rightarrow \mathcal{S}'$  is its deterministic transition function. The abstraction function  $\alpha$  will be  $\alpha(\mathfrak{A}) = \tau^{\mathfrak{A}}$  for  $\mathfrak{A} \in \mathcal{S}$ .

Using induction on  $t = 0, 1, \dots$ , we define  $\mathcal{S}'_t \subseteq \mathbb{R}$  and  $h: \bigcup_{s=0}^{t-1} \mathcal{S}'_s \rightarrow \mathbb{R}$  so that the following conditions hold:

$$\begin{aligned} \mathcal{S}'_t &= \{w \mid w = h^t(\tau^{\mathfrak{A}}) \text{ for some } \mathfrak{A} \in \mathcal{S}\}. & (5) \\ |\mathbf{E}(\tau^{\mathfrak{A}_t} \mid \mathfrak{A}_0 = \mathfrak{A}) - h^t(\tau^{\mathfrak{A}})| &= o(t) \text{ for all } \mathfrak{A} \in \mathcal{S}. & (6) \end{aligned}$$

Starting with  $t = 0$  and letting  $\mathcal{S}'_0 = \{w \mid w = \tau^{\mathfrak{A}} \text{ for some } \mathfrak{A} \in \mathcal{S}\}$  and  $h = \emptyset$ , conditions (5) and (6) hold.

Assuming (5) and (6) hold for  $t$ , for  $\mathfrak{A} \in \mathcal{S}$ , let  $\beta^{\mathfrak{A}} = \mathbf{E}(Z_1 - Z_0 \mid \mathfrak{A}_0 = \mathfrak{A})$ . For  $w \in \mathcal{S}'_t - \bigcup_{s=0}^{t-1} \mathcal{S}'_s$  and  $n \in \mathbb{N}$ , choose some  $\mathfrak{A}_n \in \mathcal{R}_n$  such that  $h^s(\tau^{\mathfrak{A}_n}) = w$  for some  $s \leq t$  (if it exists). Then extend  $h$  to  $w$  so that

$$h(w) = w + \begin{cases} \beta^{\mathfrak{A}_n} & \text{if } n \text{ is the largest } n \text{ such that } \mathfrak{A}_n \text{ exists} \\ \lim_{n \rightarrow \infty} \beta^{\mathfrak{A}_n} & \text{otherwise.} \end{cases}$$

The limit exists in the second case because of the boundedness and closure conditions stated in Definitions 8 and 9. Putting  $\mathcal{S}'_{t+1} = h(\mathcal{S}'_t)$ , condition (5) holds for  $t + 1$ , and (6) holds by the conditions in Definition 8.

Taking  $\mathcal{S}' = \bigcup_{t=0}^{\infty} \mathcal{S}'_t$ , we show that  $\mathcal{I}$  is approximated by  $\mathcal{I}'$ . Let  $\gamma \in (0, \infty)$  and  $\epsilon \in [0, 1)$ . Take  $\mathfrak{A} \in \mathcal{R}_n$  and  $w = \tau^{\mathfrak{A}}$ . Then

$$\begin{aligned} |\tau^{\mathfrak{A}_T} - h^T(w)| &\leq |Z_T - \mathbf{E}(Z_T \mid \mathfrak{A}_0 = \mathfrak{A})| + |\mathbf{E}(Z_T \mid \mathfrak{A}_0 = \mathfrak{A}) - h^T(w)| \\ &\leq \gamma T/2 + To(1) \end{aligned}$$

with probability at least  $\epsilon$ , for sufficiently large  $T$  and  $n$ , by Theorem 1 and condition (6) above.  $\square$

## 7.2 Continuous Time

We now consider discrete space and continuous time agent-based models. Let  $g^{(\mathfrak{A}, \mathfrak{A}')}$  be the transition rate function. That is, for any  $\mathfrak{A}, \mathfrak{A}' \in \mathcal{S}$ ,

$$\lim_{\Delta t \rightarrow 0} \frac{g^{(\mathfrak{A}, \mathfrak{A}')}(\Delta t) - g^{(\mathfrak{A}, \mathfrak{A}')}(\Delta t - \Delta t)}{\Delta t} = g^{(\mathfrak{A}, \mathfrak{A}')}.$$
 (7)

We will obtain versions of Theorems 1 and 2 for this case by modifying their proofs. Letting

$$q^{\mathfrak{A}}(r) = \sum_{\substack{\mathfrak{B} \in \mathcal{S} \\ \tau^{\mathfrak{B}} - \tau^{\mathfrak{A}} = r}} g^{(\mathfrak{A}, \mathfrak{B})},$$

we replace  $g$  and  $q$  by  $g'$  and  $q'$  in Definition 8, and assume the revised versions of boundedness and smoothness hold.

**Definition 10.** For any state  $\mathfrak{A} \in \mathcal{S}$ , let

$$h^{\mathfrak{A}} = \sum_r q^{\mathfrak{A}}(r)r,$$

the instantaneous rate of change of the expectation, and

$$p^{\mathfrak{A}} = \sum_{\substack{\mathfrak{B} \in \mathcal{S} \\ \mathfrak{B} \neq \mathfrak{A}}} g^{(\mathfrak{A}, \mathfrak{B})},$$

the instantaneous rate at which the state changes.

**Theorem 3.** For any  $\gamma \in (0, \infty)$  and  $\epsilon \in [0, 1)$ , for sufficiently large  $n$ , there is  $\Delta t > 0$  such that

$$\Pr \left( |\tau^{\mathfrak{A}_{\Delta t}} - \tau^{\mathfrak{A}} - h^{\mathfrak{A}} \Delta t| < \gamma p^{\mathfrak{A}} \Delta t \mid \mathfrak{A}_0 = \mathfrak{A} \right) > \epsilon.$$

*Proof.* In the following,  $T$  will be a positive integer. The theorem will follow by taking  $T$  and  $n$  large enough to satisfy certain conditions given below. Let  $\Delta t = T/p^{\mathfrak{A}}$  and  $N$  be the number of state changes that occur in the interval  $[0, \Delta t]$  when  $\mathfrak{A}_0 = \mathfrak{A}$ . For any trajectory  $\{\mathfrak{A}_t \mid t \geq 0\}$  in the state space of  $\mathcal{I}$ , there is an associated sequence  $\mathfrak{A}'_0, \mathfrak{A}'_1, \dots$  consisting of those states where a transition occurs. That is,  $\mathfrak{A}'_0 = \mathfrak{A}$ ,  $\mathfrak{A}'_1 = \mathfrak{A}_t$  where  $\mathfrak{A}_s = \mathfrak{A}_0$  for  $0 \leq s < t$  and  $\mathfrak{A}_t \neq \mathfrak{A}_0$ , and so on. Thus  $\mathfrak{A}_{\Delta t} = \mathfrak{A}'_N$ .

By the boundedness and smoothness conditions,  $p^{\mathfrak{A}'_T} \sim p^{\mathfrak{A}}$  as  $n \rightarrow \infty$ . Thus  $\mathbf{E}(N)$  is bounded below by the expectation of a Poisson random variable with mean  $p^{\mathfrak{A}}(1 - o(1))\Delta t$  and above by the expectation of a Poisson random variable with mean  $p^{\mathfrak{A}}(1 + o(1))\Delta t$ . Similarly,

$$((p^{\mathfrak{A}} \Delta t)^2 + p^{\mathfrak{A}} \Delta t)(1 - o(1)) \leq \mathbf{E}(N^2) \leq ((p^{\mathfrak{A}} \Delta t)^2 + p^{\mathfrak{A}} \Delta t)(1 + o(1)).$$

That is,

$$\begin{aligned} \mathbf{E}(N) &\sim T \text{ and} \\ \text{var}(N) &= T + o(T^2). \end{aligned}$$

Using Chebyshev's inequality and noting that  $h^{\mathfrak{A}'}/p^{\mathfrak{A}'} \leq c$ , for sufficiently large  $T$  and  $n$ , with probability at least  $\sqrt{\epsilon}$ ,  $|N - \mathbf{E}(N)| < \gamma p^{\mathfrak{A}} T / (2h^{\mathfrak{A}'})$ , and therefore and  $|N - T| < \gamma p^{\mathfrak{A}} T / (2h^{\mathfrak{A}'})$ .

