Reinforcement Learning Account of Network Reciprocity

Guzga Adrian-Dumitru[1], Luci Horatiu[1], Martis William[1], Sirbu Adriana[1]

[1]Université Libre de Bruxelles

Abstract

In social dilemma games, cooperation is not the preferred behaviour of humans over longer spans of time, according to experimental results. This contradicts evolutionary game theory which states that cooperation should be prevalent when players are linked as a network. In this paper, we make usage of a reinforcement learning model based on aspiration to achieve network reciprocity and present parameter regions of benefit to cost ratio and neighbours in which it can and cannot be observed. This way, we can visualize how network reciprocity tends to change over parameters for the Bush-Mosteller reinforcement learning model.

Introduction

Humans are considered as being the species that cooperate the most, especially when they interact with their acquaintances or relatives. They have learned from each other, cared for each other, but also competed against each other over the last million years. (Boyd and Richerson, 2009). These kind of connections between human beings create the so-called social networks, that are believed to exist since the earliest stages of human history. Therefore, researchers are highly interested in understanding how exactly the cooperation emerged and evolved in the presence of the natural selection, since the latter usually favours selfish behaviour (Nowak, 2006; Iyer and Killingback, 2016). Nowadays we can find an abundance of scientific articles concerning this topic [1] which propose a large variety of theoretical models describing the dynamics of structured populations that are defined by complex networks. By dynamics, we mean the manner in which the individuals of a population adapt their strategy.

The standard modelling method that is often used to explain the cooperative behaviour between selfish and unrelated individuals in social dilemma situations is the Evolutionary Game Theory (EGT) (Alexander, 2019; Smith and Price, 1973; Macy and Flache, 2002), where the evolutionary process is determined by strategy updating rules. These rules are categorized into two groups: imitative and non-imitative. The difference between them is that in the non-imitative rule the individuals do not consider the actions of others and the payoffs in order to adjust their own strategy, while the imitative rule is based on payoff comparison and on adjusting the individual's strategy by copying the action of neighbours that perform better than him (Dercole et al., 2019).

In order to observe the evolution of cooperation and also the network reciprocity, researchers often use a model called Prisoner's Dilemma (Yamauchi et al., 2010; Traulsen et al., 2010). This model is a social dilemma game which predicts that the levels of defection (i.e. non-cooperation) should reach 100%, while those of cooperation reach 0%. However, results from the experiments on human networks playing Prisoner's Dilemma that were conducted by Khadjavi and Lange (2013), contradict the theoretical predictions seen earlier. They observed levels of cooperation that varied from 37 to 55.6%, which demonstrates that humans have a tendency to reciprocate cooperation than to compare payoffs (Dercole et al., 2019; Coelho and McClure, 2016). After studying the same game, Nowak (2006) came to the conclusion that there are five mechanisms that promote the cooperation instead of defection and that punishment is an important factor that induces this phenomenon. These five mechanisms are the following: indirect, direct, and network reciprocity, and group and kin selection (Ohtsuki, 2018). However, according to the evolutionary biologists and evolutionary psychologists, the group selection mechanism, where the selection does not only act on individuals but also in groups, should be avoided, because of some evolutionary constraints. The network reciprocity, on the other hand, is often used in simulations (Perc et al., 2013; Hilbe et al., 2018; Dercole et al., 2019). It is known to be a simple tool that induces the cooperation. The results of the research led by Cimini and Sanchez (2015), which are based on the social networks experiments conducted by Grujić et al. (2010), demonstrate the existence of the network reciprocity within populations where the individuals ignore the payoffs when they update their strategy. Indeed, the final level of cooperation that resulted from their experiments is often

[1]About 3.220.000 results found after searching paper titles containing "evolution of human cooperation", in Google Scholar

very close to zero, which leads to the conclusion that the imitative rules are not representative of the human behaviour. The network reciprocity also depends on the structure of the network itself. Researchers defined a simple condition for network reciprocity which is $r > k$, where r is the benefit-to-cost ratio indicator and k is the average degree of a network Rand et al. (2014); Ezaki and Masuda (2017); Dercole et al. (2019).

