Community Building and Segregation through Prisoner's Dilemma

Alexandre Pires, Diogo Fouto, João Fonseca

October 2021

Abstract

The emergence of communities is a ubiquitous occurrence. For millennia, the "us vs them" mentality has resulted in the creation of small, inconsequential tribes and factions. Now, in our ultra-small world, millions of lives are subjugated to the whims of scale-free communities. Many models try to predict the emergence of such communities, but some of its details remain unknown. We provide a new model for personal interactions based on individual beliefs and group prejudice that sheds some light on these. We show that communities – and, therefore, polarization – resembling the society we live in today, emerge as a result of such interactions. This suggests that the affection for our beliefs/communities and prejudice for others play a key role in community building and, therefore, segregation.

1 Introduction

Why is there cooperation? Although the Theory of Evolution, seemingly, contradicts its existence, our species is the ultimate proof of its effectiveness.

Several models based on evolution try to explain it. [1], for instance, presents five mechanisms for its emergence: kin selection, group selection, direct/indirect reciprocity, and network reciprocity. Dawkins, [2], embraces the idea of a "gene-centred view of evolution", which states that the more individuals are genetically related, the more sense it makes for them to cooperate. These ideas may help explain cooperation among small, intimate groups, but fail to acknowledge the ultra-small world of today.

Economics, too, have their own proposals. Classical economics, which upholds the existence of rational individuals, on the one hand, tries to justify it with Game Theory. Behavioral economics, on the other hand, defends the existence of gullible and biased people, and, as such, employs Evolutionary Game Theory to help explain the ever-present cooperation among them. Both schools of thought resort to experiments such as the Ultimatum Game and Prisoner's Dilemma to showcase their theories. [3], for example, defends rational cooperation in a finitely repeated Prisoner's Dilemma when there is incomplete information about the players' options, motivations, or behavior. Another example, [4], likens the evolution of cooperation with the evolution of fairness and suggests, through the Ultimatum Game, that both are linked to the role of reputation.

These models are compelling, but we believe they don't fully explain cooperation. This, we suggest, is because humans often give much too importance to their own beliefs and fictions, and not enough to the ones of those who disagree with them.

To address this, we introduce a new model for personal interactions. Each individual in a network is born with an innate belief and, through interactions with his neighbors, that belief is either strengthened or weakened. If the strength of the belief drops under a threshold, the individual changes his mind and starts believing the opposite. Each interaction is a match of Prisoner's Dilemma; a player decides his/her tactic based on the affection he/she has for his/her belief and prejudice for the other player's belief.

Our intuition is that clusters of individuals resembling the polarized world of today will emerge when this model is put to test.

2 Relevant Work

2.1 Prisoner's Dilemma

Our model had to allow the natural emergence of cooperation and competition, as well as the like and dislike for a group, by a node. If a node has bad experiences with a particular group, then it's less likely to cooperate with it, and vice-versa; it's belief will then, in turn, affect the collective reality of its community.

The Prisoner's Dilemma is a well studied example in game theory, as well as "one of the basic framework[s] for the study of evolution and adaptation of social behavior" [5]. The game is competitive, but allows the emergence of cooperation between participants.

As Prisoner's Dilemma allows for adaptation of a node's behavior in accordance to external stimuli, as well as the natural emergence of cooperation, we found it fitting for our objective.

3 Model and Methods

3.1 Playing the Prisoner's Dilemma

Let us consider N individuals, associated with nodes in a graph. We'll focus our efforts in an undirected Scale-free network topology, generated using the Barabasi-Albert algorithm, since the emergence of hubs is to be expected in real-world Opinion-dynamics scenarios. Therefore, our network will be generated with $\gamma = 3$. Experiences were also made using the Watts-Strogatz model, but results proved to be similar.

Similarly to [6], at each time step, each individual plays a round of the Prisoner's Dilemma with all his neighbors, only unlike the Ultimatum Game, only one round has to be played per link, as the game is symmetric for each side. Table 1 shows the resulting years in prison each player will get based on their actions, collaborating or defecting.

Player 1 / Player 2	C	D
1	1/1	3/0
D	0/3	2/2

Table 1: Prisoner's Dilemma Payoff Matrix

An individual i ($i \in [1, N]$) is characterized by one unique opinion value, $O_i[0, 1]$, that defines if the player is either from opinion A, that is $O_i < 0.5$, or opinion B, where $O_i \geq 0.5$. The reason behind having a real O_i, instead of a discrete value, lies in its ability to tell not only which opinion each individual has, but also how much they believe in it, which allows us to measure polarization more precisely. Let us define the function above, that maps each O_i to the value 0 for opinion A, and 1 for opinion B, as $d(i) = round(O_i)$.

