Game Theory and Collective Phenomena: Patterns, Strategies and Clustering

Author: Antoni Domènech Borrell
*Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.**

Advisor: Josep Perelló.

Abstract: We did a behavioural study of persons playing collective games. We've seen how players interact with each other when they have to accomplish a shared goal and completed a study of how the in-game unequalities are distributed using the Gini Coefficient. Finally, we did a study (through Mutual Information) of how the betting profile of the individuals matched with others and how much of this match was due to sociodemographical traits.

I. INTRODUCTION

Trying to measure social phenomena through some physics isn't something new [1]. Here we are trying to understand the cooperation between humans under certain circumstances through statistical physics. Making an approach study of complex systems using modern tools. Our objective is, from public experiments made on citizen science [2] trying to measure how the acts of other ones influences our own.

We have used three different datasets coming from the experiments following the methodology from [4]. Two of them had the same idea behind them. 6 players starting with a total pot of 40€ each one and they are playing together. They can view what the others are betting round per round (see Fig 1). There are 10 rounds. The betting possibilities are 0, 2 or 4. And the goal is to accumulate a total collective pot of 120€ and can keep the money units not invested to their own. But if they don't reach the pot they lose all money. This game was invented in [3] As you can see, a player alone hasn't enough pot (nor time) to achieve it alone. So everyone will need to play as a group without having any kind of communication but the quantity that each player is betting. This will create some tensions between the greedy ones, the logical ones or the kind ones. We are going to look at these tensions.

The same dilemma also posed with a little diference. The starting pot is unequal. This has been made this way to detect if the ones with more are the ones who give more or not. The data coming from this variation conforms the last dataset.

Apart from the contribution for each player per round we have another amount of info. The sociodemographic of each player (gender, level of studies, age range...).

To realize the study below we: do the data cleaning, search for useful representations that give us information of whatever type and finally, model how different kind of strategies does the people follow for betting and if the shared strategies mean shared sociodemographic features.

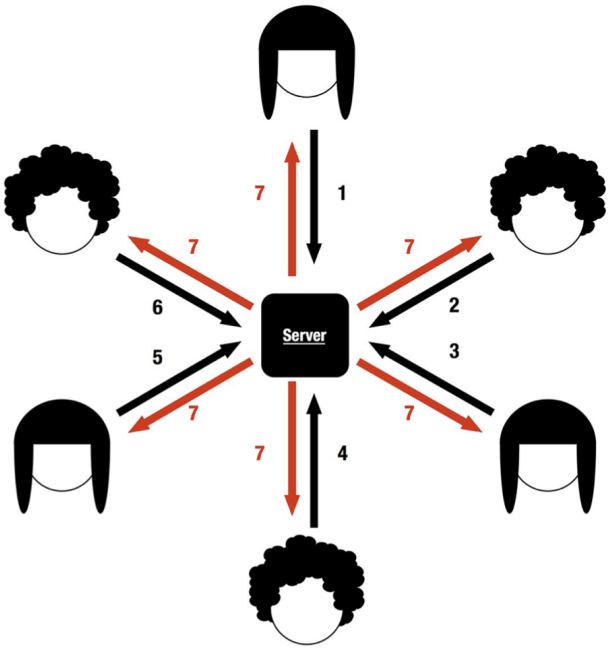

FIG. 1: Representation of how the game works. The black arrows meaning the player's bet and the red ones the output they're recieving. The output is the bets the other persons made before. Figure extracted from [4].

II. DEVELOPING SECTIONS

A. Data

We had three different games and each game was packed into different datasets containing pieces of information. The evolution for each serie, information of the players, contributions and more (see Table 1).

So, first of all, we need to do a data-preprocessing. We ended having a Pandas Dataframe per game which gave us the contribution of each player, the evolution (sum of contributions) per player, the overall money per game id, average values and extremes. We were searching for a unified way to organize the data too because we would like to use the same code to perform different algorithms. We also had to use some package like regex to clean the

*Electronic address: tonidome@gmail.com

Aigua	Games (# 30) Sociodemographics (age, studies, location...) Personal Contribution (# 180 players)
Viladecans	Games (# 21) Rounds Sociodemographics (age, studies, location...) Personal Contribution (# 126 players)
Clima	Games (# 35) Rounds Sociodemographic (age, studies, location...) Personal Contribution (# 210 players)

TABLE I: The 3 different datasets. And how is the data distributed inside of each one.

data. You can follow all this process on the Github link in the appendix.