Reinforcement learning is a popular non-imitative strategy updating rule, where the individual chooses actions that lead to satisfactory payoffs, depending on his experience (Bush and Mosteller, 1955; Foley et al., 2018; Segismundo S. Izquierdo, 2008). The paper proposed by Rand et al. (2014) that uses the reinforcement learning, demonstrates that by respecting the condition for network reciprocity stated above, we can obtain a high stable level of cooperation within fixed (i.e. static) networks, higher than in the case of shuffled networks (i.e. well-mixed populations). For this experiment, they used artificial social networks, where each individual is positioned on a ring and is connected to $\frac{k}{2}$ neighbours on each side. As in the other experiments, they choose to use the Prisoner's Dilemma game, which is played many rounds. For each round every individual can choose a single action, by either defecting (i.e. by doing nothing) or by cooperating with a cost $c = 10k$ that gives each of the k neighbours a benefit b. After every round individuals are informed about their neighbours' decisions as well as the total payoff they earned. They did this experiment for both fixed and shuffled networks. Similar results are observed in a more recent paper presented by Ezaki and Masuda (2017).

In our paper we present a model based on previous works by Ezaki and Masuda (2017). Therefore, we attempt to achieve network reciprocity by applying the reinforcement learning on small world ring networks which are based on the Watts-Strogatz model with $\beta = 0$. The Prisoner's Dilemma game is used for both fixed and shuffled networks. Moreover, we attempt to visualize the dependence of the network reciprocity on the aspiration levels.

Methods

I. The donation game on networks

The donation game is one of the special cases of the Prisoner's Dilemma Game. When two players play the donation game, they have to choose between two actions: C (standing for *cooperate*) or D (standing for *defect*). Once the two players have chosen their actions, each of them receives a payoff according to the table below, where the rows represent the first player's actions and the columns represent the second player's actions.

		Column Player: C	Column Player: D
Row Player	C	$b-c$ $b-c$	b $-c$
Row Player	D	$-c$ b	0 0

Figure 1: Payoffs of the donation game

The parameters b and c represent the benefit and the cost, more precisely, when a player chooses to cooperate, a certain cost c will be paid at the benefit b of the other player. However, if both players choose to defect, they will not receive any payoff. In this setting of the game, we have imposed that $b > c > 0$.

In our simulations, the game will be played on a network of players, with each player having a fixed number of k neighbours over a total of r_{max} rounds. More precisely, at each round of a simulation, the players (represented by the network's nodes) will play a donation game with their respective neighbours. Thus, if a player chooses to cooperate, the total cost is equal to $-kc$, whereas the total benefit is equal to the number of cooperating neighbours multiplied by the b parameter, giving a payoff defined by the difference between the total benefit and the total cost. If a player is a defector, the total payoff will be equal to the number of cooperating neighbours multiplied by the b parameter. This total payoff of each player will be averaged by the number of neighbours k. At each round, we will refer to this averaged payoff as G_r.

II. Network treatments: static and shuffled

In the simulations, we have used small world ring networks which are based on the Watts-Strogatz model with $\beta = 0$, making them regular lattices. Each node in the network has a total number of k neighbours, which is even. More precisely, on the left and on the right hand-side of each node, the next $\frac{k}{2}$ nodes represent each node's neighbours. Two graphical representations of the networks are illustrated in the two figures below.

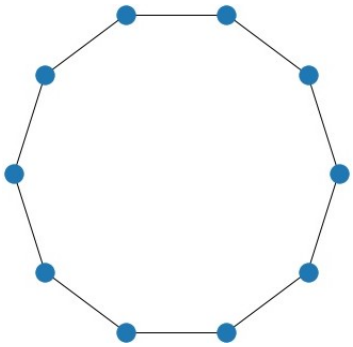

Figure 2: Network of $N = 10$ nodes and $k = 2$

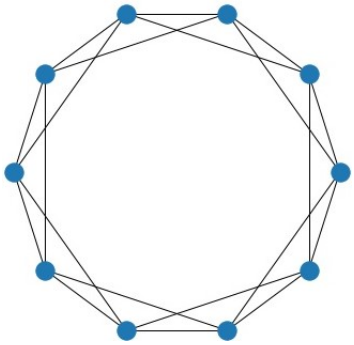

Figure 3: Network of $N = 10$ nodes and $k = 4$

As we are trying to determine the differences in inclination towards cooperation between the static and dynamically shuffled networks settings, two network treatments arise. First, in the static treatment, the nodes' and neighbours' positions are fixed throughout the whole simulation, meaning that no change of neighbours can be made. Second, in the dynamically shuffled networks, at each round, the nodes' positions and their neighbours will be shuffled at each and every round, but the lattice structure of before will be kept intact.

