Community Building and Segregation through Prisoner's Dilemma

Alexandre Pires - 92414, Diogo Fouto - 93705, João Fonseca - 92497

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 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, in [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 in 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 in the opposite belief. Each interaction is a match of Prisoner's Dilemma; a player decides its tactic based on the affection it has for its 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, competition, and prejudice traits in an individual/node. If a node has had bad experiences with a particular group/community, then the probability of its cooperating with them should decrease, and vice-versa. The node's belief and prejudice would, then, affect the collective prejudice/bias of the community to which it belongs.

The Prisoner's Dilemma is a well studied experiment in Game Theory and "one of the basic framework[s] for the study of evolution and adaptation of social behavior". [5] Since the game rewards the adaptation of a node's behavior to external stimuli and allows the emergence of cooperation (if our way of playing incites it), we adapted it for our purposes. The game also has a very important win-or-lose component that we thought was crucial for reasons we will explain in the next subsection.

2.2 Opinion Dynamics Models

There are various ways to simulate the spread and behavior of opinions throughout a network.

The simplest one is the **Voter Model**, in which a node selects a random neighbor and copies its opinion, in each time-step. [6] We found this to be too simplistic and not a good depiction of what happens in the real world, so we discarded it.

Then, there's also the **Majority Rule** model, which, in its most basic form, means that every node chooses the most widespread opinion in its neighborhood (or a random "discussion group", if we assume a mean field); this is very simplistic, but a step towards **social inertia**. [6]

A more interesting model related to Majority Rule combines the concepts of **Majority Preference** (MP) and **Minority Avoidance** (MA). In this model, MP means that a node accepts the opinion of the majority of its neighbors with a probability p. If it doesn't, the node recurs to MA and, with a probability $1-p$, removes one of its links to a node with the least widespread opinion in its neighborhood; then, with a probability ϕ, the node rewires itself to a random neighbor of one of its neighbors with the same opinion as itself, or, with a probability $1-\phi$, to a random node in the network who's not one of its nearest neighbours. [7] While this model is much more interesting than the previous two, we still found it too extreme. In the real world, people don't instantaneously change their opinion to the majority's — it's a gradual process. People also don't abruptly cut ties with people who have an unpopular opinion — they might, but it's not instantaneous.

There also were two other components that we thought were missing from the models mentioned before. First, opinions don't exist in a vacuum: they are influenced by the actions and behaviours of those who hold those opinions. Because of that, we thought that an opinion dynamics model could never be separated from a model in which nodes interact with each other at a different level, where each of them wins or loses something: social credit, a job opportunity, liking for another person, or anything that can be modelled as a win or a loss. Secondly, and this is related with node behaviour, group biases will affect the interactions between nodes — whether we like it or not, we all make assumptions about the other person not only based on our personal experience, but also based on which labels we attach to them, and those are deeply influenced by the groups we personally identify with.

To address these problems, we created our own model of opinion dynamics, which we will discuss in the next section.

3 Model and Methods

3.1 Playing the Prisoner's Dilemma

For our model, we chose an undirected scale-free network topology with N nodes, generated using the Barabasi-Albert algorithm, since the emergence of hubs is expected in real-world social scenarios, and, thus, our network will be generated with $\gamma = 3$. Experiments were also made using the Watts-Strogatz model, but the results were similar.

Similarly to [8], in each time step, every individual plays a match of Prisoner's Dilemma with all its neighbors. But, unlike the Ultimatum Game, only one match is played per link, since the game is symmetric.

Table 1 shows the punishment (this is number of years in prison in the original story) a player would suffer given its action.

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

Table 1: Prisoner's Dilemma Payoff Matrix

Each individual i ($\forall i \in \mathbb{N} : i \in [1, N]$) has an opinion attribute $O_i \in [0, 1]$ that tells us if the player believes in A, $O_i \leq 0.5$, or B, $O_i > 0.5$. We chose a continuous O_i instead of a discrete one to precisely measure an individual's belief strength and overall group polarization. We define the function that maps each O_i to the value 0 for opinion A, and 1 for opinion B, as $d(i) = round(O_i)$.

The tactic each player chooses in a match is influenced by their discrete opinion: A or B. This is decided using Table 2, which simulates the effect of the group prejudice in one's actions. Each entry holds a value $P_{ij} \in [0, 1]$ that represents the player i's perceived likelihood of player j defecting. Thus, the player i's perceived likelihood of player j cooperating is given by $1 - P_{ij}$. We experimented with different initial values for these group biases (more on this in *Results and Discussion*).