B. Behavioral Patterns

In order to start the study that follows, the best way was to simply see the contribution per game during each round.

The first feature is the probability computation of the contribution per bet. As we can see in the following figure, certainly they followed some strategies. In Fig 2

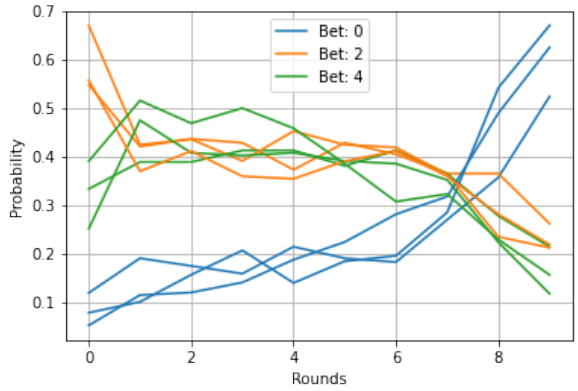

FIG. 2: Here we can observe on the y axis the probability of each betting value (0, 2 or 4) for each round on the y axis.

each one of the probability lines sharing color are from the different games. We can see that the trend is similar in all three different experimental settings. Despite the different locations or the initial game settings all three are sharing certain universalities. That's why from now on we are considering all as one. And, as we can see above, the probability of a bet with a value of 0 at the first round it's almost zero. But in the last rounds, once the goal is accomplished it clearly increases, showing a probability higher than 0.5. It's worth noticing that the most usual play is to bet 2 (a conservative one) but it's closely followed by betting 4 at the start of the game. It looks like the individuals have a rush on accomplishing the goal and show a tendency to be generous before seeing what others contributed.

If they were irrational players the betting expected would be an almost constant averaged contribution of 2 per round.

The game explains collective behaviour and this may lead to some strategies. To answer this question we proposed different approaches. The first one was to search if there was some kind of pressure in your decisions or if there were some genuine greedy players or genuine kind ones. We looked for the Gini Coefficient inside each game. The Gini Coefficient is defined as [1]:

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2n^2\bar{x}}, \tag{1}$$

where x_i is the bet of one player and x_j is the bet of the other one. In the denominator we have the number of players squared (n = 6 in this case) and the average value of the bet on each time step. This gives us an idea of the evolution of the inequalities formed by the bets through the game. This can give us an idea of the nature of

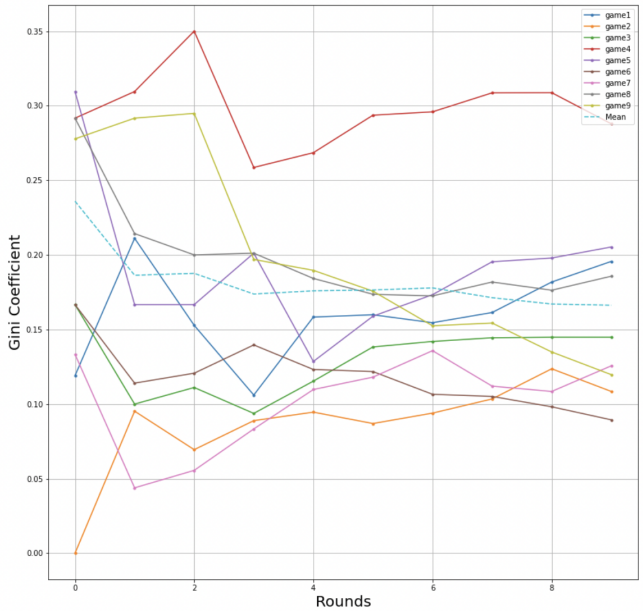

FIG. 3: Representation of the Gini coefficient of the "aigua" dataset. The y axis shows us the inequality and the x one the round where we are. To simplify this visualitzation we have used only 9 games and the mean (dashed line).

each player. In Fig 3, it's appreciable how at the start is the moment with the most uncertainty in the inequality (or Gini Coefficient). But it has a subtle trend to be the moment with more bias while beeing the round with

most inequality. But then this Gini Coefficient decreases until finding an almost stable value. This means that the low betters at the start gets cohibited by the high ones and feel some kind of pressure that pushes them to increase the amount of money in-game. The same thing happens with the high ones that they feel like they had already did what it needed to be done and reduce the intensity of the betting. The fact that the value of the Gini Coefficient is maintained constant but not zero shows that there are different types of strategies. If all of us played the same way the value of the Gini Coefficient would've been almost 0 and stable through the game. The fact we are looking to different slopes show that there are unequal fluctuations of money inside each game.

