A Logical Characterization of Individual-Based Models

James F. Lynch*
Department of Computer Science
Clarkson University
Potsdam, NY 13699-5815
email: jlynch@clarkson.edu

Abstract

Individual-based models are a relatively new approach to modelling dynamical systems of interacting entities, for example molecules in a biological cell. Although they are computationally expensive, they have the capability of modelling systems more realistically than traditional state-variable models. We give a formal definition of individual-based models, which includes state-variable models as a special case. We examine the questions of when state-variable models are sufficient for accurate modelling of a system, and when individual-based models are necessary. We define notions of abstraction and approximation, and give sufficient conditions that imply that an individual-based model can be approximated by a deterministic state-variable model. We also give negative results: examples of individual-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 modelled 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 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 modelled 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 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 [25]

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

where $x_1, \ldots, x_k$ are population sizes of k different species. Deterministic state-variable models have been the standard methodology for modelling 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

[*] 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.

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 modelled at the local level by using variants of von Neumann's cellular automaton [5]. 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 modelling methodology, for example hybrid continuous/discrete versions, and they go by several names, such as configuration models and object-oriented models, but the most widely used term is individual-based models.

In spite of the increasing reliance on individual-based modelling, progress is hindered by some of the same issues that arise in more traditional modelling. 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 individual-based models. The basic question of which aspects of the model are necessary and which can be abstracted away becomes even more difficult with individual-based models.

In this article, we propose a formal syntax and semantics for individual-based models. It is based on a term logic that defines both the state of an individual-based model at any time and the dynamics of the system. There are variants of individual-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 individual-based models to be abstracted to state-variable models. We also give some examples of individual-based models that do not satisfy these conditions and cannot be abstracted to state-variable models. We conclude with an outline of future research.

2 Examples

Due to space limitations, examples of individual-based models and proofs of our theorems are not included here. They are in the full draft of this paper [24]

3 The State Space of Individual-Based Models.

An individual-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 individual-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 [17] 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 individual-based models, such as process algebras augmented with probabilistic operators [26], Abstract State Machines [19] with probabilistic updates, CLU models with probabilistic next-state expressions [4], or relational databases with probabilistic updates. We have chosen a formalism that is simple and direct and includes all examples of individual-based models we have encountered in the scientific literature. Not surprisingly, our formalism is quite close to Abstract State Machines. In fact, the AsmL implementation of Abstract State Machines has the capability of performing random updates since it is integrated with the .NET Framework.

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 $m\colon S \to \mathbb{N}$, 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 $\mathrm{fm}(S)$ be the collection of finite multisets over S. A multiset operation on S is a function $\Gamma\colon\ \mathrm{fm}(S^k) \to S$ for some $k \in \mathbb{N}$. Γ is of arity k.

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 function 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 from $\mathrm{fm}(\mathbb{R}^k)$ to $\mathbb{R}$ and Γ is k-ary.

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 $w^{\mathfrak{A}}(a_1, \ldots, a_k) = \mathrm{undef}$ for $(a_1, \ldots, a_k) \notin \mathrm{dom}(w^{\mathfrak{A}})$.

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

The states of a classical state-variable model are a special case 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 definition of isomorphism between two models easily extends to metafinite models.

Definition 2. *Let*

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

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 $f \colon A \to B$ such that for all $w \in \mathcal{W}$ and $a_1, \ldots, a_k \in A$ where w is k-ary, $w^{\mathfrak{A}}(a_1, \ldots, a_k) = w^{\mathfrak{B}}(f(a_1), \ldots, f(a_k))$.

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

4 A Logic for Individual-Based Models

We will use a pure term calculus to express properties of individual-based models and a closely related logic to define the transition rules. The logic has 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 individual and numeric variables respectively. In any term, we assume that the type of each variable has been declared. As in first-order logic, individual 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 individual variable. Let $\mathfrak{A}$ be a metafinite model over the vocabulary $(\mathcal{W}, \mathcal{F}, \mathcal{G})$ and τ be a term with free variables among $x_1, \ldots, x_i$ and numeric variables among $y_1, \ldots, y_j$. We will define $\tau^{\mathfrak{A}}$ to be a function on $A^i \times \mathbb{R}^j$. If $a_1, \ldots, a_i \in A$ and $r_1, \ldots, r_j \in \mathbb{R}$, then $\tau^{\mathfrak{A}}(a_1, \ldots, a_i; r_1, \ldots, r_j)$ 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 τ 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 $x_1, \ldots, x_k$ are individual variables, then $w(x_1, \ldots, x_k)$ is a term with free variables $x_1, \ldots, x_k$.

