Aligning Simulation Models: A Case Study and Results

by

Robert Axtell, The Brookings Institution
Robert Axelrod, The University of Michigan*
Joshua M. Epstein, The Brookings Institution
and Michael D. Cohen, The University of Michigan

July 21, 1995

*To whom correspondence should be sent at School of Public Policy, University of Michigan, Ann Arbor MI 48109. email: Robert.Axelrod@umich.edu. Phone 313-763-0099. Fax: 313-763-9181.

The authors gratefully acknowledge financial assistance from The Brookings Institution, World Resources Institute, John D. and Catherine T. MacArthur Foundation, The Santa Fe Institute, The Program for the Study of Complex Systems, and the LS&A Enrichment Fund of the University of Michigan, and the U.S. Advanced Research Projects Administration.

Abstract

This paper develops the concepts and methods of a process we will call "alignment of computational models" or "docking" for short. Alignment is needed to determine whether two models can produce the same results, which in turn is the basis for critical experiments and for tests of whether one model can subsume another. We illustrate our concepts and methods using as a target a model of cultural transmission built by Axelrod. For comparison we use the Sugarscape model developed by Epstein and Axtell.

The two models differ in many ways and, to date, have been employed with quite different aims. The Axelrod model has been used principally for intensive experimentation with parameter variation, and includes only one mechanism. In contrast, the Sugarscape model has been used primarily to generate rich "artificial histories", scenarios that display stylized facts of interest, such as cultural differentiation driven by many different mechansims including resource availability, migration, trade, and combat.

The Sugarscape model was modified so as to reproduce the results of the Axelrod cultural model. Among the questions we address are: what does it mean for two models to be equivalent, how can different standards of equivalence be statistically evaluated, and how do subtle differences in model design affect the results? After attaining a "docking" of the two models, the richer set of mechanisms of the Sugarscape model is used to provide two experiments in sensitivity analysis for the cultural rule of Axelrod's model.

Our generally positive experience in this enterprise has suggested that it could be beneficial if alignment and equivalence testing were more widely practiced among computational modellers.

1. Introduction

1.1 Motivation

If computational modeling is to become a widely used tool in social science research, it is our belief that a process we will call "alignment of computational models" will be an essential activity. Without such a process of close comparison, computational modeling will never provide the clear sense of "domain of validity" that typically can be obtained for mathematized theories. It seems fundamental to us to be able to determine whether two models claiming to deal with the same phenomenon can, or cannot, produce the same results.

Alignment is essential to support two hallmarks of cumulative disciplinary research: critical experiment and subsumption. If we cannot determine whether or not two models produce equivalent results in equivalent conditions, we cannot reject one model in favor of another that fits data better; nor are we able to say that one model is a special case of another more general one -- as we do when saying Einstein's treatment of gravity subsumes Newton's.

Although it seems clear that there should be frequent efforts to show pairs of computer models to be equivalent, we are aware of no reports of such alignment studies.

We have identified a few cases in which an older model has been reprogrammed in a new language, sometimes with extensions, by a later author. For example, Michael Prietula has reported[1] reimplementing a model from Cyert and March (1963) and Ray

---

[1] Personal communication to Michael Cohen.

Levitt has reported a reimplementation of Cohen, March and Olsen (1972).[2] However, these procedures are not comparisons of different models that bear on the same phenomena. Rather they are "reimplementations", where a later model is programmed from the outset to reproduce as closely as possible the behavior of an earlier model. Our interest is in the more general and troublesome case in which two models incorporating distinctive mechanisms bear on the same class of social phenomena. be it voting behavior, attitude formation, or organizational centralization.

This paper therefore aims to achieve two goals: 1) to report a novel set of results from aligning two different computer models of cultural transmission; and 2) to report an informative case study of the process used to obtain these novel results.

1.2 Overview

The paper is organized into six sections. After this brief introductory section, section 2 provides more detailed background on the two models necessary for understanding the results. The third section reports our procedures in aligning the two models and in collecting information for this case report. The fourth contains results from two comparison experiments. The fifth reports our observations on the model alignment process. The conclusion is the sixth section.

2. Background on the Two Models

Our objective has been to determine if a set of results obtained in a model of cultural transmission built by Robert Axelrod (1995), could also be obtained in the different setting

[2]Personal communication to Michael Cohen.

of the Sugarscape model of Joshua M. Epstein and Robert Axtell (1995).[3] Sugarscape differs from the Axelrod model in many ways. Most notably, culture is one of many processes that can be operative in the more general Sugarscape system, which has model agents who -- among other things -- move, eat, reproduce, fight, trade, and suffer disease. The Axelrod model has much simpler agents who do none of these things, but rather occupy fixed positions on a square plane, interacting only with their immediate neighbors to the North, South, East, and West.

The two models are in this respect clear examples of distinctively different approaches to computational modeling: Sugarscape is designed to study the interaction of many different plausible social mechanisms. It a kind of "artificial world" (Lane, 1993). In contrast, the Axelrod Culture Model (ACM) was built to implement a single mechanism for a single process, with the aim of carrying out extensive experiments varying parameters of that mechanism. It much more resembles the spirit of traditional mathematical theorizing in its commitment to extreme simplicity and complete analysis of each model parameter.

We begin by describing briefly how the Axelrod model works.[4] The model studies a square array of agents all following a single rule. The agents are cultural entities which

---

[3]Axelrod's source code is approximately 1500 lines of Pascal for the Macintosh (Synamtec THINK Pascal version 4.0.1) and is available from the author. The Sugarscape source code is approximately 20,000 lines of Object Pascal and C for the 68K Macintosh (Symantec THINK Pascal version 4.0.2 and THINK C version 7.0.6 compilers). Agent objects are written in Pascal while low-level and graphics routines are primarily written in C. This code is available from Robert Axtell Executable versions of the code, configured with Axelrod's culture rule and capable of generating the data in this paper, are also available from Axtell.

