Similarity of human behavior: A geometric approach to games*

Amil Camilo Moore, Fabrizio Germano, and Rosemarie Nagel

October 25, 2024

1 Introduction

Over the last 30 years, the diversity in human behavior across strategic settings has led to the development of many solution concepts in game theory that intend to close the gap between how a conventionally rational agent ought to behave and what humans actually do (e.g., Erev and Roth, 1995; McKelvey and Palfrey, 1995; Nagel, 1995; Bolton and Ockenfels, 2000; Costa-Gomes, Crawford, and Broseta, 2001; Camerer and Ho, 2003; Camerer, Ho, and Chong, 2004; Crawford, Costa-Gomes, and Iriberri, 2013; Fudenberg and Liang, 2019; to name a few). In many cases, this gap is generated by the construction of games that challenge the real-life plausibility of established concepts at the time and demonstrate, typically in the lab, that a theoretical contribution is needed to reconcile that challenge. However, the extent to which the insights from these constructed games generalize broadly to other games is not well-known.[1] Moreover, the extent by which humans, as opposed to merely the solution concepts that are meant to emulate them, make generalizations across different games is even less understood. Understanding the external validity of solution concepts to broader sets of games as well as the extent to which humans generalize is important for assessing the relative performance of solution concepts in predicting behavior in a variety of contexts, measuring the overlap (or lack thereof) between solution concepts in their predictions, identifying the properties of games that humans actually find salient in selecting their strategies, and understanding whether and how humans perceive games differently from their peers.

*We thank Larbi Alaoui, Jose Apesteguia, Elia Benveniste, Ayah Bohsali, Luke Boosey, John Duffy, Pia Ennuschat, Evan Friedman, Alexander Frug, Malachy Gavan, Ben Golub, Duarte Gonçalves, R. Mark Isaac, Annie Liang, Raquel Lorenzo, Antonio Penta, Yaroslav Rosokha, Andrea Salvanti, Patrick Sewell, Colin Sullivan, Elias Tsakas, Robert Wojciechowski, and audiences at the BSE Summer Forum, FSU Experimental Economics Reading Group, UPF Internal Micro Seminar, and the UPF Micro Theory Reading Group, for their insightful comments. Fabrizio Germano acknowledges financial support from Grant PID2020-115044GB-I00/MICIU/AEI/10.13039/501100011033 and from the Spanish Agencia Estatal de Investigación (AEI), through the Severo Ochoa Programme for Centres of Excellence in R&D (Barcelona School of Economics CEX2019-000915-S). All mistakes are our own.

[1] One can recognize that the act itself of constructing games where established concepts fail demonstrates the existence of a boundary in the established concepts in the sense that it clearly shows a point in the broader space of games where previous insights are no longer viable. Our question is more along the lines of properly quantifying and measuring the boundaries of solution concepts in a (relatively) complete domain of games as opposed to merely demonstrating the existence of the boundaries.

In this paper, we analyze human behavior across a comprehensive set of one-shot 2×2 games wherein players obtain payoffs from {1,2,3,4} without replacement (shortened as 1to4 2×2 (*one-to-four two-by-two*) games). There are only 78 essentially unique 1to4 2×2 games, which makes having subjects play all the games in the lab tractable. Nonetheless, this set of games has a representation for every strict preference ordering of outcomes, ensuring that subjects are exposed to a relatively complete set of games.

To understand the extent to which solution concepts and human behavior generalize across games, we define a notion of similarity that allows us to classify different games as belonging to distinct classes from a purely geometric approach, as outlined in Germano (2006). In particular, for any correspondence defined on a given space of games (e.g., the Nash equilibrium correspondence defined on all 2×2 games), we define two games as being similar if there exists a path starting from one game to the other such that the correspondence remains continuous on that path. The discontinuities of the correspondence always partition the space of games into a finite number of connected components, which we refer to as similarity classes.[2] Games belonging to the same similarity class can be shown to preserve essentially the same behavior for correspondences with a finite range. The similarity classes of a correspondence reflect fundamental aspects of its logic besides providing insightful organizations of the underlying space of games.[3]The approach here is applied not only to correspondences from solution concepts (such as Nash equilibrium and level-k rules), but also to behavior in the lab as measured by aggregate frequencies of play and their respective confidence intervals. We apply this notion to the space of 1-4 2×2 games, for which we compare the similarity classes implied by the behavior in lab experiments to the similarity classes implied by the solution concepts in the literature.