C. Mutual Information

The model we tried was to see if there is some kind of information inherent from each person, meaning which are the main demographic characteristics that made people bet as they did and if that information makes them follow similar patterns when it comes to betting. Once to this point, one thing that we had to do was to choose one method to calculate this shared information. We had two options here. To calculate the well-known Pearson Correlation or to use the Mutual Information (MI) algorithm. As it's discussed in [5] the mutual information has a very distinctive point that plays in his favour. The non-linearity. This leads us to high variations on the MI for little variations on the correlation. That's why we decided to go with Mutual Information.

We define a pair of players as:

$$N_i(t) \to X = x_0, x_1, ...x_n \tag{2}$$
$$N_j(t) \to Y = y_0, y_1, ...y_n \tag{3}$$

where the subindex means the number of the move and n is the total of moves possible. In our case is 9 since each player has 10 possible bets to do. And, as we've seen before this number of bets in order has meaning. So, according to [7], we compute the Mutual Information as:

$$MI(X,Y) = \sum_y \sum_x p(x_t, y_t) log_m \frac{p(x_t, y_t)}{p(x_t)p(y_t)} \tag{4}$$

where

$$p(x_t) = \frac{\sum_{i=0}^{n} \delta_{x_t, x_i}}{n}$$

and, applying bayes theorem,

$$p(x_t, y_t) = p(x_t|y_t)p(y_t).$$

Once we made this we have a symmetrical squared matrix with a side equal to the number of players that we have where A_{ij} is the shared information between the player i and j. We can print a heatmap. That says to us the intensity of the shared information between all the players.

This heatmap resulted to be to noisy. With this in mind, we had built a mask function.

Our mask has 2 different parts. The first one consists in calculate again the MI but this time we have had shuffled the array of each player. The elements of this matrix will be called B_{ij}. All of them are shuffled randomly so the shufflings between the i player and the j array have nothing in common. Furthermore, the shuffling of i calculated with the element j it's different from the one with j+1.

And we apply Heaviside's Theta (Θ) to the difference between $(A_{ij} - B_{ij})$. To filter out automatically relevant values. The effectivity of this filter is calculated as follows:

$$\frac{\sum_i \sum_j A_{ij} \cdot \Theta(A_{ij} - B_{ij})}{\sum_i \sum_j A_{ij}} \sim 70\%.$$

So 30% of our data was noise. Also, we normalize the heatmap (See Fig 4). The normalization is going to be:

$$A'_{ij} = \frac{A_{ij}}{max(A_{ij})} \tag{5}$$

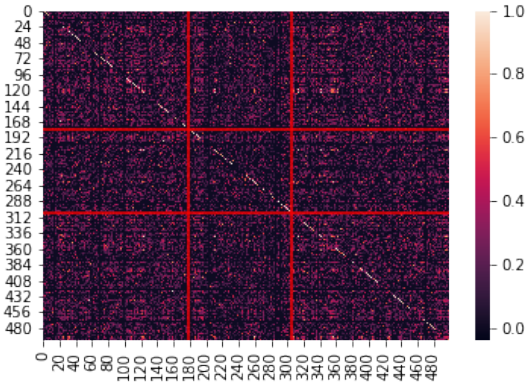

FIG. 4: Mutual Information on the three games unified matrix already filtered and normalized. The red lines are the separations between each dataset.

To increase the readability of the heatmap above we have used a Python package called "NetworkX" which has helped us to do a proper representation (Fig 5). The heatmap above it's a matrix where each point is the relation between those 2 players. Since it has been cleaned from the noise some values are zero. On the graph below, each node represents a player. We can see what was before A_{ij} as the links between the nodes. The function used to print this takes into account the minimum number of correlation necessary to draw a connection. Since our objective here it's purely a representation we have

used a minimum of 0.7. The intensity of each link the higher the value of the connection between those players. And the size of each node means more links starting at that point.