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

Table 2: Collective Opinion Table

Because we are modelling human behaviour, we base our player's choice on the idea that players should collaborate more if they share opinions, and that keeping a stable mutual collaboration network is the long-run logical strategy (and not defecting, since that's only rational for short-term success). To do that, a player i will assess how likely the other is to defect according to i's group beliefs (given by Table 2), and will decide to defect with the corresponding probability (which is the rational choice, if one believes the other player will defect). If the player thinks the other will cooperate, then it will also collaborate with a probability given by Eq. 1.

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

Thus, players with a similar opinion are more likely to engage in mutual collaboration.

3.2 Updating the beliefs

We update each player i's opinion value, O_i, considering the outcome (the payoff in Table 1) of its matches with its neighbors for each time-step t. Eq. 2 states the update rule. We assign a weight to both the previous opinion and the current experience with ω_o. We fixed ω_o at 0.95 to reduce the number of free parameters in our experiments. T_i is the sum of the add-one inverse payoffs gained from matches with neighbors of the same opinion. T'_i represents the same thing, but for neighbors with different opinions. $S_i(t)$ and $S'_i(t)$ represent the number of neighbors of the same opinion and of the opposite one, respectively, at time-step t.

This update rule allows a player to change opinions and/or believe in them more strongly based on its previous matches with neighbors and on its group's bias toward the other one.

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

$$\delta_i(t) = \begin{cases} \frac{T'_i(t)}{S'_i} & \text{if } S_i = 0 \\ -\frac{T_i(t)}{S_i} & \text{if } S'_i = 0 \\ \frac{T'_i(t)}{S'_i} - \frac{T_i(t)}{S_i} & \text{otherwise} \end{cases} \tag{3}$$

$$T_i = \sum_{j \in neigh(i):d(j)=d(i)} \frac{1}{t_j + 1} \tag{4}$$

We also update, for each time-step, the entries of P similarly to O: communities — collections of nodes with the same opinion — are processed as nodes; a community's payoff is the sum of the payoffs of all nodes of that community. Eq. 5 expresses that update for each pair of opinions, $o, p \in \{A, B\}$ in the Prisoner's Dilemma matrix. Eq. 6 gives us the **relative defection** of tag p as experienced by tag o; $D_p(t)$ and $C_p(t)$ are the defectors and cooperators, respectively, of the tag p at time-step t, and $I_{op}(t)$ the total interactions between o and p at time-step t.

This means that if the total of players from community A played with more total defectors from community B than in the previous time-step, P_{AB} – A's perceived likelihood of B's defection – will increase. The same can be said for all the other cases (players from A playing with more cooperators from community A and so on).

$$P_{op}(t) = P_{op}(t-1)\omega_o + \frac{RD_{op}(t)+1}{2}(1-\omega_o) \tag{5}$$

$$RD_{op}(t) = \begin{cases} \frac{D_p(t)-C_p(t)}{I_{op}(t)} = 2\frac{D_p(t)}{I_{op}(t)} - 1 & \text{if } I_{op}(t) \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

4 Results and Discussion

Here we present the results obtained in our experiments. We conducted several simulation runs and collected the *"bias between tags"*, *"average belief strength of each tag"*, and *"number of believers in each tag"* measurements for each time-step of each run. The simulation experiment was run with 5 different starting global biases, detailed in the following paragraphs. All the other conditions — network topology, number of nodes (N), average degree, weight (ω_o) and number of time-steps — were kept fixed. Our parameters were, then: $\omega_o = 0.95$, $N = 1000$, the number of time-steps were 200, and the average degree was 6. Each experiment was run 30 times. Let us now see each case in more detail.