The similarity classes we find from aggregate behavior are easily interpreted by the solution concepts that best explain the behavior of individual subjects when playing 2×2 games. We find that around 78% of our subjects are best explained by a variant of level-k (level-k(α)) whereas around 16% are best explained by a decision rule that selects the fairest of the Pareto efficient outcomes (near-equal split). The empirical similarity classes derived only from the behavior in the lab correspond closely to the theoretical similarity classes generated by level-k(α) and near-equal split rules. The largest class of games ($34/78 = 43.59\%$ of the 1-4 games) consists of games

[2]See Aliprantis and Border (2006) for a definition of continuity for correspondences; semi-algebraic correspondences such as the Nash equilibrium correspondence, and many others used in game theory, have a finite number of such connected components of games by the Generic Continuity or Local Triviality Theorems; see Schanuel, Simon, Zame (1991), Blume and Zame (1994).

[3]Editor's note: We've thought about trying to show this more concretely either through an example or a theoretical claim, but the best way to do so is not yet clear. One thing we can do is prove the following claim: let $G_t \subset G_{t+1} \subseteq G$ be a set of games belonging to a subset G_{t+1} of a space of games G and let R_t, R_{t+1} be correspondences defined over G such that $R_t(G_t) = R_{t+1}(G_t)$ but $R_t(G_{t+1}) \neq R_{t+1}(G_{t+1})$. Then, R_t and R_{t+1} have different similarity classes of G. In other words, if we try to model the development of solution concepts we described in the first paragraph so that the set of games in which we have studied behavior is G_t and our model (which correctly describes that behavior) is R_t, then when we consider a new set of games from which to challenge R_t, call it G_{t+1}. What our claim implies in practice is that indeed demonstrating that the human behavior (represented by R_{t+1}) is different than R_t would predict on G_{t+1} implies a discontinuity in R_{t+1}. This shows that the process of developing new solutions concepts can be analogous to simply showing the existence of new similarity classes. The reason that this claim is not in the main body is because we have not proven this result for correspondence with an uncountable range. That being said, the proof in the finite range is straightforward and can be found in the Appendix.

where all the solution concepts in our analysis agree on what action players should select, with all but one game being dominance-solvable; the recommended action is played (minimum) 68.96% to (maximum) 94.24% of the time (average 79.81%, standard error 2.6%). The second largest class of games ($21/78 = 26.92\%$) consists of games where playing level-k(α) leads to outcomes that are not Nash equilibria, with most of the games being dominance-solvable but with only one player having a dominant strategy ($18/21 = 85.71\%$). The disagreement between solution concepts in this class coincides with a wider range of playing the action prescribed by level-1(α) over the dominance-solvable solution (in any given game, at least 25% do not play the equilibrium, at most 69%, on average 31%, standard error 3.33%). The remaining eight similarity classes, which separate some of the most well-known 2×2 games (e.g., the Prisoner's Dilemma, Stag Hunt, Battle of the Sexes, and Matching Pennies, to name a few), can also be easily characterized by the disagreement between level-k(α), near-equal split, and equilibrium solution concepts. Using only aggregate frequency data and the geometry of games, we find that players can distinguish between the most well-known games but that, by and large, the most well-known games also belong to markedly small similarity classes. In addition to finding similarity classes, we also document regularities in the individual behavior of subjects, especially as to how subjects' behavior appears to combine different solution concepts. We find that subjects vary to the extent that they favor level-k(α) over near-equal split and vice-versa, but that they are consistent as to how they play 1-4 2×2 games, even after controlling for the number of games already played or the order in which we presented the games to the subjects.

Our work follows closely the work of Rapoport, Guyer, and Gordon (1976), Robinson and Goforth (2005), Bruns (2010), and Shubik (2012), all of which study the same 2×2 games we do, but classify them only in terms of traditional game theory concepts. We also contribute to a growing literature on initial play in the lab and experiments on random games. We document how subjects play 2×2 games and provide an analysis of which features of those games best predict what they play, but we further provide this analysis for a complete set of games that we think researchers would plausibly want to randomly draw from, whether that is for experiments or the training of algorithms. Much of our methodological approach to analyzing the aggregate behavior is directly inspired by Wright and Leyton-Brown's (2014), and Fudenberg and Liang's (2019) work on 3×3 games. For instance, the level-$k(\alpha)$ decision rules are extensions of Fudenberg and Liang's level-$1(\alpha)$ rule, and we assess the external validity of many of the results that they find in 3×3 toward 2×2 games.

The rest of the paper is organized as follows. Section 2 introduces the 1-4 2×2 games which we study throughout the paper. Section 3 introduces the solution concepts included in our analysis and their corresponding decision rules. Section 4 is somewhat formal and contains our definitions of similarity of games and discusses similarity classes for equilibrium and level-k decision rules. Section 5 explains the design of our experiments. Finally, Section 6 concludes the paper with a discussion of the results of our experiments. Many important graphs, tables, and formal definitions

are relegated to the Appendix.

