Rapid Weighted Random Selection in Agent-based Models of Infectious Disease Dynamics Using Augmented B-trees

Roel Bakker*‡, Tony Busker*, Richard G. White † and Sunil Choenni*
* Creating 010, Rotterdam University of Applied Sciences, Rotterdam, Netherlands
† Department of Infectious Disease Epidemiology, Faculty of Epidemiology and Population Health
London School of Hygiene and Tropical Medicine, London, UK
‡ Skardahl BV, Rotterdam, Netherlands
Email: r.bakker@hr.nl, a.l.j.busker@hr.nl, richard.white@lshtm.ac.uk, r.choenni@hr.nl

***Abstract*—Agent-based models (ABMs) are important tools for predicting infectious disease epidemics and for designing effective interventions. ABMs take into account individual differences, for instance in contact rate. The drawbacks of ABMs are high complexity and low performance. In this paper, we present a data structure - an augmented B-tree - to speed up the weighted random selection of individuals for the next transmission event in an ABM of infectious disease dynamics. An additional feature of the augmented B-tree is that it allows aggregating the force of infection for groups of simulated individuals. In short, our technique enhances the performance and simplifies the development of ABMs.**

***Keywords*—weighted random selection; ABM; agent-based modeling; infectious disease epidemics; B-tree; performance.**

I. INTRODUCTION

Agent-based models (ABMs) are important tools for predicting infectious diseases epidemics and for designing effective interventions [1]–[4].

The classic model for infectious disease dynamics is the so-called 'SIR model' formulated by Kermack and McKendrick [5] where S, I and R denote the susceptible, infected and recovered fractions of the population.

The original SIR model is a deterministic model where a set of differential equations describe the rates of change in S, I and R.

Stochastic models take both chance and the effect of population size into account, and model a population as discrete numbers of people in the S, I and R state. Each simulation run has a different outcome and evaluating a single scenario requires multiple runs and aggregating the output. These simulations take time as every single event (disease transmission, recovery) is modelled explicitly.

Agent-based models [6] are the most sophisticated type of model. In this type of model, individuals are represented as objects that differ from each other in the values of their attributes. In addition to taking chance and finite population size into account, agent-based models also take heterogeneity between individuals into account: some individuals may have higher contact rates than others, and some individuals may always recover from disease faster than others.

The main drawbacks of ABMs are high complexity and low performance. Each individual is a distinct object in software with its own attributes and life history. As in stochastic models, the time course of a simulation of an infectious disease with an ABM consists of a sequence of events (either transmission or recovery). If individuals differ in contact rate and/or recovery rate, a weighted random selection of individuals is required at each event. Selecting a single individual by iterating a list takes time proportional to the number of individuals in the list.

In this paper, we present an augmented B-tree as an efficient data structure for random selection of individuals weighted by the value of an individual attribute such as contact rate. The augmented B-tree is key for simulating epidemics in large (>100,000 individuals) populations with individual heterogeneity. The augmented B-tree can also be used to pinpoint the force of infection in a simulated population. In short, the data structure improves the performance of ABMs and makes it simpler to develop these models, which improves the tractability of this type of models.

The rest of this paper is laid out as follows. In Section II, we will provide the necessary background on the different types of SIR models. Section III describes the data structure in detail and reports performance figures. In Section IV, we will show results of using the data structure in an individual-based SIR model with different degrees in contact rate heterogeneity. Section V discusses the application of the data structure to an age structured population. In Section VI, we discuss additional features of this data structure and similar developments in the field of simulating networks of chemical reactions.

II. BACKGROUND ON INFECTIOUS DISEASES DYNAMICS

The classic model for infectious disease dynamics is the SIR model [5]. In this model, the population is subdivided into susceptible (S), infected (I) and recovered (R) categories. The following set of differential equations determines the dynamics:

$$dS/dt = -\beta \cdot c \cdot I \cdot S/N \quad (1)$$

$$dI/dt = \beta \cdot c \cdot I \cdot S/N - I/d \quad (2)$$

$$dR/dt = I/d \quad (3)$$

with

$$N = S + I + R \quad (4)$$

and

β	transmission probability per contact
c	contact rate
d	duration of infection

 ISBN: 978-1-61208-308-7

This model simulates numbers of individuals (or population fractions) as continuous variables and is deterministic: for a given set of initial values the model will always produce the same output. This type of model aims to capture the average behaviour of the epidemic.