[4]A complete account of the model structure and of results obtained from experiments can be found in Axelrod (1995).

might be thought of as 'villages'. Each agent interacts with a fixed set of neighbors, four unless the agent is located on an edge or corner of the square. Each agent has several attributes and each of those attributes can take any one of several nominal-scale values. For the work reported here we have used five attributes, each taking one of fifteen values. The initial state of each agent is determined by randomizing the value of each attribute. Attributes might be interpreted as forms of dress, linguistic patterns, religious practices, or other culturally determined features.

The central aim of the ACM is to study the effects of a simple mechanism of cultural transmission that operates as follows. An agent is selected at random to be the next one active. One of the four neighbors is selected to be that agent's next contact. An attribute is selected among the five. If the two agents have the same value for that attribute, another attribute on which they differ is selected at random, if there are any, and the active agent assumes the value for that attribute currently held by the contacted neighbor. Activity is allowed to continue until every agent differs from each of its neighbors either at every attribute or at none. At this point no further change is possible and the model run stops.

A key feature of this cultural change mechanism is that cultural change becomes more likely as two neighbors are more alike, and less likely as they differ. A central question of interest in the work with the model is whether this variability of interaction rate is itself sufficient to create stable diversity rather than eventual homogeneity -- as one would expect with a model that allowed unlike neighbors to continue interacting no matter how different they were.

While the ACM can be conveyed in a few paragraphs, and its results can be fully described in a short article, Sugarscape is a much more complex system that can be rendered fully only at book length (Epstein and Axtell, 1995). This is not because individual mechanisms of Sugarscape are complex. On the contrary, each of its mechanisms are of about the same complexity as in the Axelrod model. However, the intent of Sugarscape is to investigate the interplay of many mechanisms as they operate

simultaneously -- as happens in actual social life. In particular, Sugarscape is intended as a tool in sufficiency testing of social theories, allowing theorists to ask if a stipulated set of mechanisms and conditions (say for a market to "clear") actually will produce the predicted phenomenon.

Sugarscape therefore has processes that allow its agents to look for, move to, and eat a resource ("sugar") which grows on its toroidal array of cells. Thus while food growing cells are immobile, active agents are purposively mobile, and this is one of many fundamental differences with the Axelrod model.

Sugarscape agents also have cultural attributes. In typical studies with the model there are eleven cultural attributes, each of which takes one of two values. Cultural attributes change in Sugarscape as part of a larger cycle of agent activity.

In this model, the agents also become active in random order.[5] Each agent, when active, engages in a number of processes. For the present discussion, the most important of these is moving to a cell within its vision range that is richest in sugar. At that location an agent interacts culturally with all its neighbors. (Sugarscape agents typically do not populate all the landscape cells, so the active agent may have fewer than four other agents in its neighborhood.) In a cultural interaction, an attribute is selected at random, and if the neighbor differs from the agent the value is changed to that of the agent.

In Sugarscape, attributes are aggregated (typically by a simple majority rule) and this determines an agent's cultural type, usually labeled as either "Red" or "Blue". Cultural type then enters into many other processes in which Sugarscape agents may engage, such as trade, combat, and sexual reproduction.

Whereas the Axelrod model was designed principally for intensive experimentation with parameter variation, the intended use of the Sugarscape model is quite different in design. In it agents have many behavioral rules in addition to cultural ones, and while the

---

[5]The method is similar, but not identical, to that in ACM, as discussed below.

model may be used for exploration of parameter spaces, it has heretofore been primarily used to generate "artificial histories", scenarios that display stylized facts of interest, such as cultural differentiation driven by resource availability, or recognizable patterns of migration, trade, and combat. The principal use of the generated scenarios is for sufficiency tests, showing that the implemented individual-level mechanisms are able to produce the collective-level phenomena of interest.

It should be apparent that the two models are vastly different in many important respects. Nonetheless, they have two central features in common that suggest that they could be meaningfully compared. The first is that both are "agent-based" models. They work by specifying properties of individual actors in the system and are concerned to study the collective phenomena that result as those individuals interact --in this case in local neighborhoods of two-dimensional space. The second shared feature is that both represent cultural attributes of individual agents as strings of symbols and model cultural diffusion as a convergence process between neighbors.

3. Procedures of Our Comparison

These two strong similarities suggested to Axelrod and Cohen, as they read a draft account of the Sugarscape project, that it might be possible to "dock" the two models -- in analogy to orbital docking of dissimilar spacecraft. Thus it could be determined whether Sugarscape, under suitable conditions, would produce results equivalent to those already obtained for the ACM. Epstein and Axtell were contacted. They agreed such a test would be instructive. All four investigators believe that alignment of models will be necessary if computational modeling is to become a significant medium of theoretical expression in the social sciences. None of the four could think of a case where such an equivalence test had been reported.

3.1 Making the Comparison and Preparing the Case Report

The four investigators agreed on procedures for conducting the test, and for keeping records of the work done and problems encountered in the course of the testing. The aims were: 1) to determine if equivalent results were produced in equivalent conditions; 2) to demonstrate the effects of relaxing some of the equivalent conditions; and 3) to be able to report problems that occurred and their resolutions, thus taking first steps in establishing the practice of equivalence testing more generally in social science computational modeling.

The procedures followed were roughly analogous to those used when a second investigator in a laboratory science is attempting to reproduce results obtained in a first investigator's laboratory (Latour and Woolgar , 1979).