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

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

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

$$\begin{aligned} (\Gamma z \tau)^{\mathfrak{A}}&(a_1, \ldots, a_i; r_1, \ldots, r_j) \\ &= \Gamma(\{\!|(\tau^{\mathfrak{A}}(a_1, \ldots, a_i, b_1; r_1, \ldots, r_j), \\ \ldots, \tau^{\mathfrak{A}}&(a_1, \ldots, a_i, b_k; r_1, \ldots, r_j)) \mid b_1, \ldots, b_k \in A|\!\}). \end{aligned}$$

Term logics can define aggregate properties of individual-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 Σ 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

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

(We are using an informal syntax, where the relation $<$ is actually a binary 0-1 valued function.)

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 χ_R. Every $\mathcal{L}$ model $\mathfrak{A}$ can be expanded to a $\mathcal{L}'$ model where each χ_R is interpreted as the characteristic function of R: for $a_1, \ldots, a_k \in A$,

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

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 ϕ be a formula in $\mathcal{L}$ with free variables $x_1, \ldots, x_k, y$ and χ_ϕ 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 individual-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 $a_1, \ldots, a_k \in A \cup A'$,

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

and similarly for w'.

In general, the transition probabilities are defined 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. If we are using declarative semantics, we use a term δ to describe the transition between two states. Let $\mathcal{S}$ be any set of metafinite models over $(\mathcal{W}, \mathcal{F}, \mathcal{G})$. Fixing $\mathfrak{A} \in \mathcal{S}$ and letting $\mathfrak{A}'$ range over $\mathcal{S}$, $\delta^{(\mathfrak{A}, \mathfrak{A}')}(t, t')$ is a multivariate random variable on $\mathcal{S}$ that characterizes the transition from $\mathfrak{A}$ to $\mathfrak{A}'$. Thus, let δ be k-variate, that is, $\delta = (\delta_1, \ldots, \delta_k)$ for some $k \in \mathbb{N}$, and $\delta_i^{(\mathfrak{A}, \mathfrak{A}')} \in \mathbb{R}$ for $i = 1, \ldots, k$. We use a term $F(x_1, \ldots, x_k, t, t')$ in the logic of $(\mathcal{W}, \mathcal{F}, \mathcal{G})$ to define the conditional cumulative distribution function of the transition. For $\mathfrak{A} \in \mathcal{S}$, $r_1, \ldots, r_k \in \mathbb{R}$, and $0 \leq t \leq t'$,

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

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

$$g^{\mathfrak{A}}(r_1, \ldots, r_k, t, t') = \\ \left.\frac{\partial^k F^{\mathfrak{A}}(x_1, \ldots, x_k, t, t')}{\partial x_1 \ldots \partial x_k}\right|_{x_1 = r_1, \ldots, x_k = r_k}.$$

If the process is homogeneous, then the transition probability does not depend on t, and we can write $F(x_1, \ldots, x_k, \Delta t)$, where $\Delta t = t' - t$. If the process operates in discrete time steps, then we take $t' = t + 1$, and we can write $F(x_1, \ldots, x_k, t)$. 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 $F(x_1, \ldots, x_k)$.

If $\mathcal{S}$ is discrete, i.e., the process is a Markov chain, we can define the transition probabilities directly with a conditional probability function f:

$$f^{(\mathfrak{A}, \mathfrak{A}')}(t, t') = \Pr(\mathfrak{A}_{t'} = \mathfrak{A}' | \mathfrak{A}_t = \mathfrak{A})$$

Alternatively, we can use the transition rate function g:

$$g^{(\mathfrak{A}, \mathfrak{A}')}(t) = \lim_{\Delta t \to 0} \frac{f^{(\mathfrak{A}, \mathfrak{A}')}(t, t + \Delta t) - \alpha^{(\mathfrak{A}, \mathfrak{A}')}}{\Delta t},$$

if the limit exists, where

$$\alpha^{(\mathfrak{A}, \mathfrak{A}')} = \begin{cases} 1 & \text{if } \mathfrak{A} = \mathfrak{A}' \\ 0 & \text{otherwise.} \end{cases}$$

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

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

Update rules can also be expressed in an imperative style, as in Abstract State Machines [19]. 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 $(v_1, \ldots, v_k)$. That is, a diffusion process is a state-variable model. The update rule of a diffusion process has the form

$$\begin{aligned}(v_1, \ldots, v_k) := (v_1, \ldots, v_k) + \mu(v_1, \ldots, v_k)\Delta t \\ + (W(t + \Delta t) - W(t))\Lambda(v_1, \ldots, v_k),\end{aligned}$$

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

6 Abstractions of Individual-Based Models

All attempts at modelling 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 modellers. 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 individual-based approach makes it very easy to construct highly detailed models, but as pointed out by Grimm [18], "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 individual-based modelling.

