THE ROLE OF INSTITUTIONS IN THE EMERGENCE OF COOPERATION

REIMPLEMENTING AXELROD'S METANORMS GAME IN OSTROM'S INSTITUTIONAL ANALYSIS AND DEVELOPMENT FRAMEWORK (IAD)

Daniel Voigt Godoy[1], Sabino da Silva Pôrto Junior[2]

ABSTRACT

In this paper, we investigate the role of institutions in the emergence of cooperation and the establishment of a social norm, reimplementing a modified version of Axelrod's Metanorms Game under Ostrom's Institutional Analysis and Development (IAD) framework using an agent-based simulation implemented in NetLogo and R.

In our model, agents live in a torus and harvest a common-pool resource. Their strategies for harvesting are defined by their two attributes, boldness and vengefulness. The space of possible attributes is the cultural dimension of the model. The agents are socially connected through a scale-free network and can modify their attributes through a mechanism of cultural dissemination. They can observe other agent's actions either in their vicinity or through their social connections. They can choose to punish and meta-punish observed defections.

An institution is modeled as a special agent that lives in the cultural dimension alone and is capable of influencing other agent's attributes. It has a power and range of influence which will determine its probability of being chosen for imitation by other agents. Institutions can be either exogenous or endogenous. In the former, its values are constant throughout the whole simulation. In the latter, its values are regularly recomputed based on the attributes of the population. In our simulations, we introduce different types of institutions in populations of defectors with different densities and observe the impact the institutions have on transforming the population's attributes.

After the first 1,000 rounds, we found that a social norm, after first established, is more likely to collapse or become unstable if institutions are either weak or have a limited range of influence. For these specific cases, we extended the simulation period to 10,000 rounds. We found that exogenous institutions are unable to transform the population's attributes in the long-run. Endogenous institutions, on the other hand, are more effective in establishing a stable social norm. The existence of punishment through social control is sufficient for the establishment of a social norm in the majority of the cases. Moreover, a norm is more often established if agents have a bias towards institutional values. We also

[1]Master in Applied Economics - PPGE/UFRGS. Email: dvgodoy@gmail.com
[2]Professor - PPGE/UFRGS. Email: sabino@ppge.ufrgs.br

found that low density populations are fast to establish a social norm, yet its stability is dependent on the existence of social control, while higher population densities establish a stable social norm after a longer number of generations.

The conditions of low density population, institutions with limited power and influence, social control and agents biased towards institutional values are akin to those found in Ostrom's field experiments. Under these conditions, we found that a stable social norm is established.

Keywords: Axelrod’s Metanorms Game, Institutional Analysis and Development (IAD), Emergence of Cooperation, Agent-based Simulation, Tragedy of the Commons.

JEL Classification: C73, D02, Z13

1. INTRODUCTION

The emergence of cooperation in different types of social dilemmas and its necessary underlying conditions have been the subject of several studies in the last decades. The canonical example of social dilemma with multiple actors is the Tragedy of Commons, introduced by Garret Hardin in 1968, which addresses the trade-off between exploring and provisioning a common-pool resource. This trade-off was investigated extensively by Ostrom (1990, 2000, 2005) and Ostrom et al. (1994). In her work, it is often shown that small communities and groups of individuals often devise and enforce its own access rules to a common-pool resource, effectively addressing the social dilemma. Hence, Ostrom (1990) proposes a local regulation of access and usage of a resource, developed by those who effectively harvest it. Local monitoring and sanctioning, managed by the community itself, instead of an external authority, are key characteristics of successful communities.

Our goal in this paper is to analyze the role institutions play in the emergence of cooperation and the establishment of a social norm in the context of a social dilemma such as the Tragedy of Commons. We use Agent-Based Modeling[3] to simulate an initial population of defector agents playing a modified version of the Metanorms Model developed by Axelrod (1986) in his seminal work "*Evolutionary Approach to Norms*". Our model differs from the original one in some key aspects, though: it does not presume interactions between all agents at every turn; instead, it uses a superposition of two networks on which agents can interact with each other. In the geographic network, agents can observe its neighbors harvesting behavior and eventually punish them. In the social network, agents can learn from each other through processes of imitation and cultural dissemination. By including more components into the model dynamics, namely, biophysical conditions, attributes of the community and rules-in-use; our model endogenizes the external variables of the Institutional Analysis and Development (IAD) framework proposed by Ostrom (2005, 2007).

This paper is organized as follows: Section 2 presents a brief overview of previous works. We introduce our model and its dimensions: cultural, geographic, social and common-pool resource, detailing its implementation and parameterization for the purpose of the simulation, in Section 3. Section 4 presents the results of the simulations and Section 5 presents our conclusions and proposals for future work.

[3]The goal of Agent Based Modeling is to simulate actions and interactions between autonomous agents in order to determine its aggregated effect on a system as a whole.

2. PREVIOUS WORK

Social dilemmas such as the Tragedy of Commons are characterized by resources that are both non-excludable – restricting access to it is difficult – and rivalrous – once it has been used by one individual, it becomes unavailable to others. The harvesting of common-pool resources was investigated by Ostrom et al (1994) in field experiments. One of the key findings was the existence of multiple types of individuals, each one having a different propensity to reciprocate towards achieving the benefits of collective action. According to Ostrom, it is necessary to determine how potential cooperators can signal each other in order to develop institutions that foster conditional cooperation, instead of destroying it. The agents were classified into two different types, based on empirical studies: i) *rational and self-interested*, representatives of neoclassical economics; and ii) *norm-users*, divided into two subgroups – *conditional cooperators* and *willing to punish*. The former subgroup drives the high levels of cooperation often observed in repeated games, while the latter prevents opportunistic behavior (free-rider), creating underlying conditions for collective action.

For *norm-user* agents, its behavior and ability to learn is key. Thus, models of cultural evolution were developed (Bowles, 2004) based on studies of human populations where traits could be transmitted not only genetically but also by learning. These behavioral rules can either succeed, and spread through the population, or fail, and be confined to small niches or completely disappear. In this case, the *dramatis personae* of social dynamics are not the individuals anymore, but its behavioral rules. The actions of individuals are only important to the extent they contribute to the success or failure of the rules.

The mechanism of behavior transmission is key to Axelrod's model of cultural dissemination (1997). In his model, the author defines culture as a set of individual attributes subject to social influence – culture is something individuals learn from each other. The proposed model presumes two conditions: i) individuals have higher propensity to interact with other individuals that share more similar cultural attributes; ii) these interactions are likely to make their attributes even more similar, therefore making future interactions more likely as well. A concise description of the model's implementation can be found at Izquierdo et al. (2009).

Chang and Evans (2005) define institutions as systematic patterns of shared expectations, taken-for-granted assumptions, acceptable norms and routines of interaction,

indicating that these elements effectively shape motivations and behaviors of a group of socially connected actors. According to the authors, an institutional approach must consider how institutions shape behavior and economic results, creating an understanding of how institutions are developed and change over time. This approach goes beyond the "institutions as constraints" model, considering institutions as enabling instruments to achieve goals that require supra-individual coordination, being constitutive of the interests and world-views of economic actors. Moreover, the authors argue that agents operating under an institutional arrangement are prone to internalize institutional values and change their behavior accordingly.

The Institutional Analysis and Development (IAD) framework, developed by Elinor and Vincent Ostrom, provides a systematic approach to analyze institutions governing actions and results in a collective arrangement (Ostrom, 2007). In this framework, institutions are considered a set of prescriptions and restrictions used by individuals to organize repeated interactions, such as norms or rules (Ostrom, 2005). The IAD framework focuses on the "action arena", composed by actors and "action situations", where actors' decisions, interaction patterns and corresponding social choices happen. Moreover, the "action arena" can also be influenced by three types of external variables: i) institutions or rules in use; ii) attributes of the community; and iii) biophysical conditions (Ostrom, 2005, 2011).

The "action situation" is the social space where individuals interact with each other, exchanging goods and services, resolving issues, etc. Most of theoretical work considers the cognitive and motivational structures of individuals as given. The author proposes two refinements: i) to include factors that affect the structure of the situation; and ii) to investigate the way the "action situation" evolves over time according to previous outcomes and changes in perception and strategy (Ostrom, 2011).

The actors, while making decisions in different situations, collect information about the structure of the situation itself before choosing an action and usually get a feedback afterwards. Learning from its own mistakes, as well as from others, allows for revising one's mental models. Culture, too, can affect the mental models used by an actor in a given situation. The word "culture" is often used to describe shared values of a community and can be also seen as a reduced set of mental models shared by a group of individuals (Ostrom, 2005). As for the biophysical conditions, one can consider both the flow of resources inside the arena, as well as its attributes of excludability and rivalry (Ostrom, 2005, 2011).

Robert Axelrod is one of the first authors to make use of computer simulations to investigate the emergence of social norms that promote cooperation between agents. In his work, "*An Evolutionary Approach to Norms*" (Axelrod, 1986), the author investigates the emergence and establishment of behavioral norms in a game played by agents with bounded rationality, demonstrating the necessary underlying conditions for a norm to evolve and get established. The author chose to use the behavioral definition of a social norm: "*a norm exists in a given social setting to the extent that individuals usually act in a certain way and are often punished when seen not to be acting in this way*". Since empirical evidence suggests that norms change through trial-and-error, the author used the evolutionary approach that gives name to his work: "*...what is chosen at any specific time is based upon an operationalization of the idea that effective strategies are more likely to be retained than ineffective strategies*". The presence of agents with bounded rationality and an evolutionary approach to beliefs and norms can also be found in works from several authors (Arthur, 1999, Dawes, 1980, Kollock, 1998, Ostrom 2000).