Now consider the discrete process  $\mathfrak{A}'_0, \mathfrak{A}'_1, \dots$ , where the probability of a transition  $\mathfrak{A}'_t \rightarrow \mathfrak{A}'_{t+1}$  is

$$\frac{g^{(\mathfrak{A}'_t, \mathfrak{A}'_{t+1})}}{p^{\mathfrak{A}'_t}}.$$

Letting  $\Pr'$  and  $\mathbf{E}'$  denote the probability and expectation of events in this

space,

$$\begin{aligned}
\mathbf{E}'(\tau^{\mathfrak{A}'_N}) &= \tau^{\mathfrak{A}} + \sum_{t=0}^{N-1} \mathbf{E}'(\tau^{\mathfrak{A}'_{t+1}} - \tau^{\mathfrak{A}'_t}) \\
&= \tau^{\mathfrak{A}} + \sum_{t=0}^{N-1} \frac{h'^{\mathfrak{A}'_t}}{p'^{\mathfrak{A}'_t}} \\
&= \tau^{\mathfrak{A}} + N \frac{h'^{\mathfrak{A}}}{p'^{\mathfrak{A}}} (1 + o(1))
\end{aligned}$$

as  $n \rightarrow \infty$ . Since  $|N - T| < \gamma p'^{\mathfrak{A}} T / (2h'^{\mathfrak{A}})$ ,

$$|\mathbf{E}'(\tau^{\mathfrak{A}'_N}) - \tau^{\mathfrak{A}} - h'^{\mathfrak{A}} \Delta t| \leq \frac{\gamma T}{2}.$$

By Theorem 1, for sufficiently large  $N$  and  $n$ ,

$$\Pr'(|\tau^{\mathfrak{A}'_N} - \mathbf{E}'(\tau^{\mathfrak{A}'_N})| \leq \gamma N/4) \geq \sqrt{\epsilon}.$$

Using Chebyshev's inequality again, for sufficiently large  $T$ ,  $|N - T| \leq T$  with probability at least  $\sqrt{\epsilon}$ , and since

$$\begin{aligned}
|\tau^{\mathfrak{A}_{\Delta t}} - \tau^{\mathfrak{A}} - h'^{\mathfrak{A}} \Delta t| &= |\tau^{\mathfrak{A}'_N} - \tau^{\mathfrak{A}} - h'^{\mathfrak{A}} \Delta t| \\
&\leq |\tau^{\mathfrak{A}'_N} - \mathbf{E}'(\tau^{\mathfrak{A}'_N})| + |\mathbf{E}'(\tau^{\mathfrak{A}'_N}) - \tau^{\mathfrak{A}} - h'^{\mathfrak{A}} \Delta t| \\
&\leq \frac{\gamma T}{2} + \frac{\gamma T}{2} \\
&= \gamma p'^{\mathfrak{A}} \Delta t,
\end{aligned}$$

the theorem follows.  $\square$

By strengthening the smoothness condition, we can show convergence to a deterministic state-variable model.

**Definition 11.** We say that  $g'$  is closed if there is a continuous function  $f: \mathbb{R} \rightarrow \mathbb{R}$  such that as  $n \rightarrow \infty$ , for all  $\mathfrak{A} \in \mathcal{R}_n$

$$h'^{\mathfrak{A}} \sim f(\tau^{\mathfrak{A}}).$$

**Theorem 4.** If  $g'$  is closed, there is a deterministic state-variable model  $\mathcal{I}'$  such that for any  $\gamma \in (0, \infty)$  and  $\epsilon \in [0, 1)$ , there is  $\Delta t > 0$  such that for sufficiently large  $n$ ,  $\mathcal{I}$  can be approximated by  $\mathcal{I}'$  with accuracy  $\gamma p'^{\mathfrak{A}} \Delta t$  and confidence  $\epsilon$  with respect to  $\tau$ . Further, the evolution of the state of  $\mathcal{I}'$  satisfies a system of ordinary differential equations.

*Proof.* By Picard's method, there is a function  $F: \mathbb{R} \times \mathbb{R}^+ \rightarrow \mathbb{R}$  such that

$$\begin{aligned}
\frac{\partial F(w, t)}{\partial t} &= f'(w) \text{ and} \\
F(w, 0) &= w.
\end{aligned}$$

Then  $\mathcal{I}' = (\mathbb{R}, F)$  is a deterministic state-variable model. Since  $\frac{\partial F(w,t)}{\partial t} = f'(w)$ , by Theorem 3, for sufficiently large  $n$ ,

$$\Pr(|\tau^{\mathfrak{A}\Delta t} - F(\tau^{\mathfrak{A}}, \Delta t)| < \gamma p^{\mathfrak{A}} \Delta t$$

with probability at least  $\epsilon$ . □

## 8 Locality

In this section, we describe a canonical form for terms. It will be used in the following section to characterize conditions when an agent-based model can be approximated by a deterministic state-variable model.

First-order properties of structures are often said to be “local” in the following sense. A model-theoretic definition of distance is given, and it is shown that the truth of any first-order formula is determined by the neighborhoods of bounded radius in the model. This locality principle was used by Gaifman [25] and Hanf [34] to establish limitations on the expressive power of first-order logic. It has been extended to more powerful logics such as counting logics [41]. Of course, this notion is useful only if the neighborhoods of bounded radius are small compared to the size of the model. In the extreme opposite case, all elements are within a distance of 1 of each other, and a neighborhood of radius 1 is the whole model.

We will extend this notion of distance to metafinite models and give a canonical form for all terms in the logic of metafinite models: every term is equivalent to a multiset operation applied to the multiset of bounded neighborhoods in the model.

Let  $\mathfrak{A} = \langle A, \mathcal{W}^{\mathfrak{A}}, \mathcal{F}, \mathcal{G} \rangle$  be a metafinite model. The Gaifman graph of  $\mathfrak{A}$  is the symmetric graph  $\langle A, E \rangle$ , where

$$E = \{(a, b) \in A^2 \mid \text{for some } k\text{-ary } w \in \mathcal{W} \text{ and } a_1, \dots, a_k \in A, \\ a, b \in \{a_1, \dots, a_k\} \text{ and } w^{\mathfrak{A}}(a_1, \dots, a_k) \neq \text{undef}\}.$$

This is an obvious generalization of the Gaifman graph of a relational structure where the relations have been replaced by characteristic functions. Letting  $\gamma^{\mathfrak{A}}(a, b)$  be the length of the shortest path in the Gaifman graph between  $a, b \in A$ , since the graph is symmetric,  $\gamma$  is a metric on  $A$ .

For  $k \in \mathbb{N}$ ,  $a_1, \dots, a_k \in A$  and  $r \in \mathbb{R}$ , let

$$N_r^{\mathfrak{A}}(a_1, \dots, a_k) = \{b \in A \mid \gamma(a_i, b) \leq r \text{ for some } i = 1, \dots, k\},$$

and let  $\mathfrak{N}_r^{\mathfrak{A}}(a_1, \dots, a_k)$  be the metafinite model with universe  $N_r^{\mathfrak{A}}(a_1, \dots, a_k)$ , weight functions of  $\mathcal{W}^{\mathfrak{A}}$  restricted to  $N_r^{\mathfrak{A}}(a_1, \dots, a_k)$  and additional unary weight functions  $v_1^{\mathfrak{A}}, \dots, v_k^{\mathfrak{A}}$ , where

$$v_i^{\mathfrak{A}}(a_i) = 1, \text{ and} \\ v_i^{\mathfrak{A}}(b) = 0 \text{ for } b \neq a_i.$$

The radius of  $\mathfrak{N}_r^{\mathfrak{A}}(a_1, \dots, a_k)$  is  $r$ . We put  $[\mathfrak{N}_r^{\mathfrak{A}}(a_1, \dots, a_k)]$  for the isomorphism type of  $\mathfrak{N}_r^{\mathfrak{A}}(a_1, \dots, a_k)$ .