The process of simplifying a model is called abstraction. The two most common forms of abstraction in individual-based modelling are removing some of the weight functions and summarizing population data with aggregate functions. 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 the values of aggregate functions of the populations, and defining the state transitions in terms of these functions.

In many cases, the exact value of an aggregate function 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 individual-based model is approximated by deterministic transitions. The systems of deterministic differential equations in Section 2 are examples.

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

1. Abstracting the individual-based model to a finite number of aggregate functions.
2. Assuming that as the size of the individual-based model increases, with high probability the values of the aggregate functions 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 individual-based model with a deterministic state-variable model. That is, we give sufficient conditions for an individual-based model to be approximated in this way and show that some of the previous examples satisfy these conditions. We also examine a recent technique for equation-free modelling 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 τ_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 individual-based models over vocabularies $\mathcal{V}$ and $\mathcal{V}'$ respectively. Suppose there is a map $\alpha\colon \mathcal{S} \to \mathcal{S}'$ 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 $a_1, \ldots, a_i$ in A' and all $r_1, \ldots, r_j \in \mathbb{R}$, $w^{\mathfrak{A}'}(a_1, \ldots, a_i; r_1, \ldots, r_j) = \tau_w^{\mathfrak{A}}(a_1, \ldots, a_i; r_1, \ldots, r_j)$. We say that α is an abstraction function and $\mathcal{I}'$ is an abstraction of $\mathcal{I}$ via α.*

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 $0 \le t \le t'$, let*

$$\begin{aligned}&G^{\mathfrak{A}}(a_1, \ldots, a_i; r_1, \ldots, r_j, s, t, t') = \\ &\qquad \Pr\left(\tau_w^{\mathfrak{A}_{t'}}(a_1, \ldots, a_i; r_1, \ldots, r_j) \le s \mid \mathfrak{A}_t = \mathfrak{A}\right)\end{aligned}$$

and

$$\begin{aligned}&H^{\mathfrak{B}}(a_1, \ldots, a_i; r_1, \ldots, r_j, s, t, t') = \\ &\qquad \Pr\left(w^{\mathfrak{B}_{t'}}(a_1, \ldots, a_i; r_1, \ldots, r_j) \le s \mid \mathfrak{B}_t = \mathfrak{B}\right)\end{aligned}$$

be the conditional cumulative distribution functions of $\tau_w^{\mathfrak{A}_{t'}}(a_1, \ldots, a_i; r_1, \ldots, r_j)$ and $w^{\mathfrak{B}_{t'}}(a_1, \ldots, a_i; r_1, \ldots, r_j)$ 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* τ_w *with accuracy* γ *and confidence* ϵ *if for* $\mathfrak{A}' = \alpha(\mathfrak{A})$, *all* $a_1, \ldots, a_i \in A'$ *and* $r_1, \ldots, r_j, s \in \mathbb{R}$, *there exists* $s' \in \mathbb{R}$ *such that* $|s - s'| \leq \gamma$ *and* $|G^{\mathfrak{A}}(a_1, \ldots, a_i; r_1, \ldots, r_j, s, t, t') - H^{\mathfrak{A}'}(a_1, \ldots, a_i; r_1, \ldots, r_j, s', t, t')| \leq 1 - \epsilon$.

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 γ and confidence ϵ for the specified $\mathfrak{A}$, t and t'.

A special case of approximation implies a form of lumpability or probabilistic bisimulation [11] 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 $\{\mathfrak{A} | \tau_w^{\mathfrak{A}}(a_1, \ldots, a_i; r_1, \ldots, r_j) \leq r\}$, 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. It usually suffices to have an abstraction that approximates the individual-based model with a known degree of accuracy. A common assumption is that the accuracy of an abstraction increases relative to the time interval as population sizes increase. Further, this condition may hold only for states in a certain region of the state space. In the following definition, $\mathcal{R}$ will denote this region, and for $n \in \mathbb{N}$, $\mathcal{R}_n = \{\mathfrak{A} \in \mathcal{R} : |A| = n\}$.

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* γ *and confidence* ϵ *with respect to* τ_w *for all* $\mathfrak{A} \in \mathcal{R}_n$. *Then we say that* $\mathcal{I}$ *converges to* $\mathcal{I}'$ *with respect to* τ_w *for* $\mathcal{R}$, t *and* t'.