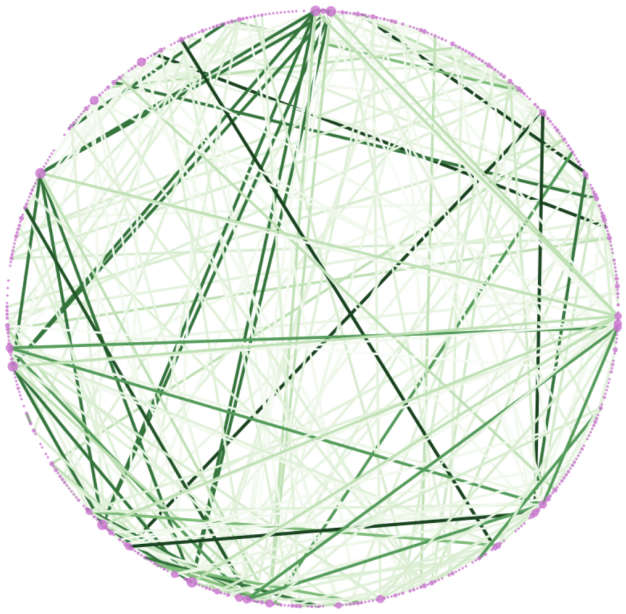

FIG. 5: Network representing all the players and it's connection. For clearness we have deleted the labels of the player. But the user1 is the one occuping the 15' of the sphere and it follows an counterclockwise order.

We tried two more things related with the Mutual Information: Compute the symbolical mutual information using markov chains registering the variation between 2 consecutive bets instead of the values of the bets we've done before. And trying to find leaders through the data. This was made by, on the MI expression [4] advancing one of the two players and computing it then. But we hadn't enough data to draw practical conclusions about this two things so we dropped them.

D. Clustering

After reading [6] chapter 9 about clusterization we have decided to go with a KMeans algorithm over a DBSCAN or an HDBSCAN. We did it following 2 main reasons: We wanted a determined number of clusters (which is the only hyperparameter needed on the KMeans). And we had a lot of hyperparameters to tune with DBSCAN or HDBSCAN. But it would be interesting to try to cluster using those algorithms.

Our points (remember that they consist only of the relation between one player and how his betting is related with others) will occupy the feature space. The KMeans algorithm would group those points around k centroids selected randomly and try to minimize the distance from each point to a centroid. Usually, and in our case, the distance used is the Euclidean one. The randomness of the process means that the final centroids won't always be the same.

The KMeans clustering needs only one hyperparameter, k. This k is the total number of clusters that we want to have. To determine k it's worthless viewing that we would have a huge accuracy having a lot of clusters but we would lose precision. So we are searching for a "k" not so big to have an overkilled number of groups but not so little (1 or 2) to take some rich conclusions.

To determine k we followed two approximations commonly known. Since the "k" value it's arbitrary we needed to make an educated guess so we used the silhouette and the elbow methods. The first one measures how similar is one point to its own cluster (cohesion) compared with other clusters (separation). So this silhouette score reaches its global maximum in the optimal k. The elbow method uses the squared error for each point and its own centroid and it sums them on what's called the WSS score. Making an elbow and having its optimal point on the minimum.

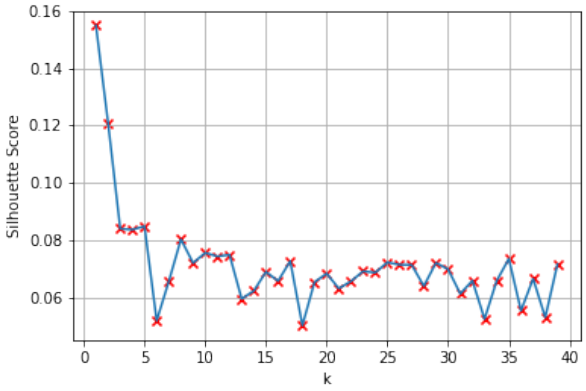

FIG. 6: Silhouette Score computed for different number of clusters (k). In the x axis we have the number of clusters. In the y axis we have the silhouette score.

With Silhouette Method (Fig 6), we can detect an obvious maximum for 1 cluster but it would have no sense to choose 1 for the number of clusters. So we are searching for other maximums (see Fig 6). The abundancy of candidates makes difficult to choose so we're using the Elbow Method too. We are searching for a drop on the WSS. But we haven't either a clear conclusion. We have an interesting drop with 4 clusters, an another one around 8. So we are finally going with this value. Since our whole dataset it's from $\sim$ 500 players this clustering gives us groups of 65 players.

Once we have made this we can plot an histogram of the sociodemographic information that we have from the players and separate them by the clusters where each one belongs.