**Definition 12.** Let  $S$  be the set of isomorphism types  $[\mathfrak{N}_r^{\mathfrak{A}}(a_1, \dots, a_k)]$ , taken over all  $k \in \mathbb{N}$ , metafinite models  $\mathfrak{A}$  over our vocabulary, and  $a_1, \dots, a_k \in A$ . A neighborhood multiset operation is a function

$$\Gamma: \text{fm}(S) \rightarrow \mathbb{R}.$$

**Definition 13.** The depth of a term is its maximum nesting of multiset operations. We define this more precisely by induction on the height of the term's parse tree. A term  $w(x_1, \dots, x_k)$  where  $w \in \mathcal{W}$  has depth 0. If the maximum depth of  $\tau_1, \dots, \tau_m$  is  $d$  and  $f \in \mathcal{F}$  is  $m$ -ary, then  $f(\tau_1, \dots, \tau_m)$  has depth  $d$ . If  $\tau$  has depth  $d$  and  $\Gamma \in \mathcal{G}$ , then  $(\Gamma y \tau)$  has depth  $d + 1$ .

We will use the function  $\rho(d) = (3^d - 1)/2$ . The key property of this function is  $\rho(d + 1) = 3\rho(d) + 1$ .

The canonical form described in the next lemma essentially says that the value of a term is determined by the number of times each isomorphism class is represented by a bounded neighborhood in the metafinite model.

**Lemma 3.** Every term  $\tau(x_1, \dots, x_k)$  of depth  $d$  is equivalent to a term  $(\Gamma y \mathfrak{N}_{\rho(d)}(x_1, \dots, x_k, y))$ , where  $\Gamma$  is a neighborhood multiset operation. That is, for every model  $\mathfrak{A}$  and every  $a_1, \dots, a_k \in A$ ,

$$\tau^{\mathfrak{A}}(a_1, \dots, a_k) = \Gamma(\{[\mathfrak{N}_{\rho(d)}^{\mathfrak{A}}(a_1, \dots, a_k, b)] \mid b \in A\}).$$

*Proof.* We use induction on the height of the parse tree of  $\tau$ . If  $\tau(x_1, \dots, x_k)$  is  $w(x_1, \dots, x_k)$  for some weight function symbol, take  $\Gamma(\{[\mathfrak{N}_0^{\mathfrak{A}}(a_1, \dots, a_k, b)] \mid b \in A\}) = w^{\mathfrak{A}}(a_1, \dots, a_k)$ . Clearly  $\Gamma$  is well-defined.

Assume the lemma holds for all terms of height less than or equal to  $h$ . Let  $\tau_1, \dots, \tau_m$  be terms of maximum height  $h$  and maximum depth  $d$  and  $f \in \mathcal{F}$  be  $m$ -ary. Without loss of generality, we can assume the free variables of each  $\tau_i$  are  $x_1, \dots, x_k$ . By the induction assumption,

$$\tau_i \equiv (\Delta_i y \mathfrak{N}_{\rho(d)}(x_1, \dots, x_k, y)),$$

where  $\Delta_i$  is a neighborhood multiset operator for  $i = 1, \dots, m$ . Then if we define

$$\Gamma(\{[\mathfrak{N}_{\rho(d)}^{\mathfrak{A}}(a_1, \dots, a_k, b)] \mid b \in A\}) = f(\Delta_1(\{[\mathfrak{N}_{\rho(d)}^{\mathfrak{A}}(a_1, \dots, a_k, b)] \mid b \in A\}), \dots, \Delta_m(\{[\mathfrak{N}_{\rho(d)}^{\mathfrak{A}}(a_1, \dots, a_k, b)] \mid b \in A\})),$$

we have proven the lemma for  $f(\tau_1, \dots, \tau_m)$ .

Lastly, let  $\tau(x_1, \dots, x_k)$  be  $(\Delta y \sigma(x_1, \dots, x_k, y))$  where  $\sigma$  is of height  $h$  and depth  $d$ . By induction,

$$\sigma(x_1, \dots, x_k, y) \equiv (\Lambda z \mathfrak{N}_{\rho(d)}(x_1, \dots, x_k, y, z)).$$

Thus,  $\tau$  is computed from the multiset of multisets  $\{\{[\mathfrak{N}_{\rho(d)}^{\mathfrak{A}}(a_1, \dots, a_k, b, c)] | c \in A\} | b \in A\}$ . We show how this multiset can be computed from the multiset  $\{\{[\mathfrak{N}_{\rho(d+1)}^{\mathfrak{A}}(a_1, \dots, a_k, b)] | b \in A\}$ . Take a fixed isomorphism type  $[\mathfrak{N}_{\rho(d)}^{\mathfrak{A}}(a_1, \dots, a_k, b, c)]$ .

**Case 1.**  $N_{\rho(d)}^{\mathfrak{A}}(c) \not\subseteq N_{\rho(d+1)}^{\mathfrak{A}}(a_1, \dots, a_k, b)$ . Then there are no edges in the Gaifman graph of  $\mathfrak{A}$  between  $N_{\rho(d)}^{\mathfrak{A}}(c)$  and  $N_{\rho(d)}^{\mathfrak{A}}(a_1, \dots, a_k, b)$ . To see this, suppose otherwise. Let  $b' \in N_{\rho(d)}^{\mathfrak{A}}(a_1, \dots, a_k, b)$  and  $c' \in N_{\rho(d)}^{\mathfrak{A}}(c)$  such that  $(b', c')$  is an edge in the Gaifman graph. But then for any  $c'' \in N_{\rho(d)}^{\mathfrak{A}}(c)$ ,

$$\begin{aligned} \gamma(\{a_1, \dots, a_k, b\}, c'') &\leq \gamma(\{a_1, \dots, a_k, b\}, b') + \gamma(b', c') + \gamma(c', c) + \gamma(c, c'') \\ &\leq \rho(d) + 1 + \rho(d) + \rho(d) \\ &= \rho(d + 1). \end{aligned}$$

Then

$$\begin{aligned} &\{\{[\mathfrak{N}_{\rho(d)}^{\mathfrak{A}}(a_1, \dots, a_k, b, c')] | c' \in A \\ &\quad \text{and } \mathfrak{N}_{\rho(d)}^{\mathfrak{A}}(a_1, \dots, a_k, b, c') \cong \mathfrak{N}_{\rho(d)}^{\mathfrak{A}}(a_1, \dots, a_k, b, c)\}\} = \\ &\{\{[\mathfrak{N}_{\rho(d)}^{\mathfrak{A}}(a_1, \dots, a_k, c')] | c' \in A \\ &\quad \text{and } \mathfrak{N}_{\rho(d)}^{\mathfrak{A}}(a_1, \dots, a_k, c') \cong \mathfrak{N}_{\rho(d)}^{\mathfrak{A}}(a_1, \dots, a_k, c)\}\} - \\ &\{\{[\mathfrak{N}_{\rho(d)}^{\mathfrak{A}}(a_1, \dots, a_k, c')] | N_{\rho(d)}^{\mathfrak{A}}(c') \subseteq N_{\rho(d+1)}^{\mathfrak{A}}(a_1, \dots, a_k, b) \\ &\quad \text{and } \mathfrak{N}_{\rho(d)}^{\mathfrak{A}}(a_1, \dots, a_k, c') \cong \mathfrak{N}_{\rho(d)}^{\mathfrak{A}}(a_1, \dots, a_k, c)\}\}. \end{aligned} \quad (8)$$