Epstein and Axtell worked with a pre-publication draft account of the ACM to do their preliminary work. They considered what steps they would have to take in order to reproduce the key results identified by Axelrod. These results show how the number of culturally identical regions that exist when stability is reached varies as a function of three parameters: the number of attributes, the number of values per attribute, and the size of the square lattice. These results included the most surprising aspect of ACM's performance: that the equilibrium number of cultural "regions" produced by the model first increases, then decreases as a function of the number of agents.[6]

Axtell and Epstein then visited Axelrod and Cohen at the University of Michigan, where a conference clarified ambiguities. Further changes were to be made to Sugarscape, and then preliminary equivalence tests run. A fuller set of tests was run and analyzed when Epstein and Axtell returned to their work site at the Brookings Institution. Epstein and Axtell then continued by relaxing some of the factors that had been made equivalent to those of ACM, in order to see what differences such changes would make.

---

[6]Axelrod defined a cultural region as a set of a contiguous sites with identical culture.

3.2 Testing Model Equivalence

A central issue was the determination of how to assess "equivalence" of the two models. The plan required an effort to show that the Sugarscape model could behave comparably to the ACM, and this entails a standard by which to assess "equivalence" of measures made on the two models. This was discussed on the telephone and via email at an early stage. The conclusion was that for this case it would suffice if Sugarscape could be shown --when using a basic cultural transmission mechanism similar to the ACM's -- to produce several distributions of measurements that were statistically indistinguishable from distributions produced by the ACM.

The four investigators agreed that this is a rather tight standard, since one might argue that Sugarscape was equivalent if it produced a set of results with the same ordinal patterns as those from the ACM. But a demanding test was felt to be appropriate since this was a first exercise of its kind, and since programming changes to Sugarscape could make its basic cultural transmission mechanism algorithmically equivalent to that in ACM. All the authors agreed that "equivalence" of models with stochastic elements must be defined in context, and further observations on this central and thorny issue are offered in the conclusions section. In particular, we expand there on the difficult problem of giving a precise statistical content to the concept "statistically indistinguishable distributions."

4. Results From the Two Experiments

We turn now to reporting our observations on the behavior of Sugarscape in comparison with that of ACM. We describe the changes made to Sugarscape in order to bring it into alignment with what were judged to be principal features determining ACM's results.

4.1 Changes Made to Dock Sugarscape with ACM

Vision range was reduced to the immediate four neighbors. Movement range was reduced to zero. The usual initialization of Sugarscape to a population sparsely distributed over its array of cells was altered to a distribution placing an agent on every cell. The toroidal topology of Sugarscape was altered to a bounded square. There was actually a discrepancy introduced in doing this, which we comment on below. The constant numbers of attributes and values per attribute in Sugarscape were made into variables that could be set to the three different levels used in the ACM runs shown in our Table 1.

One difference was deemed small and not eliminated. Sugarscape activates agents one at a time from a random permutation of the list of agents. When the list is finished, it is repermuted and activation begins again. Axelrod, as mentioned, activates a new randomly chosen agent every time. Roughly the methods correspond to sampling agents for activation without and with replacement. Thus for any given set of n agents, in Sugarscape a block of n activations will make each agent active exactly once, while in the Axelrod model most would be active once, but a few might be active either zero times or two or more times. Our decision not to eliminate this difference, small though it seemed, did have interesting consequences which we describe below.

We had decided that to reproduce Axelrod's results Epstein and Axtell should first try using exactly his rules for determining cultural change. They therefore programmed a substitute for their own cultural change rule, which took no account of inter-agent similarity in the diffusion of culture attributes among interacting neighbors, and which caused each agent to interact culturally with all its neighbors.

4.2 Sugarscape Reproduces Central Results of Axelrod's Culture Model

Table 1a, with target data from Axelrod (1995), gives the number of stable cultural regions for a 10 x 10 lattice, averaged over ten runs, as a function of the number of cultural attributes and the values per attribute. Note that, other things being equal, the number of

cultural regions present in equilibrium increases with the number of traits per feature and decreases with the number of cultural features. Of the 9 tabulated values, only four are not equal to 1.0.

A directly analogous display, Table 1b, has been generated with the Sugarscape implementation of the Axelrod cultural rule. The qualitative dependence of the number of stable cultural regions on the number of features and traits per feature is the same as in Axelrod's table. Notice that in this new table only three entries are not equal to 1.0.

Table 1 here.

Quantitative agreement between the two sets of data is clear for the five entries of 1.0 that the tables have in common. To test how well the remaining entries in the two tables agree quantitatively, non-parametric statistical comparisons were undertaken. The critical value of the two-sided Mann-Whitney U statistic at the 0.05 level of significance for samples of size 10 is 23 (Siegel, 1956). That is, one rejects the null hypothesis for a value of U at or below 23. For all comparisons between the two tables the U-statistics are greater than the critical value and thus one cannot reject the null hypothesis on nonparametric grounds. Overall, it seems very likely that the corresponding data in the two tables were drawn from the same distribution.

Figure 1 gives the target data from Axelrod (1995) on the number of stable cultural regions as a function of the lattice size for five cultural features with fifteen traits per feature. This figure has an interesting non-monotonic shape, a result discussed at some length by the author. Data for the 5 x 5, 10 x 10 and 20 x 20 lattices have been generated using the Sugarscape implementation of the Axelrod cultural rule. In each case, the sample size was 40, the same sample size used by Axelrod for these three cases. The means from the modified Sugarscape model and corresponding error bars are also displayed on Figure 1. To determine to what extent this data agrees with Axelrod's original data, we employed

the Kolmogorov-Smirnov (K-S) test of the goodness-of-fit of empirical cumulative distribution functions (c.f., Hoel, 1962) - a nonparametric test.[7]

Figure 1 here.