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

A further common simplification is the assumption that the dynamics of an individual-based model converges to a deterministic state-variable model. This is also a special case of our definition of approximation. 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 $\Pr\left(w^{\mathfrak{B}_{t'}} = s \mid \mathfrak{B}_t = \mathfrak{B}\right) = 1$. Therefore if $\mathcal{I}'$ approximates $\mathcal{I}$ with accuracy γ and confidence ϵ 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 $\mathfrak{A} \in \mathcal{R}$,

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

The following diagram illustrates the idea behind this definition.

$$\begin{tikzcd}
 & \mathfrak{A}_t \arrow[r, "\Delta t"] \arrow[d, "\alpha"'] & \mathfrak{A}_{t+\Delta t} \arrow[d, "\alpha"] & \\
\alpha(\mathfrak{A}_t) = & \mathfrak{B}_t \arrow[r, "\Delta t"] & \mathfrak{B}_{t+\Delta t} & \approx \alpha(\mathfrak{A}_{t+\Delta t})
\end{tikzcd}$$

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

We will investigate the problem of characterizing individual-based models for which this assumption is valid.

Erban et al. [13] 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 individual-based model where the behavior is approximated by deterministic evolution 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 [13]. The method consists of four steps, as ilustrated below.

$$\begin{tikzcd}
\mathfrak{A}_t \arrow[r, "\Delta t"] & \mathfrak{A}_{t+\Delta t} \arrow[d, "\alpha"] & \\
\mathfrak{B}_t \arrow[u, "\alpha^{-1}"] & \mathfrak{B}_{t+\Delta t} \arrow[r, "T"] & \mathfrak{B}_{t+\Delta t+T}
\end{tikzcd}$$

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 Δt, 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 a term to its mean-field approximation. That is, we will estimate how large an individual-based model must be in order to give a specified accuracy and confidence for the mean-field approximation. 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 individual-based modeling. In general, they help in estimating the size of a model required for simulation. More specifically, when an individual-based model converges to a state-variable model, concentration bounds can indicate the scale at which approximation by state-variable model is satisfactory. Since this size can be modest (on the order of several hundred individuals according to Gillespie [16]), 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 [2]. We assume that the state space is discrete and the Markov chain is homogeneous. We will give precise conditions that justify the use of mean-field approximations. These conditions are quite restrictive, but are satisfied by many systems studied in biology and physics, including some of those described in Section 2. We do not expect that these conditions are necessary, but we will give examples of individual-based models that fail the conditions and cannot be approximated by state-variable models. We begin with the case when time is discrete.

7.1 Discrete Time

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

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

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

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

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 $t = 0, \dots, T-1$,

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

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 individual-based models, which imply a Lipschitz condition. Our conditions are smoothness properties on the update probability function $f^{(\mathfrak{A},\mathfrak{A}')}$ and the term τ that we wish to approximate. Essentially, they say that updates change the state by a small amount, and a small change in a state $\mathfrak{A}$ results in a small change in the probability distribution of its next state.

Definition 8. *Let $\mathcal{I} = (\mathcal{S}, f)$ be an individual-based model where f is its update probability function. For any states $\mathfrak{A}$, $\mathfrak{B} \in \mathcal{S}$, and $d \in \mathbb{N}$, we put $\mathfrak{A} \xrightarrow{d} \mathfrak{B}$ if there is a sequence $\mathfrak{A} = \mathfrak{A}_0, \dots, \mathfrak{A}_d = \mathfrak{B}$ such that $f^{(\mathfrak{A}_i,\mathfrak{A}_{i+1})} > 0$ for $i = 0, \dots, d-1$.*

We say f 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 $\mathfrak{A} \xrightarrow{1} \mathfrak{B}$.

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

$$q^{\mathfrak{A}}(r) = \sum \left\{ f^{(\mathfrak{A},\mathfrak{B})} \middle| \tau^{\mathfrak{B}} - \tau^{\mathfrak{A}} = r \right\}$$

Let $\varepsilon \colon \mathbb{N} \to \mathbb{R}$ such that $\lim_{n\to\infty} \varepsilon(n) = 0$. We say that f is ε-smooth at $\mathfrak{A}$ if the following holds. For all $\mathfrak{B}$ such that $|\tau^{\mathfrak{B}} - \tau^{\mathfrak{A}}| \leq c$, and all $r \in \mathbb{R}$,

$$|q^{\mathfrak{B}}(r) - q^{\mathfrak{A}}(r)| \leq q^{\mathfrak{A}}(r)\varepsilon(|A|).$$

Let

$$\begin{aligned} \mathcal{R} = \{\mathfrak{A} \in \mathcal{S} : & f \text{ is } \varepsilon\text{-smooth at } \mathfrak{B} \\ & \text{for all } \mathfrak{B} \text{ such that } \mathfrak{A} \xrightarrow{t} \mathfrak{B},\ 0 \leq t < T\}. \end{aligned}$$

