Modeling human education data: From equation-based modeling to agent-based modeling

Yuqing Tang[1], Simon Parsons[2], and Elizabeth Sklar[2]

[1] Department of Computer Science
Graduate Center, City University of New York
365, 5th Avenue, New York, NY 10016, USA
ytang@gc.cuny.edu

[2] Department of Computer & Information Science
Brooklyn College, City University of New York
2900 Bedford Avenue, Brooklyn, NY 11210, USA
{parsons,sklar}@sci.brooklyn.cuny.edu

Abstract. Agent-based simulation is increasingly used to analyze the performance of complex systems. In this paper we describe results of our work on one specific agent-based model, showing how it can be validated against the equation-based model from which it was derived, and demonstrating the extent to which it can be used to derive additional results over and above those that the equation-based model can provide.

1 Introduction

We have been examining various sets of data related to human education. Typically, this data is collected in one of two ways: (1) very large, aggregate data sets over entire populations (like whole cities, school districts, states or provinces) or (2) very small, localized experimental samples. In both cases, the data is usually analyzed using standard statistical methods. Often, the most highly publicized statistics are the simplest, for example the mean and standard deviation of standardized test scores. These values are frequently the ones used to make policy decisions. Occasionally, analysis is performed that examines how multiple factors influence each other, such as the relationship between student-teacher ratios and test scores, dollars per student and test scores, or class size and test scores.

In this example, it is difficult to analyze and understand the relationships between these four factors (student-teacher ratios, test scores, dollars per student and class size) using standard statistical techniques; and as the set of factors increases in number and complexity, the analysis becomes even more complicated. Additionally, the statistical methods do not provide a means for examining students who fall more than one standard deviation outside the mean (either above or below). For example, maybe students who perform above the mean benefit from higher student-teacher ratios and smaller class size, while students who perform below the mean prefer lower student-teacher ratios but also smaller

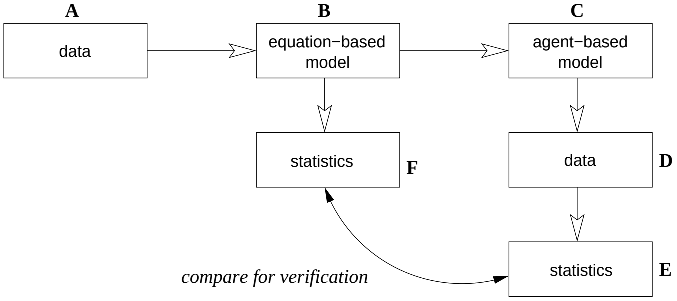

Fig. 1. Deriving an agent-based model from an equation-based model and then verifying it

class sizes. Further, the statistical methods do not provide a means for modeling the interactions between students. For example, some students may learn better in a homogeneous classroom, where all their classmates are of similar ability, while others might do better in a classroom where they can learn from social peers whose ability differs from theirs by more than a standard deviation. Our aim is to develop models that can make use of these subtle interactions, and use them to analyze the effects of education policy [10, 11]. We are using agent-based modeling to do this.

Agent-based modeling can help bridge the gap between macro and micro data sets, using both interpolation and extrapolation techniques to combine information and produce comprehensive, interactive and flexible environments for experimentation. Agent-based modeling is particularly appropriate [8] for systems in which there are many different loci of control [14], something that is a particular feature of the kinds of system that we are interested in modeling. In this paper, we describe results of our work on one specific agent-based model, showing how it can be validated against the more traditional model from which it was derived, and highlighting the extent to which it can be used to derive additional results over and above those that the traditional model can provide.

2 Agent-based modeling

Agent-based modeling contrasts with traditional approaches to simulation, which are typically built up from sets of interrelated differential equations. Such traditional models, commonly called *equation-based models* (EBMs), have been widely applied and generate useful predictions about the behavior of populations. So why use agent-based models? There seem to be four main answers [2]: (i) agent-based models are a natural way to describe systems comprised of interacting entities; (ii) agent-based models are flexible; (iii) agent-based models capture emergent phenomena; and (iv) agent-based models provide access to a greater level of useful detail. In particular, modeling interactions between entities can be much easier in agent-based systems than in EBMs, even when one is comfortable with the concepts of partial differential equations.