In the "Norms Game", agents have two attributes: boldness and vengefulness. Both attributes could have discrete values between 0 and 7 assigned to them, that is, three bits of information each. The boldness attribute (B) represents the agent's propensity to defect, while the vengefulness attribute (V), its propensity to punish an observed defection. Every round, the agents could choose between cooperation and defection. The agent's choices are influenced by its own attributes and its chance of being observed (S), a random variable drawn from a uniform distribution. Hence, whenever B > S, the agent defects, receiving a payoff corresponding to the temptation to defect (T = 3) and inflicting a negative payoff to all other agents (H = -1). A defection is observed by any other agent with a probability of S. The observer chooses to punish the defector with a probability given by its attribute V. Should the observer punish the defector, the former incurs into a cost (E = -2) and the latter, into a penalty (P = -9). A full round is complete whenever every agent had a chance to defect and observe and, eventually, punish an observed defection. A sequence of four rounds constitutes a generation. For each generation, agents with cumulative payoffs greater than one standard deviation above average are selected to have two offspring. Conversely, those agents one standard deviation below average have no offspring at all. There is also a 1% chance of mutation in each bit of an agent's attributes. The final population is adjusted to remain constant over time (20 agents), although this procedure is not detailed by the author.

We can describe Axelrod's model's spatial structure as a complete graph where every

agent is connected to every other agent. This is an important limitation of the model, as we believe the choice of a complete graph is due to the small number of agents in the simulation, given the limited computing power at that time. Moreover, the possibility of mutations allows the model to be expressed, in theory, as a Markov transition matrix, guaranteeing its convergence to a stable state.

In his simulations, Axelrod used a population of 20 agents with random initial attributes. A sequence of 100 generations constituted a "run" and the author performed five runs in total. Unsurprisingly, given the reduced number of runs, the results at the end of each run were quite different from one another. Nevertheless, it was possible to observe some common trends: there was an initial reduction in the overall level of boldness, followed by a reduction in vengefulness, which led to the resurgence of higher levels of boldness in the stable state – no norm was established. Axelrod (1986) realized that there's no incentive to punish an observed defection, since it carries a cost and then he proposed a *metanorm*, that is, the possibility of being punished for not punishing an observed defection.

The dynamics of the "Metanorms Game" are the same as in the original "Norms Game", except for the newly introduced possibility of 2^{nd} order punishment: meta-defectors get a penalty (MP = -9) and meta-punishers have a cost (ME = -2). The vengefulness attribute was used for both defectors and meta-defectors. By introducing metanorms into the game, Axelrod obtained quite different results: if the initial overall level of vengefulness was high enough, the model would converge to a stable state having high levels of vengefulness and, consequentially, high levels of cooperation – a norm was established.

Izquierdo and Galán (2005) considered a norm as established whenever the population's average boldness and vengefulness were inside the intervals [0, 2] and [5, 7], respectively. A norm was considered collapsed if those averages were lying inside the intervals [6, 7] and [0, 1].

Recent works challenge some of the assumptions used in Axelrod's model. Mahmoud et al. (2012) raise the problem of omniscient agents, as every agent in Axelrod's model had complete information about every other agent's private strategies and actions. The authors propose replacing the evolutionary approach for a learning-based approach while restricting the possibility of metapunishment to agents that observed the defection in the first place.

Moreover, the choice of network topology in Axelrod (1986) – complete graphs – and

Mahmoud et al. (2012) – lattice networks – are not representative of real interactions between individuals in a society. Thus, Galán, Latek and Rizi (2011) have analyzed the results of playing the "Metanorms Game" in different network topologies, generated using algorithms developed by Barabási-Albert, by Watts and by Erdös-Rényi. The authors found that the network structure changes the effectiveness of metanorms.

3. PROPOSED MODEL

In this paper, we propose a modified version of the "Metanorms Game" by Axelrod (1986). We introduce several dynamics into the model, namely: non-omniscient agents (Mahmout et al., 2012), two networks of interaction – geographic and social (Canova, 2011), learning by imitation (Macy, 1991), culture dissemination (Axelrod, 1997), common-pool resource dynamics (Bravo, 2011) and, more importantly, the role of institutions (Bowles, 2004, Chang and Evans, 2005).

Our model can be split into four different dimensions, each having its own dynamics: the cultural space, where learning by imitation and institutional influence processes happen; the social space, a scale-free network connecting the agents; the geographic space, a torus where agents roam; and the common-pool resource, represented by biomass that can be harvested by the agents.

Through incorporating biophysical conditions and community dynamics into the model, we are endogenizing the external variables of the Institutional Analysis and Development (IAD) framework by Ostrom (2005, 2007). Figure 1 depicts how the elements of our model relate to each other.

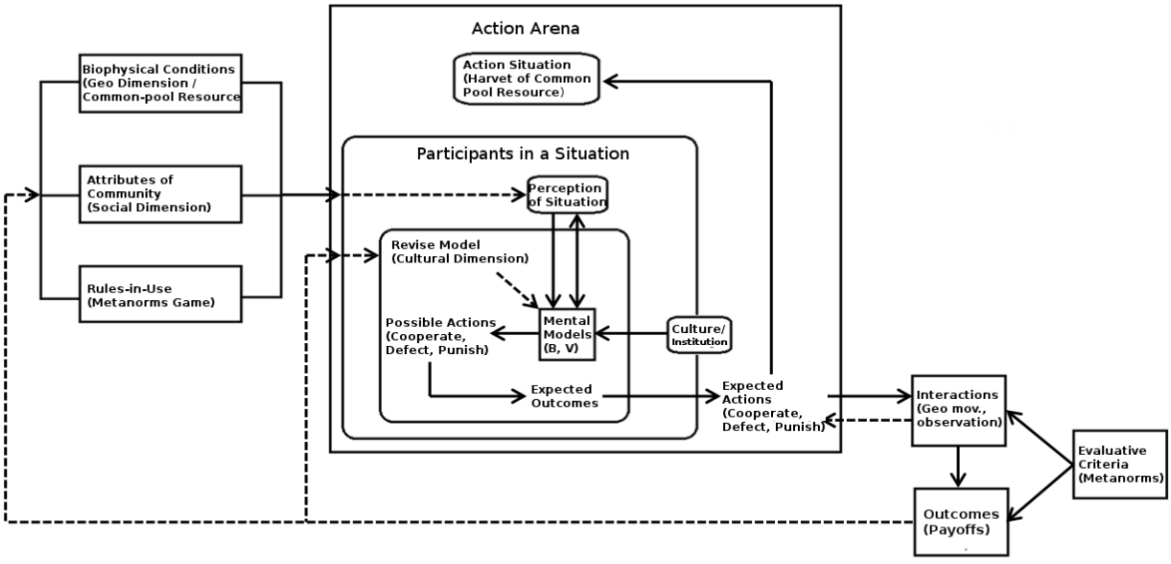

Figure 1. Elements of our model in the IAD framework. Elaborated by the authors.

We use an agent-based social simulation approach, where agents follow a set of relatively simple rules and repeatedly interact with each other and with the environment to create emergent patterns of behavior. We use open source tools – NetLogo (Wilensky, 1999) –

developed for agent based simulations – and R (R Core Team, 2010) – a statistical package that can be integrated with NetLogo.

3.1 AGENTS' ATTRIBUTES AND ACTIONS

We start by introducing the agent's attributes, the modifications to the original "Metanorms Game" and a detailed description of the dynamics of each dimension of our model, followed by its parameterization, implementation details and results.

There are N agents in our model and, as in the original "Metanorms Game", they possess two attributes: boldness (B) and vengefulness (V). Each attribute has 3 bits and may take integer values in the interval [0, 7]. An initial population of defectors is thus defined by a population of agents having values (7, 0) for its attributes (B, V) at the start of each simulation. At each round, each agent has its chance of being observed (S) randomly drawn from an uniform distribution in the interval [0, 1]. We define three column vectors of size N, for both agent's attributes and its chance of being observed:

$$B = \begin{bmatrix} b_1 \\ \vdots \\ b_N \end{bmatrix} \quad V = \begin{bmatrix} v_1 \\ \vdots \\ v_N \end{bmatrix} \quad S = \begin{bmatrix} S_1 \\ \vdots \\ S_N \end{bmatrix} \tag{1}$$

The set of possible actions and corresponding payoffs for each agent, as in the original "Metanorms Games" by Axelrod (1986), is given by its attributes and its chance of being observed. Each agent can choose to cooperate or defect and to punish or not.

Let i, j, and k be agents with attributes, respectively, (b_i, v_i), (b_j, v_j) and (b_k, v_k). Also, let S_i, S_j. s_{ij}, s_{jk}, p_j and p_k be random variables uniformly distributed in the interval [0, 1]. The actions taken by each agent in a given round can be defined as follows:

$$\begin{aligned}
Def_i &= \begin{cases} 1, if\ b_i > 7\ S_i \\ 0, otherwise \end{cases} \qquad Obs_{ij} = \begin{cases} 1, if\ s_{ij} < S_i \\ 0, otherwise \end{cases} \\
Enf_j &= \begin{cases} 1, if\ v_j > 7\ p_j \\ 0, otherwise \end{cases} \qquad Pun_{ij} = \begin{cases} 1, if\ Enf_j = 1\ \wedge Obs_{ij} = 1\ \wedge Def_i = 1 \\ 0,\ otherwise \end{cases} \\
MetaDef_j &= \begin{cases} 1, if\ Enf_j = 0\ \wedge Obs_{ij} = 1\ \wedge Def_i = 1 \\ 0, \quad otherwise \end{cases} \qquad Obs_{jk} = \begin{cases} 1, if\ s_{jk} < S_j \\ 0, otherwise \end{cases} \\
MetaPun_{jk} &= \begin{cases} 1, if\ Enf_k = 1\ \wedge Obs_{jk} = 1\ \wedge MetaDef_j = 1 \\ 0, otherwise \end{cases}
\end{aligned} \tag{2}$$

where S_i and S_j are, respectively, the probabilities of agents i and j being observed by any other agent; s_{ij} and s_{jk} are, respectively, the probabilities of agent j observe agent i and

agent *k* observe agent *j* and; p_j and p_k are, respectively, the probabilities of agents *j* and *k* being punishers. The dynamics of the Metanorms Game are depicted in Figure 2 below:

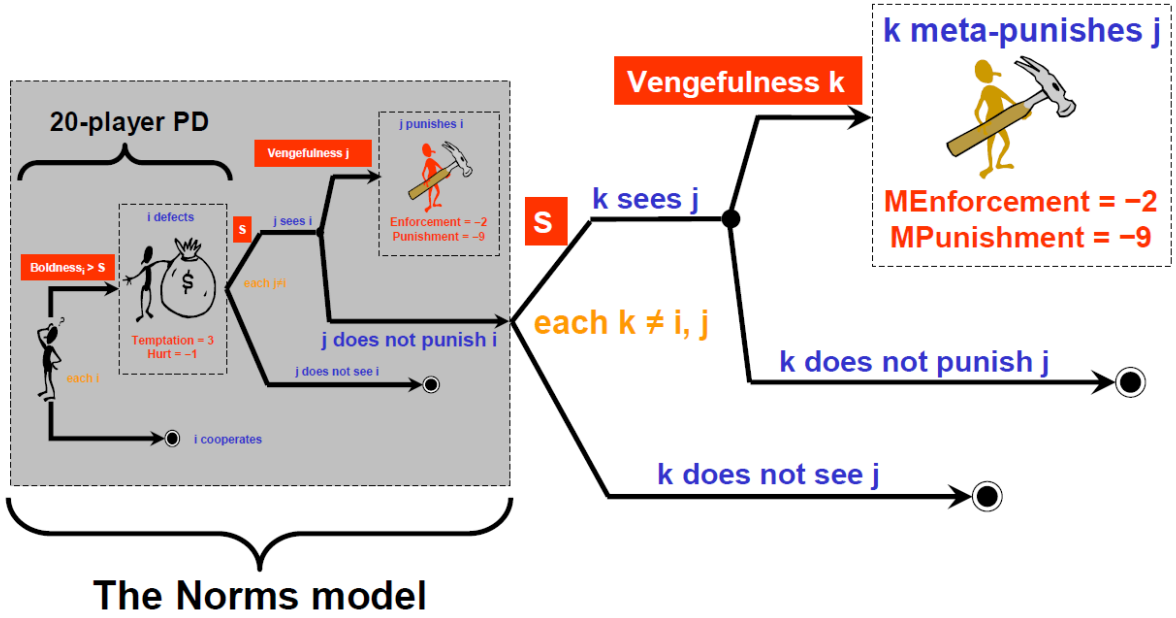

Figure 2. Metanorms Game schematics. Figure from Izquierdo e Galán (2005).

3.2 GEOGRAPHIC SPACE

The geographic space is defined as a torus-shaped lattice of size L, in order to make all cells equivalent and avoid boundary effects. Each cells can be occupied by a single agent only and the population density is given by $\rho = N / L^2$. The neighborhood of each cell is given by a Moore's neighborhood, that is, the eight cells in its vicinity. It can also be defined as the set of cells within a unit distance of Tchebycheff[4]. The geographical distance between two agents *i* and *j* is given by its Euclidean distance in the torus:

$$DistGeo_{ij} = \sqrt{(min\{|x_i - x_j|, L - |x_i - x_j|\})^2 + (min\{|y_i - y_j|, L - |y_i - y_j|\})^2} \quad (3)$$

The normalized geographical distance is then given by:

$$NDistGeo_{ij} = \frac{DistGeo_{ij}\sqrt{2}}{L} \quad (4)$$

3.3 ACTIONS AND OBSERVATIONS

In our model, agents can interact with each other in two different spaces: geographic

[4]The Tchebycheff distance is defined as the maximum distance between two points in a given coordinate system.

and social, as in Canova (2011). In the geographic space, agents can roam, harvest a common-pool resource and, more importantly, observe its neighbor's actions. Thus, we need to adjust the definition of an observed defection as follows:

$$NDG_{ij} = \begin{cases} 1, if\ NDistGeo_{ij} \leq r \\ 0, otherwise \end{cases} \quad Obs_{ij} = \begin{cases} 1, if\ s_{ij} < S_i\ \wedge\ NDG_{ij} = 1 \\ 0, otherwise \end{cases} \tag{5}$$

where $NDistGeo_{ij}$ is the normalized geographical distance, r is the radius of observation, S_i is agent i's probability of being observed and s_{ij} is a random draw from an uniform distribution in the interval [0,1] for a pair of agents i and j.

It is also possible to observe the actions of agent's who are socially connected, in a mechanism akin to "social control":

$$Social_{ij} = \begin{cases} 1, if\ agents\ i\ and\ j\ are \\ \quad socially\ connected \\ 0,\ otherwis \end{cases} \quad Obs_{ij} = \begin{cases} 1, if\ s_{ij} < S_i\ \wedge\ NDG_{ij} = 1 \\ \quad \wedge\ Social_{ij} = 1 \\ 0, otherwise \end{cases} \tag{6}$$

In both types of observation, geographic and social, the probability is given by the random variable S, uniformly distributed in the interval [0, 1], drawn at each round for each agent. The 2nd order defection (observe a defection and not punishing it), a metadefection, requires that the 2nd order punisher, the metapunisher, had observed the original defection as well. Thus, we address the issue of omniscient agents raised by Mahmoud et al. (2012). The metapunishment is expressed as follows:

$$MetaPun_{jk} = \begin{cases} 1, if\ Enf_k = 1\ \wedge\ Obs_{jk} = 1\ v\ MetaDef_j = 1\ \wedge\ Obs_{ik} = 1 \\ 0, otherwise \end{cases} \tag{7}$$

A metapunishment only happens if the agent k is a punisher and observed both agent j (metadefector, since it observed agent's i defection and did not punish it) and agent i (original defector).

3.4 COMMON POOL-RESOURCE DYNAMICS

The common-pool resource dynamics we use in our model are similar to the one proposed in Bravo (2011), a mature forest, with small modifications. In our model, each cell in the geographic space has an attribute b, indicating the amount of biomass it contains, which can be harvested by the agent in it. The attribute b can take values in the interval [0, b_{max}], where b_{max} is the maximum amount of biomass possible. At the start of a simulation, the geographic space is initialized with the representation of a mature forest, each cell containing

a value b randomly drawn from an uniform distribution in the interval [½ b_{max}, b_{max}]. The growth process, unlike Bravo (2011), is given by a logistic equation:

$$b_{growth} = b_{rate} b(1 - \frac{b}{b_{max}}) \quad (8)$$

where b_{rate} is the growth rate.

The logistic equation is commonly used to model self-limited population growth, where the growth rate is a function of both the already existing population and available resources. The first term of the equation represents the unimpeded population growth. As population grows, however, the second term slows growth down, as individuals start competing for available resources and interfere with each other's reproduction, until equilibrium is reached. The parameter b_{max} is the carrying capacity and determines the equilibrium level.

If a cell is completely deforested, that is, if $b = 0$, the probability of reforesting is given by its neighboring cells following the expression:

$$p = p^* \frac{N+1}{k+1} \quad (9)$$

where $p^* = 0{,}05$ is the basic reforesting probability, N is the number of neighboring cells with $b > 0$ and k is the total number of neighboring cells ($k = 8$ for a Moore's neighborhood). This way, we simulate forests' natural recovering ability through seed preservation on the soil or seed dispersion through animal vectors or wind, for instance. The renewed level of biomass of a reforested cell is given by the average of biomass present in its k neighboring cells.

3.5 PAYOFFS

The payoffs in our model are different from the original game, since we included common-pool resource dynamics into it. In our model, a defection is characterized by the excessive harvesting of the common-pool resource. Cooperation, on the other hand, means harvesting it at sustainable levels. Therefore, the damage inflicted by defectors is only materialized in the long-run through the degradation of the environment, instead of an immediate cost, as in the original model. The punishment of a defection, though, remains a determined and direct sanction.

The agent's payoff is given by its choice of harvesting level (defecting or cooperating) and the amount of available biomass in the cell where the agent is. At each round, every agent runs into a fixed cost (C) for harvesting. If the available biomass is greater than the fixed cost, that is, if b >= C, agents harvest it according to its choice: defecting agents harvest either an "excessive" amount (T) or all biomass, whatever is less; while cooperating agents harvest a sustainable amount (R) or all biomass, whatever is less. The harvested biomass is added to the agent's payoff. But, if the available biomass is less than the fixed cost, that is, if b < C, the agent does not receive a payoff and it tries to move to a vacant neighboring cell containing more available biomass than the fixed cost. If no neighboring cell satisfies this condition, the agents moves randomly to any vacant cell in its neighborhood.

The penalty (P) imposed for a punished defection, as well as the cost (E) incurred by the punisher, are also expressed in terms of biomass. The same holds for penalties (MP) and costs (ME) for 2nd order defections. The overall payoff of an agent in a round of simulation is given by:

$$Payoff_i = R + Def_i(T - R) + \sum_{\substack{j=1 \\ j \neq i}}^{n} Pun_{ij}E + \sum_{\substack{j=1 \\ j \neq i}}^{n} Pun_{ji}P + \sum_{\substack{k=1 \\ k \neq i}}^{n} \sum_{\substack{j=1 \\ j \neq i,k}}^{n} (1 - Pun_{jk})MPun_{ij}ME + \sum_{\substack{k=1 \\ k \neq i}}^{n} \sum_{\substack{j=1 \\ j \neq i,k}}^{n} (1 - Pun_{ik})MPun_{ji}MP - C \quad (10)$$

where $T = 0.30$, $R = 0.10$, $E = ME = -0.20$, $P = MP = -0.90$ are the payoffs of the model, $C = 0.05$ is the fixed cost for harvesting, n is the total number of agents and

$$Def_i = \begin{cases} 1, if\ agent\ i\ defects \\ 0, if\ agent\ i\ cooperates \end{cases} \quad Pun_{ij} = \begin{cases} 1, if\ agent\ i\ punishes\ agent\ j \\ 0, if\ agent\ i\ does\ not\ punish\ agent\ j \end{cases}$$

$$MPun_{ij} = \begin{cases} 1, if\ agent\ i\ metapunishes\ agent\ j \\ 0, if\ agent\ i\ does\ not\ metapunishes\ agent\ j \end{cases} \quad (11)$$

3.6 CULTURAL SPACE

The evolution of strategies, contrary to the original model, does not follow an evolutionary approach of selection and mutation. In our model, it follows a process of cultural dissemination and learning by imitation. Given each agent's social connections, they are more likely to become similar to those whom they are already culturally more similar (Axelrod,

1997) or to those whom they perceive as successful (Macy, 1991). Moreover, we consider the role of institutional values (Chang and Evans, 2005) by defining different types of institutions and its relative importance in the cultural space.