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

Lemma 1. *Let $d, t \in \mathbb{N}$ such that $d + t \leq T$. Then for all states $\mathfrak{A} \in \mathcal{R}$ and $\mathfrak{B}$ such that $\mathfrak{A} \xrightarrow{d} \mathfrak{B}$, letting $\beta = \mathbf{E}(Z_1 - Z_0|\mathfrak{A}_0 = \mathfrak{A})$,*

$$|\,\mathbf{E}(Z_t - Z_0|\mathfrak{A}_0 = \mathfrak{B}) - \beta t| \leq |\beta| t(dt + t^2)\varepsilon(|A|).$$

Corollary 1. *For $t = 1, \dots, T$, $|Y_t - Y_{t-1}| \leq |\beta|(1 + O(T^3\varepsilon(n))) + |Z_t - Z_{t-1}|$.*

Corollary 2. *For any T, for sufficiently large $\mathfrak{A} \in \mathcal{R}$, $|Y_t - Y_{t-1}| \leq 3c$.*

Theorem 1. *For any $\gamma \in (0, \infty)$ and $\epsilon \in [0, 1)$, for sufficiently large T and $\mathfrak{A} \in \mathcal{R}$,*

$$\Pr\left(|Z_T - \mathbf{E}\left(Z_T\right)| < \gamma T\right) > \epsilon.$$

Corollary 3. *For any $\gamma \in (0, \infty)$ and $\epsilon \in [0, 1)$, for sufficiently large T and $\mathfrak{A} \in \mathcal{R}$,*

$$\Pr(|Z_T - Z_0 - T\beta| < \gamma T) > \epsilon.$$

Theorem 2. *Assume $\mathcal{I}$ and τ satisfy Definition 8. 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 $\mathfrak{A} \in \mathcal{R}$, $\mathcal{I}$ can be approximated by $\mathcal{I}'$ with accuracy γT and confidence ϵ with respect to τ.*

7.2 Continuous Time

We now consider discrete space and continuous time individual-based models. Let $g^{(\mathfrak{A},\mathfrak{A}')}$ be the transition rate function. That is, for any small time interval dt, $f^{(\mathfrak{A},\mathfrak{A}')}(dt)$ is approximately $g^{(\mathfrak{A},\mathfrak{A}')}dt$. The definitions of $q^{\mathfrak{A}}(r)$ and smoothness (Definition 8) are now rephrased with f replaced by g. The ideas of Section 7.1 apply here, with the unit time interval and β replaced by dt and βdt respectively. Theorems 1 and 2 still hold in this new context, with the accuracy factor replaced by $\gamma T dt$. In addition, we have

Corollary 4. *Assume $\mathcal{I}$ and τ satisfy Definition 8. Then $\mathcal{I}$ converges to a continuous state-variable model $\mathcal{I}'$ with transitions defined by a system of ordinary differential equations.*

Proof. Let Δt be a time interval which can be arbitrarily small. For a given T, let $dt = \Delta t/T$. We regard $\mathcal{I}$ as a discrete time individual-based model with transition time interval dt. By Theorem 2, taking T and $\mathfrak{A}$ large enough, we can get any desired degree of accuracy and confindence. That is,

$$\left|\tau^{\mathfrak{A}_T} - \tau^{\mathfrak{A}_0} - T\beta dt\right| < \gamma T dt, \text{ or}$$
$$\left|\frac{\Delta w}{\Delta t} - \beta\right| < \gamma,$$

which implies

$$\frac{dw}{dt} = \beta.$$

□

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 individual-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 [15] and Hanf [20] to establish limitations on the expressive power of first-order logic. It has been extended to more powerful logics such as counting logics [23]. 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

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

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, γ is a metric on A.

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

$$\begin{aligned} N_r^{\mathfrak{A}}(a_1, \ldots, a_k) = \\ \{b \in A | \gamma(a_i, b) \leq r \text{ for some } i = 1, \ldots, k\}, \end{aligned}$$

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

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

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

Definition 9. *Let S be the set of isomorphism types $[\mathfrak{N}_r^{\mathfrak{A}}(a_1, \ldots, a_k)]$, taken over all $k \in \mathbb{N}$, metafinite models*

$\mathfrak{A}$ *over our vocabulary, and* $a_1, \ldots, a_k \in A$. *A neighborhood multiset operation is a function*

$$\Gamma\colon \operatorname{fm}(S) \to \mathbb{R}.$$

Definition 10. *The depth of a term is its maximum nesting of multiset operators. We define this more precisely by induction on the height of the term's parse tree. A term* $w(x_1, \ldots, x_k)$ *where* $w \in \mathcal{W}$ *has depth 0. If the maximum depth of* $\tau_1, \ldots, \tau_m$ *is* d *and* $f \in \mathcal{F}$ *is* m*-ary, then* $f(\tau_1, \ldots, \tau_m)$ *has depth* d*. If* τ *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 numbers of isomorphism classes that occur as bounded neighborhoods in the metafinite model.