This naturalness and ease of modeling helps to make agent-based models more flexible than EBMs. As Bonabeau argues [2], agent models are typically simple, and so are easy to understand and thus to change. It is usually easy to increase the size of a simulation, adding new agents to see if interesting effects are swamped by agent numbers, or taking agents away if interesting detail is obscured. It is also possible to look at the results of simulations at different levels of detail—at the level of a single agent, at the level of some specific group of agents, or at the level of all agents together. All these things are harder to manage in EBMs.

In addition to their inherent naturalness and flexibility, agent-based simulations allow one to identify *emergent phenomena*. Emergent phenomena result from the actions and interactions of individual agents, but are not directly controlled by the individuals. Indeed, they have an existence that is partly independent of those individuals—the classic example of an emergent phenomenon is a traffic jam, which, while caused by the actions of drivers moving in one direction, may travel in the opposite direction.

Emergent phenomena simply do not show up in EBMs, but knowing about them can be crucial. As an example, Greenwald and Kephart [3, 5] showed that while intuition suggested that frequent price updates would allow firms to steal extra profits from their competitors, in fact it would lead to damaging price wars; and [1] showed how an agent-based model identifies effects of changes in rent-control policy that are beyond the reach of EBMs. Such findings are also echoed in ecology [4, 12] where agent-based models (under the name "individual-based models") have been used for some years.

As others have described [2, 8], it is possible to generate agent-based models from more traditional models. Figure 1 shows the process by which an agent-based model is derived from an equation-based model. Presumably, the equation-based model (labeled box "B") was created after performing statistical analysis on a raw data set (box "A"). By definition, the statistical equation will be able to capture regularities in the data set and will provide a snapshot view of the environment or phenomena which it models. The agent-based model (box "C") is created by taking each of the variables in the equation-based model and the distribution of each of the variables (this would be a statistical distribution, such as a Gaussian), and then by defining agent behaviors that will produce results falling within this distribution. While single behaviors may contribute to one or two variables, the interaction between multiple behaviors can replicate the entire data set; and do so in an interactive environment that allows for experimentation.

The agent-based model can be verified by executing various scenarios iteratively, demonstrating that the parameter values stay within the expected confines and collecting statistical data on these experimental runs—the same category of values which were gathered to create the initial equation-based model. Then, statistical analysis is performed on this experimental data (box "D") to extract summary statistics (box "E") and these are then compared with the statistics derived from the original equation-based model (box "F"). If the two analyses agree, then the agent-based model has been verified. The fact that we can perform this validation is the reason that the work described here has been based on an existing model. Doing this grounds our agent-based model in reality (since the model we check it against was derived from census data), and gives us confidence that the results we obtain that go beyond mere validation are reasonable.

3 A model of human capital

The model that we consider in this paper is drawn from a paper by Kremer [6], an article that derives a linear model of the change in human capital from US census data, and analyzes the aggregate behavior of the model. The original model was derived to identify the effect of the tendency for human societies to stratify by level of education—so-called *human capital*. The reason that the model is important in our wider work on modeling aspects of the education system is that it provides a mechanism, derived from data and verified against that data in [6], by which agents choose a level of education to attain. It can therefore act as a driver for the models we have previously developed [10, 11].

The model from [6] gives the level of human capital $z_{i,t+1}$ of members of the $t+1$th generation of the ith dynasty as being:

$$z_{i,t+1} = k_{t+1} + \alpha \left(\frac{z_{i,t} + z'_{i,t}}{2} \right) + \beta \left(\frac{\sum_{j=1}^{n} z_{j,t}}{n} \right) + \epsilon_{i,t+1} \tag{1}$$

The notion of "dynasty" and "generation" that we use here are based on the definitions in [6]. Each generation of the ith dynasty has two children, one male and one female. Each is assumed to then become the spouse of an opposite sex member of another dynasty, forming a family which in turn produces one male and one female child. One family from a given generation of the ith dynasty remains in the ith dynasty, and one becomes part of another dynasty (the dynasty of the corresponding non-ith partner). Thus there is a constant number of members of each generation, and of each dynasty at each generation.