**Case 2.**  $N_{\rho(d)}^{\mathfrak{A}}(c) \subseteq N_{\rho(d+1)}^{\mathfrak{A}}(a_1, \dots, a_k, b)$ . Then

$$\begin{aligned} &\{\{[\mathfrak{N}_{\rho(d)}^{\mathfrak{A}}(a_1, \dots, a_k, b, c')] | c' \in A \\ &\quad \text{and } \mathfrak{N}_{\rho(d)}^{\mathfrak{A}}(a_1, \dots, a_k, b, c') \cong \mathfrak{N}_{\rho(d)}^{\mathfrak{A}}(a_1, \dots, a_k, b, c)\}\} = \\ &\{\{[\mathfrak{N}_{\rho(d)}^{\mathfrak{A}}(a_1, \dots, a_k, b, c')] | N_{\rho(d)}^{\mathfrak{A}}(c') \subseteq N_{\rho(d+1)}^{\mathfrak{A}}(a_1, \dots, a_k, b) \\ &\quad \text{and } \mathfrak{N}_{\rho(d)}^{\mathfrak{A}}(a_1, \dots, a_k, b, c') \cong \mathfrak{N}_{\rho(d)}^{\mathfrak{A}}(a_1, \dots, a_k, b, c)\}\}. \end{aligned} \quad (9)$$

Equations (8) and (9) imply that there is a neighborhood multiset operation  $\Gamma$  such that

$$\begin{aligned} \Gamma(\{\{[\mathfrak{N}_{\rho(d+1)}^{\mathfrak{A}}(a_1, \dots, a_k, b)] | b \in A\}\}) = \\ \Delta(\{\{\Lambda(\{\{[\mathfrak{N}_{\rho(d)}^{\mathfrak{A}}(a_1, \dots, a_k, b, c)] | c \in A\}) | b \in A\}\}), \end{aligned}$$

and therefore

$$\tau = (\Gamma y \mathfrak{N}_{\rho(d+1)}).$$

□

In the next section, we will use Lemma 3 to reduce questions about abstraction to questions about multiset operations.

## 9 Applications to Bounded Degree Structures

There are two potential applications of the general results in Section 7. The concentration bounds given in Theorems 1 and 3 could be used to support conclusions based on simulations. Theorems 2 and 4 provide justification for the common practice of abstracting agent-based models to deterministic state-variable models. In this section, we apply these general results and the locality principle of the previous section to certain classes of agent-based models. The simplest examples, e. g. models of chemical kinetics, are easy: the population sizes of the various species determine the dynamics of the system. That is, the individuals are classified according to the isomorphism type of their neighborhood of radius 0. Here, we extend this idea to systems whose dynamics is determined by populations of individuals classified according to their neighborhood of some specified radius. We give sufficient conditions for concentration bounds and approximation by deterministic state-variable model. We also give some biologically motivated examples of agent-based models that violate these conditions and cannot be approximated by any state-variable model. Our methods are based on ideas that have been applied to the analysis of expressivity of query languages. See e. g. Libkin [41].

### 9.1 Approximations of Agent-Based Models

Using our canonical representation of terms, we will give characterizations of agent-based models that satisfy the conditions of Section 7. Let  $\mathcal{I}$  be a discrete space agent-based model with state space  $\mathcal{S}$ . If there are only finitely many isomorphism types among  $\{\mathfrak{N}_1^{\mathfrak{A}}(a) : \mathfrak{A} \in \mathcal{S} \text{ and } a \in A\}$ , then we say that  $\mathcal{S}$  and  $\mathcal{I}$  are of bounded degree. This is a generalization of the graph-theoretic notion of bounded degree. It implies that for any  $r \in \mathbb{N}$ , there are only finitely many isomorphism types among  $\{\mathfrak{N}_r^{\mathfrak{A}}(a) : \mathfrak{A} \in \mathcal{S} \text{ and } a \in A\}$ . Besides assuming that  $\mathcal{S}$  is of bounded degree, we assume that the updates change a bounded number of weight function values. We begin with the case when time is discrete, i.e.,  $\mathcal{I} = (\mathcal{S}, g)$ , where  $g$  is the transition probability function as in (3).

Since the updates change a bounded number of weight function values, there is a term  $\delta(; y)$  such that  $q^{\mathfrak{A}}(r) = \delta^{\mathfrak{A}}(; r)$  for every  $\mathfrak{A} \in \mathcal{S}$  and  $r \in \mathbb{R}$ . Let  $d$  be the maximum of the depths of the observable  $\tau$  and  $\delta$ . Since  $\mathcal{I}$  is of bounded degree, there are only finitely many, say  $k$ , isomorphism classes of neighborhoods of radius  $\rho(d)$  in  $\mathcal{S}$ . Let them be numbered  $1, \dots, k$ . For every  $\mathfrak{A} \in \mathcal{S}$ , let  $\alpha^{\mathfrak{A}} = (m_1, \dots, m_k)$ , where  $m_i$  is the number of neighborhoods in  $\mathfrak{A}$  belonging to the  $i$ th isomorphism class. By Lemma 3, there are functions  $\Gamma: \mathbb{N}^k \rightarrow \mathbb{R}$  and  $\Delta: \mathbb{N}^k \times \mathbb{R} \rightarrow [0, 1]$  such that for all  $\mathfrak{A} \in \mathcal{S}$ ,  $\tau^{\mathfrak{A}} = \Gamma(\alpha^{\mathfrak{A}})$  and for all  $r \in \mathbb{R}$ ,  $\delta^{\mathfrak{A}}(; r) = \Delta(\alpha^{\mathfrak{A}}, r)$ . We also assume that the size of any state  $\mathfrak{A}$  is determined by  $\alpha^{\mathfrak{A}}$ . That is, there is a function  $\|\dots\|: \mathbb{N}^k \rightarrow \mathbb{N}$  such that  $\|\alpha^{\mathfrak{A}}\|$  is the size of  $\mathfrak{A}$ .

For any  $\overline{m} \in \mathbb{R}^k$ ,  $|\overline{m}|$  is the magnitude of the vector  $\overline{m}$ , and  $\overline{m} - \overline{m}'$  denotes vector subtraction. The next two theorems follow from Theorems 1 and 2.

**Theorem 5.** Assume  $\mathcal{I}$  satisfies the following boundedness condition: there is a constant  $c$  such that for  $\bar{m}, \bar{m}' \in \mathbb{N}^k$ ,

$$|\bar{m} - \bar{m}'| \leq 1 \Rightarrow |\Gamma(\bar{m}) - \Gamma(\bar{m}')| < c,$$

and  $\mathcal{I}$  also satisfies the smoothness condition: as  $\|\bar{m}\| \rightarrow \infty$ , for all  $\bar{m}'$  such that  $|\bar{m} - \bar{m}'| \leq 1$ ,

$$\Delta(\bar{m}; r) \sim \Delta(\bar{m}'; r).$$

Then for any  $\gamma \in (0, \infty)$  and  $\epsilon \in [0, 1)$ , for sufficiently large  $T$  and  $n$ ,

$$\Pr(|Z_T - \mathbf{E}(Z_T)| < \gamma T \mid \mathfrak{A}_0 = \mathfrak{A}) > \epsilon.$$

**Theorem 6.** Assume  $\mathcal{I}$  satisfies the boundedness condition of the previous theorem and the following closure property: as  $\|\bar{m}\| \rightarrow \infty$ ,

$$|\Gamma(\bar{m}) - \Gamma(\bar{m}')| < c \Rightarrow \Delta(\bar{m}, r) \sim \Delta(\bar{m}', r).$$

Then there is a deterministic state-variable model  $\mathcal{I}'$  such that for any  $\gamma \in (0, \infty)$  and  $\epsilon \in [0, 1)$ , for sufficiently large  $T$  and  $n$ ,  $\mathcal{I}$  can be approximated by  $\mathcal{I}'$  with accuracy  $\gamma T$  and confidence  $\epsilon$  with respect to  $\tau$ .