2 Games and perspectives

Throughout the paper, we focus on two player 2×2 games, where each player obtains payoffs from the set $\{1, 2, 3, 4\}$ with no repetitions. Since there are 4! ways to arrange the four payoffs of each player, there are $(4!)^2 = 576$ possible games. Denote by G the set of all these games. As is standard in game theory, we consider two games to be equivalent if they can be obtained one from the other by means of relabeling of the actions of a player, or by relabeling of a player, or any composition of such relabelings. Grouping together games that are equivalent in this sense, reduces the number of 576 games in G down to 78 strategically different games, of which 66 are asymmetric and 12 are symmetric games.[4] We denote by G^* the set of 78 strategically different games. As we discuss more formally in Section 4, each game in G^* is an equivalence class of games, whereby each of the 66 asymmetric games can be identified with 8 equivalent games (obtained through relabeling or compositions of relabeling of actions and players), while each of the 12 symmetric games can be identified with 4 equivalent games (obtained through relabeling or compositions of relabeling of just actions). For simplicity, we refer to these equivalence classes simply as *games* in G^*. Table 2 in the Appendix contains a list of all the (representative) games in G^*. These are numbered $1, \ldots, 78$, which also coincide with the numbers that appear beside the different nodes in many of our graphs in the paper.

We further define a *perspective* of a game as the choice problem a player faces, as either player 1 or player 2. It is immediate that asymmetric games have two perspectives and symmetric games have one. For instance, in a matching pennies type game, players 1 and 2 have different perspectives, since they face different choice problems, whereas in a standard prisoner's dilemma or chicken game, both players face the same choice problem, and so there is a single perspective associated to such a game. The 78 strategically different games in G^* thus give rise to $144 = 66 \cdot 2 + 12$ different perspectives. We denote the set of all 144 perspectives by G^{**}. Again, this is formally a set of equivalence classes of perspectives taken from the games in G, each of which, seen from the point of view of a fixed player, whether 1 or 2, are the same.

3 Behavioral rules

We introduce some minimal notation for our 2×2 games. We describe such a game by a tuple $g = (I, A, \pi)$, where $I = \{1, 2\}$ is the set of players consisting of player 1, the row player, and player 2 the column player; $A = A_1 \times A_2$, where $A_i = \{a_{i,1}, a_{i,2}\}$ is i's action space, $i \in I$; and $\pi = (\pi_1, \pi_2)$, where $\pi_i : A \to \{1, 2, 3, 4\}$ is i's payoff function.[5] We also occasionally consider mixed actions over A_i, and set $\Delta = \Delta_1 \times \Delta_2$, where $\Delta_i \equiv \Delta(A_i)$ is i's space of mixed actions.

[4] See Rapoport et al. (1976), Robinson and Goforth (2005) or Shubik (2012) for more details.

[5] Strictly speaking in our class of games with payoffs in $\{1, 2, 3, 4\}$ without repetitions, the payoff functions π_i are actually bijections from A to $\{1, 2, 3, 4\}$.

In our analysis we include a large set of rules taken from traditional game theory and over 30 years of behavioral game theory. In some cases, the concepts recommend multiple actions (e.g., Nash Equilibrium or Pareto Efficiency) or a mixed action (e.g., Nash Equilibrium or Risk Dominant Nash Equilibrium). To keep track of this, while keeping the analysis as simple as possible, we work with only pure actions and define a **rule** for player i as a map $R_i : G \to 2^{A_i}$, where 2^{A_i} is the set of subsets of A_i, including the empty subset (e.g., the concept Equal Split is empty for some games in G). The idea is that when a concept has multiple pure actions recommendations, we include all those pure actions, and when the recommendation is in mixed actions, we again include all pure actions that are in the support of the mixed action. We denote by $R = (R_1, R_2)$ the rule-profile used by both players, and will often refer to it also simply as rule.

The following lists some of the main solution concepts that we discuss and that can be used to define corresponding rules R_i for each player $i \in I$ in any game $g \in G$:

Best-response based rules:

- **Nash Equilibrium** (NE): an action profile $a^{NE} \in \Delta(A)$ is a Nash equilibrium if no player can achieve a greater payoff by deviating unilaterally from a^{NE}.
- **Risk-Dominant Nash Equilibrium** (RDNE): an action profile $a^{RDNE} \in \Delta(A)$ is a risk-dominant Nash equilibrium if a^{RDNE} is a NE and, when it is in pure strategies, it has a greater *deviation loss* than any other Nash equilibrium. The deviation loss of a (Nash) action profile is equal to the product of every player's payoff loss from deviating from that action profile.
- **Rationalizability** (RAT): an action profile $a^{RAT} \in A$ is rationalizable if it survives iterated elimination of strictly dominated actions.