Breaking down the rather simple linear model from (1) we have:

$$k_{t+1} \tag{2}$$

which is constant across dynasties, but may vary in time to capture exogenous trends in education—for example legislation that requires a certain number of years of additional schooling for given generations. This represents the basic level of education that every individual has to undergo ("education" and "human capital" are used more or less interchangeably in this model). Kremer [6] gives $k_{t+1} = 6.815$, and that constant value is what we adopt.

$$\alpha \left(\frac{z_{i,t} + z'_{i,t}}{2} \right) \tag{3}$$

1. Establish level of z based on:
   (a) Parents
   (b) Neighbors of parents
2. Establish factors that influence z for children
   (a) Spouse
   (b) Neighbors
3. Generate children

Table 1. The basic agent lifecycle.

measures the effect on the level of education of the $t + 1$th generation of the education of its parents in the t-th generation. The effect of the term is to assign to each child the average human capital of its parents, modified by α. Kremer [6] computes a baseline value of α to be approximately 0.39, based on census data. $z_{i,t}$ is clearly the human capital of a member of the previous generation of the ith dynasty, and $z'_{i,t}$ is the spouse of $z_{i,t}$.

The next term:

$$\beta\left(\frac{\sum_{j=1}^{n} z_{j,t}}{n}\right) \tag{4}$$

does something similar to (3) but based upon the level of education of the parents' neighbors rather than the level of education of the parents themselves—these are the j in the summation. Kremer [6] measures the baseline value of β to be around 0.15.

The final term in (1) is

$$\epsilon_{i,t+1} \tag{5}$$

which captures a specific "shock" to the human capital in a specific generation of a specific dynasty—for example the early death of a parent, requiring the children to curtail their education (though this value can be positive as well as negative). Once again we, follow [6] in picking $\epsilon_{i,t+1}$ from a normal distribution with mean 0 and standard deviation 1.79.

4 Agent-based simulation

We have developed an agent-based model that is based on the equation-based model given above. The agent-based model is concerned with a fixed number of agents, n in each generation, with $n/2$ dynasties, and 2 children per family. For simplicity, each family has one male child and one female child. The basic simulation loop, which executes once for each generation, has three steps given in Table 1. The result of Step 1 is fixed by (1), and Step 3 is fixed by the requirement to produce one male and one female child in each generation. Clearly the results are going to depend on the way in which Step 2 is implemented, and our model includes a number of variations.

The core of [6] is to determine, or measure, the extent to which *sorting* (that is the tendency for people to choose both spouse and neighbors with similar levels

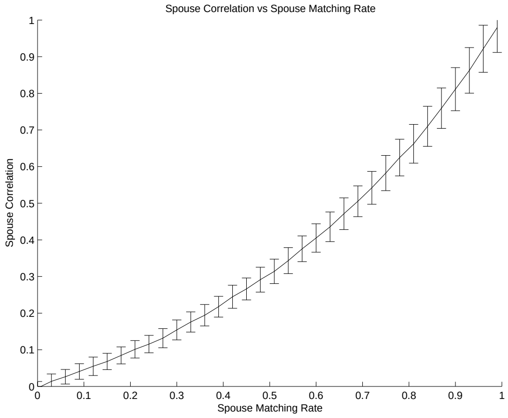

Fig. 2. The relationship between p_s and correlation between spouse agents' human capital values.

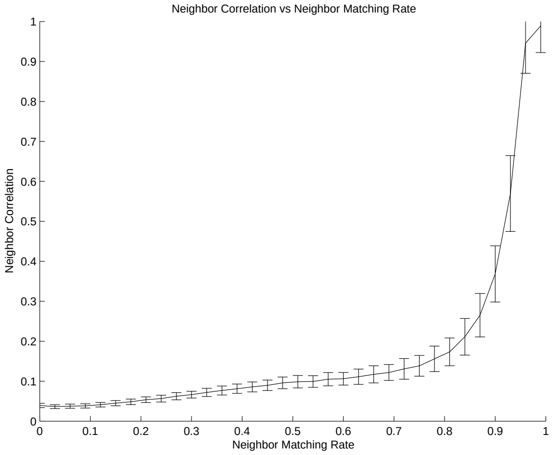

Fig. 3. The relationship between q_s and correlation between neighbors human capital values.

of human capital) affects divergence in human capital between given dynasties as generations proceed. The agent-based model includes two mechanisms by which this sorting can mimic these choices. Choice of spouse and choice of neighbor. For choice of spouse, there are three models that an agent can employ:

No sorting: Agents pick partners at random.

