Proceedings Article

Empirical Agent Based Models of Cooperation in Public Goods Games

MICHAEL WUNDER, Rutgers University
SIDDHARTH SURI, Microsoft Research, New York City
DUNCAN J. WATTS, Microsoft Research, New York City

Agent-based models are a popular way to explore the dynamics of human interactions, but rarely are these models based on empirical observations of actual human behavior. Here we exploit data collected in an experimental setting where over 150 human players played in a series of almost a hundred public goods games. First, we fit a series of deterministic models to the data, finding that a reasonably parsimonious model with just three parameters performs extremely well on the standard test of predicting average contributions. This same model, however, performs extremely poorly when predicting the full distribution of contributions, which is strongly bimodal. In response, we introduce and test a corresponding series of stochastic models, thus identifying a model that both predicts average contribution and also the full distribution. Finally, we deploy this model to explore hypotheses about regions of the parameter space outside of what was experimentally accessible. In particular, we investigate (a) whether a previous conclusion that network topology does not impact contribution levels holds for much larger networks than could be studied in a lab; (b) to what extent observed contributions depend on average network degree and variance in the degree distribution, and (c) the dependency of contributions on degree assortativity as well as the correlation between the generosity of players and the degree of the nodes to which they are assigned.

Categories and Subject Descriptors: I.6.6 [**Simulation and Modeling**]: Model Development—*Modeling methodologies*

General Terms: Human Factors, Economics

Additional Key Words and Phrases: Social Networks; Cooperation; Modeling; Agent Based Models

1. INTRODUCTION

Agent-based models (ABMs), also sometimes called "individual-based models" or "artificial adaptive agents" [Holland and Miller 1991], constitute a relatively recent approach to modeling complex systems—one that stakes out a middle ground between the highly formal but also highly abstracted approach of traditional mathematical models, which emphasizes analytical solutions of algebraic or differential equations, and the richly descriptive but also ambiguous and imprecise approach of intuitive reasoning [Bonabeau 2002]. ABMs typically assume the existence of discrete agents, whose behavior is specified by rules that depend on the states of other agents, as well as some arrangement of interactions between the agents, where both the agent rules and the interaction patterns can vary from very simple and abstract—as in cellular automata—to highly complex and realistic. On the strength of their flexibility and realism, ABMs have been extensively deployed over the past thirty years to model a wide range of problems of interest to social scientists, including neighborhood segregation [Schelling 1978], organizational problem solving [Lazer and Friedman 2007], cooperation and conflict [Axelrod 1984], opinion change [Deffuant et al. 2000], cultural evo-

This work was done while the first author was at Microsoft Research, New York City. Author's addresses: mwunder@cs.rutgers.edu, {suri, duncan}@microsoft.com.

EC'13, June 16–20, 2013, Philadelphia, USA.

lution [Epstein and Axtell 1996; Axelrod 1997], and political state formation [Cederman 1997].

The generally greater complexity of ABMs, however, also requires the modeler to make potentially many assumptions regarding (a) the amount and type of information possessed by the agents, (b) the manner in which that information is processed, and (c) the rules governing the interactions between agents. Traditionally, these modeling choices have been made on the grounds of intuitive plausibility, rather than on empirical accuracy. This reflects, in part, the philosophical position of ABMs researchers who have viewed ABMs as thought experiments intended to explicate theories and explore causal mechanisms, not as forecasting engines [Axelrod 1997; Macy and Willer 2002]. In recent years, however, the idea of grounding modeling assumptions on empirical observations of human behavior has begun to attract attention [Janssen and Ostrom 2006; Heckbert et al. 2010]. The reason being that even plausible and apparently innocuous assumptions about agent behavior can turn out not only to be mistaken but also critical to the emergent behavior of interest[1]. Even if the goal of agent-based modeling is theory explication not empirical accuracy per-se, a certain amount of empirical accuracy may be necessary in order to avoid spurious conclusions.

In this paper, therefore, we articulate an approach that we label "empirical agent based modeling" (EABM) in which candidate models are first trained and evaluated on data from human-subjects experiments, and then deployed in the same way as traditional ABMs to explore regions of the parameter space outside of those in which the original experiments were conducted.[2] Our data-oriented approach means that we motivate and evaluate our models almost exclusively in terms of how well they predict observable player actions[3], ignoring obvious criteria such as psychological interpretability or theoretical plausibility. As a consequence, our models do not map in a straightforward fashion to conventional agent-based models, which are often motivated by strategic or psychological arguments; however, as we will indicate, a number of these models are in fact behaviorally equivalent, and therefore are effectively included in our analysis. Finally, although the idea of empirical validation of ABMs is general, we illustrate the approach in the specific context of cooperation in public goods games, an important problem in social science in general, and to agent-based modeling in particular [Axelrod 1984, 1997; Macy and Willer 2002], and critically an area in which recent large-scale human subjects experiments [Suri and Watts 2011; Wang et al. 2012] have made the appropriate data available for training and testing EABMs.

2. RELATED WORK

Although empirical evaluation of ABMs is a topic that has received relatively little attention, a handful of attempts have been made, also in the context of games of cooperation. The earliest, by Deadman [1999], attempted to fit data from previously conducted common pool resource experiments with a reinforcement learning model. According to Deadman, the resulting aggregate behavior was "similar" to the empirical data, but no quantitative evaluation was performed and no alternative models were considered. Subsequently, Castillo and Saysel [2005] developed a system dynamics model of player behavior also in common pool resource games, and compared its behavior with data from field experiments involving fisherman and crab hunters from the Providence Island of Columbian Caribbean Sea. The authors assessed their model's validity predominantly in terms of its ability to display behavior that is consistent with theoretical expectations (e.g. its sensitivity to key parameters), not empirical data. Nevertheless, they showed that it was possible to find parameters for which

[1] See Mason and Watts [2012] for an example of how plausible modeling assumptions can lead to qualitatively misleading simulation results.

[2] We note that agent models could also be evaluated on data from non-experimental sources such as role-playing games, participant observation, or surveys [Janssen and Ahn 2006].

[3] Where predictive performance of competing models is close we also place some weight on parsimony.

the model could approximately replicate observed aggregate contributions, where again no quantitative evaluation was performed and no alternative models considered. Finally, and most similar to the current work, Janssen and Ahn [2006] fitted an experience-weighted attraction (EWA) model of learning [Camerer and Hua Ho 1999] to data from two earlier experiments. Fitting separate models to individual players, they identified 8 player "types," defined in terms of their best-fit parameter values, that accounted for the vast majority of the sample population.

Our contribution differs from, and builds upon, this previous work in three key respects:

(1) Whereas previous attempts have emphasized plausibility and interpretability of the candidate models over predictive accuracy, here we take a machine learning approach, similar to that adopted by Wright et al. [2012], in that we introduce a basket of models and compare their predictive performance on out-of-sample test data. We note that our approach does not rule out cognitively plausible models—indeed, as we will indicate, a number of conventional models of cooperation, including Tit for Tat and "grim trigger," are behaviorally consistent with those we propose. Because we are interested in predicting behavior, however, we are less concerned with the underlying cognitive model than with the behavior itself.
(2) We evaluate model performance more rigorously than previous work, first on average contributions over time, and second on the full round-by-round distribution of contributions—a far more challenging requirement.
(3) Finally, we go beyond simply fitting a model to the experimental data–we then deploy this model to explore parameter regimes beyond those covered by the experimental design. In other words, our approach preserves the "ABM as thought experiment" tradition of agent-based modeling, but attempts to ground it in agent rules that are calibrated to real human behavior within at least some domain.