Level-k rules:

- **Level-0** (L0): the mixed action $a_i^{L0} = (1/2, 1/2) \in \Delta(A_i)$ for i that randomizes uniformly over i's actions.
- **Level-k** (Lk): a (mixed) action $a_i^{Lk} \in \Delta(A_i)$ is Level-k for i if it is a best-response to the opponent playing a Level-(k-1) action.[6]
- **Level-$k(\alpha)$** (L$k(\alpha)$): a (mixed) action $a_i^{Lk(\alpha)} \in \Delta(A_i)$ is a Level-$k(\alpha)$ action for i if, after transforming all payoffs π_j to $\tilde{\pi_j} = \pi_j^\alpha$ for both players $j = 1, 2$, $a_i^{Lk(\alpha)}$ is a Level-k action for i, for $0 < \alpha < 1$.[7]

Efficiency and/or equity based rules:

[6] In 1-2 2×2 games, we consider k only up to 5 because of the cyclicality of the rule in this class of games. As it turns out, level-6 is the same rule as level-2, which by induction implies that level-7 is level-3 and level-8 is level-4 and so on.

[7] Note that transforming both players' payoffs π_j by raising them to the power α, for $j = 1, 2$, with $0 < \alpha < 1$ leads to unique pure actions $a_i^{Lk(\alpha)}$ for all levels $k \geq 1$ on our 2×2 games in G. Notice that when computing the Lk best replies with the transformed payoffs, both players' (i's and i's opponent's) payoffs are transformed.

- **Pareto Efficiency** (PE): an action profile $a^{PE} \in A$ is Pareto efficient if every action profile that could improve the payoff of a player compared to that achieved under a^{PE} must make someone strictly worse off. The set of Pareto efficient action profiles for a given game $g \in G$ is denoted by $PE(g)$

- **Near-Equal Split** (NES): an action profile $a^{NES} \in A$ is a Near-Equal Split if $a^{NES} \in PE(g)$ and a^{NES} achieves the lowest difference of payoffs between the two players (compared to other Pareto efficient outcomes).

  We further say that an action profile a^{sNES} is a self-favoring NES (sNES) for player i if a^{sNES} is an NES and i has a weakly greater payoff than the other player.

  Likewise, an action profile a^{oNES} is an other-favoring NES (oNES) for player i if a^{oNES} has a weakly lower payoff than the other player.

  Finally, an action profile a^{ES} is an Equal Split (ES) if a^{ES} is an NES and it achieves a payoff difference of zero. (This can be empty-valued for some games.)

- **Max-Max** (MM): an action $a_i^{MM} \in A_i$ is a Max-Max for i if it can lead to the outcome with the highest payoff for player i.

- **Soc-Max** (SM): an action profile $a^{SM} \in A$ is a Soc-Max if $a^{SM} \in \arg\max_{a \in A} \sum_{i \in I} \pi_i(a)$.

Hybrid rules:

- **Pareto-Dominant Nash Equilibrium** (PDNE): an action profile a^{PDNE} is a Pareto-dominant Nash equilibrium if $a^{PDNE} \in \Delta(A)$ is a NE and it Pareto dominates every other NE. Therefore, a unique NE is always PDNE.

In the following sections, we use these rules to analyze individual and aggregate behavior, but also, importantly, to define the similarity of classes of games.

4 Similarity classes of games

In this section, we introduce a notion of similarity that groups games according to the continuity a rule maintains through the space of games. For that notion of similarity to be coherent, we need to formalize how G^* simplifies G into equivalence classes and define a topology that allows us to measure the proximity of games to one another in G.

For any game g, one can identify the game also by its associated payoff matrix, defined by

$$\Pi = \begin{bmatrix} a, e & b, f \\ c, g & d, h \end{bmatrix},$$

which can also be written as a tuple $(a, b, c, d, e, f, g, h) \in \{1, 2, 3, 4\}^8 \subset \mathbb{R}^8$, and where it is understood that $a = \pi_1(a_{1,1}, a_{2,1}), \ldots, d = \pi_1(a_{1,2}, a_{2,2})$ and $e = \pi_2(a_{1,1}, a_{2,1}), \ldots, h = \pi_2(a_{1,2}, a_{2,2})$.[8]

[8]There is a slight abuse of notation here and the following paragraph in that g and a are also used as payoff entries in Π. In general, g denotes a game and a denotes an action profile unless otherwise noted.

Moreover, as mentioned before, there are a total of 7 more games represented by their respective payoff matrix, that are essentially the same as the one above, namely:

$$\begin{bmatrix} c,g & d,h \\ a,e & b,f \end{bmatrix}, \begin{bmatrix} b,f & a,e \\ d,h & c,g \end{bmatrix}, \begin{bmatrix} e,a & g,c \\ f,b & h,d \end{bmatrix}, \begin{bmatrix} d,h & c,g \\ b,f & a,e \end{bmatrix}, \begin{bmatrix} g,c & e,a \\ h,d & f,b \end{bmatrix}, \begin{bmatrix} f,b & h,d \\ e,a & g,c \end{bmatrix}, \begin{bmatrix} h,d & f,b \\ g,c & e,a \end{bmatrix}$$