Lemma 2. *Every term* $\tau(x_1, \ldots, x_k)$ *of depth* d *is equivalent to a term* $(\Gamma y \mathfrak{N}_{\rho(d)}(x_1, \ldots, x_k, y))$*, where* Γ *is a neighborhood multiset operator. That is, for every model* $\mathfrak{A}$ *and every* $a_1, \ldots, a_k \in A$,

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

9 Applications to Bounded Degree Structures

The first step in abstracting an individual-based model to a state-variable model is finding a finite set of terms whose values characterize the states of the individual-based model. In the simplest cases, e. g. models of chemical kinetics, this is easy: the population sizes of the various species determine the dynamics of the system. In this section, we generalize this idea to a class of individual-based models that can be characterized by population sizes of bounded neighborhoods of the individuals. This class is a generalization of the class of bounded degree structures. We give sufficient conditions for approximation by a state-variable model. We also give some biologically motivated examples of individual-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 [23].

Let $\mathcal{I} = (\mathcal{S}, f)$ be a discrete space individual-based model. If there are only finitely many isomorphism types among $\{\mathfrak{N}^{\mathfrak{A}}_1(a) : \mathfrak{A} \in \mathcal{S} \text{ and } a \in A\}$, 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 $\{\mathfrak{N}^{\mathfrak{A}}_r(a) : \mathfrak{A} \in \mathcal{S} \text{ and } a \in A\}$.

9.1 Approximation of Individual-Based Models by Deterministic State-Variable models

Using our canonical representation of terms, we will give characterizations of individual-based models of bounded degree with regions that can be approximated by deterministic state-variable models. Some of the best known examples of individual-based models satisfy these conditions.

Let τ be a term in the logic of $\mathcal{I}$ and d be the maximum of the depths of f and τ. Since $\mathcal{I}$ is of bounded degree, there are only finitely many, say k, isomorphism classes among the neighborhoods of radius d in $\mathcal{I}$. Let $\mathcal{S}' = \mathbb{N}^k$, and for every $\mathfrak{A} \in \mathcal{S}$, let $\alpha(\mathfrak{A}) = (n_1, \ldots, n_k)$, where n_i is the number of neighborhoods in $\mathfrak{A}$ belonging to the ith isomorphism class. By Lemma 2, there are functions $\Gamma\colon \mathbb{N}^k \to \mathbb{R}$ and $\Delta\colon \mathbb{N}^{2k} \to \mathbb{R}$ such that for all $\mathfrak{A} \in \mathcal{S}$,

$$\tau^{\mathfrak{A}} = \Gamma(\alpha(\mathfrak{A})),$$

and for all $\mathfrak{A}' \in \mathcal{S}$,

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

Applying Lemma 2, Theorem 2, and Corollary 4, we have

Theorem 3. *Assume* Γ *and* Δ *satisfy the following Lipschitz condition. There is a constant* c *such that for any* $n_1, \ldots, n_k, n'_1, \ldots, n'_k \in \mathbb{N}$,

$$\Delta(n_1, \ldots, n_k, n'_1, \ldots, n'_k) > 0 \implies |\Gamma(n_1, \ldots, n_k) - \Gamma(n'_1, \ldots, n'_k)| \leq c.$$

Let $\mathcal{R} \subseteq \mathcal{S}$ *be any region in which* $\mathcal{I}$ *satisfies the smoothness condition.*

If $\mathcal{I}$ *is a discrete time individual-based model, 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* $\mathfrak{A}$*,* $\mathcal{I}$ *can be approximated by* $\mathcal{I}'$ *with accuracy* γT *and confidence* ϵ *with respect to* τ.

If $\mathcal{I}$ *is continuous, it converges to a continuous deterministic state-variable model whose transitions are defined by a system of ordinary differential equations.*

Of the examples in Section 2, the models of chemical kinetics satisfy the conditions of Theorem 3, for regions where the population sizes are fixed or increase without bound. This also holds for the models with spatial information where spatial relations are represented by a lattice of bounded degree. Similar reasoning applies to the ecological models. The trophic models that do not include spatial information satisfy the theorem, as do the patch-occupancy models [22] provided each patch has a finite number of states. The individuals in the behavioral models do not have bounded degree because of their spatial attributes, and the theorem does not apply.

In general, the graph growth models do not satisfy the conditions of Theorem 3 because they have unbounded degree. However, in some cases, e.g., [3, 6, 21], some parameter settings result in graphs with finite average degree, and regions containing such graphs satisfy the theorem.

Dalvi, Miklau, and Suciu [8, 9] 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 Theorem 3, this may be a topic worth exploring.