Stochastic models take the random nature of transmission events between discrete individuals into account. A stochastic model simulates discrete numbers of people in the S, I or R state and produces different output for each simulation run. A stochastic equivalent of the deterministic SIR model simulates the transition events from the $S \rightarrow I$ state and the $I \rightarrow R$ state for discrete individuals, with:

$$r_{S \rightarrow I} = -dS/dt = \beta \cdot c \cdot I \cdot S/N \quad (5)$$

$$r_{I \rightarrow R} = dR/dt = I/d \quad (6)$$

The direct method [7] is an algorithm for stochastic models of chemical reaction kinetics that can be used to simulate the dynamics of this system:

1) sum the event rates
2) draw a random number x between 0 and the sum of the rates i.e., uniform on $(0, r_{S \rightarrow I} + r_{I \rightarrow R})$
3) determine whether infection or recovery will occur: infection if $0 < x < r_{S \rightarrow I}$ and recovery if $r_{S \rightarrow I} <= x < r_{S \rightarrow I} + r_{I \rightarrow R}$
4) draw a value for Δt from an exponential distribution with rate $r_{S \rightarrow I} + r_{I \rightarrow R}$
5) move the time forward to t+Δt and execute the event
6) go to step 1

Deterministic and stochastic models are relatively simple to implement although running stochastic models may be time consuming - especially for large populations - as every individual state transition is simulated.

Stochastic models do not take consistent heterogeneity between individuals into account. Stochastic models model numbers of molecules of a species or numbers of individuals in a certain state. Although it is possible to categorise a population into subgroups with different contact rates (e.g., the core group model for gonorrhea in the US [8]), an arbitrary and continuous distribution of contact rates (and/or recovery rates) within a population requires modelling at the level of the individual.

In the stochastic SIR model, the infection and recovery event rates are easily calculated from (5) and (6) and moving the simulation forward in time using the direct method is straightforward. In an agent-based SIR model with heterogeneity in both contact rate and duration of infection, the event rates are given by (assuming proportionate mixing):

$$r_{S \rightarrow I} = \beta \cdot \sum_{j=1}^{I} c_j \cdot \frac{\sum_{j=1}^{S} c_j}{\sum_{j=1}^{N} c_j} \quad (7)$$

$$r_{I \rightarrow R} = \sum_{j=1}^{I} \frac{1}{d_j} \quad (8)$$

In (7) the summed contact rate $\sum_{j=1}^{I} c_j$ of infected individuals replaces the product of contact rate and numbers of infected individuals $c \cdot I$ of(5). As we assume proportionate mixing, the summed contact rates of susceptible divided by the summed contact rate of all individuals $\sum_{j=1}^{S} c_j / \sum_{j=1}^{N} c_j$ in (7) replaces the fraction of contacts with susceptibles S/N of (5).

The principle of the direct method still works, but now we do not only have to determine which *type* of event occurs but also which *individual* should be selected for the transition event. Therefore step 5 in the algorithm is replaced by:

5) move the time forward to t+Δt and execute the event:
   a) if infection:
      - draw a random number y uniform on $(0, \sum_{j=1}^{S} c_j)$
      - iterate over all susceptibles subtracting c_j from y until $y < 0$
      - select that individual and execute the infection event
   b) if recovery:
      - draw a random number y uniform on $(0, \sum_{j=1}^{I} 1/d_j)$
      - iterate over all infected subtracting $1/d_j$ from y until $y < 0$
      - select that individual and execute the recovery event

The random selection of an individual weighted by contact rate (or recovery rate) in steps 5a and 5b performs poorly if individual rates are stored in a simple data structure such as an array or a list: the time complexity for iterating an array or list is $O(n)$ (i.e., the required time increases proportionally with the number of individuals). In the next section we present a more efficient data structure.

III. An augmented B-tree for weighted random selection of individuals

A. Data structure

To move an agent-based SIR model forward in time requires summing the individual infection and recovery rates, drawing a time till the next event and selecting the event type and the individual. As the individual rates may differ, the selection of an individual is a weighted random selection. After an individual has been selected, he or she is moved to the next state.