3. BACKGROUND ON EXPERIMENTAL SETUP AND DATA

Before we define and analyze our models, we first briefly describe the experiments used to gather our data, which were conducted using Amazon Mechanical Turk[4] (AMT), and were originally reported by Suri and Watts [2011] (hereafter referred to as SW). The experiments were a variant of a linear public goods game [Ledyard 1995], a game of cooperation that is widely studied in laboratory settings. Each game comprised 10 rounds, where in each round each participant i was allocated an endowment of $e = 10$ points, and was required to contribute $0 \leq c_i \leq e$ points to a common pool. In standard public goods games, participants' contributions are shared equally among members of the same group. SW, however, studied a variant in which participants were arranged in a network, so they shared their contributions with their neighbors. To reflect this change, players' payoffs were given the payoff function $\pi_i = e_i - c_i + \frac{a}{k+1} \sum_{j \in \Gamma(i)} c_j$, where in place of the summation over the entire group of n players, payoffs are instead summed over $\Gamma(i)$, the network neighborhood of i (which we define to include i itself), and k is the vertex degree (all nodes in all networks have the same degree). Therefore, i's contributions were, in effect, divided equally among the edges of the graph that are incident on i, where payoffs are correspondingly summed over i's edges. From this payoff function it is easy to show that when $1 < a < n$, players face a social dilemma in that all players contributing the maximum amount maximizes social welfare, but individually players are best off if they contribute nothing, thereby free-riding on the contributions of others.

SW chose networks that spanned a wide range of possible structures between a collection of four disconnected cliques at one extreme, and a regular random graph at the other, where all networks comprised $n = 24$ players, each with constant vertex degree $k = 5$.

[4] `http://www.mturk.com`

SW conducted a total of 74 networked experiments on AMT over a period of 6 months, including the following treatments which we analyze in this work[5]:

(1) *All Human,* 23 *games*. All 24 players were human subjects.
(2) *Altruistic Dummies,* 13 *games*. Four positions were played by computer, which contributed the full endowment each round. The dummies were arranged so that each human player was adjacent to precisely one dummy (i.e. the dummies constituted a covering set for the graph)
(3) *Free Riding Dummies,* 17 *games*. Same as for altruistic dummies, but the dummies contributed zero in each round.
(4) *Neighboring Altruistic Dummies,* 20 *games*. Same as for altruistic dummies, but the four dummies were arranged in two pairs, such that some human player were adjacent to two dummies, while others were adjacent to zero.

Surprisingly, SW found that network topology had no significant effect on contributions in any of the experimental treatments. From the Altruistic and Free Riding Dummy conditions, they established that players were behaving as conditional cooperators, hence contributions in neighborhoods with high local clustering were more correlated than those with low clustering; however, the symmetrical nature of conditional cooperation effectively led positive and negative effects to cancel out. Moreover, from the concentrated dummies condition, they also established the absence of multi-step contagion of positive effects, although they did not rule out negative contagion.

4. DETERMINISTIC MODELS

In this section we first define and then evaluate a collection of models that we refer to as *deterministic*, meaning that the output of a model is the expected contribution for the next round. As we will show later, all the deterministic models that we consider suffer from a major shortcoming in predicting the full distribution of contributions. Nevertheless, we begin with them for three reasons: first, they are relatively simple and intuitive; second, they perform reasonably well at predicting average contributions; and third, they are frequently invoked both in agent-based models of cooperation [Axelrod 1984] and also in previous attempts to replicate empirical data [Deadman 1999; Castillo and Saysel 2005; Janssen and Ahn 2006].

4.1. Model Definitions

Linear Self-factor Model. Perhaps the simplest model one might imagine captures the commonly observed empirical regularity that players who contribute a lot (respectively, a little) in the previous round are more likely to contribute a lot (respectively, a little) in the current round [Wang et al. 2012]. Formally, the model predicts $c_{i,t}$, player i's contribution on round t, to be a linear function of player i's contribution in the previous round $c_{i,t-1}$:

$$\hat{c}_{i,t} = \beta_1 c_{i,t-1}$$

Linear Neighbor-factor Model. A second simple model is motivated by the notion of conditional cooperation [Fischbacher et al. 2001]—that the more player i's neighbors contribute, the more player i is likely to contribute. Specifically, $c_{i,t}$ is predicted by the weighted average of player i's neighbors' contribution in the previous round, $\bar{c}_{i,t-1}$.

$$\hat{c}_{i,t} = \beta_2 \bar{c}_{i,t-1}$$

Linear Two-factor Model. Next, we combine these two single-parameter models in a two-factor model that predicts $c_{i,t}$, player i's contribution on round t, as a weighted linear

[5]In section5.3 we make use of an additional set of 15 related experiments conducted after the publication of SW. Because they were not described in SW, however, we do not use them for our main results.

combination of (a) player i's contribution in the previous round $c_{i,t-1}$, and (b) the average contribution in round $t-1$ of the local neighbors of player $i, \bar{c}_{i,t-1}$:

$$\hat{c}_{i,t} = \beta_1 c_{i,t-1} + \beta_2 \bar{c}_{i,t-1}$$

The coefficients β_1 and β_2 therefore capture the relative importance of player's previous actions versus his neighbors' previous actions, where we note that models of this general form ("place some weight on my own intrinsic inclination to contribute and some weight on my neighbors' contributions") generate behavior that is consistent with conditionally-cooperative models such as Tit-for-Tat [Axelrod 1984], and even more complicated strategic models such as that proposed by Kreps et al. [1982].

Triangle-shaped Model. Motivated by Fischbacher et al. [2001], who observed that some players contribute proportional to their neighbors up to about 50% of the total endowment, after which their contributions decline in proportion to their we propose the following "triangle" model:

$$\hat{c}_{i,t} = \beta_1 c_{i,t-1} + \beta_2 \bar{c}_{i,t-1} + \beta_3 (5 - |5 - \bar{c}_{i,t-1}|)$$

Threshold Model. Previous theoretical models [Glance and Huberman 1993; Lopez-Pintado and Watts 2008] have posited that players will contribute to a public good only when the average neighborhood contribution is above a certain threshold. We capture the essence of these "threshold models" using a logistic function, which maps a continuous variable onto the $[0, 1]$ range and does so with a gradual probabilistic change between binary options. This function can represent rapid changes in behavior as a threshold and is written as:

$$\sigma(\bar{c}_{i,t-1}) = \frac{1}{1 + e^{-\lambda(\bar{c}_{i,t-1} - \theta)}}$$

Note this function has two parameters: θ, which is the midpoint where an average neighbor contribution of $\bar{c}_{i,t-1} = \theta$ leads to a probability equal to 0.5; and λ, which indicates how rapidly the function changes around the midpoint (i.e. as λ increases, the threshold approaches a step function). The resulting model is as follows:

$$\hat{c}_{i,t} = \beta_1 c_{i,t-1} + \beta_2 \sigma(\bar{c}_{i,t-1}) = \beta_1 c_{i,t-1} + \frac{\beta_2}{1 + e^{-\lambda(\bar{c}_{i,t-1} - \theta)}}$$

where we note that the so-called "grim trigger" strategy ("cooperate until someone defects and then defect forever") translates roughly [6] to a threshold model with $\beta_1 \approx 0$, $\lambda \gg 1$, and $\theta = \theta^*$, where θ^* determines the position of the trigger.