Sorting: An agent with human capital z attempts to pick a partner with a human capital value in $[0.9z, 1.1z]$. If there are no such agents that are unmarried, the original agent picks the eligible agent with the highest human capital.

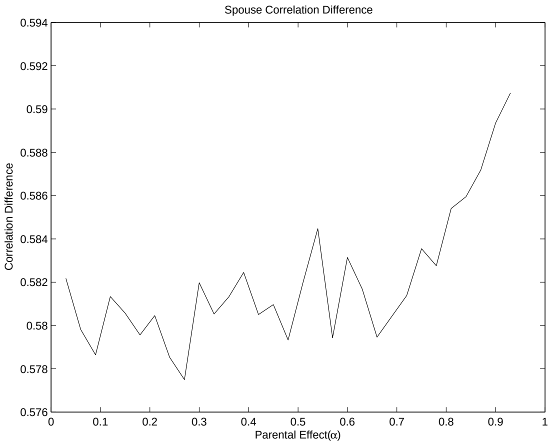

Fig. 4. The relationship between the parental effect α and relative change in the standard deviation of the human capital distribution when sorting is increased

- **Max-matching:** Agents pick as their partner the agent with the closest human capital value.

In our experiments we need to be able to manipulate the correlation between married agents' human capital values. We achieve this by setting the probability p_s that a given agent uses a sorting method to choose a spouse. If $p_s = 0$, then, all agents will pick a partner at random. If $p_s = 1$, then every agent will use one of the sorting methods to pick a spouse. Figure 2 shows how varying p_s changes the correlation between spousal human capital. As elsewhere in this paper, the error bars indicate one standard deviation above and below the mean value. Here, and throughout the paper, the choice the agent makes with p_s is between no sorting and max-matching.

Given that the model in [6] is based upon census data, and that this has built into it a geographic notion of neighborhood, that is the kind of neighborhood used in the agent-based model[3]. Each dynasty has a unique location. Initial positions for dynasties are picked randomly, and as each generation goes through Step 2(a), the female child stays in the dynastic location, and the male child "moves" to the position of the spouse. The dynastic location is allowed to change between generations, modeling "sorting" between neighborhoods. Again we have three possibilities:

- **No sorting:** Step 2(b) involves no operation—dynasties do not move relative to one another.
- **Sorting:** Step 2(b) allows the families established in Step 2(a) to move to the neighborhood with the highest human capital value that has room.

[3] As opposed, for example, to a "social neighborhood" based on the acquaintances of the parents, which might not coincide with the geographical neighbors.

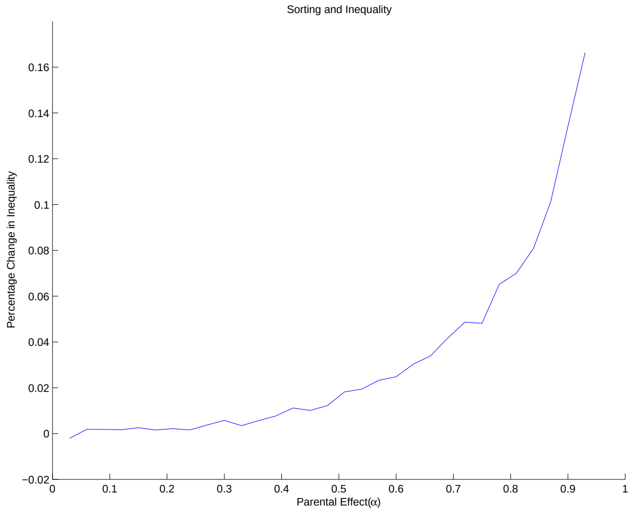

Fig. 5. The relationship between the parental effect α and the percentage change in inequality.

Max-matching: Dynasties move to the neighborhood that has the human capital value closest to the parental average and has room.