To prevent $O(n)$ time complexity, we would like an alternative to simply iterating over a list with individual rates. The main requirements for an alternative data structure or algorithm are:

- rapid weighted random selection of elements
- quick and easy insertion and removal of elements (or of updating the rates)

A data structure that fulfils these requirements is a balanced search tree with nodes that are augmented [9] with a record of the sum of an attribute of the child nodes. We chose a standard B-tree [10] as the basis. B-trees are widely used in relational databases and have $O(\log(n))$ time complexity for search, delete and insert actions [10]. Fig. 1 illustrates a B-tree

 ISBN: 978-1-61208-308-7

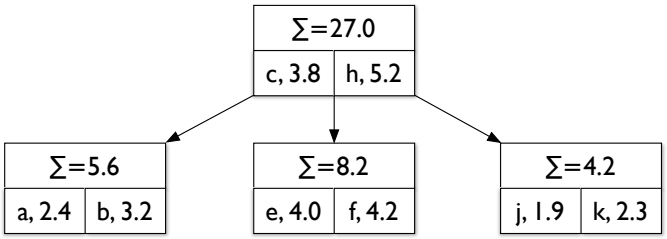

Fig. 1. Augmented B-tree for rapid selection weighted by individual rate. Each node contains *i.* the sum of the values of the elements in that node and in all subtrees of that node, *ii.* an ordered list of key, value pairs (e.g., a, 2.8 for smallest element in the tree), and *iii.* pointers to child nodes with keys intermediary between the keys of the elements left and right of the pointer. In this example, the tree contains contact rates (the values) of individuals identified by single character as key.

(of order 2, i.e., with either 1 or 2 elements per node) in which the nodes have been augmented with the sum of the values of the elements in that node and all subtrees of that node. In this example, each element would represent an individual denoted by a key (single lowercase character) and a value (e.g., contact rate). To select a random individual weighted by contact rate, we proceed as follows. First, a random number x is drawn from a uniform distribution between 0 and the sum of all values in the tree. Next, x is compared with the sum of the leftmost child node, and if the sum of that node is less than x, subtract that sum from x, and continue to the value of the leftmost element in the root node, the sum of middle child node etc. As soon as the current value of x is less than the value of an element or child node, the element or child node will be selected. Suppose the sum of the values is 27.0 (see Fig.1) and we have drawn $x = 10$; as the sum of the left child node $5.6 < 10$, $x \leftarrow 4.4$; as the value of $c, 3.8 < 4.4$, $x \leftarrow 0.6$ and we descend to the middle child node; as the value of e, $4.0 > x$, e is selected.

B. Expected performance

For each level in the tree, at most 4 comparisons (2 values within the node itself and 2 out of 3 subtree sums referenced by pointers) are needed to either find the required element or find the subtree containing the element. Thus, the number of comparisons for weighted random selection increases linear with the number of levels whereas the number of elements in a tree increases exponentially with the number of levels of the tree. Therefore, time complexity of weighted random selection is $O(\log(n))$. Time complexity of insert, delete and update actions in an augmented B-tree is also $O(\log(n))$ as only the sums of the nodes on the path leading to the element need updating for these actions. Selecting by key is the same as in standard B-trees, i.e., $O(\log(n))$.

As for standard B-trees [10], the space complexity for the augmented B-tree is $O(n)$; the pointer structure is identical to that of a B-tree but additional space is required for storing the sums. The space for storing the sums decreases with increasing number of elements per node and can be tuned. Note that the values do not have to be stored in the tree if the values can be referenced through the key.

TABLE I. PERFORMANCE OF A JAVA ARRAYLIST VS. AN AUGMENTED B-TREE FOR WEIGHTED RANDOM SELECTION. TIME IN μSECS PER SELECT (AVERAGE±SEM OF 10 RUNS OF 5,000 SELECTS EACH). ALL DIFFERENCES BETWEEN ARRAYLIST AND AUGMENTED B-TREE WERE SIGNIFICANT AT P < 1E-6 (STUDENT T-TEST).