All of these can be obtained from Π by either switching rows (i.e., relabeling 1's actions), switching columns (relabeling 2's actions), transposing and switching the off-diagonal entries (relabeling player 1 as 2 and player 2 as 1), or some combination of these operations. And so, although they represent different tuples in $\mathbb{R}^8$, they all represent the same strategic interaction up to relabeling of actions and players. Formally, the operations of relabeling actions and players and their combinations are defined as **symmetry operations**, which are linear maps $\psi : \mathbb{R}^8 \to \mathbb{R}^8$ that include the identity map on $\mathbb{R}^8$, so that the set of all such symmetry operations Ψ, contains exactly eight different linear maps, one being the identity map and the remaining seven being the ones mapping Π to one of the seven transformed payoff matrices above. We say the associated games are all **equivalent**, and in particular, they are equivalent to the game associated with the payoff matrix Π above.

This notion of equivalence allows us (and others in the literature) to reduce the number of games from 576 (in G) to 78 (in G^*). This reduction can be done by brute force using graph theoretic methods. As is standard, we use $\mathbf{\Gamma} = (\Gamma, E)$ to denote a graph with vertices Γ and edges $E \subseteq \Gamma \times \Gamma$. We define $E(A)$ as the edges induced by the adjacency matrix A so that $(g, g') \in E(A)$ implies that $A(g, g') \neq 0$. Finally, we say that $\mathbf{\Gamma}_c$ is a component of $\mathbf{\Gamma}$ if it is a connected subgraph of $\mathbf{\Gamma}$ that is not part of any larger connected subgraph of $\mathbf{\Gamma}$.[9]. If one defines the 576×576 adjacency matrix A_Ψ so that for $g \neq g'$

$$A_\Psi(g, g') = \mathbb{1}\{\exists \psi \in \Psi : g = \psi g'\},$$

and $A_\Psi = 0$ otherwise, we show in the Appendix (see figure A1) that there are 78 components of the graph $(G, E(A_\Psi))$, meaning that there are only 78 games that are unique up to symmetry operations.[10]

Our notion of two games being neighbors is directly taken from Robinson and Goforth (2006):

Definition 1 (Neighboring games)**.** *Let $g, g' \in G$. We say that g' is a* neighbor *of g (or $g' \in N(g)$) if g' can be obtained from g by swapping two entries in the payoff matrix of g for one of the players only, and the two entries differ 1. $N(g) \subset G$ is the set of all neighbors of g, including g itself.*

For example, $\begin{bmatrix} 4,4 & 2,1 \\ 1,2 & 3,3 \end{bmatrix}$ is a neighbor of $\begin{bmatrix} 4,4 & 1,1 \\ 2,2 & 3,3 \end{bmatrix}$, but is not a neighbor of $\begin{bmatrix} 4,4 & 1,2 \\ 2,1 & 3,3 \end{bmatrix}$.

Notice that $N(g)$ coincides with the set of all games in G with Euclidean distance $d \leq \sqrt{2}$ from

[9] A graph (Γ, E) is connected if for any $g, g' \in \Gamma$, there exists a path from g to g', that is, a sequence $(g_1, \ldots, g_n)$ such that $g_1 = g, g_n = g'$ and $(g_k, g_{k+1}) \in E$. A graph (Γ', E') is a subgraph of (Γ, E) if $\Gamma' \subseteq \Gamma$ and $E' \subseteq E$

[10] $\mathbb{1}\{\cdot\}$ is short-hand for the indicator function.

g. The Robinson-Goforth topology can be represented by a graph, $(G, E(A_N))$ where A_N is a 576×576 adjacency matrix such that for $g \neq g'$

$$A_N(g, g') = \mathbb{1}\{g' \in N(g')\},$$

and $A_N = 0$ otherwise (see figure A2).

Finally, we define our notion of similarity which follows Germano (2006) and defines two games $g, g' \in G$ as similar according to a rule R if g and a game equivalent to g' can be joined by a path of neighboring games in G, and the rule R always prescribes the same action or set of actions for all games along the path. This ensures that g and a game equivalent to g' belong to the same connected component of games in G. Formally:

Definition 2 (Similarity of games)**.** *Let R be a rule. We say that g is* similar *to g' according to R (written as $g \sim_R g'$) if and only if there exists a path $(g_1, g_2, \ldots, g_n)$ such that $g = g_1$, $g' = \psi g_n$, for some symmetry operation $\psi \in \Psi$, and, for any two consecutive games $g_\nu, g_{\nu+1}$, along the path, $g_{\nu+1} \in N(g_\nu)$ and $R(g_\nu) = R(g_{\nu+1})$, for $\nu = 1, \ldots, n-1$.*

If we are given a finite set of rules $\mathcal{R}$, then we say $g \sim_{\mathcal{R}} g'$ if the same conditions hold as with one rule, but where the condition $R(g_\nu) = R(g_{\nu+1})$ now holds for all $R \in \mathcal{R}$.