The null hypothesis is that the corresponding data points are drawn from the same distribution. At the 5% significance level, the two-tailed critical value of the K-S statistic with forty observations is 0.304. That is, if the actual K-S value exceeds this critical value then the null hypotheses is rejected.

For the two sets of data corresponding to the 5 x 5 lattice the K-S statistic is 0.225. Therefore the null hypothesis cannot be rejected. In the 10 x 10 case the K-S statistic is 0.175, and so once again the null hypothesis cannot be rejected. Finally the 20x20 lattice size reveals that the K-S statistic is $0.5 > 0.304$ and thus the null hypothesis is rejected - the data for this parameter value appear likely drawn from different distributions. The ACM mean in this case was 16.25. The modified Sugarscape mean is 9.23.

1) In what sense may the computational models still be called "equivalent"? The modified Sugarscape model produces results that are numerically identical to those from ACM in some cases. It produces distributions of results that cannot be distinguished statistically from ACM distributions in eleven of the twelve comparions. In one case it produces a distribution that can be distinguished, although the mean is in the desired relationship to the other means. That is, the 20x20 lattice has a mean number of regions less than the 10 x 10 case. This non-monotonicity was the important character of the result in Axelrod's view of his own results. In our conclusion we argue that these are three natural categories of model equivalence, which we call 'numerical identity', 'distributional

---

[7]The Kolmogorov-Smirnov test was not used for the comparisons in Table 1 because it has low power for small sample sizes.

equivalence', and 'relational equivalence'. We discuss implications of these distinctions in Section 6.

2) What is the likely cause of the observed difference? Because we had brought so many aspects of the two models into algorithmic agreement, we were surprised when this discrepancy occurred. But not all aspects of the two models agreed, and the statistically significant difference indicates that this mattered in the 20x20 case. We believe the difference arises from our decision not to convert the Sugarscape activation method to the ACM method. The Sugarscape method does not allow for the same agent to be occasionally active several times before other agents have had their "fair" share of influence. This additional uniformity of influence appears to be sufficient to induce greater ultimate convergence in cultures.[8] When we convert the activation code in Sugarscape to the "sampling with replacement" method of ACM, 20x20 case no longer causes a problem. And when all the cases are rerun in Sugarscape with random activation, each one of them gives data that are indistinguishable from the ACM.[9]

---

[8]A possibly related result is obtained below in Section 4.3, where it is shown that allowing agents to mix with non-neighbors also reduces the eventual equilibrium number of cultures.

[9]The Sugarscape with random activation gives means and non-zero standard deviations as follows. For table 1, reading across 1.2 ± 0.4, 4.10 ± 1.3, 18.8 ± 9.7, 1.0, 1.0, 1.9 ± 1.0, 1.0, 1.0, 1.0. The Mann-Whitney U statistics for these 9 sets of data of ten data points each do not reject the null hypothesis that these data are drawn from the same underlying distributions as Table 1 a, at the 0.05 level of significance. For Figure 1, the data are 9.83 ± 2.75 (for 5x5 case) , 20.40 ± 7.93 (for 10x10 case), and 14.80 ± 7.01 (for 20x20 case). The Kolmogorov-Smirnov statistics for these 3 data sets of 40 points each do not reject the hypothesis that these data have the same distributions as the corresponding distributions from Axelrod's Culture Model shown in Figure 1.

3) What is the correct null hypothesis for statistical testing of equivalence? We have conformed in our statistical testing to the usual logic that formulates the problem as rejection of a null hypothesis of distributional identity. But the alert reader may have noticed that this is not entirely satisfactory in the special circumstances of testing model equivalence. With one exception discussed earlier, we have concluded that we cannot reject, at conventional confidence probabilities, the null hypothesis that the distributions are the same.

The unsatisfactory aspect of this approach is that creates an incentive for investigators to test equivalence with small sample sizes. The smaller sample, the higher the threshold for rejecting the null hypothesis, and therefore the greater the chance of establishing equivalence. We have resisted this temptation and used sample sizes typical of simulation studies. We feel satisfied in this case that, with the one exception noted, the models behave equivalently. In the long run, however, we see a need to formulate a more appropriate statistical logic.[10]

---

[10]Our current view of the most promising direction is to reverse the usual null hypothesis formulation and ask whether we can confidently reject the claim that the distributions are different. However, there are two complications in this approach. First, with stochastic models it will be extremely hard to conclude that all the observed difference in sample means is due to sampling fluctuation. This suggests that it will be necessary to use a null hypothesis such as "the two distributions differ by no more than X percent", with X chosen by convention or to be appropriate within the referent context. Second, with such a reversed and non-simple null hypothesis, and with no solid reason to assume a convenient (e.g. Gaussian) form of the underlying distributions, it is unlikely that there will be a manageable analytic method of obtaining confidence levels for the statistics. This suggests that the problem will have to attacked with computational statistical tools, such as the bootstrap approach of Efron and Tibshirani (1993).

4.3 Sensitivity Analysis of Agent-Based Models

The literature on sensitivity analysis in agent-based models is, as yet, quite small. How do alterations in local rules affect emergent macroscopic structures, such as cultural patterns? Dockings of the sort we have reported facilitate this new kind of sensitivity analysis. Here, we conduct two experiments involving agent movement rules.

4.3.1 A Mobility Experiment

As noted earlier, "agents" in the ACM occupy fixed positions on a square lattice, while in Sugarscape, agents are mobile. One natural question, therefore is: what happens to the equilibrium number of cultures in the ACM if agents are permitted to move around the Sugarscape interacting with neighbors, with interaction governed by the ACM cultural transmission rule? Will we see greater equilibrium cultural diversity or less? In the ACM, there is zero probability that non-neighboring agents will directly interact, while in Sugarscape, depending on the landscape topography, any two agents might eventually interact directly. Since the effect of movement is therefore to "mix" the population, we would expect that eventually there would be less diversity than without movement. This is what we find.