4.1 No Prejudice Start

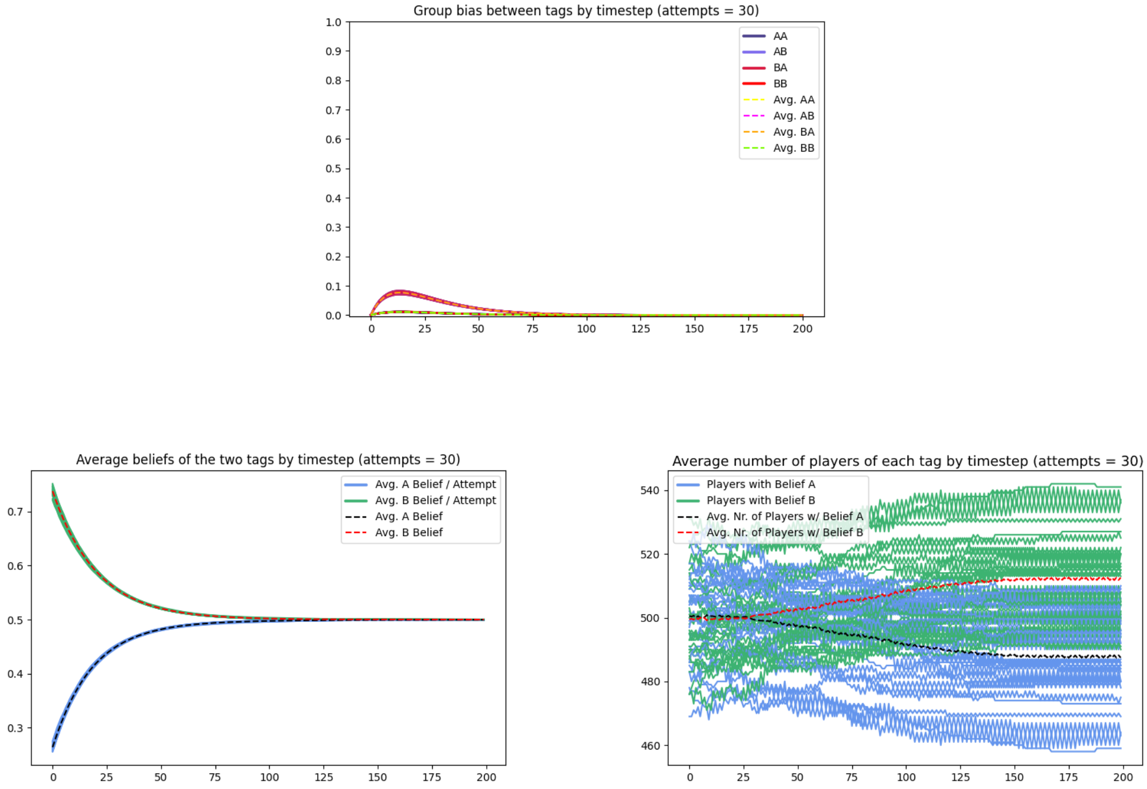


Figure 1: Results for $P = [[0, 0], [0, 0]]$. Top figure: Each community's bias by time-step; Left figure: Average individual belief by time-step; Right figure: Average number of players of each tag, by time-step.

When we start with no prejudice, we can see that the variance is tiny (about 20-40 nodes in each direction, in a network with 1000 nodes), regarding the total number of players of each tag, in each run. Some of those nodes are constantly changing tag. The group bias between each tag tends to converge to 0 after 50 time-steps, and the average belief of each tag's population is very similar to each other after that time-step, as well.

The high belief in cooperation each player has for their neighbours, as shown by the group bias graph, results in a very low chance of believing the other will defect, which causes the node to not choose defection right away. In turn, as the average beliefs of each tag's population get higher or lower (for A and B respectively) because of their good experiences, there's a higher chance of cooperation since the distance between each tag's belief lowers, resulting in more and more cooperation. All of this makes it so that players don't cling to a particular tag.

What will happen if we introduce prejudice from one tag to the other?