III. Bush-Mosteller reinforcement learning model

Each player's update of the action to take over each round is governed by an adapted Bush-Mosteller model, which is a model of reinforcement learning based on stimuli. More precisely, each player has a certain probability p_r of choosing to cooperate at round $r \in \{1, 2, ..., r_{max}\}$. In the first round, each player has a probability $p_1 = 0.8$ to choose to cooperate. Afterwards, for each round and each player, the cooperation probability is computed by taking into account the player's previous action and the *stimulus*, denoted as a_{r-1} and s_{r-1} respectively. The formulas according to which the probability of cooperation is updated are given by the following equation:

$$p_r = \begin{cases} p_{r-1} + (1 - p_{r-1})s_{r-1} & (a_{r-1} = C, s_{r-1} \geq 0) \\ p_{r-1} + p_{r-1}s_{r-1} & (a_{r-1} = C, s_{r-1} < 0) \\ p_{r-1} - p_{r-1}s_{r-1} & (a_{r-1} = D, s_{r-1} \geq 0) \\ p_{r-1} - (1 - p_{r-1})s_{r-1} & (a_{r-1} = D, s_{r-1} < 0) \end{cases} \tag{1}$$

The stimulus in the above equation is defined as a value which is upper bounded by the maximum and minimum values that the hyperbolic tangent can take, without actually touching those bounds as the argument of the function will never be close to $-\infty$ or $+\infty$. Mathematically, this translates to $s_{r-1} \in (-1, 1)$. The formula of the stimulus is given by the following equation:

$$s_{r-1} = tanh(\beta(G_{r-1} - A)), \tag{2}$$

where A represents the aspiration level of the players, G_{r-1} represents the total gains of the player in the previous round and β represents the sensitivity parameter and is a real number greater than 0. Thus, a previous action's probability of choice will be increased from one round to another if and only if the difference between the payoff and the aspiration level is greater than 0 ($G_{r-1} - A > 0$). Following the same principle, an action's probability of choice will diminish from one round to the other if and only if the previous round's payoff has not exceeded the aspiration level, meaning that ($G_{r-1} - A < 0$). We should state that the equations above guarantee that the values of all p_r with $r \in \{1, 2, ..., r_{max}\}$ belong to the interval $(0, 1)$, thus respecting the properties of probabilities.

As an addition to the previously presented model, another parameter, ϵ, has been added, which corresponds to the probability of misimplementation. More precisely, this parameter represents the probability that a player chooses the opposite action of his intention. For example, if the player's intent is to cooperate, it would defect. Thus, the probability with which a player chooses to cooperate is given by the following equation:

$$p_r = p_r(1 - \epsilon) + (1 - p_r)\epsilon \tag{3}$$

Results

The following results have been obtained by averaging cooperation levels within ring networks of $N = 100$ nodes, over each of the 50 rounds for 1000 trials.

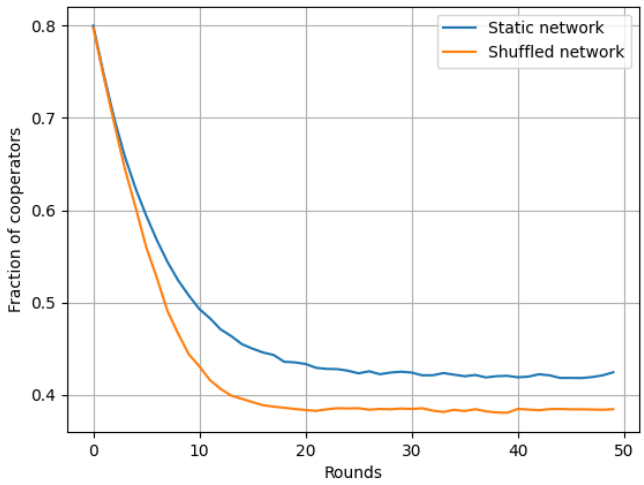

Figure 4: Cooperation level over time $k = 2$, $\frac{b}{c} = 6$

We start by observing the proportion of cooperators over time for a small amount of neighbors $k = 2$, with an aspiration level $A = 1$, the sensitivity $\beta = 0.2$ and an implementation error probability $\epsilon = 0.05$. First, by setting a benefit to cost ratio $b/c = 6$, we note a decrease in cooperation for both types of networks that stabilizes around 25 rounds. The static network stays at approximately 42% of cooperators whereas the shuffled network stays at a lower 37% of cooperators. Since both networks stabilize around the same time but the shuffled network's stable proportion of cooperators is lower, its drop is steeper but reaches stability a bit faster.