Number of elements	ArrayList	Augmented B-tree
10,000	6.6±0.1	0.32±0.01
20,000	13.9±0.1	0.50±0.04
50,000	35.0±0.3	0.65±0.01
100,000	71.5±0.5	0.88±0.01
200,000	174±1	1.13±0.01
500,000	522±2	1.40±0.01
1,000,000	1073±3	1.71±0.02

C. Measured performance

Table 1 shows the performance of weighted random selection using a Java ArrayList versus an augmented B-tree. The time required for 5,000 weighted random select actions was determined for increasing numbers of elements in the data structure. Each test was performed 15 times. To allow the Java VM to warm up, only the final 10 runs were used for calculating the average and SEM. All tests were performed on a MacBook Pro with 8 GB RAM and a 2.8 GHz Intel Core i7 processor on OS X 10.8.4 using the Java SE 6 runtime. Minimum and maximum Java heap space was set to 768 MB. A Java software library including the augmented B-tree will be published as open source on www.skardahl.com, the website of Skardahl BV, before October 27, 2013.

The figures in Table 1 show that the augmented B-tree has much better performance than the Java ArrayList, even when just 10,000 elements are present in the data structure. In addition, the table show that the time complexity is about $O(n)$ for the ArrayList (about a 10-fold increase in time from 100,000 to 1,000,000 elements) whereas for the augmented B-tree the increase is about proportional to $\log(n)$. The amount of memory required for the augmented B-tree was about six times that of a Java ArrayList: for 1 million elements 147 MB for the augmented B-tree vs. 25 MB for the ArrayList.

IV. APPLICATION OF THE AUGMENTED B-TREE TO AN INDIVIDUAL-BASED SIR MODEL

Fig. 2 shows the dynamics of an agent-based SIR model with and without heterogeneity in contact rate. Heterogeneity decreases variability and causes an earlier and higher peak in the number of infecteds whereas the fraction susceptible remaining after the epidemic has died out is the same (not shown). The data shown in fig. 2 were generated by 20 simulation runs each modelling a population of size 100,000 with individuals with either the same or different contact rates. The 20 simulation runs took 5 seconds in total. When an ArrayList was used for weighted random selection the 20 runs took 12 minutes in total. Using the augmented B-tree therefore caused a speed-up of a factor 140. For larger population sizes the difference would even be larger.

V. USE OF THE AUGMENTED B-TREE FOR EPIDEMICS IN AGE STRUCTURED POPULATIONS

So far, we focused on random selection of elements weighted by the value of the element. Although the elements

 ISBN: 978-1-61208-308-7

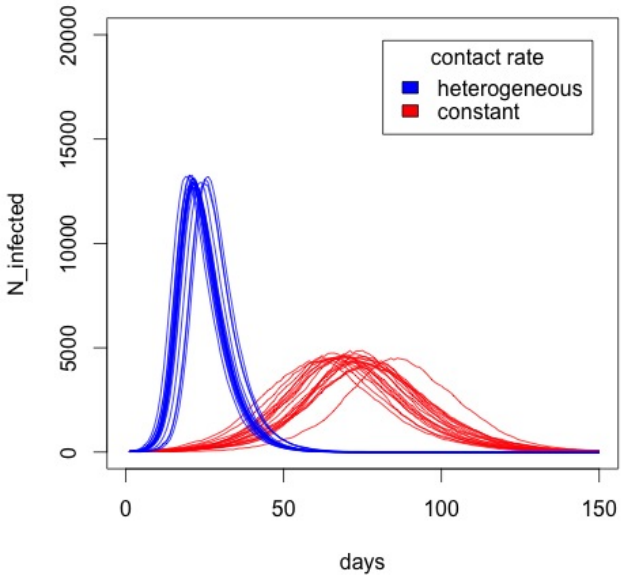

Fig. 2. Dynamics of a SIR model with 100,000 individuals, an average effective contact rate (equal to transmission probability per contact multiplied with contact rate) of 0.35 per day, and a disease duration of 4 days. The parameters are equivalent to R0 = 1.4. The red lines (peaking around 75 days) show the time course of the number of infected individuals for a model where all individuals have the same effective contact rate and the blue lines (that peak around 25 days) show the same for heterogeneous individual contact rates which were drawn from an exponential distribution with mean 0.35 per day.

are ordered by key, the ordering was irrelevant for the simple individual based SIR model.

When modeling SIR dynamics in an age-structured population, the obvious key to use is age (or birthdate). When using age as a key, the augmented tree allows summing the individual effective contact rates of the infecteds by age range. For each age range, the summed rate can be distributed over age ranges of susceptibles according to a contact matrix.