The value of a neighborhood is the average value of the human capital of the agents located in that neighborhood.

Again, we control the sorting effect probabilistically, with each dynasty having a probability q_s of moving at a given generation. $q_s = 1$ means that all dynasties will move, and $q_s = 0$ means no dynasty will move. This probability, just like p_s, can be used to manipulate the correlation between the human capital of neighbors, and this relationship is plotted in Figure 3. For all the experiments in this paper, q_s chooses between no sorting and max-matching.

The impact of these different sorting policies will clearly depend on the nature of neighborhoods. We incorporated two types of neighborhood in the model:

1. **Moore neighborhood:** The neighborhood for each dynasty is the set of locations directly around that dynasty—hence each dynasty has its own neighborhood, and these neighborhoods overlap.
2. **Fixed neighborhood:** The whole area we simulate is carved up into fixed neighborhoods, so several dynasties share the same neighborhood, and neighborhoods do not overlap.

For the experiments described in this paper, we only used fixed neighborhoods.

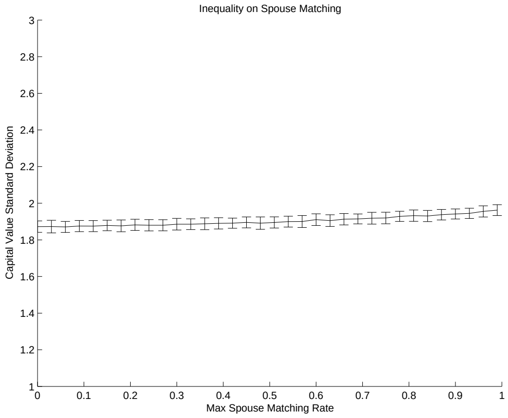

Fig. 6. The relationship between p_s, the probability of agents picking partners based on capital value, and the percentage change in inequality.

5 Experiments

We implemented the model described in the previous section in REPAST [9], a Java-based Swarm-like [13] tool developed at the University of Chicago for agent-based modeling in social science applications. We handled the geographic aspects by placing agents on an $n \times n$ grid, where at most one dynasty "lives" in a single grid-square. By varying the size of the grid and number of agents we can create environments of differing population density and have modeled communities of up to 10,000 dynasties.

5.1 Verification

Having constructed an agent-based model of human capital from the equation-based model in [6], we first need to "complete the loop" (as in Figure 1) by performing a statistical analysis of the results from the agent-based model, obtained when using the parameter values assumed in the paper, to show that our agent-based model will achieve the same results as the equation-based model we started with. This verification step is needed in order to justify the further experimental results with the model.

The central result of [6], and the only quantitative result from [6] that we can use to check the model against, is the prediction that increasing sorting—which the paper takes to mean increasing the correlation between the human capital values of the parent agents of a generation—will only cause an increase in inequality—which the paper takes to mean that the standard deviation of the human capital distribution grows generation by generation—when the value of α is large. [6] demonstrates this by showing the effect of changing correlation from 0.6 to 0.8 for various values of α. This result can be established though a steady-state analysis of (1), and this is done in full detail in [6]. Since the latter

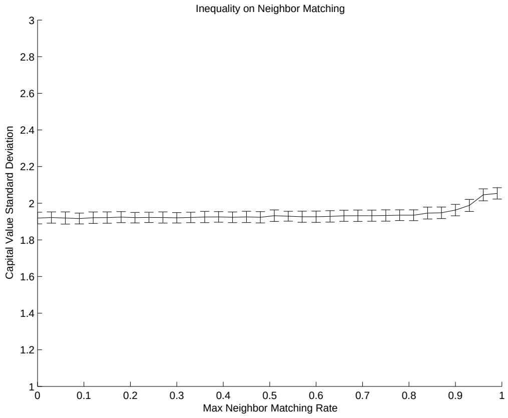

Fig. 7. The relationship between q_s, the probability of agents picking location based on capital value, and the percentage change in inequality.