4.2 One-sided Inter-tag Prejudice Start

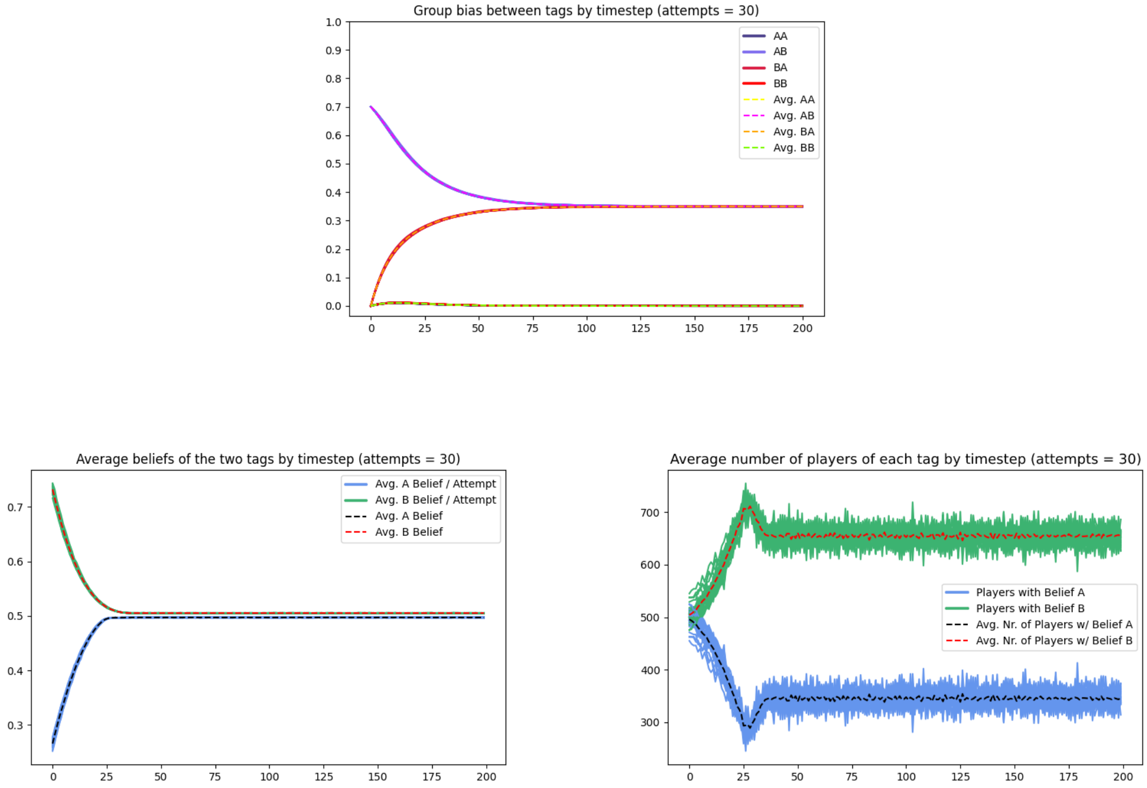


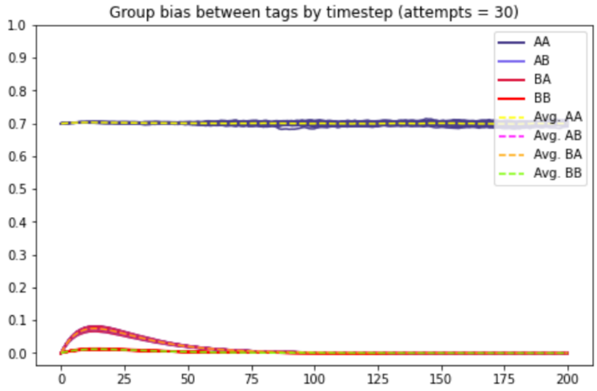
Figure 2: Results for $P = [[0, 0.7], [0, 0]]$. Top figure: Each community's bias by time-step; Left figure: Average individual belief by time-step; Right figure: Average number of players of each tag, by time-step.

In this case, even though only the tag A has prejudice against the tag B, the tag B gains prejudice by playing with As, since they experience a great amount of defections. The prejudice A has for B lowers over time, because of their good experience with Bs. Eventually, this mutual prejudice converges and remains constant. This means that prejudice emerges when distrust is created between communities, and then remains stable if nothing is done against it.

It's worth noting that the number of As decreased very rapidly (minimum at $t \approx 25$), because they had large amount of good experiences with B.

Now that we tested prejudice from a tag to the other, what happens when a group doesn't trust its own members?