The cultural space is defined by the Cartesian product of both attributes: boldness and vengefulness. Since these attributes can only take integer values in the interval [0, 7], the cultural space is a lattice of size 7. Analogously to the geographic space, we can define the cultural distance between two agents in the cultural space using its Euclidean distance:

$$DistCult_{ij} = \sqrt{(b_i - b_j)^2 + (v_i - v_j)^2} \tag{12}$$

The normalized cultural distance is then given by:

$$NDistCult_{ij} = \frac{DistCult_{ij}}{7\sqrt{2}} \tag{13}$$

Both dynamics, of cultural dissemination and learning by imitation, happen inside the cultural space. Therefore, agents "move" in the cultural space whenever they modify their own attributes through one of the two aforementioned processes. This movement happens in two steps: first, by selecting a target agent, then by choosing to "move" (or not) in the cultural space. The move itself can represent an imitation, meaning, reducing the cultural difference between the agents; or a rejection, increasing it.

The probability of selecting a target agent as "cultural role model" to be either imitated or rejected is a function of how "visible" the latter is to the former. We define this visibility between two agents as proportional to their differences in payoff and inversely proportional to their geographic distance. We assume that very successful agents (large positive payoff difference) are more likely to be select for imitation and, conversely, very unsuccessful agents (large negative payoff difference) are more likely to be selected for rejection. In both cases, agents with larger absolute payoff differences are more visible. Similarly, agents that are geographically closer (lower distance) are more visible than those farther away. Thus, the visibility of agent B, from agent A's perspective, is then given by:

$$Vis_{AB} = \frac{P_B - P_A}{NDistGeo_{AB}} \tag{14}$$

where P_A and P_B are the agents' payoffs and $NDistGeo_{AB}$ is the normalized geographic distance between them.

The set of target candidates, to be either imitated or rejected, is comprised of both agents within a geographic radius and socially connected to the subject agent that is having its attributes updated. The probability of selection assigned to each one of the target candidates is proportional to its relative importance, given the overall sum of computed visibilities for all target candidates. Moreover, there is a small chance (ω) of selecting, by mistake, an unsuccessful agent to imitate or a successful agent to reject.

Let's first consider the probability of selecting a successful target agent to be imitated, which is given by:

$$Prob(j) = \begin{cases} \dfrac{(1-\omega)Vis_{ij}max\{1, Social_{ij} + Geo_{ij}\}}{\sum_{k=1}^{n} Vis_{ik}\,(Vis_{ik} > 0)max\{1, Social_{ik} + Geo_{ik}\}}, if\ Vis_{ij} > 0 \\ \dfrac{\omega}{\sum_{k=1}^{n}(Vis_{ik} < 0)max\{1, Social_{ik} + Geo_{ik}\}}, otherwise \end{cases} \quad (15)$$

where ω is the chance of selecting the wrong type of target agent, Vis_{ij} is the visibility of agent j from agent i 's perspective, r is the geographic radius and

$$Social_{ij} = \begin{cases} 1, if\ agents\ i\ and\ j\ are\ socially\ connected \\ 0, otherwise \end{cases} \quad (16)$$

$$Geo_{ij} = \begin{cases} 1, if\ NDistGeo_{ij} \leq r \\ 0, otherwise \end{cases} \qquad where\ r\ is\ the\ neighborhood's\ radius$$

The same expression holds for the probability of selecting an unsuccessful target agent to be rejected, except for flipping the inequalities in both denominators of equation (15).

Once a target agent is selected, the second step involves defining the probability of the subject agent "moving" or not in the cultural space to become more culturally similar or dissimilar to a target agent to be imitated or rejected, respectively. Gino, Ayal and Ariely (2009) showed that individuals imitate their peers, even when they engage in dishonest behavior. Axelrod (1997) proposed that culturally similar agents are more likely to interact with each other and become even more similar in a virtuous cycle. Hence, we assume this probability to be inversely proportional to the normalized cultural distance between the agents, thus considering effects of both group identity and dissemination of culture. A function commonly used to model social learning is the sigmoid function:

$$\frac{1}{1+\exp(-\beta(P_j - P_i))} \quad (17)$$

where $\beta \geq 0$ is the strength of imitation, indicating how strongly the agent's decision is based on the payoff difference (P_j - P_i). For $\beta \rightarrow 0$ (or $P_j = P_i$), the probability of imitation is purely random. For small values of β, payoff-based imitation is weak and more successful agents are just slightly more likely than others to be imitated. For $\beta \rightarrow \infty$, a more successful agent is always imitated and, a less successful one, never (Sigmund et al., 2010).

In our model, however, we replace the payoff-based imitation with a cultural-based one such that the probability of imitation is given by:

$$\frac{1}{1+\exp(\beta(NDisCult_{ij}-0{,}5))} \tag{18}$$

where $|\beta|$ represents the strength of the imitation and positive and negative values of β are used for imitation and rejection, respectively. Figure 3 illustrates both processes.

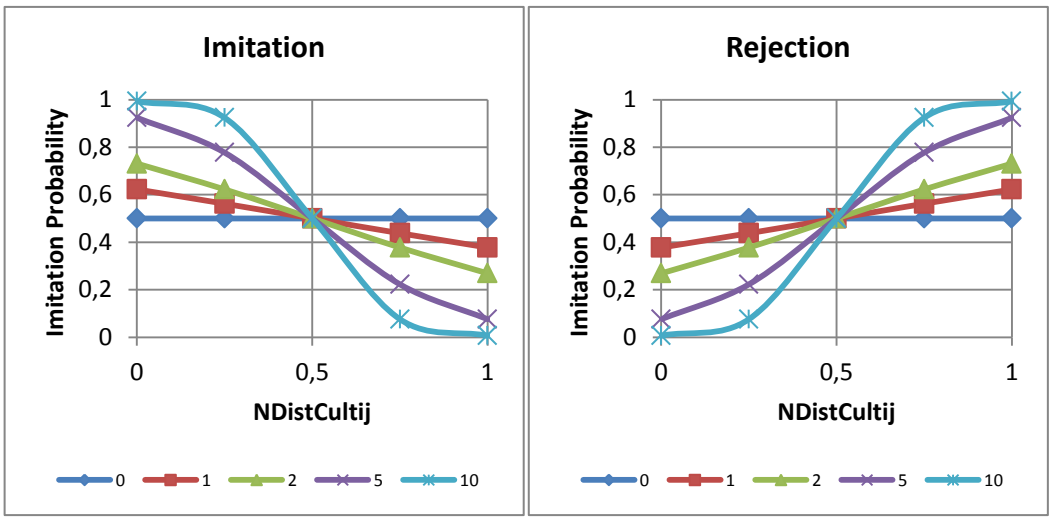

Figure 3. Probabilities of imitation and rejection given a strength of imitation (β) between 0 and 10. For β = 0, the probability of imitation or rejection is 0.5, regardless of the normalized cultural distance between the two agents. Elaborated by the authors.

The cultural space dynamics follow a combination of best-response and conformism mechanisms, as in Bowles (2004). A bold agent living in a neighborhood of vengeful agents is likely severely punished and thus will change its cultural attributes in order to reduce its punishments and increase its payoff (best response), imitating other agents in its neighborhood (conformism). The imitation process aims at reducing the largest difference observed among the two agents' attributes, according to the following expressions:

$$\Delta b_i = \begin{cases} \frac{b_j - b_i}{|b_j - b_i|}, se\ |b_j - b_i| \geq |v_j - v_i| \\ 0, \qquad\quad se\ |b_j - b_i| < |v_j - v_i| \end{cases}$$

$$\Delta v_i = \begin{cases} \frac{v_j - v_i}{|v_j - v_i|}, se\ |v_j - v_i| \geq |b_j - b_i| \\ 0, \qquad\quad se\ |v_j - v_i| < |b_j - b_i| \end{cases}$$

$$m_i = (\Delta b_i, \Delta v_i)$$

(19)

where (b_i, v_i) and (b_j, v_j) are the agents' attributes and m_i is agent *i*'s vector of movement in the cultural space. By swapping all terms with indices *i* and *j*, we obtain the rejection process as well.

In our model, we used Izquierdo and Galán (2005) definitions for defining a norm as established or collapsed, as follows:

	Average Attributes of the Population	
Norm	Boldness	Vengefulness
Established (N)	[0, 2]	[5, 7]
Collapsed (C)	[6, 7]	[0, 1]
Undefined (U)	otherwise	

In the Figure 4, three agents (A, B and C) are positioned in the cultural space and the distances among them are given by d_{AB}, d_{BC} and d_{AC}. The first two are highlighted, indicating that those agents are also connected, socially or geographically and, therefore, can participate in an imitation or rejection process. The vectors m_A and m_C represent agent B's possible movements towards becoming more similar to agents A and C, respectively.