We can obtain similar results for continuous time agent-based models. Letting  $\mathcal{I} = (\mathcal{S}, g')$ , where  $g'$  is the transition rate function, we redefine  $\Delta$  so that

$$q'^{\mathfrak{A}}(r) = \Delta(\alpha^{\mathfrak{A}}, r).$$

Then using the methods of Theorems 3 and 4,

**Theorem 7.** Assume  $\mathcal{I}$  satisfies the conditions of Theorem 5 with the modified  $\Delta$ . Then for any  $\gamma \in (0, \infty)$  and  $\epsilon \in [0, 1)$ , for sufficiently large  $n$ , there is  $\Delta t > 0$  such that

$$\Pr(|\tau^{\mathfrak{A}\Delta t} - \tau^{\mathfrak{A}} - h'^{\mathfrak{A}}\Delta t| < \gamma p'^{\mathfrak{A}}\Delta t \mid \mathfrak{A}_0 = \mathfrak{A}) > \epsilon.$$

**Theorem 8.** Let  $\mathcal{I}$  satisfy the boundedness condition of Theorem 7. Also assume that there is a continuous function  $f: \mathbb{R} \rightarrow \mathbb{R}$  such that as  $\|\bar{m}\| \rightarrow \infty$ ,

$$\Delta(\bar{m}, r) \sim f(\Gamma(\bar{m}), r).$$

Then there is a deterministic state-variable model  $\mathcal{I}'$  such that for any  $\gamma \in (0, \infty)$  and  $\epsilon \in [0, 1)$ , for sufficiently large  $n$ , there is  $\Delta t > 0$  such that  $\mathcal{I}$  can be approximated by  $\mathcal{I}'$  with accuracy  $\gamma p'^{\mathfrak{A}}\Delta t$  and confidence  $\epsilon$  with respect to  $\tau$ . Further, the evolution of the state of  $\mathcal{I}'$  satisfies a system of ordinary differential equations.

## 9.2 Examples

Of the examples in Section 2, those of bounded degree are the models of chemical kinetics (trivially since all their weight functions are unary), chemical reaction models with spatial information represented by lattices, patch-occupancy models in ecology [40] provided each patch has a finite number of states, and the trophic models that do not include spatial information. Some of the models of behavioral ecology in Section 2.5 do not have bounded degree because their individuals have spatial attributes, and our theorems do not apply to them.

In general, the graph growth models do not satisfy the conditions of Theorems 5 through 8 because they have unbounded degree. However, in some cases, e.g., [5, 11, 38], some parameter settings result in graphs with finite average degree, and there are large regions of the state space that satisfy the theorems.

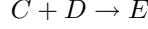
Dalvi, Miklau, and Suciu [15, 16] have studied the logic of random databases where the average number of edges is bounded. Although they do not consider the random evolution of databases, their probability distributions generate structures of bounded degree. Since database operations often satisfy the conditions of Theorems 5 through 8, applications to database theory may be worth exploring.

We will give some specific examples of agent-based models that satisfy the concentration bounds (Theorems 5 and 7) and the closure properties that imply approximation by a deterministic state-variable model (Theorems 6 and 8). We also give related examples of agent-based models that cannot be approximated by deterministic state-variable models. These models are not intended to be biologically realistic; they are simplified models of fundamental processes of molecular biology that illustrate some of the ideas described above.

The classic models of chemical kinetics described in subsection 2.2 assume the molecules are in a rapidly mixing solution. Therefore they do not need spatial attributes, and the state is completely described by the numbers of molecules in each species. Thus the state is a metafinite model with a single unary weight function which labels each agent with its species, and all terms in the logic are multiset operations on the sizes of the isomorphism classes of radius 0. Using the notation above,  $\alpha^{\mathfrak{A}}$  is the vector of population sizes. Since the transition probabilities are instances of the law of mass action [61], they are polynomial functions of  $\alpha^{\mathfrak{A}}$ . By defining  $\|\bar{m}\| = \min(\bar{m})$ , the closure property holds, and if  $\Gamma$  obeys the boundedness condition, the theorems apply.

If the solution is not rapidly mixing, then spatial attributes must be associated with the molecules, and we have a model of a reaction-diffusion system. There are many ways of implementing this. We shall describe a simple patch-occupancy model similar to StochSim 1.4 [53]. Space is represented as a regular lattice; StochSim 1.4 supports three kinds of 2-dimensional lattice models, where each vertex has three, four, or six nearest neighbors. We shall use an  $n \times n$  square lattice on a torus, where every vertex has four nearest neighbors. This architecture avoids boundary effects, which could be added at the cost of a more complicated model. We consider one of the simplest types of reaction, where two molecules of different species combine to form one molecule of a third

species:



Each vertex  $v$  is labeled with a symbol  $l(v)$  indicating its state:

**blank** if vacant

$C$  if occupied by a molecule of species  $C$

$D$  if occupied by a molecule of species  $D$

$E$  if occupied by a molecule of species  $E$

Discrete and continuous time transitions can be considered. In discrete time, each update is described by the following pseudo-code:

choose a random occupied vertex  $v$

choose a random nearest neighbor  $u$  of  $v$

**if**  $(l(v) = C \text{ and } l(u) = D)$  or  $(l(v) = D \text{ and } l(u) = C)$

**then** with probability  $p$ ,

$l'(v) = \mathbf{blank}$

$l'(u) = E$

**if**  $(l(v) = C \text{ and } l(u) = \mathbf{blank})$  or  $(l(v) = D \text{ and } l(u) = \mathbf{blank})$

**then** with probability  $q$ ,

$l'(v) = \mathbf{blank}$

$l'(u) = l(v)$

We could easily augment the rule to allow diffusion of  $E$  and the reverse reaction



Let  $\tau$  be an observable of depth  $d_\tau$ . By Lemma 3,  $\tau^{\mathfrak{A}}$  is determined by the population of neighborhoods in  $\mathfrak{A}$  of radius  $\rho(d_\tau)$ . Since a transition affects only one vertex  $v$  and one of its nearest neighbors  $u$ , the transition can change only those neighborhoods of radius  $\rho(d_\tau)$  that include  $v$  or  $u$ . Therefore the transition is completely described by its effect on the neighborhood of radius  $d = 2\rho(d_\tau) + 1$  around  $v$ .

Let the isomorphism classes of neighborhoods of radius  $d$  be numbered  $1, \dots, k$ . Using the notation above, for  $i = 1, \dots, k$  we can define a function  $\pi_i: \mathbb{R} \rightarrow [0, 1]$  such that for  $r \in \mathbb{R}$ ,  $\pi_i(r)$  is the conditional probability that  $\tau^{\mathfrak{A}'} - \tau^{\mathfrak{A}} = r$ , given that  $v$  was selected and  $\mathfrak{N}_d^{\mathfrak{A}}(v)$  belongs to isomorphism class  $i$ . Let  $S$  be the set of  $i \in \{1, \dots, k\}$  such that the center of a neighborhood in isomorphism class  $i$  is labeled  $C$  or  $D$ . Then

$$\Delta(\bar{m}, r) = \sum_{i \in S} \frac{m_i}{\sum_{j \in S} m_j} \pi_i(r).$$

if we define  $\|\bar{m}\| = \min(m_i : i \in S)$ , the smoothness condition holds, and if  $\Gamma$  obeys the boundedness condition, Theorem 5 and 7 apply.

In general,  $\tau^{\mathfrak{A}}$  does not determine the probability distribution of neighborhoods of radius  $d$ , and additional assumptions are needed for closure. As noted by Erban et al. [22], finding a closed set of observables usually comes from extensive experimentation before it is justified theoretically. In the next subsection, we show that no finite set of observables can approximate the dynamics of our example. That is, for any abstraction of our example to a state-variable model, there are observables  $\tau$  such that the state-variable model cannot approximate our example with respect to  $\tau$ .