In order to carry out the experiment the Sugarscape was configured as a 50 x 50 grid having a single (Gaussian) "sugar mountain" in the center. One hundred mobile agents were given random initial locations on this landscape. Each agent engages in purposive behavior as follows: 1) it searches locally for the lattice location having the most sugar; 2) it moves to the nearest best site and 3) it gathers (eats) the sugar on that site. The agent population is heterogeneous with respect to its vision, i.e., how far each agent can "see" locally in the principal lattice directions (north, south, east and west). In these runs vision was uniformly distributed between 5 and 10 in the agent population. After moving

agents engage in cultural exchange--here according to Axelrod's cultural exchange rule--with one of their neighbors. One important difference between the Sugarscape and the ACM is that agents may have anywhere from 0 to 4 neighbors on the Sugarscape while (non-boundary) agents always have exactly 4 neighbors in the ACM. Once sugar is "harvested" by the agents it grows back at unit rate. The "termination criterion" employed had to be modified somewhat for this run. In the case of fixed agents, cultural transmission terminates when all neighboring agents are either completely identical or completely different. With mobile agents it is necessary to check whether all agents are either completely the same or completely different, independent of whether or not they are neighbors. This "global" stopping criterion is computationally more expensive than the "local" one appropriate for fixed agents.

Since we expected movement to reduce the number of cultures present it seemed natural to test it using the parameters which yielded the most cultures in the ACM. In the case of 100 agents (10 x 10 grid) having 5 cultural features and 15 traits/feature, the ACM produced an average of 18.5 distinct cultures, while the Sugarscape version of the ACM (fixed agents) yielded 20.4. The introduction of movement dramatically reduces the number of cultures. Over a sample of 10 runs the average was 1.1. When the experiment was repeated for the case of 5 features-30 traits/feature, under the expectation that this larger "cultural space" would yield more distinct cultures in equilibrium, the average number of cultures present increased somewhat to 2.2.

4.3.2 A "Soup" Experiment

Movement mixes the population. The extreme form of this is the so-called "soup", in which agents are paired at random regardless of location, and then interact under the ACM rule. Since this results in more thorough mixing than movement, we would expect the "culturally homogenizing" effect to be even stronger. And it was.

For 100 agents having 5 cultural features and 15 traits/feature, in 10 runs there was never a case in which more than 1 culture remained. When the number of traits/feature was increased to 30, a sequence of 10 runs yielded 7 runs which ended up with a single culture, 2 instances of 2 distinct cultures and a single case of 3 equilibrium cultures; an average of 1.4 overall. Essentially, most of the ACM's cultural diversity disappears in soup.
In summary, the more well-mixed the society, the lower is the equilibrium number of distinct cultures. Relatedly, multi-cultural equilibria in the ACM require that the probability of interaction between completely different agents be literally zero. If there is any probability of interaction (or if there is any point mutation rate) the long-run attractor is one culture. The above points concern the number of equilibria only; can we say anything about the rates at which these set in?

Recall the basic dynamic of the ACM: the greater the similarity between neighboring agents, the more rapidly does their similarity grow. Once similarity reaches a certain state, convergence is rapid--almost as if a phase transition occurs. Now, the counterintuitive result is that the more well-mixed the society, the later is this "phase transition." In the ACM model local clusters of neighboring agents develop similarities. Their high spatial correlation permits these agents to arrive at "agreement" very quickly, while in the Sugarscape mobility case, agents "hop away" before agreement is possible; and in the extreme soup, where spatial correlation is zero, the "phase change" is later still. In summary, the lower the spatial correlation the later is the onset of rapid convergence to equilibrium, and the lower is the equilibrium number of cultures.

5. Results on the Docking Process Itself

5.1 The Docking Process

When Epstein and Axtell visited the University of Michigan they brought their model with them on a portable computer. A portion of the work needed for the equivalence

testing was done prior to their arrival. This encompassed most of the changes described in Section 4.1.

Fortunately, Sugarscape was programmed in Object Pascal and with considerable attention to generality. It was therefore possible to make most of these changes as substitutions of parameter values or by "throwing switches".

On their arrival in Ann Arbor, a meeting was held to resolve several ambiguities that remained on the basis of reading Axelrod's text. We note an implication of this: under current standards of reporting a simulation model it will often not be possible to resolve all questions for an alignment exercise. Thus it will be necessary either to contact the author of the target model, to have access to the source code, or to have access to a documentation of the target model more complete than is generally provided in accounts published in contemporary journals.

The meeting also determined what steps were to be taken next. Axtell spent an evening doing additional programming. The next day it was possible to run a number of cases that would be needed to build a Sugarscape version of the Axelrod results.

Two months later while preparing to write up the results, Axtell realized that another change was necessary for the docking. Whereas the ACM altered the active agent when a cultural borrowing happened, the original Sugarscape model altered the agent's neighbor. This made a subtle difference because agents on the edge of the territory have fewer neighbors than those in the interior. To be sure that every site had the same chance to change, the ACM method is needed. When this was realized, Axtell made the necessary change to the Sugarscape implementation, and generated the data shown in Table 1 and Figure 1.

5.2 Total Time Required

The various tasks entailed in this docking exercise and the experimental extensions of the ACM are listed in the Appendices. These two appendices provide a description of the

specific tasks undertaken by Axelrod and Axtell respectively, along with the times required for each. All told, the work took about 23 hours for Axelrod and 37 hours for Axtell.[11]

5.3 Factors Making This Case Relatively Easy

There are at least four factors that can be identified as contributing to the relative ease with which the equivalence test was accomplished. First, the Sugarscape program was written from the outset with the objective of maximizing generality and ease of change. These goals are not especially easy to attain in practice, but object-oriented programming certainly did help.[12]

A second positive contributing factor is the extreme simplicity of the Axelrod model. This allowed the prose description to be essentially complete, and had ACM contained as many processes as Sugarscape it would have been considerably more difficult to bring the two so fully into alignment.