An additional feature of the augmented tree (not related to scheduling transmission events) is that we can get a quick response to queries for the sum of attribute values in a certain key range. For example, in an agent-based SIR model we could find out which age group (or birth cohort) causes most new infections by using birthdate as key and the effective contact rate as attribute value and querying different age ranges of infecteds. In the same way we could find out which age group is most subject to new infections by querying aggregate contact rates in age ranges of susceptibles. As for selection and update, the time complexity of key range queries is $O(\log(n))$.

VI. Discussion

We have described a data structure that can speed up agent-based simulations of infectious disease dynamics in large populations by a factor of 100 or more. The augmented B-tree that we have presented enables more realistic modeling of individual heterogeneity (e.g., in contact rate) while at the same time offering the option to easily aggregate the individual rates by key range thereby providing insight into the groups causing and experiencing the force of infection.

A similar algorithm as presented here has been described for the simulation of chemical reaction kinetics. Gibson and Bruck [11] developed a logarithmic scaling version of the SSA (stochastic simulation algorithm) using an indexed priority queue or binary tree for a more efficient way to select the chemical reaction that will fire next. Slepoy et al. [12] present a constant-time kinetic Monte Carlo algorithm that could in principle also be used for agent-based models of epidemics. However, the composition-rejection algorithm presented in [12] requires sorting the rates in categories to prevent too many rejections, and is not easy to implement. Also, it does not offer the option to aggregate rates by key range.

We believe that our augmented B-tree is useful in all agent-based models where random selection weighted by individual risk (i.e., rate, susceptibility, etcetera) is required. In addition, the augmented B-tree is useful to aggregate individual attribute values (e.g., rates), optionally by key range.

Acknowledgment

The first author would like to thank Marijn Bom for support and inspiration, and Sake de Vlas, Luc Coffeng and Richard Steen for stimulating discussions.

References

[1] K. K. Orroth, E. E. Freeman, R. Bakker, A. Buvé, J. R. Glynn, M.-C. Boily, R. G. White, J. D. F. Habbema, and R. J. Hayes, "Understanding the differences between contrasting hiv epidemics in east and west africa: results from a simulation model of the four cities study," *Sexually transmitted infections*, vol. 83, no. suppl 1, pp. i5–i16, 2007.

[2] J. A. Hontelez, S. J. de Vlas, F. Tanser, R. Bakker, T. Bärnighausen, M.-L. Newell, R. Baltussen, and M. N. Lurie, "The impact of the new who antiretroviral treatment guidelines on hiv epidemic dynamics and cost in south africa," *PloS one*, vol. 6, no. 7, p. e21919, 2011.

[3] N. M. Ferguson, D. A. Cummings, S. Cauchemez, C. Fraser, S. Riley, A. Meeyai, S. Iamsirithaworn, and D. S. Burke, "Strategies for containing an emerging influenza pandemic in southeast asia," *Nature*, vol. 437, no. 7056, pp. 209–214, 2005.

[4] J. M. Epstein, "Modelling to contain pandemics," *Nature*, vol. 460, no. 7256, pp. 687–687, 2009.

[5] W. Kermack and A. McKendrick, "A contribution to the mathematical theory of epidemics," *Proc. R. Soc. Lond. A.*, vol. 115, no. 772, pp. 700–721, 1927.

[6] V. Grimm and S. F. Railsback, *Individual-based modeling and ecology*. Princeton university press, 2005.

[7] D. T. Gillespie, "Exact stochastic simulation of coupled chemical reactions," *The journal of physical chemistry*, vol. 81, no. 25, pp. 2340–2361, 1977.

[8] H. W. Hethcote and J. A. Yorke, *Gonorrhea transmission dynamics and control*. Springer Berlin, 1984, vol. 56.

[9] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to algorithms*. The MIT press, 2001.

[10] R. Bayer and E. McCreight, *Organization and maintenance of large ordered indexes*. Springer, 2002.

[11] M. A. Gibson and J. Bruck, "Efficient exact stochastic simulation of chemical systems with many species and many channels," *The journal of physical chemistry A*, vol. 104, no. 9, pp. 1876–1889, 2000.

[12] A. Slepoy, A. P. Thompson, and S. J. Plimpton, "A constant-time kinetic monte carlo algorithm for simulation of large biochemical reaction networks," *The journal of chemical physics*, vol. 128, p. 205101, 2008.

 ISBN: 978-1-61208-308-7