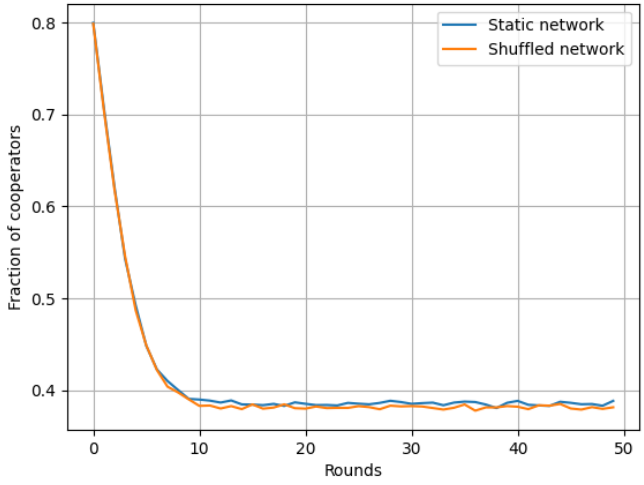

Figure 5: Cooperation level over time $k = 2$, $\frac{b}{c} = 2$

Next, we repeat the procedure but with a smaller ratio of benefit to cost $b/c = 2$. The plot in the figure 5 shows the same decrease and stabilization behaviour as previously, however in an accelerated fashion, but no significant distinctions can be made. However, even though both networks stay around a value of 37%, the static network stays at a slightly higher value than the shuffled network after 12 rounds. These results match the results that can be found in the work of Ezaki and Masuda (2017) for the two figures above.

In order to verify that these results are reproducible (in maybe varying effects) over parameter values and that these 2 parameter sets are not special cases, we repeated the procedure over 10 000 combinations of aspiration levels A and ϵ for different k values : 2, 4, 6, 8 on both types of networks. The x axis represents aspiration levels A going from -1 to 5 and the y axis implementation error ϵ from 0 to 0.5. Each axis has 100 values evenly spaced between their extremes. A pixel in the following heat-maps represents the average value over 25 rounds according to its legend. The first row of heat-maps shows the effect of parameter variation for the static networks whereas the second row illustrates the same effect for the shuffled network.

As we can see, for a benefit to cost ratio of $b/c = 2$, cooperation levels are low (less than 50%) over almost the entire parameter region and this phenomenon is observable for all the values of k. The highest level can be seen in the bottom left corner(70%), but it quickly fades when going north and/or east on the map. The same can be said in the case of a benefit to cost ratio of $b/c = 6$, but here the highest values represent a much larger region, spreading further along the x axis, showing us that in this setting, aspiration is more robust than implementation error for this parameter space. The third row illustrates the differences in cooperation levels on the entire parameter space, between both types of treatments. When the benefit to cost ratio is $b/c = 2$, no significant difference is visible but for a higher benefit to cost ratio of $b/c = 6$, we observe a large difference in the x axis and a small difference in the y axis. This indicates that A is more robust in the static network than in the shuffled one. However, this difference vanishes as the value of k increases and is almost null at $k = 8$.

Ezaki and Masuda (2017) define the *assortment* at round r as the value representing the difference between P(C|C, r) and P(C|D, r). Concretely, the probability that a neighbour of a cooperator is a cooperator is P(C|C, r), whereas the probability that a neighbour of a cooperator is a defector is given by the value of P(C|D, r). For the same parameter space, we have averaged the assortment over 25 rounds. The fourth rows in the figures below represent the assortment of the static network. In the case of a benefit to cost ratio $b/c = 2$, we observe higher assortment

levels at the bottom left corner. Moving further from the lower left corner, we can observe a bigger region in the parameter space which reveals negative assortment values. This phenomena is more visible the more the number of neighbours k increases. Otherwise, the assortment levels are not significant in the rest of the parameter space. On the other hand, in the case of a benefit to cost ratio $b/c = 6$, we observe higher values for most of the parameter space. However, when the aspiration level gets bigger than 3, the assortment level diminishes.