It can be checked that, for any rule R, the similarity relation $\sim_R$ uniquely partitions the set of games G into a finite number of components, which we refer to as the **similarity classes** in G relative to R.

Computing the similarity classes for the rules in Section 3 is feasible by combining adjacency matrices through matrix operations, as we show in the following theorem. Let $A_R(g, g') = \mathbb{1}\{R(g) = R(g')\}$ and denote the Hadamard (or element-by-element) product of two matrices A and B by $A \odot B$ with $(A \odot B)(g, g') = A(g, g')B(g, g')$. Finally, we'll say that R is relabeling invariant if $R(g) = R(g')$ implies that for any $\psi \in \Psi$, $R(\psi g) = R(\psi g')$.[11]

Theorem 1 (Computation of Similarity Classes)**.** *Let R be relabeling invariant. Then, $g \sim_R g'$ if and only if g and g' belong to the same component of $(G, E(A_\Psi + (1 - A_\Psi) \odot A_N \odot A_R))$. Thus, the similarity classes according to R over G correspond precisely to the components of $(G, E(A_\Psi + (1 - A_\Psi) \odot A_N \odot A_R))$.*

The proof is in the Appendix. The theorem makes operational the idea that similar games are connected either by the symmetry operations (thus being in the same component defined by A_Ψ) or through the neighborhood topology by a constant rule recommendation given R (thus being in the same component defined by $(1 - A_\Psi) \odot A_N \odot A_R$). The theorem states that these components characterize the similarity classes for the given rule R.

We now present the similarity classes for some key rules as examples. To simplify the figures, we visualize only the similarity classes over the 78 games in G^*. We refer to the table in Appendix

[11] It is easy to show that all the decision rules in Section 3 satisfy this property.

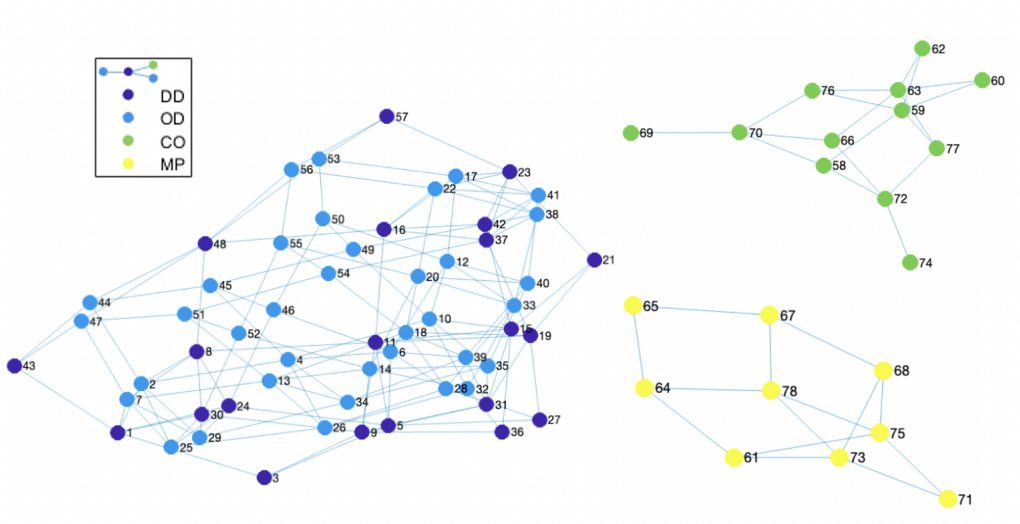

Figure 1: Graph of the similarity classes in G^* implied by Nash equilibrium.

A for a list of all the games in G^* (including the corresponding empirical frequency of play from the experiments described in Section 5). In describing the similarity classes, we make use of the following standard types of games distinguished in the game theory literature, which partition the space G^*:

- **Double-sided dominance games** (DD): These are games with a unique pure NE and where both players have a strictly dominant action (dark blue nodes; 21 games);
- **One-sided dominance games** (OD): These are games with a unique pure NE and where only one of the players has a strictly dominant action (light blue nodes, 36 games);
- **Coordination type games** (CO): These are games with two pure NE (green nodes; 12 games);
- **Matching pennies type games** (MP): These are games with a unique mixed NE and no pure NE (yellow nodes; 9 games in G^*).

These four basic types of games will be used throughout the paper and will be also further refined.

Example 1. (Nash Equilibrium similarity classes) If $R : G \to 2^A$ is the rule based on Nash Equilibrium (NE), then it partitions the games in G (and hence G^*) into three similarity classes as follows (see Figure 1):

- **Games with a unique pure NE** (DD $\cup$ OD): These games are all dominance solvable and thus containing all DD and OD games (dark blue and light blue nodes; 57 games);
- **Games with two pure NE** (CO) These are all the coordination type games (green nodes; 12 games);
- **Games with zero pure NE** (MP) These are all the matching pennies type games (yellow nodes; 9 games).