Time Discounted Models. A final relevant concept is "future discounting:" the idea that people prefer payoffs today to larger payoffs tomorrow [Williams 1938]. Assuming that i's contribution in the present round serves as an investment in keeping i's neighbors in a generous state, and setting $0 \leq \delta \leq 1$ as the discount rate, we can derive time-discounted versions of the two-factor linear and threshold models as follows.

$$\hat{c}_{i,t} \;=\; \beta_1 c_{i,t-1} + \beta_2 \sum_{\tau=t}^{T} \delta^{\tau - t} \bar{c}_{i,\tau}$$

where T is the total number of rounds in the game and δ is the discount rate. A player may have realized from prior play that his neighbors contributions levels decline with time. So

[6] Technically grim trigger is defined for a two-player repeated prisoner's dilemma, so the translation to a multiplayer public goods game is necessarily imperfect.

we can model the above equation as:

$$\hat{c}_{i,t} \;=\; \beta_1 c_{i,t-1} + \beta_2 \sum_{\tau=t}^{T} \delta^{\tau-t}\theta^{\tau-t}\bar{c}_{i,t-1}$$

where $0 \leq \theta \leq 1$. Setting $\gamma = \delta\theta$ and simplifying the geometric series gives:

$$\hat{c}_{i,t} \;=\; \beta_1 c_{i,t-1} + \beta_2 \bar{c}_{i,t-1}\left(\frac{1-\gamma^{T-t+1}}{1-\gamma}\right)$$

Thus we obtain the following models.

$$\text{Discounted Two-Factor Model:} \qquad \hat{c}_{i,t} = \beta_1 c_{i,t-1} + \beta_2 \frac{1-\gamma^{T-t+1}}{1-\gamma}\bar{c}_{i,t-1}$$

$$\text{Discounted Threshold Model:} \qquad \hat{c}_{i,t} = \beta_1 c_{i,t-1} + \beta_2 \sigma\left(\frac{1-\gamma^{T-t+1}}{1-\gamma}\bar{c}_{i,t-1}\right)$$

where σ is the logistic function described in the Threshold Model subsection.

4.2. Predicting Average Contributions

Having defined our basket of models, we now proceed to evaluate them on their ability to predict the next action in the game for player i ($c_{i,t}$, the contribution at time t) given the current level of personal ($c_{i,t-1}$) and average neighbor contributions ($\bar{c}_{i,t-1}$). Consistent with previous work [Janssen and Ahn 2006], we will define and perform two different types of evaluation based on predicting individual contributions: a *homogenous* population evaluation, which assumes that all players act the same way; and a *heterogeneous* population evaluation, in which each player is allowed to behave differently–sometimes very differently. Previous studies of public goods experiments [Fischbacher et al. 2001; Janssen and Ahn 2006] have observed that behavioral data is better explained by allowing for heterogeneous types; however, homogenous strategies allow us to use more data to fit and evaluate each model, so we consider both.

4.2.1. Homogenous Population Evaluation. As just noted, we begin by assuming a homogenous population, where all players are described by the same set of model parameters. Each model is then fit using regression or parameter search where appropriate. For evaluation, we use the leave one out method; that is, for a total of g games we train on $g-1$ games, and test on the gth game, where every game gets exactly one chance at being the test set. We then evaluate each model's performance using root mean squared error (RMSE), a simple, intuitive measure of predictive accuracy.[7]

Table I shows the results of this evaluation. The two single factor models do the worst, where the self-factor model beats the neighbor-factor model, indicating that the contribution of player $i, c_{i,t-1}$, has more predictive power than the average contribution of player i's neighbors, $\bar{c}_{i,t-1}$. The linear 2-factor model, which uses both player $c_{i,t-1}$ and $\bar{c}_{i,t-1}$, has better predictive accuracy than either single factor model alone; thus there is predictive power in using both $c_{i,t-1}$ and $\bar{c}_{i,t-1}$. In general, the linear 2-factor, discounted 2-factor, triangle, and threshold models are comparable in performance. Because simple linear 2-factor has an error close to the other models with more parameters, it is a good tradeoff between parsimony and predictive accuracy.

[7]Note that using log-likelihood and max-likelihood to fit the models is a common technique in these domains. However, we decided that the risk of over-fitting individual behavior (see next section) was significant due to sparsity of available data in many cases. We found regression to robustly fit the models.

Table I. Homogenous population evaluation where we leave-one-game-out. The errors are the average RMSE for predicting individual contributions. Standard errors are all less than ±0.02.

Model name	# of Params	All Human	Altruistic Dummies	Free-Riding Dummies	Neighbor Dummies	Mean
Self-factor	1	2.37	2.19	2.09	2.36	2.25
Neighbor-factor	1	3.16	3.40	2.72	3.37	3.16
Linear 2-Factor	2	2.27	2.14	2.02	2.31	2.18
Discounted 2-factor	3	2.26	2.12	2.02	2.3	2.18
Triangle-shaped	3	2.26	2.11	2.00	**2.27**	2.18
Threshold	4	2.25	2.12	**1.99**	2.29	2.16
Discounted Threshold	5	**2.23**	**2.07**	2.00	2.26	**2.14**

Table II. Heterogeneous Agent Model: RMSE results for several behavioral models, based on learning a custom one for each player.

Model name	# of Params	All Human	Altruistic Dummies	Free-Riding Dummies	Neighbor Dummies	Mean
Self-factor	1	2.05	1.87	1.74	1.96	1.91
Group-factor	1	2.36	2.24	1.81	2.37	2.20
Linear 2-Factor	2	**1.97**	1.87	**1.57**	1.89	1.83
Discounted 2-factor	3	1.98	**1.80**	**1.57**	1.92	1.82
Triangle-shaped	3	2.11	1.93	1.67	**1.75**	1.87
Threshold	4	1.98	1.87	1.58	1.76	**1.80**
Discounted Threshold	5	2.02	1.86	1.59	1.87	1.83

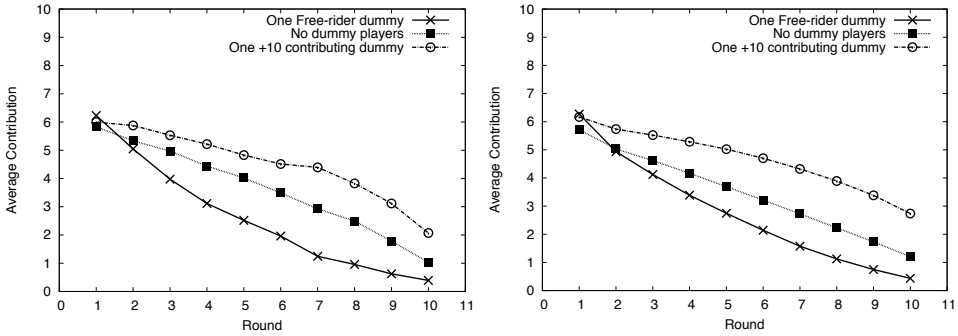

Fig. 1. Average contributions per round for (a) experimental results from Watts and Suri [2011] and (b) simulated results using the time-discounted two-factor model.

4.2.2. Heterogenous Population Evaluation. Analogous to the homogeneous case, we train each model on the majority of a player's games, keeping a hold-out set of 20% of the total or a single game, whichever is larger[8]. We then evaluate the model on the hold-out set, repeating this procedure with a rotating hold-out set until all games are tested. We compute the RMSE on the test set and average those across all players weighted by their experience. The results of this analysis are shown in Table II.