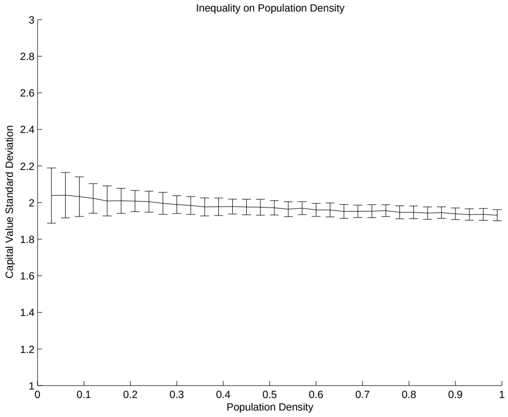

Fig. 8. The relationship population density and the percentage change in inequality.

paper is based on census data, we take this as the experimentally determined truth against which we compare the predictions of our agent-based model.

Our agent-based model does not give us direct control of the correlations, but as we have already shown, we can, rather imprecisely, change the value of the correlations by changing the value of p_s. Running experiments on a 50×50 grid—which allows us to deal with a population that is considerably larger than the 1500 individuals analyzed in [6]—we find that our model gives good agreement with the predictions made in [6].

First, we plot the value of α against the change in the standard deviation of the human capital distribution (expressed as a fraction of the standard deviation) caused by switching from $p_s = 0.75$ (which is a correlation between parental capital of 0.6) to $p_s = 0.88$ (a correlation between parental capital of 0.8). This

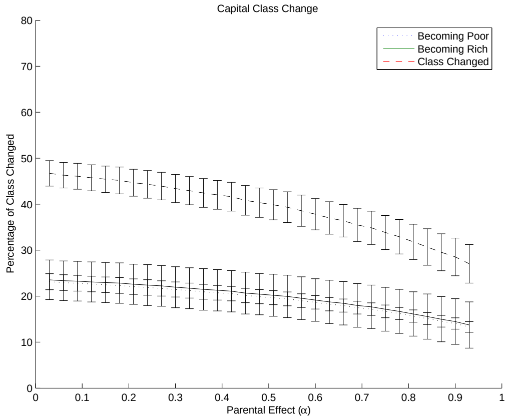

Fig. 9. The relationship between the percentage of dynasties that change "class" and α.

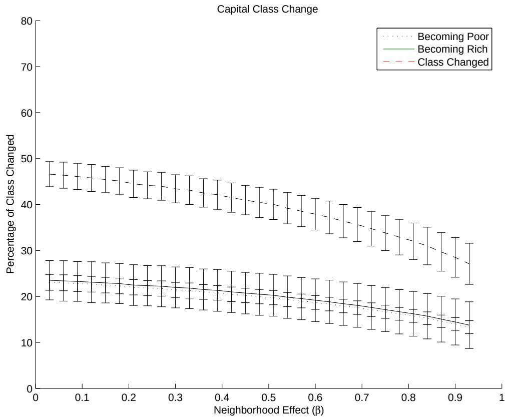

Fig. 10. The relationship between the percentage of dynasties that change "class" and β.

gives us Figure 4, which shows that the increase in standard deviation of the human capital distribution, and hence inequality, that is caused by increased sorting doesn't start to grow until α exceeds 0.8. We can also plot the effects in terms of the percentage change in inequality (as defined in [6]) rather than the increase in standard deviation of the human capital distribution. For $p_s = 0.88$, we get the relationship between α and inequality plotted in Figure 5.

To check that this change in inequality was really due to the change in α, and not due to some other parameter in the model, we examined how inequality changes when we vary such parameters. Figures 6–8, for example, show that for α held at 0.39 and β held at 0.15, there is no significant change in inequality if we change p_s, q_s and population density.

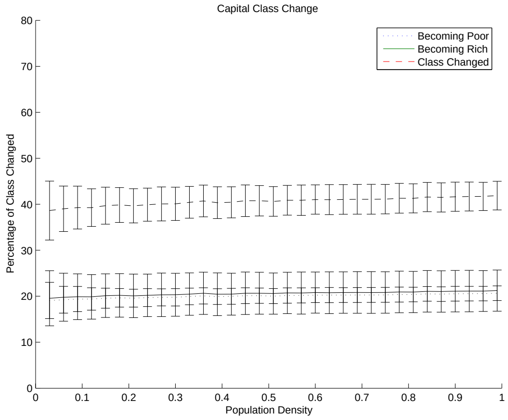

Fig. 11. The relationship between the percentage of dynasties that change "class" and population density.