A third factor, was the recency of the ACM project. The statistical comparison required the full 210 points underlying the results reported in the original article.[13] These were relatively easy for the original investigator to provide, and this might not be so in other cases.

---

[11]These numbers do not include the time spent by all four participants in writing this report.

[12]It should also be said that Axtell, the lead programmer on Sugarscape, is a relatively skilled practitioner. He does not have experience producing commercial quality code, but his training did include doctoral level course work in computer science.

[13]There were 10 runs for each of the 9 cases in Table 1, and 40 runs from each of the 3 cases used for comparison from Figure 1.

A fourth factor, already mentioned briefly, is the underlying agreement of the two models on a basic, "agent-oriented" framework. In the absence of that, the architectures of the two models might have been so different as to make the project inconceivable.

5.4 Factors Making This Case Relatively Hard

On the other side of the ledger, there are several factors in the situation of this case study that probably made the exercise more difficult than future cases might be. Foremost among these is the probability that in the future models may be built with a prospect of equivalence testing clearly in view. ACM did not exist when Sugarscape was designed. Thus demonstrating equivalence to the ACM was not among its design specifications. If it had been, the equivalence testing could have been simpler still.

Also, one can plausibly imagine that there may someday be a number of more standardized code modules available which are reused in successive modeling projects. Random number generators meet this criterion today, and more substantive model elements may do so in the future. This too could substantially decrease costs of equivalence testing.

Overall, we would say that we did not find it completely straightforward to align the two models. But we were able to accomplish it in the end. And while the difficulties we encountered in reconciling them may seem disquieting, we should recall that they are not without precedent. Differential and integral calculus produced different results in the hands of different investigators until the foundations were solidified in the 19th century by the work of Cauchy and Weierstrass (Kramer, 1970). And what is the alternative to confronting these difficulties, to look away and rest our theorizing on unverified assumptions of equivalence?

6. Observations on the Value and Difficulties of Alignment

We conclude with some further observation on three matters: whether the face-to-face meeting we used in this alignment effort is likely to be typical; how we might label different approaches to defining "equivalence" ; and a brief proposal for the use of equivalence tests in evaluations made of journal submissions and research funding proposals.

There is one point at which the process we report might not be typical of alignments that would be done in the future, if this kind of analysis were to become more common. It is that a meeting, such as Epstein and Axtell had with Axelrod, might not be necessary in general. The meeting that was held served two functions: establishing details about the procedure of alignment and clarifying ambiguous aspects of the ACM. If the situation were one of comparison to a published model situated in an established line of research the former issues might be decided entirely by the author of a new model who seeks to establish its equivalence to an older one. This situation is one that we imagine might become more usual.

The second function of the meeting, resolving ambiguities about the construction of the target model, is not one that we imagine as likely to go away. On the contrary, many target models will be considerably more complex than the ACM. However, it may also be true that those attempting to show a new model equivalent to an old one will have source code for the old one -- a resource which was deliberately not employed in this case. It may also be true that the criteria of equivalence may be looser than they were in this case, a point we discuss below.

Considering all these factors, our impression is that good alignments can be made without actual meetings of model authors. This will be all the more likely if authors who report their models begin to assume that alignments may later be tried and thus become careful about providing information that may be essential to such efforts. We emphasize that 1) a precise, detailed statement of how the model works is critical, and 2) that

distributional information about reported measurements is necessary if statistical methods to test equivalence are to be employed by a later investigator.

As we noted above, the problem of specifying what will taken as "equivalent" model behavior is by no means trivial. Our reflections on it suggest that there are at least two categories of equivalence beyond the obvious criterion of numerical identity, which will not be expected in any models that have stochastic elements. We call these two categories "distributional" and "relational" equivalence. By distributional equivalence we mean a showing that two models produce distributions of results that cannot be distinguished statistically. This is the level of equivalence we eventually chose to test for in our case. By "relational equivalence" we mean that the two models can be shown to produce the same internal relationship among their results. For example, both models might show a particular variable is a quadratic function of time, or that some measure on a population decreases monotonically with population size.

Clearly, relational equivalence will generally be a "weaker", less demanding, test. But for many theoretical purposes it may suffice. And distributional equivalence may sometimes be possible only with alignment of parametric details of the two models that would be quite laborious to achieve.

Finally, our generally positive experience in this enterprise has suggested to us that it could be beneficial if alignment and equivalence testing were more widely practiced among computational modellers. It can be done within the reasonable effort level of a few days or weeks work -- possibly less if it is planned for from the outset. And the consequences are quite large. The Sugarscape group can now say with confidence that their model can be modified to reproduce the ACM results, and they point to specific mechanisms of Sugarscape which are sufficient to change the effect of the ACM transmission mechanism. This begins to build confidence that other results with Sugarscape may be robust over potential variation in the specifics of its cultural transmission process.

Readers of papers on Sugarscape and the ACM can now have a clearer conception of how they related to each other. And future modellers of cultural transmission will have a clearer understanding the likely consequences of different transmission mechanisms. In short, the interested community obtains from such an exercise an improved sense of the robustness, the range of plausibility, of model results. And points of difference have been established which could allow empirical evidence to discriminate between the models. These are major hallmarks of cumulative disciplinary theorizing that are unavailable without alignment of models.