Although each model is now fit with much less data than in the homogeneous case, we find that in general errors are reduced by learning individually customized models. We also see varying performance in the different treatments. For example, including a triangle strategy hurts performance when predicting the case with free riders, but helps when there are multiple high contributors present. Finally, Fig. 1 shows graphically, for the special case of the discounted two-factor model, how the predicted average contributions (right panel) compare with the empirically observed contributions from Suri and Watts [2011],

[8]Any player with fewer than three games is excluded on the basis that there is not enough training data for that individual.

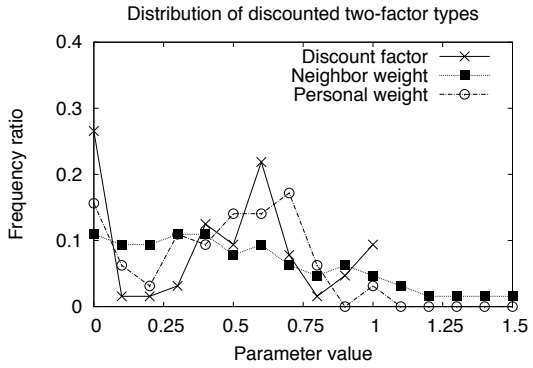

Fig. 2. The distribution over types in the heterogeneous discounted two-factor model.

Table III. Frequency of type by player, low, medium, or high discount, low, medium, or high personal weight, low or high neighbor weight. One could easily ascribe cognitive motivations to these values.

Neighbor weight (Reciprocation)	Low ($\beta_2 < 0.5$)			High ($\beta_2 \geq 0.5$)			
Personal weight	Low ($\beta_1 < .25$)	Medium ($.25 \leq \beta_1 < .75$)	High ($\beta_1 \geq .75$)	Low	Medium	High	Sum for γ
Discount factor γ							
Low ($\gamma < .25$)	0.07	0.07	0.02	0.07	0.10	0.00	0.33
Medium ($.25 \leq \gamma < .75$)	0.02	0.14	0.11	0.10	0.12	0.00	0.49
High ($\gamma \geq .75$)	0.00	0.09	0.02	0.02	0.05	0.00	0.18
Sum for β_1	0.09	0.30	0.15	0.19	0.27	0.00	
Sum for β_2		0.54			0.46		

for the three main treatments: all human, altruistic dummies, and free riding dummies. Visually the curves, which are generated via the method described in section 4.3, are hard to distinguish, indicating the that quantitative performance measures in Table II correspond to qualitatively meaningful agreement.

4.2.3. Analysis of Types. The superior performance of the heterogenous models in spite of their more limited data suggests that players use a variety of strategies that is not being captured by the homogeneity assumption. Fig. 2 confirms this intuition, showing that the distributions of the three parameters in the discounted two-factor linear model, β_1, β_2, and γ, all have broad support. Interestingly, Fig. 2 also shows that the distributions of β_1 and γ are effectively tri-modal, while the distribution of β_2 is close to uniform. Motivated by this observation, we partition the population into "types" as follows: for β_1 we allocate players to "low" ($\beta_1 < 0.25$), "medium" ($0.25 \leq \beta_1 < 0.75$), and "high" ($0.75 \leq \beta_1$); for β_2, we have "low" ($\beta_2 < 0.5$) and "high" ($0.5 \leq \beta_2$); and γ, low, medium, and high as per β_1. As Table III shows, this partition corresponds to 18 cells, or "types", of which 9 have between 7% and 14% of the population, where these 9 types account for nearly 90% of the population[9]. In addition, we note that 90% of the population lies in the medium ranges of γ and β_1, which constitutes only half of the parameter space, while there is a near even split between highly reciprocating players with high β_2 (46%) and those with low β_2 (54%). We might describe players with high γ as forward thinkers, and those with high β_2 as conditional cooperators.

4.3. Predicting Full Distribution of Contributions

The model evaluation of the previous section seems promising and is also consistent with previous attempts to validate models empirically, which have also focused on average con-

[9]Interestingly, Janssen and Ahn [2006] found a similar result using a different methodology, finding that a similar majority of players were accounted for by 8 out of 16 possible types.

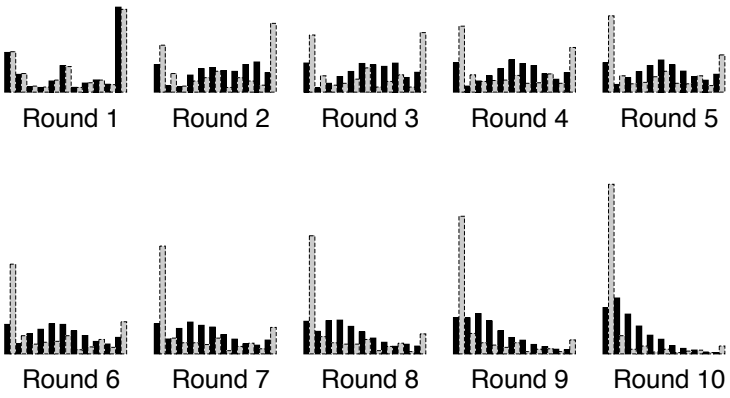

Fig. 3. Actual population behavior compared to the deterministic discounted 2-factor model.

tributions over time [Deadman 1999; Castillo and Saysel 2005; Janssen and Ahn 2006]. In light of this history, however, it is important to realize that the average contribution is potentially an extremely poor proxy for the full distribution of contributions. The reason is that contributions in public goods games are strikingly bimodal, with extreme actions of zero and ten appearing as the two modes, and a minority playing the actions between one and nine [Ledyard 1995; Suri and Watts 2011]. Over time the number of players at the maximum contribution declines while those who contribute zero increases significantly, but the bimodality persists. Clearly it is possible to accurately predict the average of a bimodal distribution while completely misrepresenting the underlying distribution. Yet also clearly it would be desirable for any agent-based model to replicate the full distribution as well as the average.

Thus motivated, we now evaluate the same models in terms of their ability to predict the full empirical distributions, training one instance of each model per player on half of the data in each treatment, and testing against the distribution of the other half. Specifically, we first construct a simulated population of agents in the following manner: if player i is in the test set and the training set, we put the model for player i in the simulated population in proportion to its experience in the test set; and if player i is in the test set but not the training set we select at random from the training set chosen weighted by that player's experience in the training set. For each simulated population we then run a simulated game by sampling 24 players from the population and running their models using first round contributions chosen from the distribution of actual first round contributions in the test set. We repeat this process 20 times to get 20 simulated games which is roughly the number of actual games we had for each experimental treatment.

The result for the discounted 2-factor model is illustrated graphically in Figure 3, from which it is evident that the distribution of the model's predictions clearly distinguishable from the bimodal distribution of the empirical data. Expressing this qualitative observation quantitatively, we repeat this process 100 times, where in each instance we find the Kullback–Leibler (KL) divergence, a standard measure for the extra information needed for a model to represent some original distribution, between the simulated and empirical distributions of rounds 2-10. For example, as shown in Table IV, the linear 2-factor model with low RMSE has a relatively high KL divergence value above 1, meaning that, on average, the log-odds ratio of the two distributions is off by a factor of 3 or greater. In general, Table IV shows relatively poor performance for all the deterministic models. The reason is that in spite of their differences, all the deterministic models predict that high contributing agents will reduce their contributions steadily over time—a tendency that leads the initially bimodal distribution to become increasingly unimodal—whereas empirically, human agents

Table IV. Evaluation of the distribution of the population's contribution for the deterministic models trained on half the experiments and tested on the other half. KL divergence measures (non-symmetrically) the difference between the true data and the model output. Lower KL divergence represents higher accuracy.