The next example is a model of the formation and degradation of molecular complexes. A complex is a connected set of agents, which can be atoms or molecules, joined by undirected edges, which represent chemical bonds. Our example has only one type of agent, which we will call a vertex. Each vertex can be joined to at most  $k$  other vertices by an edge. Thus the state of the system is a random graph of bounded degree. The degree sequence of the graph is  $m_0, \dots, m_k$ , where each  $m_i$  is the number of vertices of degree  $i$ . In model-theoretic terms, the degree sequence is the multiset of neighborhoods of radius 1. The transition rule for the discrete time version of the system generalizes the evolution of random graphs (the degree of vertex  $v$  is  $d_v$ , and  $E$  is the characteristic function of the edge relation):

```

choose a random vertex  $v$ 
choose random  $p \in [0, 1]$ 
if  $p \leq p_{d_v}$ 
  then
    choose a random vertex  $u$  not adjacent to  $v$ 
    with probability  $q_{d_v, d_u}$ ,  $E'(v, u) = 1$ 
  else
    choose a random vertex  $u$  adjacent to  $v$ 
    with probability  $r_{d_v, d_u}$ ,  $E'(v, u) = 0$ 

```

Using methods of [44], it can be shown that, when all graphs with a given degree sequence are equally likely, the asymptotic probability of a transition for almost all graphs depends only on their degree sequence. Therefore for every  $n \in \mathbb{N}$ , there is a large region of the state space such that the transition probability is approximated by a rational function of the degree sequence. This function is a sum over the various cases that depend on  $d_v$  and  $d_u$ . For example, taking  $k \geq 3$ , the probability that the edge between  $v$  of degree 3 and  $u$  of degree 2 is broken is asymptotic to

$$\frac{m_3}{n} (1 - p_3) \left( 3 \frac{m_2}{n} - 3 \left( \frac{m_2}{n} \right)^2 + \left( \frac{m_2}{n} \right)^3 \right) r_{3,2},$$

where  $n = \sum_{i=0}^k m_i$ . Then the degree sequence is a closed set of observables, and consequently this agent-based model can be approximated by a deterministic state-variable model.

In the case of continuous time, the transition probabilities of the various cases are multiplied by rate terms. Since the probabilities are continuous functions

of the degree sequence, if the rate terms are also continuous functions, then the members of the degree sequence are a closed set of observables, and the system can be approximated by a deterministic state-variable model.

### 9.3 Agent-Based Models That Cannot be Approximated by State-Variable models

If a bounded-degree agent-based model can be approximated by a state-variable model, then roughly speaking, its states are described by a finite set of terms in a counting logic. It is well-known from database theory that counting logics cannot define certain topological properties such as connectedness [41]. Thus it should not be surprising that there are agent-based models that cannot be approximated by any state-variable model. In fact, the seeming inability of existing state-variable models to capture important behavioral features of systems is one of the main reasons for the growing acceptance of agent-based models. We give two examples of agent-based models that cannot be approximated by state-variable models.

Suppose our first example  $\mathcal{I}$  (the lattice model from the previous subsection) is abstracted via  $\alpha$  to a state-variable model  $\mathcal{I}'$ , say its state set is  $\mathcal{S}' \subseteq \mathbb{R}^k$ . We put  $\bar{w}_t = (w_{1,t}, \dots, w_{k,t})$ , where each  $w_{i,t} \in \mathbb{R}$ , for the state of  $\mathcal{I}'$  at time  $t$ . Taking the case when both  $\mathcal{I}$  and  $\mathcal{I}'$  run in discrete time (continuous time is similar), let  $\delta': \mathcal{S}' \times \mathbb{R}^k \rightarrow [0, 1]$  define the transitions of  $\mathcal{I}'$ : for  $\bar{w} = (w_1, \dots, w_k) \in \mathcal{S}'$  and  $\bar{r} \in \mathbb{R}^k$ ,

$$\delta'(\bar{w}, \bar{r}) = \Pr(\bar{w}_{t+1} - \bar{w}_t = \bar{r} | \bar{w}_t = \bar{w}).$$

Since  $\delta'$  is also an observable, by Lemma 3, there is some  $d$  such that the dynamics of  $\mathcal{I}$  is approximated by a multiset operation on the neighborhoods of radius  $d$ .

For each  $n \in \mathbb{N}$ , we construct two states  $\mathfrak{A}_n$  and  $\mathfrak{B}_n$  such that  $\|\mathfrak{A}\|, \|\mathfrak{B}\| \geq n$ . Let  $k$  be the number of isomorphism classes of neighborhoods of radius  $d$ .  $\mathfrak{A}_n$  consists of three disconnected toroidal lattices  $\mathfrak{A}_{n,1}, \mathfrak{A}_{n,2}, \mathfrak{A}_{n,3}$  with labeled vertices:

$\mathfrak{A}_{n,1}$  is a  $(k(2d+1)n) \times ((2d+1)n)$  lattice containing at least  $n^2$  copies of each isomorphism class of neighborhoods of radius  $d$ .

$\mathfrak{A}_{n,2}$  is a  $((2d+1)n^2) \times ((2d+1)n^2)$  lattice containing  $n^4$  vertices labeled  $C$ , all of which are  $2d+1$  apart, and all other vertices are labeled **blank**.

$\mathfrak{A}_{n,3}$  is similar to  $\mathfrak{A}_{n,2}$  except its non-**blank** vertices are labeled  $D$ .

$\mathfrak{B}_n$  consists of two disconnected toroidal lattices  $\mathfrak{B}_{n,1}, \mathfrak{B}_{n,2}$ :

$$\mathfrak{B}_{n,1} = \mathfrak{A}_{n,1}.$$

$\mathfrak{B}_{n,2}$  is a  $(2(2d+1)n^2) \times ((2d+1)n^2)$  lattice containing  $n^4$  vertices labeled  $C$  and  $n^4$  labeled  $D$ , all of which are  $2d+1$  apart, and all other vertices are labeled **blank**.

Since  $\mathfrak{A}_n$  and  $\mathfrak{B}_n$  have the same number of neighborhoods in each isomorphism class of radius  $d$ ,  $\alpha(\mathfrak{A}_n) = \alpha(\mathfrak{B}_n)$ . Therefore, if  $\mathcal{I}'$  approximates  $\mathcal{I}$ , then with high probability  $(\alpha(\mathfrak{A}_n^T) - \alpha(\mathfrak{B}_n^T))/T$  can be made arbitrarily small for sufficiently large  $T$  and  $n$ . But from the theory of random walks, as  $n \rightarrow \infty$ , with high probability  $C$  vertices in  $\mathfrak{B}_n$  will react with  $D$  vertices, producing  $E$  vertices, but this cannot happen in  $\mathfrak{A}_n$ . Therefore with high probability  $(\alpha(\mathfrak{A}_n^T) - \alpha(\mathfrak{B}_n^T))/T$  will be bounded below by a positive vector.

Our second example is a simplified model of transcription, the process where one chain of molecules (DNA) serves as a template for generating another chain (mRNA). The template is read from beginning to end by an enzyme called RNA polymerase (RNAP), which outputs the mRNA chain one link at a time, in much the same way that a finite state transducer generates an output string from an input string. Translation, the process where an mRNA chain serves as a template for the production of a protein chain, is conceptually similar, but three links (amino acids) in the protein chain are generated for every link in the mRNA chain, and a molecular complex known as a ribosome plays the role of the RNAP. Since our simple version of transcription cannot be modeled by a state-variable model, this negative result also holds for these more complicated systems.

The states of our agent-based model are directed labeled graphs whose components are chains, the transcription enzymes, and the chain/enzyme complexes. Template chains are of the form SA...AF, where S and F are noncoding links signifying the start and finish of the chain. There are also defective template chains that lack either an S or an F. Output chains are of the form B...B. A transcription enzyme is a vertex labeled R or R'. Initially, there are no R' vertices, all R vertices are isolated, there are no B vertices, but there are template chains (possibly defective).

In any state, if there are no R' vertices, then one R vertex is randomly selected and relabeled with R'. Then it is joined to a randomly chosen S vertex by an edge. This begins the process of growing a B-chain which is joined to the R' vertex. At each step, the outgoing edge from the R' vertex is broken and rejoined to the next A vertex in the template chain, and a new B vertex is inserted at the end of the B-chain joined to the R' vertex. This is illustrated in Figure 3. These steps are repeated until the F link is reached. Then the R' is relabeled with R, its ingoing and outgoing edges are deleted, thus releasing the R and the completed B-chain. If a template chain is defective, then it cannot generate an output chain, either because in the absence of the S, the process cannot start, or in the absence of the F, the process cannot finish.