We are led to the suggestion that it might be valuable if authors of computational models knew they would receive credit for having made such alignments. If reviewers of journal and research grant submissions were encouraged to give substantial positive weight to such demonstrations, and authors knew of this policy, the effects could be dramatic. Among other things, this would create an incentive to establish a model in an area of inquiry that could readily serve as a "benchmark" for comparisons by later models. The result might well be families of computational models displaying an explicit and clear network of relations to each other, rather than the current situation in which virtually every model has been constructed entirely de novo.

Computational modeling offers a striking opportunity to fashion miniature worlds, and this appeals to powerful creative impulses within all of us. William Blake expressed this deep need writing in his Jerusalem (1804/1974, pl.10, 1.20).

"I must Create a System, or be enslav'd by another Man's; I will not Reason and Compare: my business is to Create."

But if these wonderful new possibilities of computational modeling are to become intellectual tools well-harnessed to the requirements of advancing our understanding of social systems, then we must overcome the natural impulse for self-contained creation and carefully develop the methodology of using them to "Reason and Compare".

References

Axelrod, R. (1995), "The Convergence and Stability of Cultures: Local Convergence and Global Polarization," Santa Fe Institute working paper 95-03-028.

Blake, William, (1804/1974). Jerusalem, the emanation of the giant Albion. B. Quarich: London.

Cohen, M. D., J. G. March and J. P. Olsen (1972), "A Garbage Can Model of Organizational Choice." Administrative Science Quarterly, volume 17, pp. 1-25.

Cyert, R. M., and J. G. March (1963), A Behavioral Theory of the Firm. Prentice-Hall: Englewood Cliffs, New Jersey.

Efron, B. and R. J. Tibshirani (1993), An Introduction to the Bootstrap. Chapman and Hall: New York.

Epstein, J. M, and R. Axtell (forthcoming 1995), Growing Artificial Societies: Social Science From the Bottom Up. The Brookings Institution: Washington, D.C.

Hoel, P.G. (1962.), Introduction to Mathematical Statistics. Third ed. Wiley: New York.

Kramer, E. E. (1970), The Nature and Growth of Modern Mathematics, Volume 2. Fawcett, Greenwich, Connecticut.

Lane, D. (1993), "Artificial Worlds and Economics, Parts 1 and 2." J. Evol. Econ. 3. 89-107 and 177-197.

Latour, B. and S. Woolgar (1979), Laboratory Life : the Social Construction of Scientific Facts Sage Publications: Beverly Hills.

Siegel, S. (1956), Nonparametric Statistics for the Behavioral Sciences. McGraw-Hill: New York.

Appendix 1. Axelrod's Work Log

A1.1 Design of the Replication Study

1. Discussion with Cohen about the general idea of the replication experiment, including the suitability of my cultural model and Sugarscape for this purpose. (hours:minutes. 3:00)

2. Writing letter to Axtell and Epstein specifying what we came to call the docking experiment, including the choice of data points to be compared. (Cohen had already discussed the idea with them in Vienna.) (1:00)

3. Trip arrangements for Axtell and Epstein (1:00)

4. Discussion among all four of us of the docking experiment and its motivation, especially the importance of doing what we came to call distributional equivalence, rather relational equivalence (1.00).

5. Discussion among all four of us about the details of Axelrod's cultural model. This included discussion about the sentence that said "the chance of interaction is proportional to the cultural similarity two neighbors already have" where cultural similarity is the proportion of attributes which have the same value. Axtell had implemented this literally, but I pointed out that I actually used a more efficient (and equivalent) method, namely to allow interaction if a single randomly chosen attribute has the same value. (1:00)

6. Preliminary specification of what became the mobility experiment. See Section 4.3.1. (2:00).

Subtotal: 9:00

A1. 2 Data Analysis

1. Extraction of key raw data from my old computer output for Axtell to use in comparing my data to his. (1:00)

2. Communications with Axtell about receiving his data, and updating it after he corrected for changing the agent rather than the neighbor. See Section 5.1. (2:30)

3. Discussions with Cohen and a statistical consultant, Pat Guire, on proper statistical testing (3:00)

4. Putting Axtell's data in a format comparable to mine, and calculating basic statistics (2:00)

5. Consideration of alternative possible reasons why the original attempt at docking did not succeed for the 20x20 case. Development of tests of these possibilities (e.g. a bug in my code or Axtell's code), and identification of the likely cause as differences in the activation methods See Section 4.2. (5:30).

Subtotal: 14:00

Grand Total: 23:00

Appendix 2. Axtell's Work Log

A 2.1 Code changes accomplished in Ann Arbor:

1. Generalize culture representation from type BOOLEAN
to an enumerated type (Hours: Minutes, 0:10)

2. Change agent initialization:

    A. Fill lattice densely with agents (0:20)

    B. Give agents random initial cultures (0:20)

3. Implement a version of Axelrod's culture rule (01:00)

4. Draw boundaries between agents not culturally identical (0:30)

5. New stopping criterion (0:15)

6. Count distinct cultures (surrogate for counting regions) (0:15)

7. Switch landscape from torus to square negligible (<0:01)

8. Turn-off all other Sugarscape rules negligible (<0:01)

9. Debugging all of above (1:00)

Sub-total: 3:50

A 2.2 Subsequent code modifications:

1. Modify neighborhood representation so that agents on the border of the lattice
do not attempt to interact with non-existent (NIL) neighbors (0:30)

2. Represent regions as social networks and then use
'clique_size' of social network object to count regions
(this obviated the need for #6 above) (0:30)

3. File output of number of cultural regions (0:10)

Sub-total: 1:10

A 2.3 Running the model:

1. Make executable files for various parameter settings (0:40)

2. 90 runs for comparison to Axelrod's ata on features and values/feature. See Table 1. (2:00)