Model name	# of Params	All Human	Altruistic Dummies	Free-Riding Dummies	Neighbor Dummies	Mean
Linear 2-factor	2	1.32	2.40	1.15	4.82	2.42
Discounted 2-factor	3	**0.84**	1.36	**0.93**	4.17	1.83
Threshold	4	0.96	**1.10**	1.87	**2.12**	**1.52**
Discounted Threshold	5	3.29	2.17	1.07	6.16	3.17

tend to jump from very high to very low contributions almost discontinuously, spending very little time in the middle of the distribution and thereby preserving the bi-modality of the distribution even as the mean decreases. Predicting average contributions, in other words, is no guarantee of having captured the underlying behavioral dynamics.

5. STOCHASTIC MODELS

Motivated by this observation that individual contributions are not well represented by the expectation, we now introduce a method for constructing stochastic models that builds on the successful aspects of the deterministic models, but more accurately captures the bi-modality of the empirical distribution. Our general approach is that for each of the deterministic models in the previous section, we can define a corresponding stochastic model that invokes the deterministic model as a subroutine. Rather than predicting an expected contribution, however, the stochastic model instead makes use of the deterministic model to predict that player i will make the same contribution they did in the last round with probability ϕ. In addition, the stochastic model also predicts that a player will make a strategy uniformly distributed in the range $[1, 10]$ with probability ϵ, which is estimated directly from the data and reflects the empirical observation that some agents to play the non-extremal actions. And finally, it predicts that the player will make a contribution of 0 with probability $1 - \phi - \epsilon$.

We illustrate our method for generating a stochastic model from a deterministic one using the linear 2-factor model described above. From Section 4.1 we see that the two factor model predicts the next round's contribution via

$$\hat{c}_{i,t} \;=\; \beta_1 c_{i,t-1} + \beta_2 \bar{c}_{i,t-1}$$

Conditioned on $c_{i,t-1} > 0$ we can rewrite this as

$$E[\hat{c}_{i,t} \mid c_{i,t-1} > 0] \;=\; c_{i,t-1}\left(\beta_1 + \beta_2 \frac{\bar{c}_{i,t-1}}{c_{i,t-1}}\right)$$

Next we show how one can interpret this expectation as a value times the probability of a player contributing that value. Since contributions generally decrease, it is most often the case that $c_{i,t} \leq c_{i,t-1}$. In addition, contributions are always at least 0. Thus, we can interpret $\phi(c_{i,t-1}, \bar{c}_{i,t-1}) = \beta_1 + \beta_2 \frac{\bar{c}_{i,t-1}}{c_{i,t-1}}$ as a probability of playing $c_{i,t-1}$ again during round t. Players may, of course, choose not to contribute the same as they did last round. It is possible that players increase their contributions or contribute some amount between 1 and 10. To capture these cases we let ϵ be the probability of contributing a random amount $\Pr[c_{i,t} = \mathcal{U}[1, 10]] = \epsilon$. Figure 3 shows that players often contribute 0. So we let the remaining probability, $1-\epsilon-\phi(c_{i,t-1}, \bar{c}_{i,t-1})$ be the probability of contributing 0. Combining all of this gives

$$E[\hat{c}_{i,t} | c_{i,t-1} > 0] = c_{i,t-1}(1 - x)\phi(c_{i,t-1}, \bar{c}_{i,t-1}) + 5.5\epsilon,$$

where $x = \frac{5.5\epsilon}{\phi c_{i,t-1}}$ corrects the upward bias in the expectation caused by the uniform random variable in $[1, 10]$. Observe that if we plug x into the above equation we get:

$$E[\hat{c}_{i,t}|c_{i,t-1} > 0] = c_{i,t-1}\phi(c_{i,t-1}, \bar{c}_{i,t-1}),$$

which shows that the stochastic model will output the same prediction, in expectation, as the deterministic model. Although, we shall see that the actual distribution of predictions is much closer to the experimental data. The above describes the model for when $c_{i,t-1} > 0$. When $c_{i,t-1} = 0$, players most often play 0 for the rest of the game, but occasionally they do increase their contributions. To capture this we say that a player might contribute an amount uniformly distributed in $[1, 10]$ with probability ϵ_0, giving

$$E[\hat{c}_{i,t} \mid c_{i,t-1} = 0] = 5.5\epsilon_0$$

In this case we can fit ϵ_0 to the data so that we can ensure that the expected prediction of the stochastic model is the same as the prediction of the probabilistic model.

Recall that ϕ was defined in terms of the linear 2-factor model. The other parameters, ϵ and ϵ_0 were fit to the data. Thus this stochastic model is determined by

$$\Pr[c_{i,t} = c_{i,t-1} \mid c_{i,t-1} > 0] = \phi(c_{i,t-1}, \bar{c}_{i,t-1}) = \beta_1 + \beta_2 \frac{\bar{c}_{i,t-1}}{c_{i,t-1}}$$

The general technique described here can similarly be applied to each of the models defined in Section 4.1.

5.1. Baseline Stochastic Models

Although our recipe for generating a stochastic version of each of the previously defined deterministic models yields a corresponding collection of stochastic models, it is clearly not the only way of generating a plausible stochastic model. To check that the deterministic component of our stochastic models is contributing to their performance in a meaningful way, therefore, we also define two unrelated baseline models that are also stochastic in nature but derive their probabilities in different ways.

5.1.1. Simple Stochastic Model. The first baseline model is extremely simple. Again, let ϕ be the probability of a player contributing the same in round t as in $t-1$. But, this model estimates ϕ directly from the training data and does not use a deterministic model to do so. Let ϵ be the probability of contributing some amount uniformly distributed in the range $[1, 10]$. Again, ϵ is estimated from the training data. Finally, let $1 - \epsilon - \phi$ be the probability of contributing 0. Thus, this model is given by

$$E[c_{i,t}] = c_{i,t-1}\phi + 5.5\epsilon$$

Since this model estimates ϕ directly from the data, comparing the stochastic models that estimate ϕ using a deterministic algorithm to it, will allow us to understand how much predictive accuracy using a deterministic model adds.

5.1.2. Experience-Weighted Attraction. A different type of stochastic model, a version of which has been used previously to model agent behavior in public goods games [Janssen and Ahn 2006], is motivated by the notion of Experience-Weighted Attraction (EWA), proposed by Camerer and Hua Ho [1999], as a way to represent gradual learning in response to payoffs. The EWA model keeps track of two variables for every player: the number of observations N_t, and the A_{jt}, attraction of action j after period t. These attraction values represent the current tendency of playing the corresponding actions, and can therefore be converted directly into a probabilistic strategy. Updating has two steps. In step one, the experience is updated as $N_t = \rho N_{t-1} + 1$, where ρ is an experience decay parameter.

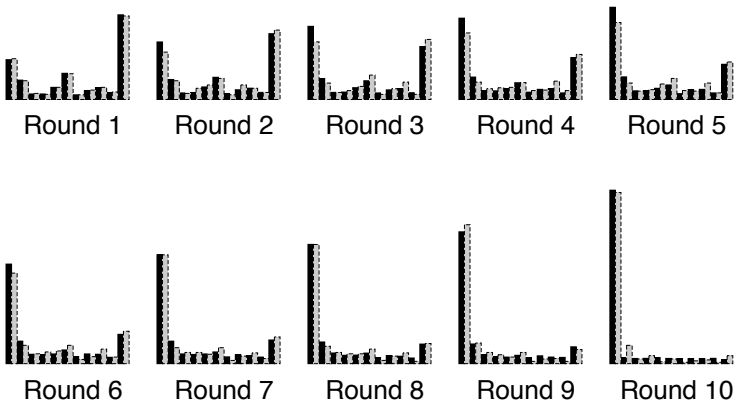

Fig. 4. Average action frequencies of the actual population (gray bars) versus the heterogeneous population, 2-factor discounted stochastic model (black bars).