4.3 Intra-tag Prejudice Start



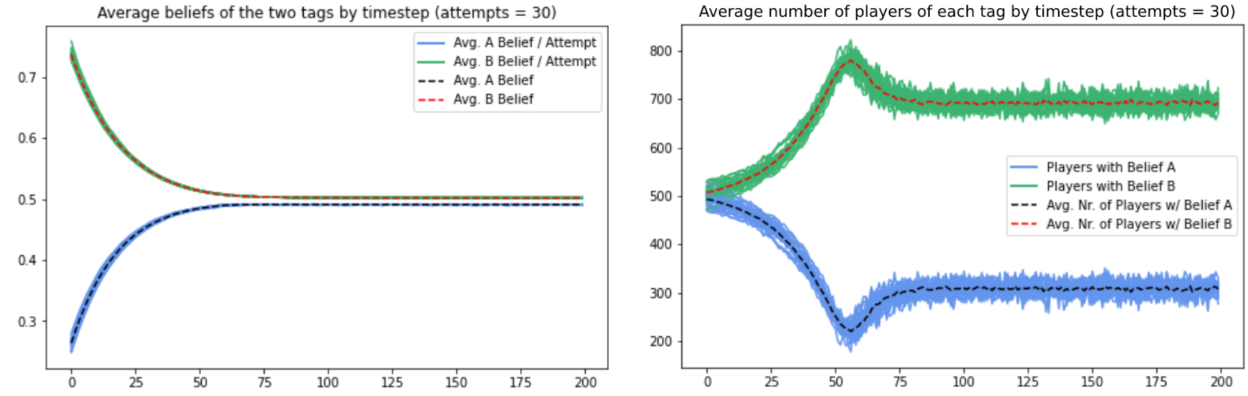


Figure 3: Results for $P = [[0.7, 0], [0, 0]]$. Top figure: Each community's bias by time-step; Left figure: Average individual belief by time-step; Right figure: Average number of players of each tag, by time-step.

When there's distrust inside a community, we can see that no other prejudice is generated, and the intra-tag prejudice is shown to be stable. That leads to a rapid decrease of the total members of that tag (with a minimum at $t \approx 50$), as the experiences with the other tag are better. The decrease, although rapid, is not as rapid as the one seen in the previous simulation, with one-sided prejudice.

An interesting thing to note is that the total number of members of A stabilizes at, approximately, 300, lower than in the other simulations. This suggests that, for a community where distrust exists to remain a community, it needs to have a lower number of members.

Now, let's see what happens to communities who don't trust each other.

4.4 50% Prejudice Between Tags Start [1]