Let  $\tau$  be the number of B vertices of outdegree 0. Since there are no B vertices initially, the value of  $\tau$  at some time  $T$  is the number of output chains that have been generated up to that time, which can be positive only if there are non-defective template chains at time 0. It can be shown that the existence of non-defective template chains is not expressible in our term logic. The proof is similar to that of the well-known fact that SQL cannot express the property of connectedness in graphs [41]. The proof that our agent-based model cannot be approximated by a state-variable model is based on the same ideas.

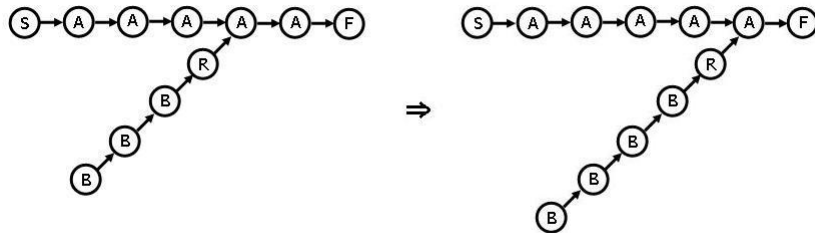


Figure 3: One step in the growth of the B-chain

Assuming this agent-based model is abstracted via  $\alpha$  to a state-variable model let  $r$  be the maximum depth of all the terms that correspond to weight functions in the state-variable model. For arbitrarily large  $n$ , let  $\mathfrak{A}$  have the following components:

$n$  chains  $SA^{2r}$

$n$  chains  $A^{2r}F$

$n$  chains  $A^{2r}$

$n$  chains  $SA^{2r}F$ .

Let  $\mathfrak{B}$  have  $2n$  components  $SA^{2r}$  and  $A^{2r}F$ . Again by Lemma 3,  $\tau^{\mathfrak{A}} = \tau^{\mathfrak{B}}$ , but with positive probability,  $\tau^{\mathfrak{A}T} = \Theta(T)$ , while  $\tau^{\mathfrak{B}T} = 0$ .

## 10 Conclusions and Open Problems

It should be evident that discrete, stochastic interactions are an essential feature of many dynamical systems. As related by Wilkinson [61], the original version of SBML (Systems Biology Markup Language), which is intended to be the standard for describing complex biochemical reaction systems, was designed for continuous deterministic modeling. Later versions that included the capability of agent-based modeling have had limited acceptance. Perhaps it will be necessary to include this capability as a basic feature of future modeling and simulation languages.

Most software tools in systems biology are aids to specifying and simulating biochemical networks. A further step, which is already underway, is to develop verification tools similar to those used in software and hardware design [12, 18, 50]. A temporal logic for agent-based models could be developed for the purpose of model checking. This would also require developing methods for approximating agent-based models with finite state systems. The methods of Desharnais et al. [21] may be useful here.

More specific problems are to improve the concentration bounds in Section 7 and to extend the scope of applicability of the methods outlined in Section 9. The bounds on  $n$ ,  $T$ , and  $\Delta t$  given in the theorems of Section 7 are not at all tight because the assumptions in the theorems are so general. Better bounds may be possible for specific classes of agent-based models.

Since the results of Section 9 apply only to structures of bounded degree, an obvious problem is to extend them to graph growth models and models with spatial information that have unbounded degree. Models that include spatial information, e. g. reaction-diffusion systems, often use terms with numeric variables, and they are approximated by partial differential equations. Extensions of Theorem 8 to this class of models might be useful in automating the development of software for equation-free simulation [22].

Theorems 6 and 8 apply to regions of state space where the sizes of the neighborhood isomorphism classes are fixed or increase without bound. The fluctuations of small populations of certain molecules can have a strong effect on the behavior of cells. Arkin et al. [3] have modeled genetic networks that include some of these effects. Perhaps some kind of hybrid system could approximate these systems. State space would be partitioned into regions determined by those individuals that are present in small numbers. Changes in these numbers would be modeled by discrete transitions to other regions, and changes to the sizes of the large populations would be approximated by state-variable changes, as in Theorems 6 and 8. Important questions about stability—when is the system resistant to small perturbations of its state, and when can small perturbations lead to bifurcations in its behavior—could be formalized and studied.

As pointed out by Firth and Bray [24], if the number of conformational states of the molecules is very large, then describing the possible reactions with a state-variable model becomes impractical. Theorems 6 and 8 give sufficient conditions for approximating an agent-based model with a state-variable model, but it does not address the question of how many variables are needed by the state-variable model. Can this number be reduced, or are there agent-based models for which this is optimal?

We have mentioned the models of very sparse random databases. They may be regarded as states of an agent-based model whose rules are randomly selected database updates. To our knowledge, this viewpoint has not been investigated. Characterizing these rules and extending our results to evolving databases may have applications to very large databases.

Our approximations result in deterministic state-variable models. Are there agent-based models that can be approximated by nondeterministic state-variable models but not deterministic ones? In particular, can they be approximated by stochastic differential equations but not deterministic differential equations?

## 11 Acknowledgements

The author thanks Vincent Danos, Prakash Panangaden, and the other participants at the Bellairs Workshop on Computational Modelling of Biological

Systems, Barbados, March 16–March 20, for many insightful questions and observations about some of the ideas in this article.

## References

- [1] W. Aiello, F. Chung, and L. Lu. Random evolution of massive graphs. In J. Abello, P. M. Pardalos, and M. G. C. Resende, editors, *Handbook of Massive Data Sets*, pages 97–122. Kluwer Academic Press, 2002.
- [2] M. Ander, P. Beltrao, B. D. Ventura, J. Ferkinghoff-Borg, M. Foglierini, A. Kaplan, C. Lemerle, I. Tomás-Oliveira, and L. Serrano. SmartCell, a framework to simulate cellular processes that combines stochastic approximation with diffusion and localisation: analysis of simple networks. *Syst. Biol.*, 1:129–138, 2004.
- [3] A. Arkin, J. Ross, and H. H. McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage  $\lambda$ -infected *Escherichia coli* cells. *Genetics*, 149:1633–1648, 1998.
- [4] K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 68:357–367, 1967.
- [5] N. Berger, C. Borgs, J. T. Chayes, R. M. D’Souza, and R. D. Kleinberg. Competition-induced preferential attachment. In *Proc. 31st Int. Colloquium on Automata, Languages and Programming, Lecture Notes in Computer Science*, volume 3142, pages 208–221. Springer, 2004.
- [6] D. Bernstein. Simulating mesoscopic reaction-diffusion systems using the Gillespie algorithm. *Phys. Rev. E*, 71:041103 (13 pages), 2005.
- [7] C. K. Birdsall and A. B. Langdon. *Plasma Physics Via Computer Simulation*. CRC Press, 2004.
- [8] R. E. Bryant, S. K. Lahiri, and S. A. Seshia. Modeling and verifying systems using a logic of counter arithmetic with lambda expressions and uninterpreted functions. In *Computer-Aided Verification (CAV ’02), Lecture Notes in Computer Science*, volume 2404, pages 78–92. Springer, 2002.
- [9] A. W. Burks. Von Neumann’s self-reproducing automata. In A. W. Burks, editor, *Essays on Cellular Automata*, pages 3–64. University of Illinois Press, Urbana, 1970.
- [10] K. Burrage, J. Hancock, A. Leier, and D. V. Nicolau Jr. Modelling and simulation techniques for membrane biology. *Briefings in Bioinformatics*, 8:234–244, 2007.
- [11] D. S. Callaway, J. E. Hopcroft, J. M. Kleinberg, M. E. J. Newman, and S. H. Strogatz. Are randomly grown graphs really random? *Phys. Rev. E*, 64:041902 (7 pages), 2001.
- [12] L. Cardelli. Brane calculi-interactions of biological membranes. In *Proc. Int. Conf. Computational Methods in Systems Biology, Lecture Notes in Computer Science*, volume 3082, pages 257–280. Springer, 2005.
- [13] F. Chung, L. Lu, T. G. Dewey, and D. J. Galas. Duplication models for biological networks. *J. Comput. Biology*, 10:677–687, 2003.
- [14] I. D. Couzin, J. Krause, N. R. Franks, and S. A. Levin. Effective leadership and decision-making in animal groups on the move. *Nature*, 433:513–516, 2005.
- [15] N. N. Dalvi. Query evaluation on a database given by a random graph. In *Proc. 11th Int. Conf. on Database Theory, Lecture Notes in Computer Science*, volume 4353, pages 149–163. Springer, 2007.
- [16] N. N. Dalvi, G. Miklau, and D. Suciu. Asymptotic conditional probabilities for conjunctive queries. In *Proc. 10th Int. Conf. on Database Theory, Lecture Notes in Computer Science*, volume 3363, pages 289–305. Springer, 2005.