The decision each player takes is based on their discrete opinion, that is, either A or B. The way this is decided is using a table similar to Table 2, simulating through an individual's opinion, the collective opinion's believes. Each entry will hold a value $P_{ij} \in [0, 1]$, which defines the possibility of a player i, of opinion $d(O_i)$, collaborates (holds silence) against a player j, of opinion $d(O_j)$. Consequentially, the probability that player i will defect (betrays) j is given by $1 - P_{ij}$. Each entry in P will be started with 0.5 for an unbiased start.

At each round, a player i will generate a random probability value which will be compared with P_{ij} (where j is the opponent) in order to decide if he shall collaborate, or defect, and so will player

Player 1 / Player 2	A	B
A	P_{AA}	P_{AB}
B	P_{BA}	P_{BB}

Table 2: Collective Opinion Table

j, comparing against P_{ji}. In order to enable mutual collaboration, which isn't the rational action if a player knows the other will collaborate, there is a chance of collaboration, P_c, if a player predicts another will collaborate as well. This chance is given by Eq. 1. If the prediction is that the other player will defect, the player will defect too.

$$P_c(i,j) = |O_i - O_j|^2 \tag{1}$$

3.2 Updating the beliefs

Each player i's opinion value, O_i, shall be updated considering the total sum of years the player received from his neighbors of the same opinion at time-step t, $S_i(t)$, and the from the opposite opinion, $S_i^-(t)$. using Eq. 2, where $delta(t)$ is given by Eq.3, W_o is a weight that states how important the past opinion, and, inversely, the current experience, was. In order to reduce the number of free parameters in our experiments, we fixed W_o at 0.95. In Eq.3, T_i is given by the inverse sum of the total payoff of i in round t that was given by players of the same tag, and T_i' the same, but of different tags; S_i is given by the total number of neighbours with the same tag as i, and S_i' is similar, but for different tags.

$$O_i(t) = O_i(t-1) * W_o + \frac{delta_i(t)+1}{2} * (1 - W_o) \tag{2}$$

$$delta_i(t) = \frac{T_i'(t)}{S_i'} - \frac{T_i(t)}{S_i} \tag{3}$$

That means that a player could, potentially, change from opinion A to B or vice-versa, or believe more strongly in his opinion, all based on his personal experiences with players of the same or opposing opinions, even if acting through the believes of the collective opinion's perception of what another player will do, based on his opinion.

Each entry in P will also suffer updates after each time step, similarly to O, but considering that every player of the same tag is only one, with the sum of all constituent player's experiences, that is, the total amount of collaborators and defectors that the players of an opinion met, and the payoff obtained by them. This is done for each possible pair of opinions, following equation 4, where $o, p \in [A, B]$. W_o is still used here, as we're considering a community to behave like an individual. In Eq.5, T_{op}, T'_{op}, S_{op} and S'_{op} are similar to their individual opinion counterparts, but considered for the entire community.

$$P_{op}(t) = P_{op}(t-1) * W_o + \frac{delta_{op}(t)+1}{2} * (1 - W_o) \tag{4}$$

$$delta_{op}(t) = \frac{T'_{op}(t)}{S'_{op}} - \frac{T_{op}(t)}{S_{op}} \tag{5}$$

For example, if all A's put together experienced a bigger amount of defectors when playing with B's, they will decrease the believe that they collaborate, in this case, by increasing value of the corresponding entry in P_{AB}. This will be done for all pairs of opinions, resulting on either acceptance, where individuals of different opinions believe in collaboration, or prejudice, where individuals of different opinions believe the others to be defectors. Notice that the same behaviors can happen with individuals of the same opinion, leading to distrust inside opinion sharers. Another important factor is that P needs not to be symmetric, leading to cases where prejudice or acceptance are more prominent in one opinion than in the other.

4 Results and Discussion

Case for no bias between any pair of opinions: AA=0;AB=0;BB=0;BA=0 -When no bias is introduces before hand, we see that very little variation exists in regards to the total number of players of each tag. This is expected as this is a symmetric situation. -It is also shown that if no bias exists,

this leads to little to no bias, in the long run, between any pair of opinions. -Opinions tend to merge to a central point, as it is more beneficial for them to collaborate, as no bias exists.