9.2 Individual-Based Models That Cannot be Approximated by State-Variable models

If a bounded-degree individual-based model can be approximated by a state-variable model within some region, then roughly speaking, the states in the region 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 [23]. Thus it should not be surprising that there are individual-based models that cannot be approximated by any state-variable model in certain regions. 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 individual-based models. We give two examples of individual-based models that cannot be approximated by state-variable models in certain nontrivial regions. They are simplified models of fundamental aspects of molecular biology.

The first example illustrates the use of a membrane to control molecular interactions. We will model this in the style of StochSim. The universe of each state consists of n^2 individuals, for $n \in \mathbb{N}$, which we will call sites, arranged in a square lattice. Each site is labelled with a symbol 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

M if occupied by a membrane molecule

For an individual a, we put $l(a)$ for its label. Time is discrete, and transitions are determined locally. At each step, a site is randomly selected. If the site is labelled C, D, or E, and the four nearest neighbors are vacant, then the symbol C, D, or E, can be shifted to one of these neighbors or remain in place, with equal probabilities for all choices. If the site is labelled C, and at least one of its four nearest neighbors is labelled D, then one of them is randomly chosen and changed to blank, and the C is changed to E. All other possible neighborhoods of the selected site remain the same. In particular, sites constituting the membrane are fixed. This is intended to be a simple model of reaction-diffusion controlled by a membrane. Let τ be a term whose interpretation is the number of molecules of type E.

Let $\mathcal{R}$ be any region satisfying the following. There are arbitrarily large states $\mathfrak{A}$ and $\mathfrak{B}$ such that

- $|A| = |B|$.
- $|\{a \in A : l(a) = C\}| = |\{a \in B : l(a) = C\}| = \Theta(n^2)$.
- $|\{a \in A : l(a) = D\}| = |\{a \in B : l(a) = D\}| = \Theta(n^2)$.
- $|\{a \in A : l(a) = M\}| = |\{a \in B : l(a) = M\}| = \Theta(n^2)$.
- $|\{a \in A : l(a) = E\}| = |\{a \in B : l(a) = E\}| = 0$.
- For all $a, b \in A$, if $l(a) \neq$ blank and $l(b) \neq$ blank, then $\gamma^{\mathfrak{A}}(a, b) > d$, where d is determined below.
- For all $a, b \in B$, if $l(a) \neq$ blank and $l(b) \neq$ blank, then $\gamma^{\mathfrak{B}}(a, b) > d$.
- The M sites in both $\mathfrak{A}$ and $\mathfrak{B}$ divide their universes into two parts called left and right.
- All $a \in A$ such that $l(a) =$ C or $l(b) =$ D are in the left part of A.
- All $a \in B$ such that $l(a) =$ C are in the left part of B, and all $a \in B$ such that $l(a) =$ D are in the right part of B.

Suppose the above individual-based model $\mathcal{I}$ is abstracted via α to a state-variable model $\mathcal{I}'$. Let $\mathcal{V} = (\mathcal{W}, \mathcal{F}, \mathcal{G})$ and $\mathcal{V}' = (\mathcal{W}', \mathcal{F}', \mathcal{G}')$ be their respective vocabularies. Then for every $v \in \mathcal{W}'$, there is a term τ_v in the logic of $\mathcal{V}$ that corresponds to v.

We will show that $\mathcal{I}$ does not converge to $\mathcal{I}'$ with respect to τ for any sufficiently large time interval T. Let d be the maximum depth of all the terms τ_v for $v \in \mathcal{W}'$. If we assume that $\mathcal{I}$ converges to $\mathcal{I}'$ with respect to τ, then $\tau = \tau_w$ for some $w \in \mathcal{W}'$. From the theory of random walks, with positive probability, $\tau^{\mathfrak{A}_T} = \Theta(T)$. By Lemma 2, $\tau^{\mathfrak{A}} = \tau^{\mathfrak{B}}$. But $\tau^{\mathfrak{B}}$ will always remain 0. Therefore our assumption leads to a contradiction.

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 modelled by a state-variable model, this negative result also holds for these more complicated systems.

The states of our individual-based model are directed labelled 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 labelled R. Initially, all R vertices are isolated, and there are no B vertices, but there may be template chains (possibly defective).

At each step, a vertex is randomly selected. If it is an R vertex of outdegree 0, then a second vertex is randomly selected. If the second is an S of indegree 0, then an edge is added from the R to the S. If the first vertex is an R with an edge to an S, the R breaks its attachment to the S and reattaches to the first A in the template chain. This begins the process of concatenating a B to the end of the growing output chain and moving the attachment of the R to the next A molecule. This is repeated each time the R is selected until the F link is reached, and then the R and the completed output chain are released. 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 τ be the number of B vertices of outdegree 0. Since there are no B vertices initially, the value of τ 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 [23]. The proof that our individual-based model cannot be approximated by a state-variable model is based on the same ideas.