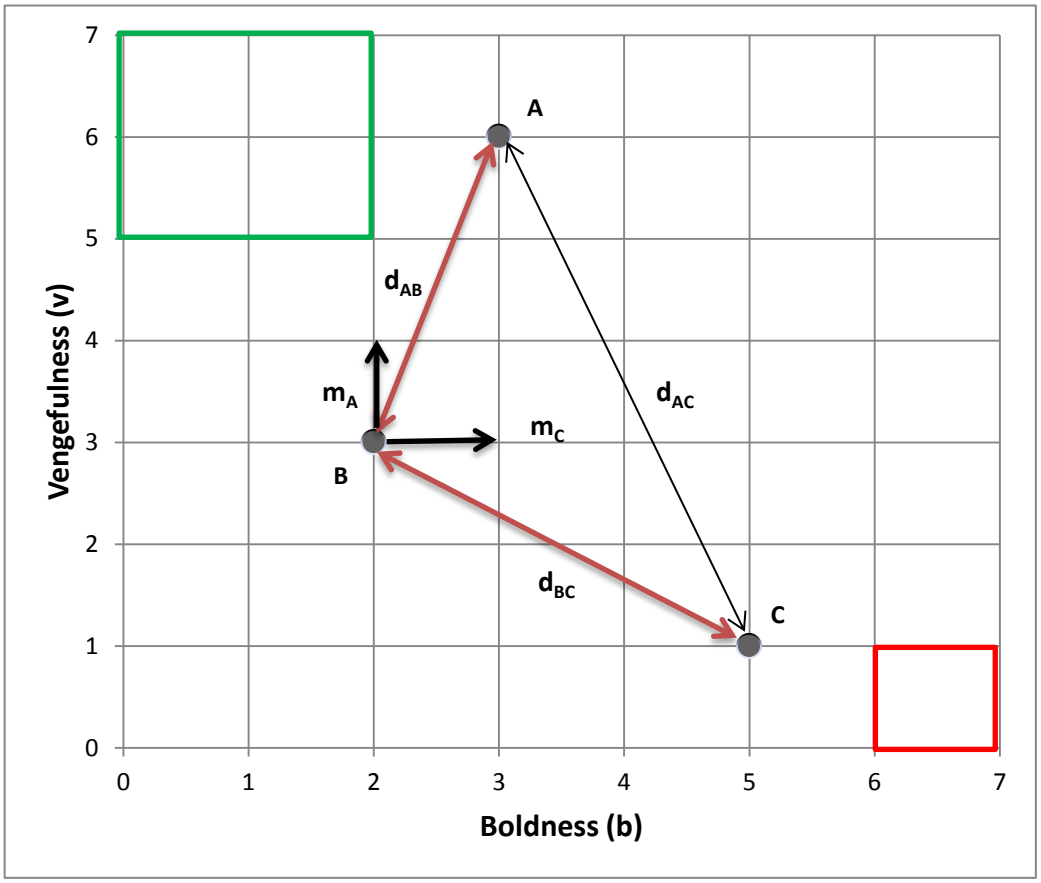

Figure 4. Cultural space defined by the Cartesian product of agent's attributes. The green and red squares are the regions where a norm is considered, respectively, established and collapsed. Three agents, A, B and C are depicted. Agent B is connected to A and C and may move towards one of them, according to vectors m_A and m_C. Elaborated by the authors.

We also included a mechanism akin to the mutation in Axelrod (1986) to introduce some exogenous variability into the process, namely, a behavioral innovation given by a random movement in the cultural space. The mechanism works through a parallel shock over all agents in both attributes, boldness and vengefulness, which will be modified by one unit each in a random direction with probabilities x_1* and x_2*, respectively.

The cultural space dynamics happen at the end of each generation, that is, every four rounds.

3.7 INSTITUTIONS

Next, we introduce an institution in the cultural space. We want to determine how effective different institutional arrangements are as tools for enabling the establishment of a social norm in a population. According to North (1995) there is a strong relationship between

individuals' beliefs and the institutional framework under which they operate. Beliefs and other informal mechanisms influence human behavior both directly, shaping the perception of the world and the actions perceived as appropriate in a given situation; and indirectly, affecting the structural development of institutions.

In our model, we spouse a "culturalist" vision of institutions, namely, as a set of values or a worldview (Chang and Evans, 2005). Hence, an institution is defined as a special agent with coordinates (b_{inst}, v_{inst}) in the cultural space. Since an institution does not have payoffs, we need to compute its visibility to other agents differently. To that purpose, we define the institution's power or weight (w_{inst}) as the maximum probability it has to be selected as "role model" to be imitated. We also define an institution's range (r_{inst}), its power's rate of decay, as a function of its normalized cultural distance towards the agent ($NDistCult_{iinst}$). Thus, the effective probability of an institution being selected as target for imitation is given by:

$$Prob(inst) = w_{inst} - r_{inst} NDistCult_{iinst} \tag{20}$$

The parameter r_{inst} can also be seen as a measure of plurality or diversity of opinions in the population of agents. For larger values of r_{inst}, up to $r_{inst} = w_{inst}$, agents situated farther away in the cultural space will be less frequently influenced by institutional values. On the other hand, for the extreme case of $r_{inst} = 0$, every agent has the same probability of being influenced by institutional values.

Given the presence of an institution with attributes (b_{inst}, v_{inst}) and parameters w_{inst} and r_{inst}, we need to adjust Formula 2 accordingly, resulting in the expression given by:

$$Prob(j) = \begin{cases} \dfrac{(1 - w_{inst} + r_{inst} NDistCult_{i\,inst})(1-\omega) Vis_{ij} max\{1, Social_{ij} + Geo_{ij}\}}{\sum_{k=1}^{n} Vis_{ik}\,(Vis_{ik} > 0) max\{1, Social_{ik} + Geo_{ik}\}}, if\ Vis_{ij} > 0 \\ \dfrac{(1 - w_{inst} + r_{inst} NDistCult_{i\,inst})\omega}{\sum_{k=1}^{n} (Vis_{ik} < 0) max\{1, Social_{ik} + Geo_{ik}\}}, otherwise \end{cases} \tag{21}$$

The possibility of imitating institutional values is compatible with a hierarchical modeling of a population and multi-level selection as proposed by Bowles (2004). In Bowles (2004), the process of cultural transmission may happen simultaneously in multiple levels – not only individuals attributes can be replicated, but also institutional ones. The probability of imitating an institution follows the same cultural-based process previously described for

imitating agents. The parameter β used to model the strength of the imitation can be seen, in the context of institutions, as an agent's bias (or opinion) towards institutional values.

Our model's definition of an agent's movement in the cultural space is in line with Chang and Evans (2005) description of agents internalizing institutional values and changing their behavior accordingly.

We have defined an institution as a set of values or attributes, but we haven't yet defined how these attributes came to be. Institutions can be either exogenous or endogenous. Exogenous institutions have immutable sets of values defined *a priori*, with no regard to the characteristics of the population. Endogenous institutions, on the other hand, have its set of values determined by the attributes of the agents in the population, that is, a two-way causality.

In our model, exogenous institutions are defined as having its values B and V set to 0 and 7, respectively, thus representing a norm through values of cooperation and punishment against defections. The endogenous institutions' values, on the other hand, are updated following a simple majority rule given by the median values of both attributes of the population. Since our agents do not have memory, the updates happen at regular intervals, namely, four generations (16 rounds).

The mechanisms in place in the cultural space can be summarized as follows: i) imitation / rejection of the more / less successful agents according to their cultural similarity; ii) imitation of institutional values according to institution's power and range, as well as cultural similarity and; iii) institutional change through regular updates based on population's attributes (endogenous institutions only).

3.8 SOCIAL SPACE

The last dimension of the model, its social space, is defined by a sociogram of its agents, that its, the adjacency matrix of a directed graph, where the agents are represented by its nodes and the social connections are represented by its edges. The social network of agents in our model is given by a scale-free network that is randomly generated at the beginning of each simulation using Barabási and Albert's (2002) algorithm with a degree of -1 for the preferential attachment. The social connections remain unchanged over the time, thus making it an exogenously defined element of our model.

The characteristics of the generated networks depend upon the population density of agents, as shown in the Table 1:

Table 1. Networks' Characteristics

Population Density	Average Degree of Nodes	Average Path Length		
		Minimum	Mean	Maximum
20%	1.900	3.11	3.92	5.19
50%	1.960	4.37	5.65	7.62
80%	1.975	5.39	6.60	8.10

Source: Elaborated by the authors.

3.9 SIMULATION

In our model's simulations, time and space are represented as discrete variables. At the beginning of each round, both agents and cells (forest dynamics) execute their actions synchronously. A full simulation run involves several steps, from variable initialization to actions' execution at every round, following these steps:

1. Initialization
    a. Geographic Space
        i. Create a torus of size L = 10 (100 cells);
    b. Agents
        i. A number of agents N = [20, 50, 80] is randomly placed on the cells of the torus, no two agents can occupy the location;
        ii. Every agent has its attributes B and V set to 7 and 0, respectively, characterizing a population of defectors;
    c. Social Space
        i. The agents are randomly connected to each other in a scale-free network using Barabási and Albert's algorithm;
    d. Common-Pool Resource
        i. Every $L^2=100$ cell on the torus has its biomass randomly drawn from a uniform distribution in the interval [½ b_{max}, b_{max}], characterizing a mature forest.

2. At every round:
    a. Agents
        i. Random variables *S*, *s* and *p* are randomly drawn from an uniform distribution in the interval [0, 1];
        ii. …**choose their harvesting roles** as **defectors** or **cooperators**, according to variables *B* and *S*;
        iii. …have a **fixed cost** for harvesting (C = 0.05);
        iv. …**harvest the common-pool resource**, obtaining their role's corresponding payoff (R = 0.10 or T = 0.30) and, **in case of biomass scarcity, move** to a different cell on the torus;
        v. …**choose their punisher roles** (punisher, metadefector, metapunisher) according to variables *S* and *s*;
        vi. …**observe each other** in their geographic neighborhood (*k = 8*, *r* =

1) and social connections;

vii. …**punish defections and/or are punished** according to their chosen roles, receiving the corresponding payoffs (E = ME = -0.20, P = MP = -0.90).

b. Common-Pool Resource

i. **Cells containing biomass** ($b_{xy} > 0$) **grow** according to the logistic equation ($b_{max}=1$ and $b_{rate}=0,5$);

ii. **Deforested cells are randomly reforested** ($p^* = 0.05$ and $k = 8$).