In Fig 7 it's visible, how each cluster (diferentiated by color) reigns in one or another range of sociodemographic

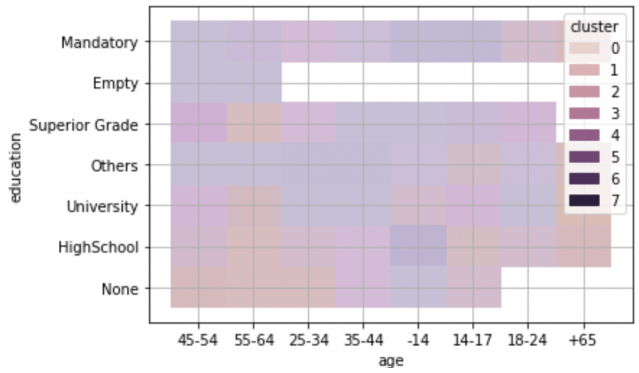

FIG. 7: Histogram containing the demographic information of all the players. On x axis we can see the age. On y the education. And the color shows the cluster predominating each quarter.

traits. So we have some cluster that is coherent with the age/education graph. If we dive deeper here we can see that the level of education is more important than the age range where you can be found.

III. CONCLUSIONS

Humans, while playing collectively follow strategies. There are different kind of strategies. Kinder ones, greedy others. We have seen that in Fig 2 showing the value of the bet probabilities of each round of the game. The first round have quite a lot more importance because the disparity there it's when its higher (Fig 3) and this, in general trends will play a significant role (but not determinant) on the rest of the game.

However, following with Fig 3, we can see how the essence of each player gets truncated by what the others do. The fact to see what others are doing influences our next decission. We have seen on the Fig 3 how the inequality would decrease through the game.

For the following conclusions we think the size of our datasets and the simplicity of the game had strongly influenced them. It would be interesting to build a more complex game or maybe to increase the data of this one.

It's possible to compute a Mutual Information algorithm with this kind of data and it would give us really different values. In fact, as we have seen in Fig 5, once the data has been cleaned it gives use a considerable network. And we strongly think that with the NetworkX package we should have an amount of tools that could bring this study further.

We indeed can see how the Mutual Information method can be used for clusteritzation and, as we have seen in figure 7 it is possible to abstract some conclusions from the sociodemographics of each player. We are considering here only sex, education and age. We could study in deep here. Obtain more sociodemographical data and increase our accuracy. As said before, the size of our data prevents us from being categorical with this conclusion. Either way, we have achieved some accurated distinctions between the clusters made off the betting strategies.

This distinctions gave more importance to the education than the age. But our sample is unbalanced so with this conclusion we have to say the same that we had said before. We should get a more balanced set.

IV. APPENDIX

All the advance in this study has been made using GitHub and properly pushing it. You can see all the historical of the project and more figures here: `https://github.com/menektoni/TfG`. You can read also a copy of the conclusions and some more figures in the "readme" section. There are the instructions to how to read the code to.

Acknowledgments

This thesis wouldn't have been possible without the indefatigable advice and pattience of Prof. Perelló, the punctual but altruistic help of PhD Student Ferran Larroya and the Physics Faculty of the UB.

I would like to thank my colleagues in physics, who made this time lighter and, off course, my family for supporting me since day one and being by my side independently what was happening around. DOG.

[1] Sánchez, A. "Physics of human cooperation: experimental evidence and theoretical models" IOP (2018).

[2] Vicens, J and Perelló, J. "Resource heterogeneity leads to unjust effort distribution in climate change mitigation " Plos One (2018).

[3] Milinski M., Sommerfeld, R. D.,Krambeck, H. J.,Reed, F. A., Marotzke, J.(2008) "The collective-risk social dilemma and the prevention of simulated dangerous climate change." Proceedings of the National Academy of Sciences, 105(7), 2291–2294.

[4] Vicens, J and Perelló, J. "Citizen Social Lab: A digital platform for human behavior experimentation within a citizen science framework." Plos One (2018).

[5] Taleb, N. "Fooled by Correlation: Common Misinterpretations in Social "Science", `https://www.academia.edu/39797871/Fooled_by_Correlation_Common_Misinterpretations_in_Social_Science_`

[6] A. Burkov, *The hundred page Machine Learning Book*, (Andriy Burkov, 2019. 1st. ed.).

[7] Gutierrez-Roig, M. et al. "Mapping individual behavior in financial markets: synchronization and anticipation". EPJ Data Science (2019).