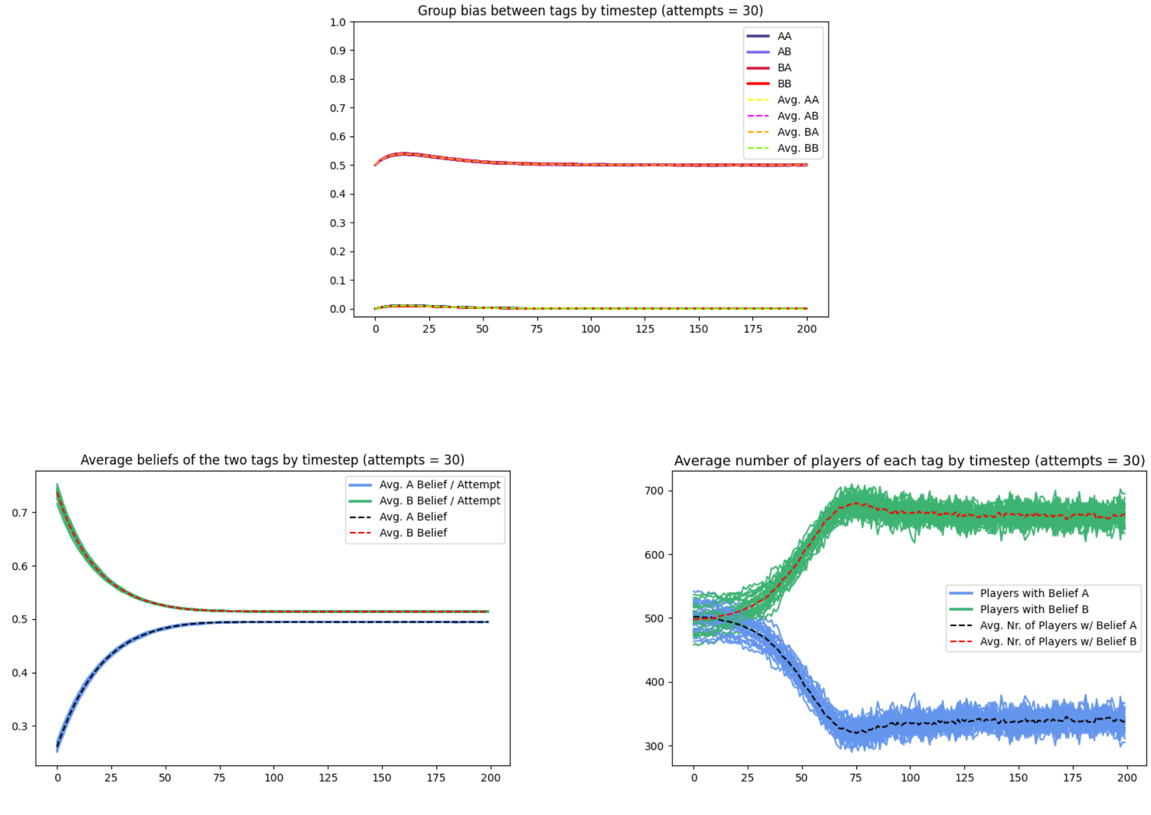


Figure 4: Results for $P = [[0, 0.5], [0, 0.5]]$. Top figure: Each community's bias by time-step; Left figure: Average individual belief by time-step; Right figure: Average number of players of each tag, by time-step.

[1] Tag A, is shown to always become a minority in this case. This could be a result of a bias in our testing system or in our model.

As we can see, the average beliefs of each tag's population continues to converge, and gets very similar, but not as similar as in, for example, experiment 1 or 2. The inter and intra-tag group bias remained relatively constant.

The lower similarity between average beliefs is possibly due to the prejudice each tag has for the other, causing a lower amount of good experiences with the other tag, and hence each player has less reasons to change their belief. Because there's less reasons to change their belief, the number of members of each tag is more stable.

The inter-tag prejudice remains stable possibly for the same reason there's a lower number of tag changes — if each tag believes the other has a 50% chance of defection, then each player of such tag will defect at least 50% of the times they play with the other tag. This makes it so that at least 50% of the interactions each tag has with the other will be defections, and hence the bias remains approximately the same.

One last question remained: what would happen to a community that had internal distrust and prejudice against another community?

4.5 One-sided Intra and Inter-tag Prejudice Start

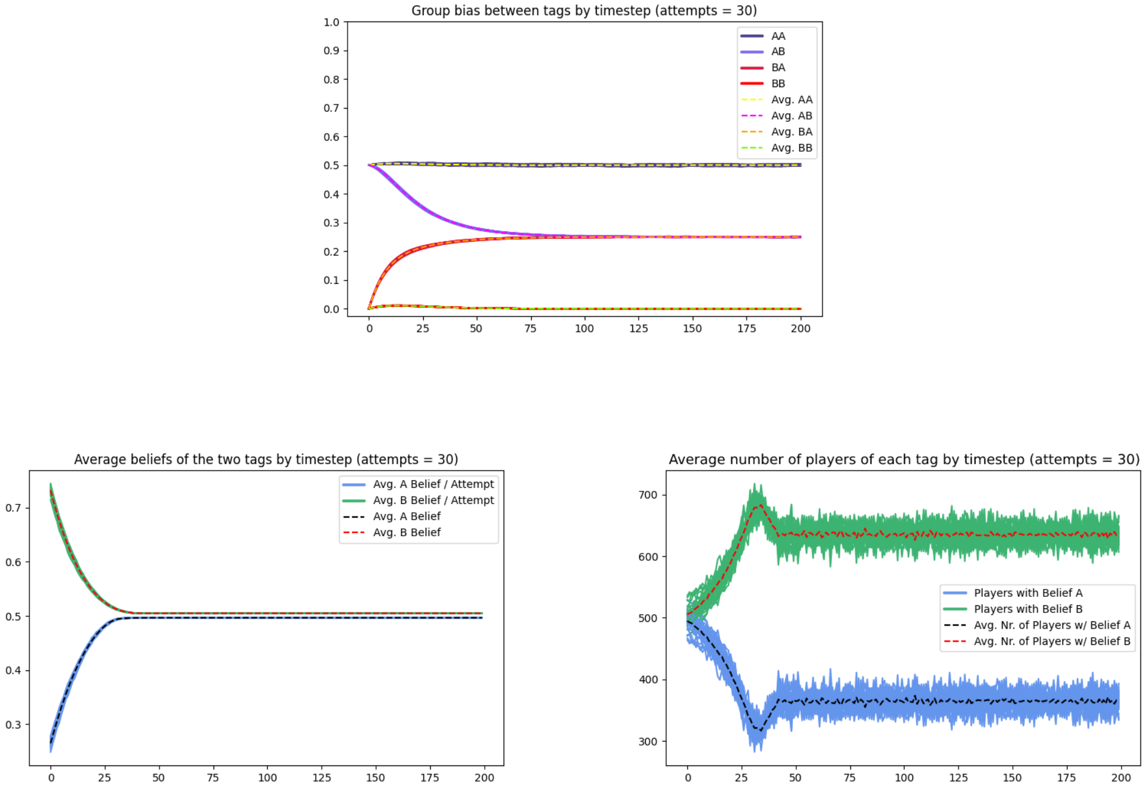


Figure 5: Results for $P = [[0.5, 0.5], [0, 0]]$. Top figure: Each community's bias by time-step; Left figure: Average individual belief by time-step; Right figure: Average number of players of each tag, by time-step.