In step two, the attractions are changed:

$$A_{j,t} = \frac{1}{N_t}(\phi N_{t-1} A_{j,t-1} + [\delta + (1-\delta) I(s_i, s_j)] U_i(c_{i,t}, \bar{c}_t))$$

where U is the utility function over the actions of the players in the neighborhood and I indicates whether the strategy was used at time t. The values ϕ and δ are parameters of the model respectively representing the decay of previous attractions and a calibration of actual versus imagined effects.

To convert the attraction values to a strategy, a logit function is typically used, which has its basis in the quantal response function and uses a temperature parameter λ:

$$P_{j,t+1} = \frac{e^{\lambda A_{j,t}}}{\sum_{k=1}^{M} e^{\lambda A_{k,t}}}.$$

Along with the experience decay ρ, this model contains four parameters that must be set by exhaustive brute-force search. Extra parameters are sometimes added to represent temporal decay or modify the utility function that might be shifted towards considering other players' utilities. Unfortunately, the entire parameter space must be searched simultaneously because of the ways that each parameter interacts with and depends on the others. As a result, fitting this model is exponential in the number of parameters.

5.2. Predicting the Full Distribution of Contributions

We now test the ability of the stochastic models to predict the distribution of the population's contributions using the same method described in Section 4.3; that is, we trained our models on the no-dummy treatment and compared to the human behavior data across each treatment in order to test for transferability across experiments. Figure 4 shows the results for the stochastic version of the discounted two-factor model, where as before we generated 20 independently generated populations, each playing one game with 24 players. Visually, the match is much better than for the deterministic case, an impression that is confirmed quantitatively in Table V, which shows the KL divergence between the true population behavior and the actions output by the simulated model. Clearly, the performance of the stochastic models is strikingly better than their deterministic counterparts. Moreover, the stochastic models using the deterministic subroutines outperform both the simple stochastic baseline model and also the EWA model, which performs relatively poorly. These results, in other words, justify our approach to constructing stochastic models: clearly the information

Table V. Evaluation of the simulated stochastic models' output distribution of contributions where individual models are trained on half the experiments and tested on the other half. KL divergence measures (non-symmetrically) the difference between the true data and the model output. Lower KL divergence represents higher accuracy.

Model name	# of Params	All Human	Altruistic Dummies	Free-Riding Dummies	Neighbor Dummies	Mean
Simple Stochastic	3	0.44	**0.47**	0.61	0.83	0.59
Stochastic 2-factor	4	0.34	0.68	0.53	0.81	0.59
Stochastic Discounted 2-factor	5	**0.20**	0.53	0.47	0.72	**0.48**
Stochastic Threshold	5	**0.20**	0.65	**0.43**	**0.71**	0.50
Stochastic Discounted Threshold	6	0.24	0.63	0.64	1.11	0.66
EWA	4	0.70	1.22	1.21	1.34	1.12

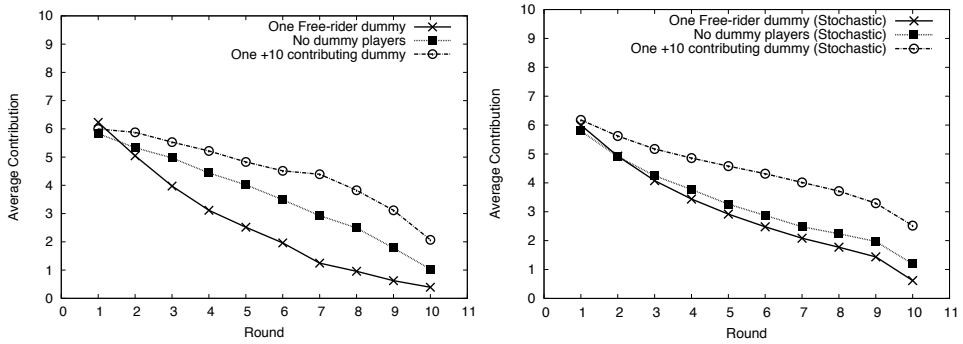

Fig. 5. Average contributions per round for the stochastic discounted two-factor model compared with empirical data.

contained in the deterministic predictions is useful. However, converting them to stochastic generative processes dramatically improves their ability to replicate the full distribution while slightly decreasing RMSE performance.

5.3. Selecting a Model

Table V shows that the stochastic discounted two-factor model exhibits the best overall performance with respect to the KL divergence. In addition, Figure 5 shows that this simulated model generates average aggregate contributions over time that are again visually similar to those from Suri and Watts [2011] and comparable to those generated by the deterministic version of the same model.[10] Finally, Table VI shows the transfer learning performance of each model; i.e. where we train each model on the all-human data and then evaluate it on a distinct experimental treatment. To maintain a fair comparison between all treatments, the test set for the all-human treatment that we use here is a second set of all-human experiments conducted by Suri and Watts several months after the experiments reported in SW [2011]. This set of experiments differed from the original all-human experiments in two respects: first, given the lapse in time relative to the churn rate of workers on AMT, the subject pool was largely distinct; and second, subjects were informed not only of the contributions and payoffs of their immediate network neighbors (the original treatment), but also those of their neighbors' neighbors, along with the connections between them. For

[10]Because the stochastic model makes predictions about the probability of a move, not the actual contribution, it is unclear how to evaluate its performance using the RMSE tests from the previous section. On the one hand, evaluating the expected contribution yields performance very close to the deterministic models, where the only effective difference lies in the additional noise term. On the other hand, first generating the full distribution of simulated moves and then scoring each move results in much higher RMSE. This is because the stochastic models predict extreme values and RSME penalizes heavily when one of these predictions is wrong. Since we are interested primarily in replicating the distribution of moves, and because the average of this distribution is also close the empirical average, we omit the RMSE tests.

Table VI. Evaluation of the simulated stochastic models' ability to transfer experience across different experimental setups. Actual behavioral data is compared to simulated output distribution of the population's contribution where individual models trained on the all human treatment and tested on the other treatments, including previously left-out all human experiments. Lower KL divergence represents higher accuracy.

Model name	# of Params	All Human	Altruistic Dummies	Free-Riding Dummies	Neighbor Dummies	Mean
Simple Stochastic	3	0.19	0.21	0.32	0.22	0.24
Stochastic 2-factor	4	0.13	0.24	0.19	0.17	0.18
Stochastic Discounted 2-factor	5	**0.08**	**0.16**	0.28	**0.15**	**0.17**
Stochastic Threshold	5	0.13	0.32	**0.17**	0.18	0.20
Stochastic Discounted Threshold	6	0.14	0.32	0.38	0.19	0.26
EWA	4	0.44	0.40	1.09	1.13	0.77

both reasons, we consider this set of all-human experiments to be a true out-of-sample test set, hence the all-human results in Table VI can be compared naturally with those of other treatments. Based on both within treatment (Table V) and between treatment (Table VI) performance, therefore, we select the stochastic discounted two-factor model as our preferred model for conducting the agent-based simulations, to which we turn next.