Case for a symmetric bias between opposite opinions: AA=0;AB=0.5;BB=0;BA=0.5 -When a symmetric bias is introduced, it is shown that populations tend to keep their biases. Coincidentally, even if extremism still perishes, opinions do not merge as strongly as in the case for no bias. -One of the tags, A, is also shown to always turn into the minority in this case. This could be a result of a bias in our testing systems.

Case for an one-sided bias between opposite opinions: AA=0;AB=0.7;BB=0;BA=0 -TODO

Case for an one-sided bias between the same opinion, that is, an unstable community: AA=0.7;AB=0;BB=0;BA=0 -When a community is shown to be unstable, that is, have a bias towards itself, we see that no other bias is generated, and the inside bias is shown to be stable. That, however, leads to a rapid decrease of that opinion, as the other leads to more collaborations.

Case for an one-sided bias between the same opinion, from an unstable community: AA=0.5;AB=0.5;BB=0;BA=0 -Similarly to the results for an unstable community with no outsider bias, the unstable community sees their population size reduced, as they move to a more stable and collaborative community. Insider bias seems to always stay stable, which might be a result of a lack of interactions after the population numbers decay. -This time, we see a bias being originated by B against A, ending up merging with A's decaying prejudice against B.

Since both the update of the collective opinion table and of the individual belief is done at the same time, it is often unclear if one caused the change in the other. TODO CONTINUAR

It is observable that no matter how a collective's bias towards themselves is initialized, that is, P_{AA} or P_{BB}, this value will stay relatively stable throughout time. TODO HYPOTESE

It is also noteworthy that if starting at the same values, both P_{AB} and P_{BA} will also tend towards the same value. If the population is unbalanced in the starting prejudice, either $P_{AB} \geq P_{BA}$ or vice versa, both will converge to the same value, which is the case since one will start of by collaborating against a highly defecting opponent, therefore the collaborators will adopt a higher bias, leading to less collaborations. The other collective responds by lowering their bias (TODO WHY), until a balance is reached.

Extremism, that is, a strong belief in one opinion, is also shown to dissolve in the long run, as it is more beneficial to opt for a more common opinion, as mutual collaboration occurs more frequently the smaller the gap between each player's opinion is. More mutual collaboration leads to a better payoff, so players end up tending to more central opinions, that is, around 0.5.

An important point to be made, is that since the updates depend on the interacting between all opinion combinations, whenever an opinion becomes more prevalent than the other, interactions between the opinion in minority become scarce, leading to a stagnation in the update of the collective bias of that minority. Curiously, the opinion with the majority will also have the same bias, so P_{AB} and P_{BA} will always be similar. TODO HIPOTESE

5 Conclusion

6 Future work

Several additions could be made to the model, as in to better simulate the different components of opinions. Such addition would be to give more weight to the experience of nodes with a stronger opinion when updating the collective table, instead of giving equal weight to every player.

Other possible results could be obtained by experimenting with different starting values in P, or with other values of W_o.

The team also experimented with a forgetting factor, where if $delta()$ was zero, that is, a player/collective was neither worst of better in a time step, the belief/bias would tend towards a neutral position, 0.5/0 for the player and the collective respectively. The idea was abandoned as in not to further complicate the analysis of the results, but it is expected that it would better model the dynamics of a social system.

Finally, various other equations could be used to update the opinions, biases and the mutual collaboration probability.

References

[1] M. A. Nowak, "Five rules for the evolution of cooperation," *Science*, vol. 314, no. 5805, pp. 1560–1563, 2006.

[2] R. Dawkins, *The Selfish Gene*. Oxford university press, 1976.

[3] D. M. Kerps, P. Milgrom, J. Roberts, and R. Wilson, "Rational cooperation in the finitely repeated prisoner's dilemma," *The Economic Journal*, vol. 27, pp. 245–252, 1982.

[4] M. A. Nowak, K. M. Page, and K. Sigmund, "Fairness versus reason in the ultimatum game," *Science*, vol. 289, no. 5485, pp. 1773–1775, 2000.

[5] A. Barrat, M. Barthelemy, and A. Vespignani, *Dynamical processes on complex networks*. Cambridge university press, 2008.

[6] R. Sinatra, J. Iranzo, J. Gomez-Gardenes, L. M. Floria, V. Latora, and Y. Moreno, "The ultimatum game in complex networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2009, no. 09, p. P09012, 2009.