3. 120 runs for comparison to Axelrod's data on lattice size. See Figure 1. (8:00)

Sub-total: 10:40

A 2.4 Statistical comparison:

1. .Development of Mann-Whitney U test in Mathematica (2:00)

2. Analysis of data using the Mann-Whitney U test (1:00)

3. Development of Kolmogorov-Smirnov (K-S) analysis routines in Mathematica (4:00)

4. Analysis of data using K-S test (2:00)

Sub-total: 9:00

A 2.5 Mobility experiment: See Section 4.3.1

1. Modify the stopping criteria to consider agent interactions with the entire population (0:10)

2. Time series plot for the distinct number of cultures (1:00)

3. Instantiate a standard version of the Sugarscape with the Epstein-Axtell culture rule replace by Axelrod's (0:10)

4. Make executable file (0:05)

5. Perform multiple runs of this model (1:00)

Sub-total: 2:25

A 2.6 'Pure soup' experiment: See Section 4.3.2

1. Instantiate soup version of the Sugarscape with Axelrod's culture rule (0:10)

2. Make executable file (0:05)

3. Perform multiple realizations of this model (1:00)

Sub-total: 1:15

A 2.7 Re-docking: See Section 4.2

1. Change agent activation from sequential to random (0:10)
2. Re-run the model (40 runs) (8:00)
3. Analysis of new data (0:20)

Sub-total: 8:30

Grand total: 36:50

Table 1

Average Number of Stable Regions

a. Axelrod's Cultural Model

		Values/Feature		
		5	10	15
	5	1.0	3.2 ± 1.8	20.0 ± 10.1
Features	10	1.0	1.0	1.4 ± 0.5
	15	1.0	1.0	1.2 ± 0.4

b. Sugarscape Implementation

		Values/Feature		
		5	10	15
	5	1.0	5.3 ± 3.9	21.3 ± 12.5
Features	10	1.0	1.0	1.5 ± 0.7
	15	1.0	1.0	1.0

Note: Each cell is based on ten replications. Standard deviations are shown when they are not equal to zero. All data are for territories of 10x10 sites.

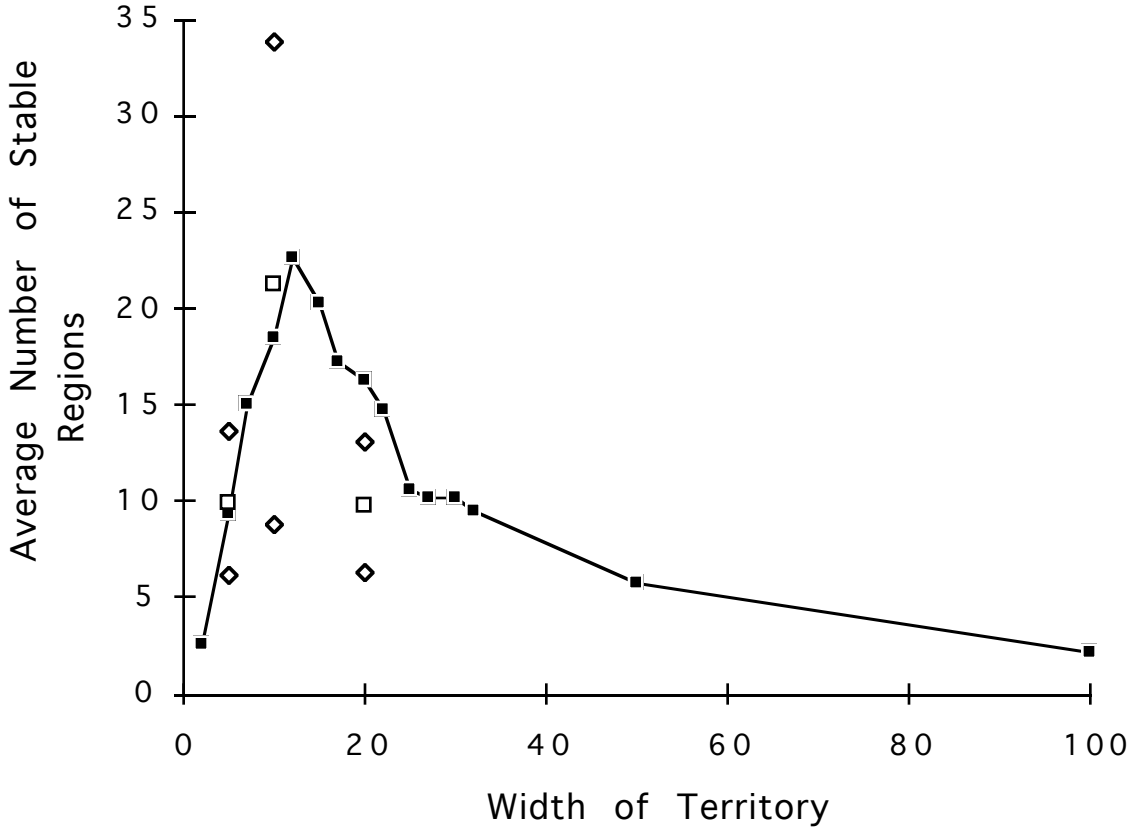

Figure 1

Average Number of Stable Regions

in Axelrod's Cultural Model and Sugarscape Implementation

Legend:

Solid squares represent the target ACM data for 5 cultural features and 15 traits per feature. Each territory size was replicated 40 times, except the territories with 50x50 sites and 100x100 sites which were replicated 10 times.

Open squares represent the Sugarscape data for the same number of features and traits per feature. Each territorial size (5x5, 10x10 and 20x20) was replicated 40 times.

Open diamonds represent Sugarscape means plus and minus a standard deviation.