6. SIMULATING EMPIRICAL AGENT-BASED MODELS

Having selected the stochastic discounted two-factor model (SD2F) model as our candidate empirical agent-based model, we now return to our original motivation of deploying this model in the traditional manner of ABMs, namely as thought-experiments designed to generate new theoretical insights. Specifically, we first fit a customized model for all players, then construct a model population from which we draw agents to participate in a series of games, where now other parameters of the situation, such as the network size or structure, or the arrangement of player types to nodes in the network, can be varied systematically. In this way, we can explore a much broader range of the parameter space than would be possible with human subjects experiments.

6.1. Network Size

Recall that the main result of SW was their surprising finding that network topology had no significant impact on contributions. Because, however, the networks in question were relatively small ($N = 24$) it is possible that the lack of effect was due simply to insufficient variation in the path lengths, which for the connected networks varied only between 2 and 2.5. If true, then running the same experiments on much larger networks would allow for greater variation in the underlying structural features, and hence greater impact of structure on contribution. To test this hypothesis, we simulate our model populations on networks of increasing size, ranging from $N = 24$ to $N = 2400$. Interestingly, Figure 6A shows no dependency on size for the three fully connected topologies studied by SW: the connected cliques, the small-world network, and the random regular graphs. Figure 6B shows similar findings for three other natural topologies—an Erdös-Renyi random graph, a random graph with an exponential degree distribution, and a scale-free random graph[11]—suggesting that the conclusion of SW is robust with respect to network size.

6.2. Network Density and Variance of Degree

Another possible explanation for the absence of dependence on network structure in the SW experiments is that all players had equally sized neighborhoods, thus overlooking two additional sources of variation in network structure: the average degree d of the network; and the variance $var(d)$. Testing these dependencies, Figure 7 shows that although varying d has no impact (Figure 7A), increasing the variance of degree leads to lower contributions

[11]The exponential and scale-free random graphs were constructed using the configuration method [Newman 2003]

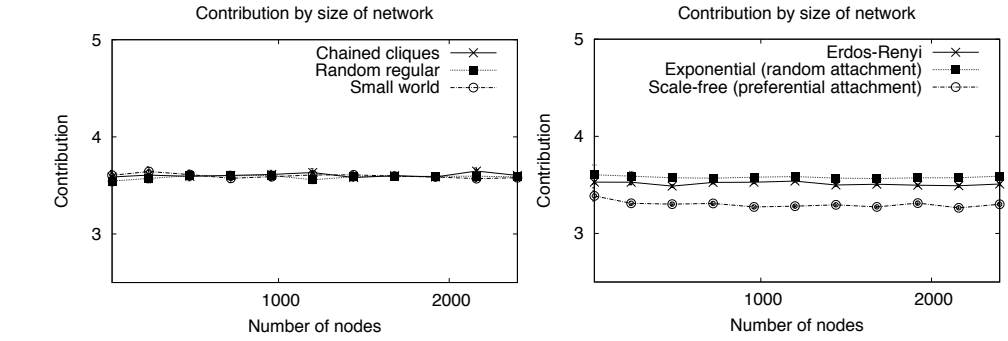

Fig. 6. Average game contributions vs. N for (a) the connected clique, small-world and random regular topologies studied by Suri and Watts [2011], and (b) Erdös-Renyi, exponential, and scale-free random graphs.

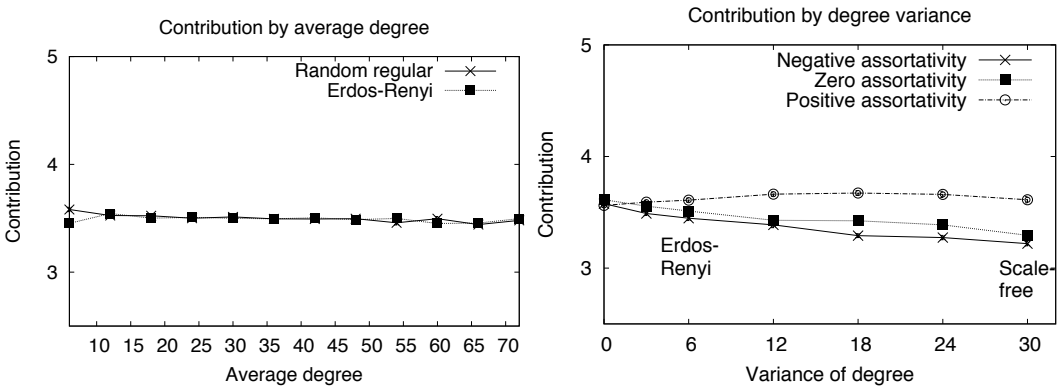

Fig. 7. Average game contributions for random graphs of size $N = 240$ vs (a) average degree k and (b) variance $var(k)$ of the degree distribution.

(Figure 7B, solid squares), consistent with the scale-free results from Figure 6B. On reflection, these results make sense: as explained in section 3, the network version of the public goods game effectively splits a player's contribution equally among its edges, hence all else equal nodes with many partners contribute less per partner than nodes with few partners. As long as all players have the same number of partners, the dependency on degree is symmetrical, hence the average density has no effect in the case of zero variance. Increasing the variance, however, breaks this symmetry, creating winners (high degree nodes) and losers (low degree nodes), where the latter are thereby more inclined to lower their contributions. Following this reasoning, we should expect networks with high variance to yield somewhat lower average contributions, as indeed we find in our simulations.

For similar reasons, we might also expect that contributions should depend on "degree assortativity" α, the tendency of high (low) degree nodes to be adjacent with other high (low) degree nodes [Newman 2003]. Indeed, Figure 7B shows that as α changes from negative (crosses) to positive (open circles), the dependency on variance decreases. Most of this effect, however, is due to the positive assortativity: that is, when high-degree players are more likely to be neighbors with each other (likewise for low-degree nodes), contributions increase, mitigating the effects of degree variance.

6.3. Correlations between Player Type and Node Degree

The dependency of contributions both on degree variance and also assortativity raises an additional possible source of dependency—namely that assigning more (or less) generous

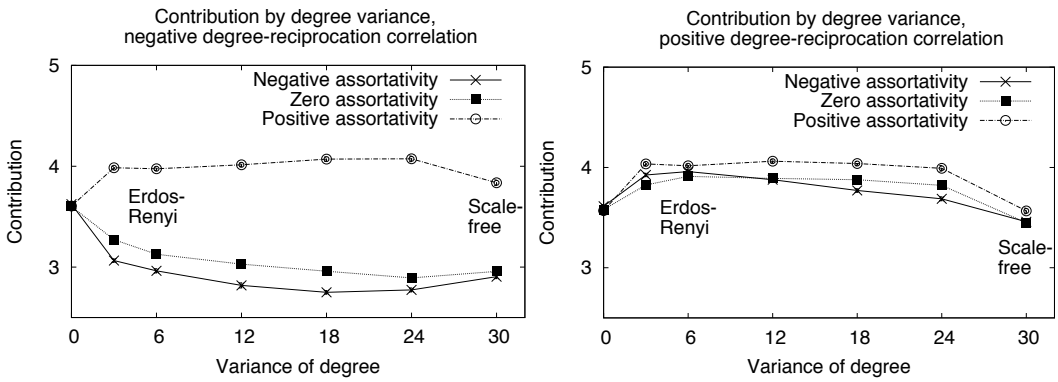

Fig. 8. Average game contributions vs. degree variance where player generosity and node degree are (a) negatively correlated, and (b) positively correlated.

players to nodes with higher (or lower) degrees might mediate or alternatively exacerbate the effects of breaking the degree symmetry. To check this hypothesis, we first define a new parameter $\rho = corr(\frac{\beta_1+\beta_2}{1-\gamma}, d)$, which quantifies the correlation between the overall generosity of agents in the SD2F model (as measured by their respective parameters) and the degree of a node in the network. As ρ is varied, that is, high degree nodes become either more ($\rho > 0$) or less ($\rho < 0$) likely to be generous. Figure 8 shows the same results as Figure 7B except where ρ is now strongly negative (left panel) and strongly positive (right panel) respectively. Interestingly, in networks with negative or no assortativity (α), a negative ρ lowers contributions further, while a positive ρ can by and large reverse the effects of negative assortativity. Positive assortativity, moreover, appears to compensate for increasing variance regardless of ρ. Overall, we conclude that both positive ρ and positive α can reverse the negative contributory effects of an unequal network, while negative values cause low contributions in an unequal network to drop still further.