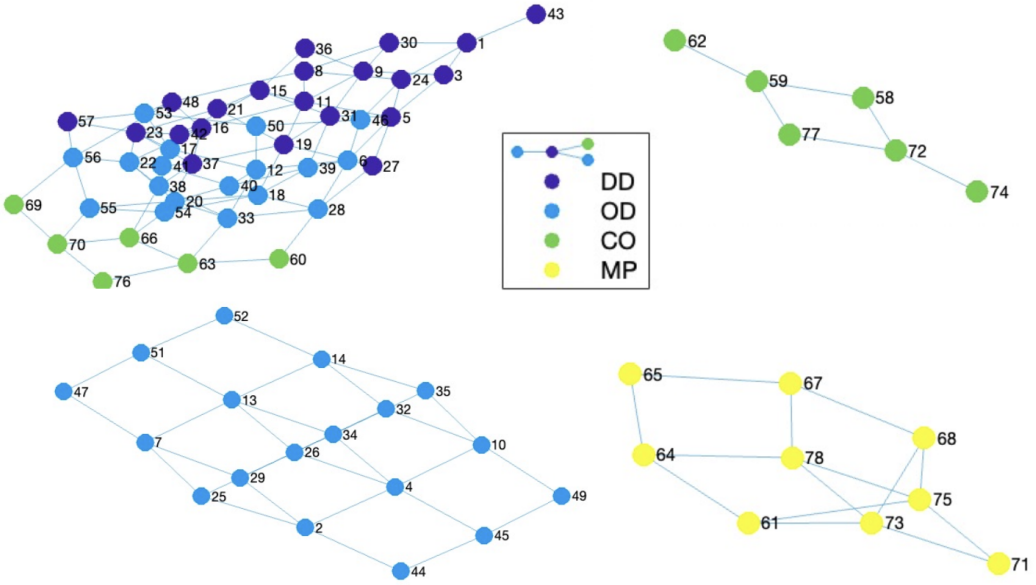

Figure 2: Graph of the similarity classes in G^* implied by Level-$k(\alpha)$ rules, $k = 1, \ldots, 5$.

The same similarity classes arise if one considers the rule based on Rationalizability (RAT). In particular, the games in MP and the ones in CO are separated, although all four action profiles are rationalizable in both classes. The reason these two classes (yellow and green nodes) are separated is that within the topology of neighboring games, one has to go through games with a unique pure NE (dark and light blue nodes) to move from a game in one class (say MP games) to a game in the other (say, CO games) (see figure A3 and figure A4 in the Appendix). ■

The next example shows the similarity classes for the Level-$k(\alpha)$ rule that plays a key role in our experimental results.

Example 2. (Level-$k(\alpha)$ similarity classes) If we use rules based on Level-$k(\alpha)$ (L$k(\alpha)$, $k = 1, \ldots, 5$), we get different similarity classes from Nash. Taking all L$k(\alpha)$ rules for $k = 1, \ldots, 5$, as a set $\mathcal{R}$ of rules, leads to four similarity classes being distinguished, see Figure 2, namely:

- **Single outcome games** (DD $\cup$ OD1 $\cup$ CO1) These are games where Lk(α) actions span only one outcome, converging to an equilibrium in one step. This includes all the DD games as well as some OD and CO games, henceforth referred to as OD1 and CO1 (dark blue, light blue, and dark green nodes in Figure A5; 45 games).
- **Two outcomes games** (OD2) These are games where L$k(\alpha)$ actions span two outcomes, with the solution converging to an equilibrium in $k = 2$ steps. This consists only of OD games, henceforth referred to as OD2 (cyan in Figure A5; 18 games).
- **Four outcomes games with two pure NE** (CO2) These are games where L$k(\alpha)$ actions span all outcomes and thus both pure strategy equilibria. This consists only of CO games, henceforth referred to as CO2 (light green in Figure A5; 6 games).

- **Four outcome games with zero pure NE** (MP) These are games where L$k(\alpha)$ cycles over the four possible outcomes. This coincides precisely with MP (yellow in Figure A5; 9 games).

How these similarity classes are connected is shown in the Appendix (figure A6 and figure A7). The reason why level-k(α) separates these classes is that each class has a different degree to which level-k(α) cycles through the solutions of the game. MP games cycle through all action profiles, which is different to how the nearest classes to MP cycle: DD $\cup$ OD1 $\cup$ CO1 converges at $k = 1$ step to an action profile whereas OD2 converges at $k = 2$ steps. Likewise, CO2 games never converge to an action profile and thus are separated from its nearest classes (again, DD $\cup$ OD1 $\cup$ CO1 and OD2). ■

The four classes above form the basis for characterizing the empirical similarity classes obtained in Section 5. In the Appendix, we discuss other similarity classes, such as the ones obtained from Risk Dominant Nash Equilibrium or the rules based on Pareto efficiency or Near Equal Split.