3. At every $\boldsymbol{\tau_e} = 4$ rounds (1 generation):

a. Agents

i. …move in the cultural space, **imitating or rejecting cultural attributes** (B and V) of other socially connected **agents** or in their geographic neighborhood according to their difference in payoffs, geographical and cultural distances, strength of imitation ($\beta_c = 3$) and probability of selection mistake ($\omega=0.10$);

ii. …move in the cultural space, **imitating institutional cultural attributes** (B and V) according to institution's power ($w_{inst} = [0.05, 0.20]$) and range ($r_{inst} = [0.03, 0.05, 0.15, 0.18, 0.20]$), as well as the strength of imitation ($\beta_i = [0, 2]$).

iii. …receive a **random shock in their cultural attributes** (B and V) with probabilities $x_1^* = 0.03$ and $x_2^* = 0.03$.

4. At every $\boldsymbol{\tau_a} = 16$ rounds (4 generations):

a. Agents

i. If the **institution is endogenous**, agents **vote** to update institutional values.

3.10 IMPLEMENTATION

Our model was implemented in NetLogo (Wilensky, 1999), an open source software developed for agent-based simulations, and integrated with R (R Core Team, 2010), a statistical package, using the NetLogo-R (Thiele and Grimm, 2010) extension, which allows NetLogo to execute commands in R and retrieve the corresponding results. The model and accompanying R code can be found at https://github.com/dvgodoy/netlogo-role-of-institutions.

The user interface was designed in NetLogo to allow for easier interaction and parameter setting. The geographic and common-pool resource elements of our model are handled by NetLogo, while the remaining elements leverage the power of the R engine for faster computations.

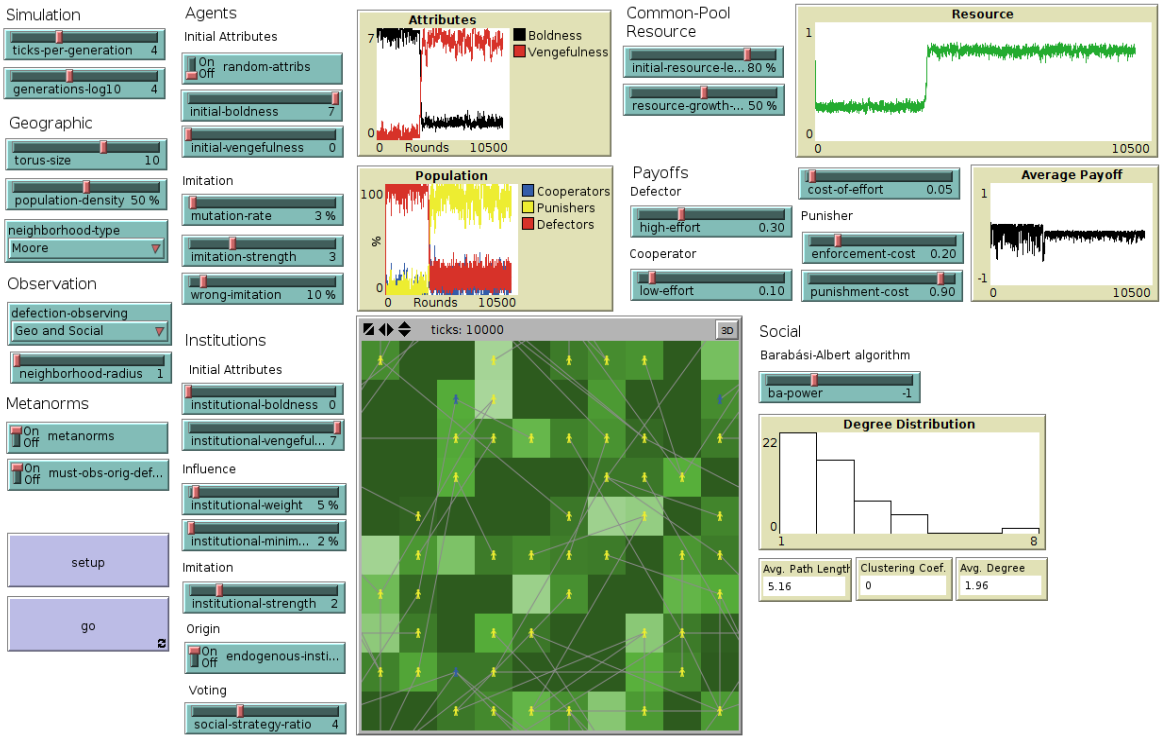

Figure 5. Sample run of our model in NetLogo and its configured parameters. The torus representing the geographic space and its forest is depicted in the center. Agents are colored according to its role – yellow for punishers, blue for cooperators and red for defectors. Their social connections are represented by the gray lins. There are also four line plots depicting the evolution over time for the average attributes of the population, the percentage of population playing each role, the resource level and the average payoff received by all agents.

Our primary focus is the cultural space, where institutions may exert their influence, so we investigate more thoroughly the effects of the following parameters of the model:

- NN – population density (number of agents)
- WW – power (weight) of the institution;
- M – minimal power of the institution, given by the difference between its power and range ($m = w - r$);
- O – origin of institution, either exogenous (X) or endogenous (E);
- D – type of observation mechanism, either geographic only (G) or both geographic and social (S).
- S – strength of imitation;

These six parameters yield a total of 144 combinations for evaluation. Each combination will be referenced to as a sequence of its parameters in the following format: NN.WW.M.S.O.D. For every combination, we simulated 10 runs of 1,000 rounds each, totaling 1.44 million rounds. For a selected subset of 72 combinations of interest, we simulated yet another 10 runs of 10,000 rounds each, totaling 7.2 million rounds.

4. RESULTS

For every run of our simulation, we stored the evolution of the average of population's attributes over the simulated time. Then we used Izquierdo and Galán (2005) definitions for norm establishment and collapse (see Section 3.6 for details) to compute the norm status over time, as shown in Figure 6:

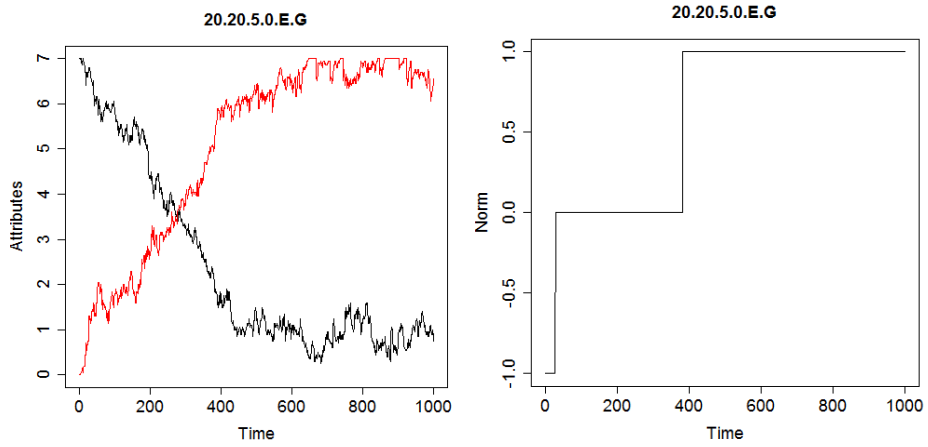

Figure 6: Left: evolution of the average of population's attributes, boldness (black) and vengefulness (red). Right: corresponding norm status, assuming one of three values: norm established (1), norm collapsed (-1) or undefined (0). In the example, a norm is initially collapsed (boldness greater than 6, vengefulness less than 1), then it rapidly goes into an undefined situation for about 400 rounds, becoming established (boldness less than 2 and vengefulness greater than 5) for the remainder of the simulation.

Next, we average the resulting norm status over all 10 runs sharing the same combination of parameters, as shown in Figure 7:

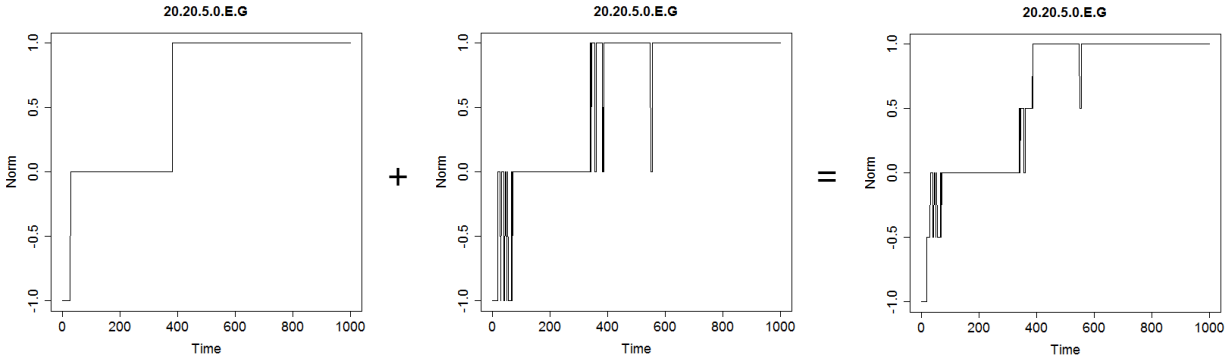

Figure 7. Averaging two norm status for the combinations 20.20.5.0.E.G (left side) and the resulting average norm status (right side). In this example, the average norm was only definitively established around 600 rounds.

For the results presented in the tables that follow, we always refer to the average norm status over the 10 runs of simulation executed for each combination being considered.

The results for the first batch of simulations are shown in Table 2. For each combination, it lists the number of times a norm was established (N), collapsed (C) or remained undefined (U) according to the average values of the population's attributes at the

end of a simulation (see Section 3.6 for details). It also shows the average percentage of time the requirements for norm establishment were fulfilled during the simulation.