7. CONCLUSIONS AND FUTURE WORK

We conclude by noting that our approach to constructing empirical agent-based models has both advantages and limitations. Among its advantages, we have presented a method for selecting among a collection of competing models, that is, by training and testing predictive models on experimental data. We have also demonstrated how questions left unresolved by the experiments in question can be investigated systematically by conducting simulations across a much broader range of parameters than is practical in human subjects experiments. In the current work, we confined our hypothetical exploration to varying parameters associated with the network, and relatedly the allocation of player types to nodes, assuming in effect that the population of players mirrored that of the experiments. In future work, one could easily do the opposite, fixing the network structure and altering the distribution of types in the population as represented by the frequencies in the cells of Table III.

A possible limitation of our approach, however, relates to our emphasis on empirically accurate models of agent behavior over the traditional emphasis among ABM researchers on cognitive plausibility. Aside from interpretability, cognitively plausible models would seem to have the advantage of generalizability—that is, one might expect the same model to work not only in the exact conditions tested in a given experiment, but across a broad range of conditions. By contrast, a cognitively implausible or otherwise uninterpretable model seems less likely to apply to novel conditions, even if it performs well on the training data. For example, our finding in the previous section that contribution levels do not change with network density seems highly dependent on the assumption—implicit in the behavioral

model—that the marginal per-capita return (MPCR) defined in the payoff function does not depend on degree. How would player behavior change if that assumption were violated? Because we have no model of how the agent is thinking about the game, or evaluating its utility, we cannot say.

Although the issue of generalizability is an important one, we would also note that even cognitively plausible models can fail in exactly the same way. Most obviously, this can happen when the circumstances are varied in a way not imagined by the modeler, but as noted in Section 1, models can fail even under the precisely the conditions imagined simply because humans agents violate the model assumptions in subtle but consequential ways. Thus while interpretability seems a desirable feature for ABMs, all else equal, we would continue to advocate empirical calibration, where the challenge of generalizability can be reframed as one of conducting the appropriate range of experiments.

This last point therefore motivates a need a for tighter integration between agent-based modeling and behavioral experiments. In the current work, that is, we have used data from behavioral experiments to identify an empirically accurate ABM. We then used this empirical agent-based model to explore the behavior of hypothetical human agents across a much broader parameter space than was possible in the experiments. A natural next step is to view these results as new hypotheses—about the effect of assortativity, for example, or lack of effect of density—to be tested in future experiments. These experiments, in turn, would no doubt lead to more accurate and generalizable EABMs, which could then be used to perform still more general simulations, followed again by more hypotheses and more experiments. In short, we advocate that future work should attempt to close this "hypothesis-testing loop" thereby allowing behavioral experiments and EABMs to complement and reinforce one another over time.

REFERENCES

Axelrod, R. 1984. *The Evolution of Cooperation.* Basic Books.

Axelrod, R. 1997. *The complexity of cooperation: Agent-based models of competition and collaboration.* Princeton University Press.

Bonabeau, E. 2002. Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences of the United States of America 99,* Suppl 3, 7280–7287.

Camerer, C. and Hua Ho, T. 1999. Experience-weighted attraction learning in normal form games. *Econometrica 67*, 827–874.

Castillo, D. and Saysel, A. 2005. Simulation of common pool resource field experiments: a behavioral model of collective action. *Ecological economics 55,* 3, 420–436.

Cederman, L. 1997. *Emergent actors in world politics: how states and nations develop and dissolve.* Princeton University Press.

Deadman, P. 1999. Modelling individual behaviour and group performance in an intelligent agent-based simulation of the tragedy of the commons. *Journal of Environmental Management 56,* 3, 159–172.

Deffuant, G., Neau, D., Amblard, F., and Weisbuch, G. 2000. Mixing beliefs among interacting agents. *Advances in Complex Systems 3,* 01n04, 87–98.

Epstein, J. and Axtell, R. 1996. *Growing artificial societies: social science from the bottom up.* MIT press.

Fischbacher, U., Gachter, S., and Fehr, E. 2001. Are people conditionally cooperative? evidence from a public goods experiment. *Economics Letters 71,* 3, 397–404.

Glance, N. S. and Huberman, B. A. 1993. The outbreak of cooperation. *Journal of Mathematical sociology 17,* 4, 281–302.

Heckbert, S., Baynes, T., and Reeson, A. 2010. Agent-based modeling in ecological economics. *Annals of the New York Academy of Sciences 1185,* 1, 39–53.

Holland, J. and Miller, J. 1991. Artificial adaptive agents in economic theory. *The American Economic Review*, 365–370.

Janssen, M. and Ahn, T. 2006. Learning, signaling, and social preferences in public-good games. *Ecology and Society 11,* 2, 21.

Janssen, M. and Ostrom, E. 2006. Empirically based, agent-based models. *Ecology and Society 11,* 2, 37.

Kreps, D. M., Milgrom, P., Roberts, J., and Wilson, R. 1982. Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic theory 27,* 2, 245–252.

Lazer, D. and Friedman, A. 2007. The network structure of exploration and exploitation. *Administrative Science Quarterly 52,* 4, 667–694.

Ledyard, J. 1995. Public goods: A survey of experimental research. In *Handbook of Experimental Economics*, J. H. Hagel and A. E. Roth, Eds. Princeton University Press, Princeton, NJ, 111–194.

Lopez-Pintado, D. and Watts, D. J. 2008. Social influence, binary decisions and collective dynamics. *Rationality and Society 20,* 4, 399–443.

Macy, M. and Willer, R. 2002. From factors to actors: Computational sociology and agent-based modeling. *Annual review of sociology*, 143–166.

Mason, W. and Watts, D. 2012. Collaborative learning in networks. *Proceedings of the National Academy of Sciences 109,* 3, 764–769.

Newman, M. E. 2003. The structure and function of complex networks. *SIAM review 45,* 2, 167–256.

Schelling, T. 1978. *Micromotives and Macrobehavior*. WW Norton.

Suri, S. and Watts, D. J. 2011. Cooperation and contagion in web-based, networked public goods experiments. *PLoS One 6,* 3.

Wang, J., Suri, S., and Watts, D. 2012. Cooperation and assortativity with dynamic partner updating. *Proceedings of the National Academy of Sciences 109,* 36, 14363–14368.

Williams, J. B. 1938. *The Theory of Investment Value*. Harvard University Press.

Wright, J. R. and Leyton-Brown, K. 2012. Behavioral game-theoretic models: A bayesian framework for parameter analysis. *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems*.