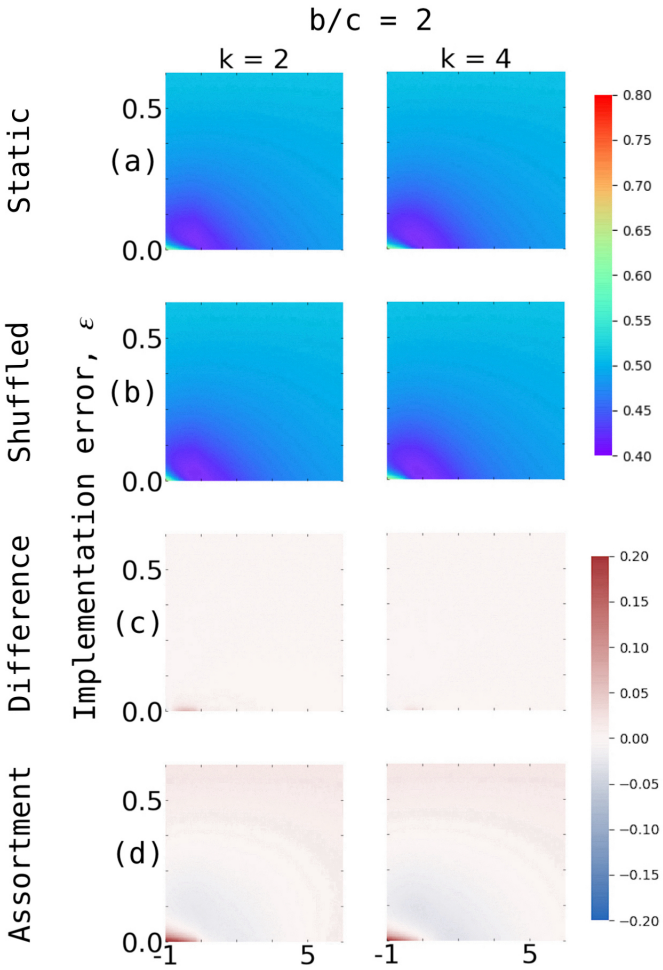

Figure 6

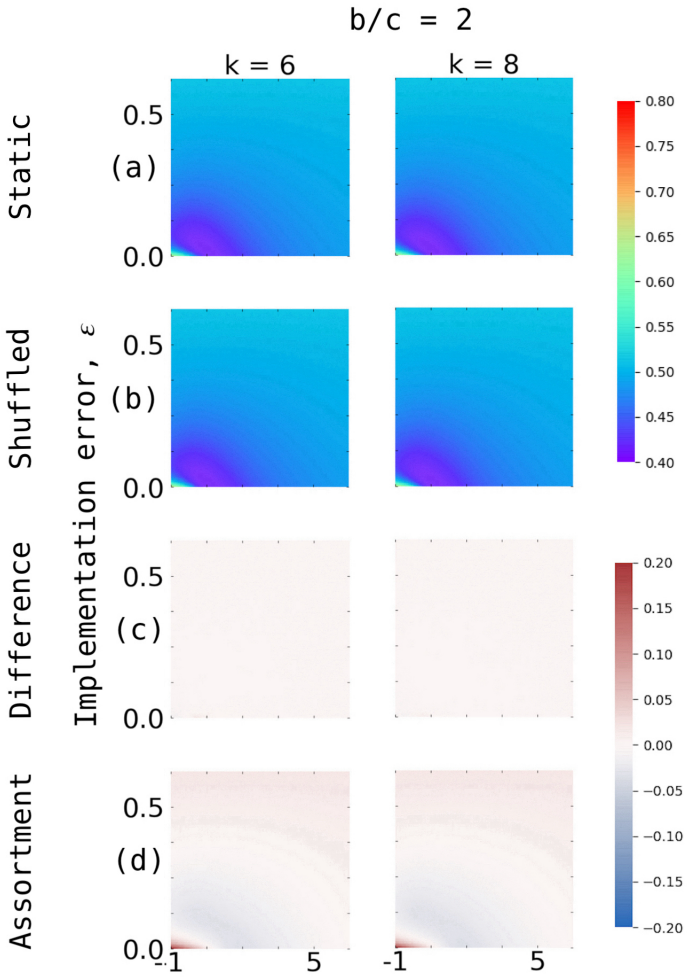

Figure 7

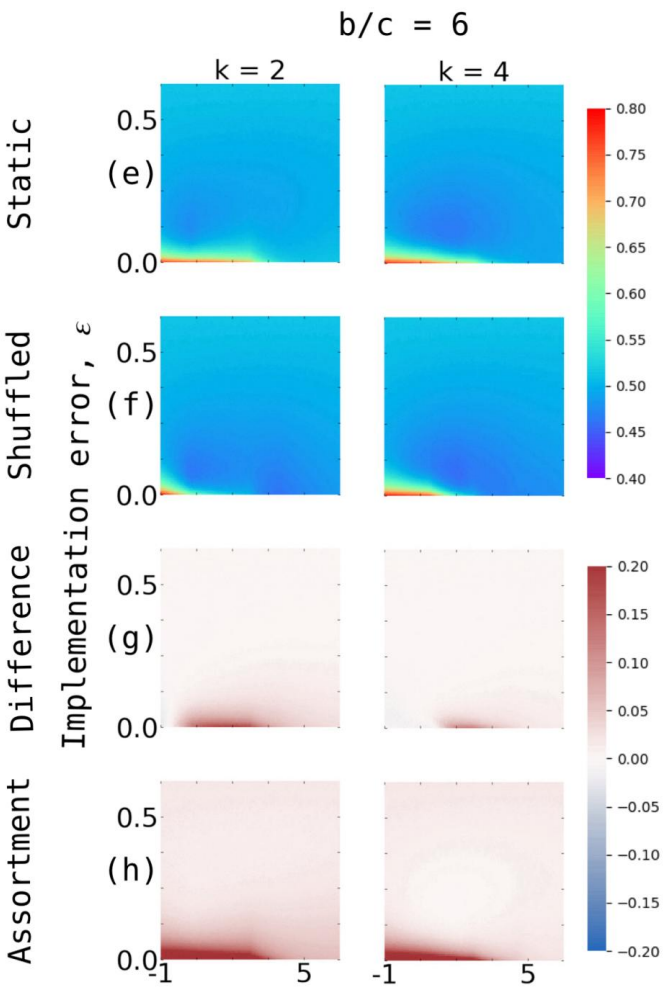

Figure 8

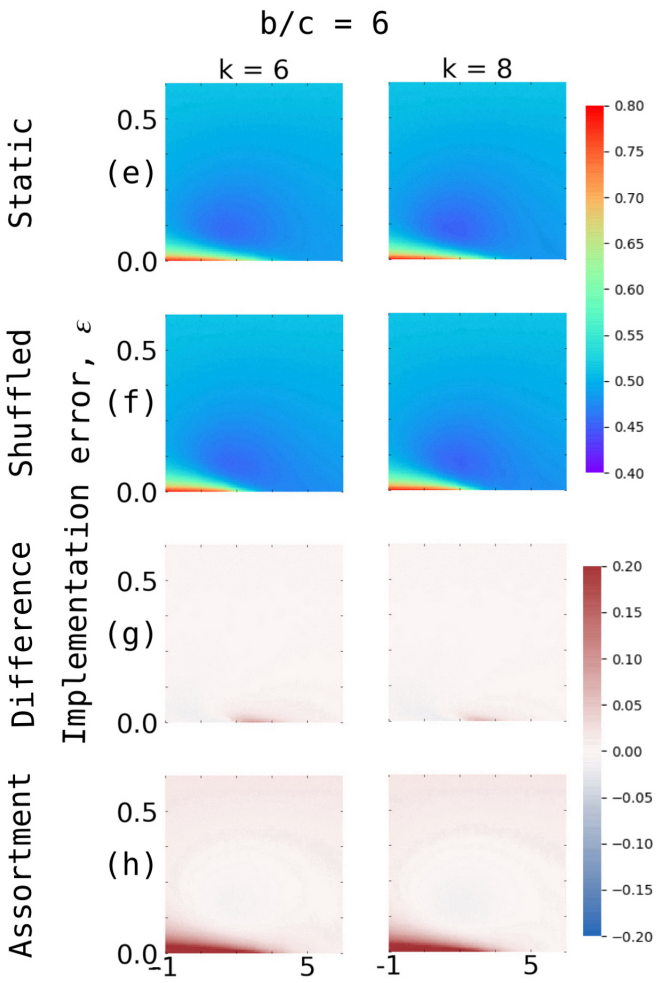

Figure 9

Discussion

According to the results presented in the above section, we have been able to match previous experimental results regarding network reciprocity. This has been achieved by using the Bush-Mosteller reinforcement learning model on static and shuffled ring networks (Watts-Strogatz). Additionally, the conditional cooperation cannot be explained by evolutionary game theory.