Note that the changes in inequality that we observe due to changes in α hinge on the value of $\epsilon_{i,t+1}$, the term in (1) that does not depend on the capital values of parents or neighbors. If we run our model with $\epsilon_{i,t+1}$ set to zero for all dynasties and all generations, then inequality does not grow. Indeed, the standard deviation of the capital distribution falls over time until all agents have the mean value. This "seeding" effect of $\epsilon_{i,t+1}$ is another prediction that can be made from the analysis of (1).

Together these results—where statistics that can be extracted from the original, equation-based, model match against the predictions made by the agent-based model—suggest that the agent-based model we have constructed adequately replicates the essence of the model it was designed to capture.

5.2 Identifying new features

As we discussed above, one of the advantages that agent-based models have over equation-based models is that one can examine the model in greater detail. Whereas equation-based models can only really be studied in terms of broad statistical features—such as the results from [6] examined above—we can probe agent-based models in considerable detail, discovering what happens to individuals as well as to classes of individual. We have carried out such an investigation into the human capital model.

The main result from [6], replicated by our agent-based model, is that *on average* inequality in terms of human capital grows over generations. The widening standard deviation of the human capital distribution suggests that rich dynasties get richer and poor dynasties get poorer. However true this may be at a population level, it is interesting to ask whether it is true for all (or even most) individual dynasties, or whether there is some mobility between dynasties with different levels of human capital. It turns out that such mobility exists.

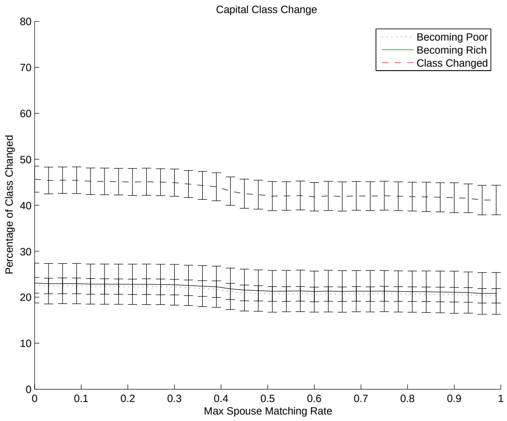

Fig. 12. The relationship between the percentage of dynasties that change "class" and p_s, the probability that a given agent chooses a partner by human capital value.

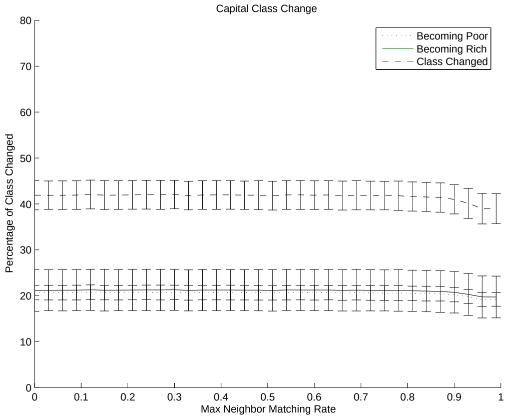

Fig. 13. The relationship between the percentage of dynasties that change "class" and q_s, the probability that a given dynasty chooses its location by human capital value.

We divided our dynasties up into three "classes"—the quotes reminding us that this terminology, while convenient, conflates human capital, basically years of formal schooling, with monetary capital and social status. We call dynasties that fall within one standard deviation above or below the average human capital for the population *middle class*, we call those more than one standard deviation below average *poor*, and those more than one standard deviation above average *rich*. We then examined whether dynasties moved between classes.

The results are given in Figures 9 and 10, which show the way that the number of dynasties that are mobile in this sense changes for two different values of α and β, respectively. When α changes, β is held constant and vice-versa. These graphs show the total percentage of dynasties that move, and the percentage that become richer and poorer. They show that, no matter what the value of α and β, there is some mobility (at least 25% of the population, and as much as 45% of the population changes class). Furthermore this change is symmetrical.