Similarly to previous results, the community where there's internal distrust experiences a rapid decline in its members, at a pace akin to experiment 2. It's also similar to that experiment in the emergence of prejudice from B to A.

The size of the community A is approximately the same as the one achieved in experiment 2. This suggests that communities with internal distrust can have a larger amount of members if there's also distrust between different communities.

4.6 Remarks

Here are some insights we extracted from our results:

- Naive communities (those with zero distrust towards members of their and other communities) tend to have members that constantly switch between tags.

- No matter how intra-group prejudice (P_{AA} or P_{BB}) is initialized, this value will remain stable throughout time.

- Communities with internal distrust tend to be smaller than communities with no internal distrust. The only way that a community with internal distrust to increase its numbers is to have inter-community distrust.

- Internal distrust is the fastest way for a community to decrease in number of members.

- If internal distrust is present, inter-community distrust slows down the decrease in members of such communities.

- If starting at the same values, both P_{AB} and P_{BA} will remain constant.

- If the inter-tag prejudice is asymmetrical ($P_{AB} \geq P_{BA}$ or vice-versa), both will converge to the same value. An explanation of this phenomena is that one will start of by collaborating against a highly defecting opponent, therefore one will adopt a higher prejudice towards the other tag, leading to less collaboration. Through a similar thought-process, the other group responds by lowering their bias. This happens until an equilibrium is reached.

- Extremism, i.e., a strong belief in one opinion, is also shown to dissolve in the long run. It is more beneficial to opt for a wide-spread 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, around 0.5.

- Since the updates depend on the interaction between all tag combinations, whenever a tag 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. Interestingly, the most wide-spread opinion will have the same bias inter-group bias as the minority, so P_{AB} and P_{BA} will always be similar.

5 Conclusion

We presented a new model for personal interactions that addresses the lack of importance given by previous ones to the individual's beliefs and prejudice's role in community building, segregation, and individual action. We also developed a progressive way of updating a node's beliefs based on its personal experience.

We conducted simulation experiments to test our hypothesis and found that this new model helps explain the laudable cooperation and vile segregation our species persistently displays. We observed that: i) minorities are overwhelmed by big communities; ii) group prejudice from A to B creates group prejudice from B to A; iii) one's extremism withers in favor of cooperation with neighbors.

These findings present new challenges and suggest some paths for further work. Here is an example: considering that belief and prejudice build and break communities, and that they are often incorrect and harmful, is it possible to build robustness to bias the same way we build robustness to link-removals in a network? Such innovative thinking would help solve some, if not most, of the ever-so-present segregation and polarization witnessed in the world of today.

Several extensions could be added to our model in the future. For instance, a heavier weight could be assigned to nodes with a stronger belief when updating the group biases. This would account for the role of extremists in our society. Another idea would be to try to run more simulation experiments with different starting group biases. This could help find the exact bias strength present in our world. One could also try adding some sort of "gossip", in which nodes "share" their experiences with each other and, therefore, influence one another. Another possible extension would be the existence of more than two tags. Experimenting with a forgetting factor for each node could also yield interesting results. One way to go about implementing this would be to nudge the bias/belief strength towards a neutral position as $\delta(t)$ approached 0. Finally, an effort could also be made to generalize this model to a higher number of tags, since our current model is limited in this regard by the equations used.

To conclude, we helped shed some light on the importance of beliefs and prejudice in our communities, but there is still much research to be conducted. We can only hope our work will be useful for those at the front lines of opinion dynamics.

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] C. Castellano, S. Fortunato, and V. Loreto, “Statistical physics of social dynamics,” *Reviews of modern physics*, vol. 81, no. 2, p. 591, 2009.

[7] F. Fu and L. Wang, “Coevolutionary dynamics of opinions and networks: From diversity to uniformity,” *Physical Review E*, vol. 78, no. 1, p. 016104, 2008.

[8] 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.