Indeed, as mentioned in the introduction of this paper, a benefit to cost ratio which is greater than the number of neighbours of the network led to increased cooperation, meaning that more clusters of cooperators are formed when the aforementioned condition is met. The reciprocity observed in our simulations is the result of the usage of reinforcement learning on the way the players update their actions in the donation game. Therefore, the capacity of a reinforcement learning model based on aspiration to explain the cooperative behaviour observed in animals is greater

than the one of evolutionary game theory.

References

Alexander, J. M. (2019). Evolutionary Game Theory. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2019 edition.

Boyd, R. and Richerson, P. J. (2009). Culture and the evolution of human cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1533):3281–3288.

Bush, R. R. and Mosteller, F. (1955). *Stochastic models for learning.* John Wiley & Sons Inc, Hoboken.

Cimini, G. and Sanchez, A. (2015). How Evolutionary Dynamics Affects Network Reciprocity in Prisoner's Dilemma. *Journal of Artificial Societies and Social Simulation*, 18(2):22.

Coelho, P. R. P. and McClure, J. E. (2016). The evolution of human cooperation. *Journal of Bioeconomics*, 18(1):65–78.

Dercole, F., Della Rossa, F., and Piccardi, C. (2019). Direct reciprocity and model-predictive rationality explain network reciprocity over social ties. *Scientific Reports*, 9(1):5367.

Ezaki, T. and Masuda, N. (2017). Reinforcement learning account of network reciprocity. *PLOS ONE*, 12(12):e0189220.

Foley, M., Forber, P., Smead, R., and Riedl, C. (2018). Conflict and convention in dynamic networks. *Journal of The Royal Society Interface*, 15(140):20170835.

Grujić, J., Fosco, C., Araujo, L., Cuesta, J. A., and Sánchez, A. (2010). Social Experiments in the Mesoscale: Humans Playing a Spatial Prisoner's Dilemma. *PLoS ONE*, 5(11):e13749.

Hilbe, C., Chatterjee, K., and Nowak, M. A. (2018). Partners and rivals in direct reciprocity. *Nature Human Behaviour*, 2(7):469–477.

Iyer, S. and Killingback, T. (2016). Evolution of Cooperation in Social Dilemmas on Complex Networks. *PLOS Computational Biology*, 12(2):e1004779.

Khadjavi, M. and Lange, A. (2013). Prisoners and their dilemma. *Journal of Economic Behavior & Organization*, 92:163–175.

Macy, M. W. and Flache, A. (2002). Learning dynamics in social dilemmas. *Proceedings of the National Academy of Sciences*, 99(Supplement 3):7229–7236.

Nowak, M. A. (2006). Five Rules for the Evolution of Cooperation. *Science*, 314(5805):1560–1563.

Ohtsuki, H. (2018). Evolutionary Dynamics of Coordinated Cooperation. *Frontiers in Ecology and Evolution*, 6:62.

Perc, M., Gómez-Gardeñes, J., Szolnoki, A., Floría, L. M., and Moreno, Y. (2013). Evolutionary dynamics of group interactions on structured populations: a review. *Journal of The Royal Society Interface*, 10(80):20120997.

Rand, D. G., Nowak, M. A., Fowler, J. H., and Christakis, N. A. (2014). Static network structure can stabilize human cooperation. *Proceedings of the National Academy of Sciences*, 111(48):17093–17098.

Segismundo S. Izquierdo, L. R. I. a. N. M. G. (2008). Reinforcement Learning Dynamics in Social Dilemmas.

Smith, J. M. and Price, G. R. (1973). The Logic of Animal Conflict. *Nature*, 246(5427):15–18.

Traulsen, A., Semmann, D., Sommerfeld, R. D., Krambeck, H.-J., and Milinski, M. (2010). Human strategy updating in evolutionary games. *Proceedings of the National Academy of Sciences*, 107(7):2962–2966.

Yamauchi, A., Tanimoto, J., and Hagishima, A. (2010). What controls network reciprocity in the Prisoner's Dilemma game? *Biosystems*, 102(2-3):82–87.

Appendix