- [17] V. Danos, J. Feret, W. Fontana, and J. Krivine. Scalable simulation of cellular signaling networks. In Z. Shao, editor, *Proc. 5th Asian Symp. Programming Languages and Systems, Lecture Notes in Computer Science*, volume 4807, pages 139–157. Springer, 2007.
- [18] V. Danos and C. Laneve. Formal molecular biology. *Theoretical Computer Science*, 325:69–110, 2004.
- [19] H. de Jong. Modeling and simulation of genetic regulatory systems: A literature review. *J. Computational Biol.*, 9:67–103, 2002.
- [20] J. Desharnais, A. Edalat, and P. Panangaden. Bisimulation for labelled Markov processes. *Information and Computation*, 179:163–193, 2002.
- [21] J. Desharnais, V. Gupta, R. Jagadeesan, and P. Panangaden. Approximating labelled Markov processes. *Information and Computation*, 184:160–200, 2003.
- [22] R. Erban, I. G. Kevrekidis, and H. G. Othmer. An equation-free computational approach for extracting population-level behavior from agent-based models of biological dispersal. *Physica D: Nonlinear Phenomena*, 215:1–24, 2006.
- [23] P. Erdős and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci.*, 5:17–61, 1960.
- [24] C. A. J. M. Firth and D. Bray. Stochastic simulation of cell signalling pathways. In J. M. Bower and H. Bolouri, editors, *Computational Modeling of Genetic and Biochemical Networks*, pages 263–286. MIT Press, 2001.
- [25] H. Gaifman. On local and non-local properties. In *Proc. of the Herbrand Symposium, Logic Colloquium '81*. North-Holland, 1982.
- [26] J.-L. Giavitto, C. Godin, O. Michel, and P. Prusinkiewicz. Computational models for integrative and developmental biology. Technical Report 72-2002, Université d'Evry Val d'Essone, Evry, France, Mar. 2002.
- [27] M. A. Gibson and E. Mjolsness. Modeling the activity of single genes. In J. M. Bower and H. Bolouri, editors, *Computational Modeling of Genetic and Biochemical Networks*, pages 1–48. MIT Press, 2001.
- [28] D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Computational Physics*, 22:403–434, 1976.
- [29] D. T. Gillespie. A rigorous derivation of the chemical master equation. *Physica A*, 188:404–425, 1992.
- [30] E. Grädel and Y. Gurevich. Metafinite model theory. *Information and Computation*, 140:26–81, 1998.
- [31] V. Grimm. Ten years of individual-based modelling in ecology: what have we learned and what could we learn in the future? *Ecological Modelling*, 115:129–148, 1999.
- [32] V. Grimm and S. F. Railsback. *Individual-Based Modeling Ecology*. Princeton University Press, 2005.
- [33] Y. Gurevich. Sequential abstract state machines capture sequential algorithms. *ACM Trans. on Computational Logic*, 1:77–111, 2000.
- [34] W. Hanf. Model-theoretic methods in the study of elementary logic. In J. W. Addison, L. Henkin, and A. Tarski, editors, *The Theory of Models*, pages 132–145. North-Holland, 1965.
- [35] M. Huston, D. DeAngelis, and W. Post. New computer models unify ecological theory. *BioScience*, 38:682–691, 1988.
- [36] S. A. Kauffman. *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press, New York, 1993.
- [37] R. C. Kennicutt, F. Schweizer, and J. E. Barnes. *Galaxies: Interactions and Induced Star Formation*. Springer, 1998.

- [38] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Ufal. Stochastic models for the web graph. In *Proc. 41st IEEE Symp. on Foundations of Computing Science*, pages 57–65, 2000.
- [39] D. P. Landau and K. Binder. *A Guide to Monte Carlo Simulations in Statistical Physics*. Cambridge University Press, 2005.
- [40] S. A. Levin, T. Powell, and J. H. Steele, editors. *Patch Dynamics*. Springer, 1993.
- [41] L. Libkin. Expressive power of SQL. *Theoretical Computer Science*, 296:379–404, 2003.
- [42] A. Lindenmayer. Mathematical models for cellular interaction in development, Parts I and II. *J. Theoretical Biology*, 18:280–315, 1968.
- [43] A. J. Lotka. *Elements of Physical Biology*. Williams and Wilkins, Baltimore, 1925.
- [44] J. F. Lynch. Convergence law for random graphs with specified degree sequence. *ACM Trans. Computational Logic*, 6:727–748, 2005.
- [45] R. M. May. *Theoretical Ecology: Principles and Applications*. Saunders, Philadelphia, 1976.
- [46] H. H. McAdams and L. Shapiro. Circuit simulation of genetic networks. *Science*, 269:650–656, 1995.
- [47] R. Milner, J. Parrow, and D. Walker. A calculus of mobile processes (I and II). *Information and Computation*, 100:1–77, 1992.
- [48] E. Mjolsness and G. Yosiphon. Stochastic process semantics for dynamical grammars. *Ann. Math. Artif. Intell.*, 47:329–395, 2006.
- [49] C. Priami. Stochastic pi-calculus. *Computer Journal*, 38:578–589, 1995.
- [50] C. Priami, A. Ingólfssdóttir, B. Mishra, and H. R. Nielson, editors. *Trans. Computational Systems Biology VII, Lecture Notes in Computer Science*, volume 4230. Springer, 2006.
- [51] H. Reuter and B. Breckling. Selforganization of fish schools: an object-oriented model. *Ecological Modelling*, 75/76:147–159, 1994.
- [52] W. L. Romey. Individual differences make a difference in the trajectories of simulated schools of fish. *Ecological Modelling*, 92:65–77, 1996.
- [53] T. S. Shimizu, N. L. Novère, M. D. Levin, A. J. Beavil, B. J. Sutton, and D. Bray. Molecular model of a lattice of signalling proteins involved in bacterial chemotaxis. *Nat. Cell. Biol.*, 2:792–796, 2000.
- [54] I. Shmulevich, E. R. Dougherty, and W. Zhang. From Boolean to probabilistic Boolean networks as models of genetic regulatory networks. *Proc. of the IEEE*, 90:1778–1792, 2002.
- [55] E. H. Snoussi and R. Thomas. Logical identification of all steady states: The concept of feedback loop characteristic states. *Bull. of Math. Biology*, 55:973–991, 1993.
- [56] L. Tesfatsion. *Agent-Based Computational Economics: Modeling Economies as Complex Adaptive Systems*. Elsevier Science Inc., New York, 2003.
- [57] C. Tofts. Using process algebra to describe social insect behaviour. *Trans. Soc. Computer Simulation*, 9:227–283, 1993.
- [58] A. M. Turing. The chemical basis of morphogenesis. *Philos. Trans. Roy. Soc. London B*, 237:37–72, 1952.
- [59] V. Volterra. Fluctuations in the abundance of a species considered mathematically. *Nature*, 118:558–560, 1926.
- [60] G. von Dassow, E. Meir, E. M. Munro, and G. Odell. The segment polarity network is a robust developmental module. *Nature*, 406:188–192, 2000.
- [61] D. J. Wilkinson. *Stochastic Modelling for Systems Biology*. CRC Press, Boca Raton, FL, 2006.