Table 2. Statistics for norm establishment (1,000 rounds)

# Rounds					1,000																		1,000					
Pop. Density					20%						50%						80%						20%		50%		80%	
Institution			Origin		Exog.			Endog.			Exog,			Endog.			Exog.			Endog.			X	E	X	E	X	E
					Norm			Norm			Norm			Norm			Norm			Norm								
#	WW	M	S	O	N	U	C	N	U	C	N	U	C	N	U	C	N	U	C	N	U	C	% *rounds* Norm (Avg)					
1	5	0	0	G	1	**9**	0	**5**	4	1	1	0	**9**	0	0	**10**	2	1	**7**	0	0	**10**	9	11	0	0	8	0
2	5	0	0	S	**6**	4	0	**10**	0	0	1	3	**6**	0	2	**8**	1	2	**7**	0	1	**9**	24	31	6	1	5	0
3	5	0	2	G	3	**7**	0	3	**7**	0	0	0	**10**	0	1	**9**	1	0	**9**	0	1	**9**	8	14	0	0	3	0
4	5	0	2	S	**8**	2	0	**6**	2	2	0	1	**9**	1	1	**8**	1	1	**8**	0	2	**8**	37	22	0	1	8	0
5	5	2	0	G	**5**	**5**	0	4	**6**	0	**6**	4	0	**6**	3	1	**5**	2	3	**6**	3	1	18	18	29	19	25	25
6	5	2	0	S	**10**	0	0	**9**	1	0	**5**	4	1	**6**	3	1	**7**	2	1	**5**	3	2	48	56	24	29	21	15
7	5	2	2	G	1	**9**	0	**5**	**5**	0	2	**4**	**4**	3	3	**4**	**4**	2	**4**	1	2	**7**	14	23	9	9	17	5
8	5	2	2	S	**9**	1	0	**10**	0	0	**7**	2	1	**4**	3	3	**5**	3	2	**4**	**4**	2	38	48	28	16	18	18
9	5	5	0	G	**6**	4	0	**5**	**5**	0	**10**	0	0	**9**	1	0	**10**	0	0	**10**	0	0	38	33	67	61	61	64
10	5	5	0	S	**10**	0	0	**10**	0	0	**10**	0	0	**10**	0	0	**10**	0	0	**9**	1	0	72	74	69	64	62	70
11	5	5	2	G	**7**	3	0	**5**	**5**	0	**10**	0	0	**10**	0	0	**10**	0	0	**9**	1	0	37	39	52	57	49	50
12	5	5	2	S	**10**	0	0	**10**	0	0	**10**	0	0	**9**	1	0	**9**	1	0	**10**	0	0	69	66	57	48	56	55
13	20	0	0	G	**10**	0	0	**9**	1	0	2	0	**8**	**7**	1	2	**8**	2	0	**8**	0	2	53	46	11	25	36	47
14	20	0	0	S	**10**	0	0	**10**	0	0	**7**	1	2	**8**	0	2	**8**	1	1	**9**	0	1	70	63	45	45	38	59
15	20	0	2	G	**6**	4	0	**9**	0	1	**6**	0	4	**8**	1	1	**9**	0	1	**5**	0	**5**	31	53	26	25	46	29
16	20	0	2	S	**10**	0	0	**10**	0	0	**6**	0	4	**8**	0	2	**6**	1	3	**8**	0	2	65	68	13	32	30	43
17	20	2	0	G	**10**	0	0	**10**	0	0	**10**	0	0	**10**	0	0	**10**	0	0	**10**	0	0	67	61	55	70	60	73
18	20	2	0	S	**10**	0	0	**10**	0	0	**10**	0	0	**10**	0	0	**10**	0	0	**10**	0	0	72	78	78	69	65	78
19	20	2	2	G	**10**	0	0	**10**	0	0	**9**	1	0	**9**	0	1	**10**	0	0	**10**	0	0	45	63	38	56	62	72
20	20	2	2	S	**10**	0	0	**10**	0	0	**9**	0	1	**10**	0	0	**10**	0	0	**10**	0	0	73	72	60	56	51	54
21	20	5	0	G	**10**	0	0	**10**	0	0	**10**	0	0	**10**	0	0	**10**	0	0	**10**	0	0	78	67	84	86	84	87
22	20	5	0	S	**10**	0	0	**10**	0	0	**10**	0	0	**10**	0	0	**10**	0	0	**10**	0	0	83	84	85	81	86	83
23	20	5	2	G	**10**	0	0	**10**	0	0	**10**	0	0	**10**	0	0	**10**	0	0	**10**	0	0	65	60	72	72	79	77
24	20	5	2	S	**10**	0	0	**10**	0	0	**10**	0	0	**10**	0	0	**10**	0	0	**10**	0	0	77	81	76	79	79	80

Source: Elaborated by the authors.

The results indicate that a norm rarely collapses[5] in low population densities (NN = 20%) and is likely established whenever institutions are strong (WW = 20, rows 13 to 16), even if we allow for some agents completely ignoring institutional values (M = 0). In higher population densities, though, the power of the institution is not sufficient for establishing a norm. In those cases, the norm collapse is more likely to be prevented only if all agents are within range of influence from institutional values (rows 17 to 24).

For weaker institutions (WW = 5, rows 1 to 12) in low population densities, the results

[5] A low population density makes the dissemination of defecting behavior difficult, if there are already punisher agents, thus preventing the collapse of the social norm.

oscillate between a norm being either established or remaining undefined. For higher population densities, the norm is often collapsing and its collapse is likely prevented if institutional values can influence all agents (W = M = 5).

The inclusion of the social type of observation (D = S) seems to act as catalyst, facilitating the transition from an undefined to an established norm in the case of low population densities (NN = 20%). Since this effect is not observed in higher population densities, this mechanism seems to be compensating the fact that a sparse population makes a direct, geographical observation only, more unlikely to happen.

There does not seem to be a discernible difference in the results regarding the origin of the institutions (O), be it endogenous or exogenous. It is also not clear how the strength of imitation (S), the bias towards institutional values, is affecting the results.

The 72 highlighted combinations (rows 1 to 8 and 13 to 16) are particularly interesting, as their outcome regarding norm establishment is unclear. Thus, they were selected to be further investigated in a second batch of simulations. Since most of these runs exhibited an undefined norm for the majority of the simulated time in the first batch, we chose to extend the duration of each simulation to 10,000 rounds, while keeping the number of simulations for each combination (10) constant.

Table 3. Statistics for norm establishment (10,000 rounds)

# Rounds					10,000																	10,000					
Pop. Density					20%						50%						80%					20%		50%		80%	
Institution Origin					Exog.			Endog.			Exog,			Endog.			Exog.			Endog.		X	E	X	E	X	E
					Norm			Norm			Norm			Norm			Norm			Norm								
#	WW	M	S	O	N	U	C	N	U	C	N	U	C	N	U	C	N	U	C	N	U	C	% *rounds* Norm (Avg)					
1	5	0	0	G	0	**10**	0	**6**	4	0	0	2	**8**	**7**	1	2	0	2	**8**	**7**	0	3	6	42	1	49	0	44
2	5	0	0	S	2	**8**	0	**10**	0	0	0	0	**10**	**10**	0	0	0	**6**	4	**8**	2	0	19	82	0	62	3	54
3	5	0	2	G	0	**9**	1	**7**	3	0	0	4	**6**	**8**	0	2	0	2	**8**	**7**	2	1	5	51	0	52	1	44
4	5	0	2	S	1	**9**	0	**10**	0	0	0	2	**8**	**10**	0	0	0	3	**7**	**10**	0	0	19	89	1	57	1	67
5	5	2	0	G	1	**9**	0	**8**	2	0	0	1	**9**	**9**	1	0	0	3	**7**	**10**	0	0	5	52	1	90	0	86
6	5	2	0	S	2	**7**	1	**8**	2	0	0	2	**8**	**9**	1	0	0	2	**8**	**9**	1	0	21	89	1	91	1	91
7	5	2	2	G	0	**8**	2	**7**	3	0	0	1	**9**	**10**	0	0	0	2	**8**	**10**	0	0	4	58	0	83	1	76
8	5	2	2	S	1	**6**	3	**10**	0	0	0	2	**8**	**10**	0	0	0	0	**10**	**10**	0	0	23	91	0	87	0	93
9	20	0	0	G	0	**10**	0	**10**	0	0	0	2	**8**	**10**	0	0	0	**8**	2	**10**	0	0	3	95	2	91	6	93
10	20	0	0	S	1	**5**	4	**10**	0	0	0	**5**	**5**	**10**	0	0	2	**7**	1	**10**	0	0	13	97	1	94	7	95
11	20	0	2	G	1	**8**	1	**10**	0	0	0	4	**6**	**10**	0	0	2	**6**	2	**10**	0	0	5	91	2	85	19	91
12	20	0	2	S	2	**7**	1	**10**	0	0	0	**6**	4	**10**	0	0	1	**8**	1	**10**	0	0	10	96	1	87	9	91

Source: Elaborated by the authors.

The results for the second batch of simulations are shown in Table 3. The extended simulation time brought some new and interesting insights. First, there is a clear advantage for endogenous institutions – they are effectively establishing a norm in the majority of the cases considered – while exogenous institutions are more likely to lead to collapsed norms (for higher population densities) or an undefined situation (for low population densities).

Given an endogenous institution, allowing for the social type of observation (D = S, even rows) helps avoiding collapsed norms (for higher population densities) or an undefined situation (for low population densities). It is also noteworthy the effect it has on the average percentage time of an established norm, jumping from 50% to almost 90%.

The strength of imitation (S), the bias towards institutional values, seems to slightly improve the chance of a norm being established, while also generally increasing the average percentage of time a norm is considered established in weaker institutional arrangements (rows 1 to 4).