Note that this effect is separate from the growing inequality—because "middle class" is always defined in terms of the *current* standard deviation, if inequality was the only effect, the percentage of dynasties changing class would be lower than the figure we find. What we see here is the result of mixing, that is, individuals choosing partners or neighbors who are sufficiently far above or below them in human capital terms that their offspring move from one class to another.

We can follow up this investigation with a subsidiary one, checking to see whether additional factors have an effect on the class mobility of dynasties. One of the factors that we can imagine having an impact on the results we obtain in the model is the *density* of the agent population. In terms of the model, population density relates to the number of agents that are placed on the grid. Since the neighbor effect is based upon a geographic notion of neighborhood, and since neighbors certainly have an effect on class mobility, for example (as shown in Figure 10), then one might imagine that changing the density of the population might have some effect on class mobility as well. However, this is not the case. As Figure 11 shows, population density has no systematic effect on class mobility. Carrying out similar investigations for the effects of p_s and q_s, Figures 12 and 13 respectively, again show no systematic effect on class mobility.

6 Summary

This paper set out to construct an agent-based model from a traditional, equation-based, model, and to show that (i) this model could be verified against the predictions make by the equation-based model; and this model could identify new predictions that could not be obtained directly from the equation-based model. Both these aims have been achieved.

This work fits into our wider effort to model aspects of the education system [10, 11], with the overall aim of being able to establish the impact of, changes in education policy (rather as [1] does for the case of rent control). As described in [11], we have developed a number of models, including a model of interactions in classrooms [10]—which, for example, shows the effects of different pedagogical techniques to overcome absenteeism—and a model of school districts—which, for example, shows the effect of policies like "No child left behind". Our current work is to tie these models together, and, more ambitiously, to tie them into a comprehensive simulation of the way that education fits into the economy. This latter can be done, for example, by using the model in [7], a model that relates education and student ability with their lifetime productivity.

Acknowledgments This work was partially supported by funding from NSF #REC-02-19347. Many thanks to the anonymous referees for their helpful comments.

References

1. R. N. Bernard. Using adaptive agent-based simulation models to assist planners in policy development: The case of rent control. Working Paper 99-07-052, Sante Fe Institute, 1999.
2. E. Bonabeau. Agent-based modelling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Science*, 99(3):7280–7287, May 2002.
3. A. Greenwald and J. Kephart. Shopbots and pricebots. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 506–511, Stockholm, Sweden, August 1999.
4. V. Grimm. Ten years of individual-based modeling in ecology: what have we learned and what could we learn in the future? *Ecological Modeling*, pages 129–148, 1999.
5. J. Kephart and A. Greenwald. Shopbot economics. *Autonomous Agents and Multi-Agent Systems*, 5(3):255–287, 2002.
6. M. Kremer. How much does sorting increase inequality? *The Quarterly Journal of Economics*, 112(1):115–139, Feb. 1997.
7. J. Laitner. Earnings within educational groups and overall productivity growth. *The Journal of Political Economy*, 108(4):807–832, August 2000.
8. H. V. D. Parunak, R. Savit, and R. L. Riolo. Agent-based modeling vs. equation-based modeling: A case study and users' guide. In *Proceedings of the Workshop on Multiagent-based Simulation*, pages 10–25, 1998.
9. `http://repast.sourceforge.net`.
10. E. Sklar and M. Davies. Multiagent Simulation of Learning Environments. In *Proceedings of the Fourth International Conference on Autonomous Agents and MultiAgent Systems (AAMAS-2005)*, 2005.
11. E. Sklar, M. Davies, and M. S. T. Co. SimEd: Simulating Education as a MultiAgent System. In *Proceedings of the Third International Conference on Autonomous Agents and MultiAgent Systems (AAMAS-2004)*, pages 998–1005, 2004.
12. R. V. Sole, J. G. P. Gamarra, M. Ginovart, and D. Lopez. Controlling chaos in ecology: From deterministic to individual-based models. *Bulletin of Mathematical Biology*, 61:1187–1207, 1999.
13. `http://www.swarm.org/`.
14. M. Wooldridge, N. R. Jennings, and D. Kinny. The gaia methodology for agent-oriented analysis and design. *Autonomous Agents and Multi-Agent Systems*, 3(3):285–312, 2000.