Assuming this individual-based model is abstracted via α to a state-variable model let d be the maximum depth of all the terms that correspond to weight functions in the state-variable model. Let $\mathfrak{A}$ have the following components:

n chains SA^{2d}

n chains $\mathrm{A}^{2d}\mathrm{F}$

n chains A^{2d}

n chains $\mathrm{SA}^{2d}\mathrm{F}$.

Let $\mathfrak{B}$ have $2n$ components SA^{2d} and $\mathrm{A}^{2d}\mathrm{F}$. Again by Lemma 2, $\tau^{\mathfrak{A}} = \tau^{\mathfrak{B}}$, 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 [28], 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 modelling. Later versions that included the capability of individual-based modelling have had limited acceptance. Perhaps it will be necessary to include this capability as a basic feature of future modelling 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 [7, 10, 27]. A temporal logic for individual-based models could be developed for the purpose of model checking. This would also require developing methods for approximating individual-based models with finite state systems. The methods of Desharnais et al. [12] may be useful here.

We have mentioned the models of very sparse random databases. To our knowledge, the rules for evolving such databases have not been investigated. Characterizing these rules and extending our results to evolving databases may have applications to very large databases.

Since our results 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. They are approximated by partial differential equations. Can Theorem 3 be extended to this class of models?

Theorems 2 and 3 apply to regions of state space where the sizes of the neighborhod 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. [1] have modelled 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 modelled 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 2 and 3. 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 [14], if the number of conformational states of the molecules is very large, then describing the possible reactions with a state-variable model becomes impractical. Theorem 3 gives sufficient conditions for approximating an individual-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 individual-based models for which this is optimal?

Our approximations result in deterministic state-variable models. Are there individual-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?

References

[1] A. Arkin, J. Ross, and H. H. McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected *Escherichia coli* cells. *Genetics*, 149:1633–1648, 1998.

[2] K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 68:357–367, 1967.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] N. N. Dalvi, G. Miklau, and D. Suciu. Asymptotic condtional probabilities for conjunctive queries. In *Proc. 10th Int. Conf. on Database Theory, Lecture Notes in Computer Science*, volume 3363, pages 289–305. Springer, 2005.

[10] V. Danos and C. Laneve. Formal molecular biology. *Theoretical Computer Science*, 325:69–110, 2004.

[11] J. Desharnais, A. Edalat, and P. Panangaden. Bisimulation for labelled Markov processes. *Information and Computation*, 179:163–193, 2002.

[12] J. Desharnais, V. Gupta, R. Jagadeesan, and P. Panangaden. Approximating labelled Markov processes. *Information and Computation*, 184:160–200, 2003.

[13] R. Erban, I. G. Kevrekidis, and H. G. Othmer. An equation-free computational approach for extracting population-level behavior from individual-based models of biological dispersal. *Physica D: Nonlinear Phenomena*, 215:1–24, 2006.

[14] 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.

[15] H. Gaifman. On local and non-local properties. In *Proc. of the Herbrand Symposium, Logic Colloquium '81*. North-Holland, 1982.

[16] D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Computational Physics*, 22:403–434, 1976.

[17] E. Grädel and Y. Gurevich. Metafinite model theory. *Information and Computation*, 140:26–81, 1998.

[18] 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.

[19] Y. Gurevich. Sequential abstract state machines capture sequential algorithms. *ACM Trans. on Computational Logic*, 1:77–111, 2000.

[20] 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.

[21] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. In *Proc. 41st IEEE Symp. on Foundations of Computing Science*, pages 57–65, 2000.

[22] S. A. Levin, T. Powell, and J. H. Steele, editors. *Patch Dynamics*. Springer, 1993.

[23] L. Libkin. Expressive power of SQL. *Theoretical Computer Science*, 296:379–404, 2003.

[24] J. F. Lynch. A logical characterization of individual-based models. Draft, http://people.clarkson.edu/˜jlynch/ibms.pdf.

[25] R. M. May. *Theoretical Ecology: Principles and Applications*. Saunders, Philadelphia, 1976.

[26] C. Priami. Stochastic pi-calculus. *Computer Journal*, 38:578–589, 1995.

[27] C. Priami, A. Ingolfsdottir, B. Mishra, and H. R. Nielson, editors. *Trans. Computational Systems Biology VII, Lecture Notes in Computer Science*, volume 4230. Springer, 2006.

[28] D. J. Wilkinson. *Stochastic Modelling for Systems Biology*. CRC Press, Boca Raton, FL, 2006.