Since our results so far only consider the status at the end of a simulation, let's investigate further how fast a norm is established for the first time in a simulation and, after it happens, the proportion of time it is still considered established, as shown in Table 4:

Table 4. Statistics speed and stability of norm establishment (10,000 rounds)

	# Rounds				10,000						10,000					
	Pop.Density				20%		50%		80%		20%		50%		80%	
	Inst. Origin				X	E	X	E	X	E	X	E	X	E	X	E
#	WW	M	S	O	% *rounds* until first norm (avg)						% *rounds* with norm after first norm (avg)					
1	5	0	0	G	21	13	79	48	100	53	8	49	4	95	0	93
2	5	0	0	S	17	7	81	36	81	42	23	88	2	97	13	93
3	5	0	2	G	27	14	92	47	95	55	6	60	4	99	11	99
4	5	0	2	S	15	6	85	42	99	32	22	94	5	99	43	98
5	5	2	0	G	22	6	86	7	99	12	6	55	4	97	11	98
6	5	2	0	S	16	4	81	7	91	7	25	93	3	97	12	98
7	5	2	2	G	51	11	97	16	96	23	7	65	2	99	14	100
8	5	2	2	S	17	5	99	13	94	7	28	97	11	99	1	100
9	20	0	0	G	65	5	81	9	64	7	8	99	9	100	17	100
10	20	0	0	S	20	3	74	6	56	5	17	100	5	100	16	100
11	20	0	2	G	58	9	90	15	60	9	11	100	20	100	47	100
12	20	0	2	S	46	4	76	13	66	9	19	100	5	100	26	100

Source: Elaborated by the authors.

For endogenous institutions, we can observe that a norm is established relatively fast

for the majority of the combinations (left part of the table), though the process is particularly slower for higher population densities and weaker institutional arrangements (rows 1 to 4). These results suggest that low density populations can also work as catalysts for establishing a norm. Interestingly, this pattern seems to hold for exogenous institutions as well.

Regarding the stability of an established norm (right part of the table), we find that higher population densities exhibit more stable norms, with values close to a 100%. Low population densities show more unstable norms in general, even though they were faster to establish it. In these cases, the social type of observation improves substantially the norm stability.

Exogenous institutions not only take a long time to establish a norm for the first time (if at all), but they also produce unstable norms in general.

5. CONCLUSION

Our goal was to investigate the role of institutions in the emergence of cooperation in a Tragedy of Common situation using an agent-based simulation. We based our model on the Metanorms Game and Dissemination of Culture models proposed by Axelrod (1986, 1997) and expanded it incorporating elements from Ostrom's Institutional Analysis and Development (IAD) framework (2005), endogenizing several aspects of the latter.

We focused our attention on the parameters that most affected the dynamics of institutional influence over the attributes and behaviors of the population of agents, namely: institution's power and range of influence, its origin (endogenous or exogenous), the agents' bias towards its values (strength of imitation), the existence of social control (agents observing others through their social connections) and population density.

An initial population of defectors was chosen as the worst case scenario. Could an institution effectively transform the population's attributes and establish a social norm, thus making its members cooperators? The ideal institution would have diametrically opposed values, those of cooperation, and it would work as a “beacon”. This was straightforward for exogenous institutions, as we could set its values *a priori*. Endogenous institutions, on the other hand, had as initial values the average of the population's values: defector values. Only through behavior innovation, introduced using small random shocks to population's attributes, some defectors could be slowly pushed towards a more vengeful behavior and, eventually, become punishers that, in turn, would coerce the defectors into cooperating.

The first batch of simulations, comprised of 10 runs of 1,000 rounds each for each combination of parameters showed us that exogenous institutions slightly outperformed endogenous institutions, establishing a social norm and achieving the goal of having a population of cooperators more often. Nonetheless, in half of the parameter combinations, the outcome was unclear: some runs would end up in a norm being established, others in a norm being collapsed (in higher population densities), and many more would remain in an undefined situation for most of the simulated time (low population density). These combinations represented weak institutional arrangements with limited range of influence.

The second batch of simulations extended the simulated time to 10,000 rounds, in order to determine if situations deemed undefined would come to a resolution in any way – establishing or collapsing the social norm. The most noteworthy difference was the fact that,

given more time, endogenous institutions vastly outperformed its exogenous counterparts which, somewhat surprisingly, could not sustain the established norm in the long-run.

An endogenous arrangement itself could not guarantee the norm establishment, though. In higher population densities, the norm would still collapse sometimes, unless the possibility of social control was available to the agents. For low population densities, on the other hand, the norm did not collapse at all and undefined situations were remedied once again by the possibility of social control. Additionally, we found that an agent's bias towards institutional values can help preventing norm collapse, especially in very weak institutional arrangements with lack of social control.

We also found distinct patterns regarding how fast a norm is established depending on population density. Higher densities were associated with slower and more stable processes, while low density populations were much faster to establish a social norm, but its stability was highly dependent on the possibility of social control.

In summary, in the context of our simulations, we found results that support the following hypotheses:

1) endogenous institutional arrangements are more effective for promoting changes in cultural values;
2) when considering weak endogenous institutions with limited range of influence, the results also show that:
    a) the possibility of social control is sufficient for establishing a norm;
    b) medium or large communities are capable of establishing stable norms;
    c) in small communities, the stability of a social norm is dependent on the existence of social control; and
    d) indifference to the institutional values is an obstacle.

The results we found for small communities and institutions with limited power and range of influence, capable of fast establishing a social norm, yet depending on the existence of social control for keeping its stability (rows 1 to 4 in Table 2) are compatible with those found by Ostrom (1990) in her field experiments.

In this work, we focused on a small subset of possible parameterizations of our model, namely, the parameters that describe how agents are affected by institutional values. We found some promising results in our simulations concerning the dynamics of small communities and

future work could benefit from exploring it further, either by testing other parameterizations or by trying to understand more deeply its underlying dynamics.

There are also innumerous possible improvements to be yet incorporated into the model, ranging from using different network structures, or even dynamic ones, to socially connect the agents; up to changes in population dynamics, allowing for its growth; and including modifying agents' behavior by giving them memory or different learning speeds.

6. REFERENCES

ALBERT, R.; BARABASI, A. Statistical mechanics of complex networks. Reviews of Modern Physics. Vol.74, 47-97, 2002.

ARTHUR, W. B. Complexity and the economy. Science. Vol. 284, 107-109, 1999.

AXELROD, R. An evolutionary approach to norms. American Political Science Review. Vol.80, N.4, 1095-1111, 1986.

AXELROD, R. The dissemination of culture: A model with local convergence and global polarization. Journal of Conflict Resolution. Vol. 41, N.2, p. 203-226, 1997.

BOWLES, S. Microeconomics: Behavior, Institutions and Evolution. Princeton: Princeton University Press, 2004.

BRAVO, G. The evolution of institutions for common-pool resource management: an agent based model. Rationality and Society. Vol. 23, 117-52, 2011.

CANOVA, G. A. Jogos evolutivos: efeitos de difusão em redes complexas. Monografia de conclusão de curso. Porto Alegre, UFRGS, 2011.

CHANG, H.J.; EVANS, P. The role of institutions in economic change. In: Reimagining Growth. Londres: Zed Press, 2005.

DAWES, R. M. Social Dilemmas. Annual Review of Psychology. Vol. 31, 169-93, 1980.

EPSTEIN, J.M.; AXTELL, R. Growing artificial societies. Washington D.C.: Brookings Institution Press, 1996.

GALAN, J.M.; LATEK, M.M.; RIZI, S.M.M. Axelrod's metanorm games on networks. PLoS ONE Vol. 6, N.5, 2011. Disponível em < http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0020474>

GINO, F.; AYAL, S.; ARIELY, D. Contagion and differentiation in unethical behavior. Psychological Science. Vol. 20, N.3, 393-8, 2009.

HARDIN, G.R. The tragedy of the commons. Science. Vol. 162, p. 1243-48, 1968.

KOLLOCK, P. Social dilemmas: the anatomy of cooperation. Annual Review of Sociology. Vol. 24, 183-214, Ago/1998.

MACY, M.W. Chains of cooperation: Threshold effects in collective action. American Sociology Review. Vol. 56, n. 12, p.730-47, 1991.

MAHMOUD, S.; GRIFFITHS, N.; KEPPENS, J.; LUCK, M. Overcoming omniscience for norm emergence in Axelrod's metanorm model. Coordination, Organizations, Institutions, and Norms in Agent System VII. Vol. 7254, 186-202, 2012.

NORTH, D. Institutions, Institutional Change and Economic Performance. Cambridge: Cambridge University Press, 1995.

OSTROM, E.; JANSSEN, M.; ANDERIES, J. A framework to analyze the robustness of social-ecological systems from an institutional perspective. Ecology and Society. Vol. 9, N. 1,

2004. Disponível em < http://www.ecologyandsociety.org/vol9/iss1/art18/>

OSTROM, E. Collective action and the evolution of social norms. Journal of Economic Perspectives. Vol. 14, N.3, 137-158, 2000.

OSTROM, E. Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge: Cambridge University Press, 1990.

OSTROM, E. Institutional analysis and development: Elements of the framework in historical perspective. In Historical Developments and Theoretical Approaches in Sociology: Vol. 2. Oxford: EOLSS Publishers, 2010.

OSTROM, E. Institutional rational choice: an assessment of the Institutional Analysis and Development Framework. In Theories of the Policy Process. Boulder, CO: Westview Press, 2007.

OSTROM, E. Understanding Institutional Diversity. Princeton: Princeton University Pressa, 2005.

OSTROM, E.; GARDNER, R.; WALKER, J. Rules, games and common-pool resources. Ann Arbor: The University of Michigan Press, 1994.

R CORE TEAM. R: A language and environment for statistical computing. Viena: R Foundation for Statistical Computing, 2010. Disponível em < http://www.R-project.org/>

SIGMUND, K.; DE SILVA, H.; TRAULSEN, A.; HAUERT, C. Social learning promotes institutions for governing the commons. Nature. Vol. 466, p. 861-3, 2010.

THIELE, J.C.; GRIMM, V. NetLogo meets R: Linking agent-based mdels with a toolbox for their analysis. Environmental Modelling and Software, Vol. 25, Issue 8, 972-974, 2010.

WILENSKY, U. NetLogo. Evanston, IL Center for Connected Learning and Computer-Based Modeling, Northwestern University, 1999. Disponível em <http://ccl.northwestern.edu/netlogo/>