To conclude this section, we show how the above concepts are easily adapted to study our main object, which is the similarity classes of games, obtained from the empirical behavior of the subjects. For this, define an **empirical rule** $\hat{R} : G \to 2^{\Delta}$, as a map assigning the relevant confidence intervals around the frequency of play in game g for both players (1 and 2). This map can be obtained from either bootstrapping the sample proportion or calculating the intervals implied by the binomial test. The same definition of similarity can be used with the only difference that the condition that rules ought to stay constant is replaced with one that the confidence intervals (for some confidence level, typically 99%) around the frequencies of play in g overlaps with the one for g'. This is necessary since looking only at empirical frequencies and requiring exact equality of these, as in the original definition, would be a too stringent (and statistically unreasonable) condition for evaluating the similarity of empirical behavior.[12] Formally:

Definition 3 (Empirical similarity of games)**.** *Let $\hat{R} : G \to 2^{\Delta}$ be an empirical rule. We say that g is* empirically similar *to g' according to $\hat{R}$ (written as $g \sim_{\hat{R}} g'$) if and only if there exists a path $(g_1, g_2, \ldots, g_n)$ such that $g = g_1$, $g' = \psi g_n$, for some symmetry operation $\psi \in \Psi$, and, for any two consecutive games $g_\nu, g_{\nu+1}$, along the path, $g_{\nu+1} \in N(g_\nu)$ and $\hat{R}(g_\nu) \cap \hat{R}(g_{\nu+1}) \neq \emptyset$, for $\nu = 1, \ldots, n-1$.*

From the above definition, we obtain **empirical similarity classes**, which again partition the space G (or equivalently G^*) into a finite number of components depending on the empirically

[12] Editor's note: one could argue that our notion of empirical similarity could actually be too lenient as it requires completely separation between confidence intervals as opposed to the estimated difference between proportions to be statistically significantly non-zero. The results are not substantially different under the latter requirement though it does make more separations than the former. That being said, one can parameterize the degree of leniency in perceiving two games as being similar by computing a p-value for every pair of games for the null hypothesis that $||\hat{R}(g) - \hat{R}(g')|| < \Delta$ where $\Delta > 0$ is a parameter chosen by the researcher that represents how big differences need to be in order to reject the hypothesis that two games are similar. In this different notion of similarity, we would replace the overlap condition for continuity with a condition that continuity requires non-rejection over a path of games. The empirical similarity classes we get in our results are identical to the case when $\Delta = 0.02$ at the 0.001 significance level.

observed behavior of the subjects. In the next section, we compare theoretical similarity classes above with the empirical ones.

5 Experimental design

The objective of our experiment is to estimate the joint frequency distribution of the actions subjects selected in every game as well as to analyze the individual behavior of the subjects across the entire set. Our experiment has four stages: in stage 1, subjects play all the 1-4 2×2 games. In stage 2, we have subjects complete a risk-elicitation task. Finally, in stage 3, we have an exit questionnaire where subjects can detail their experience of participating in the experiment.

5.1 Stage 1: Games

Given the high number of games, we asked subjects to play, we wanted to ensure that subjects took every game seriously by minimizing the extent to which subjects could predict features of the games they would play in the future. To do this, we designed 15 different treatments that differed in the sequence of games subjects encountered, with each treatment's sequence having a quasi-random order of games and random labeling of each game through symmetry operations.

While we could have generated an order of games purely randomly, we elected to provide some structure to avoid random bunching of similar games that would make subjects expect future games to have similar payoff matrices as previous ones. For instance, if subjects saw perspective 1 and then perspective 2 from table 2 one after the other, they may expect future games to have an outcome where both players get the maximum payoff, which could bias subjects to play the near-equal split profile.

With this in mind, we defined a numerical code for every perspective in G^{**} that corresponded to the payoff matrix of the perspective so that if $\Pi = (a, b, c, d, e, f, g, h)$, the code of the perspective would be "abcdefgh" (for example, the Prisoner's Dilemma which has $\Pi = (4, 2, 3, 1, 1, 2, 3, 4)$ would have a code of "42311234" (see perspective 57 of 2). We then ordered the numerical codes in increasing order and binned the codes into 16 quantile bins. Each of the 16 bins consists of 9 perspectives in G^{**} represented by their numerical codes.

We then generated 15 sequences of perspectives for every treatment according to the following procedure.

1. Randomly generate a random order of quantile bins (e.g., (1, 5, 9, 14, 4, 16, 10, 12, 6, 8, 11, 2, 3, 7, 15, 13) would be one such order).

2. Randomly draw perspectives from each quantile bin without replacement in the order described in step 1 (e.g., the 1st perspective belongs to bin 1, the 2nd from bin 5, the 3rd from bin 9, and so on...) and append this to a sequence.

3. Repeat step 1 and 2 until the sequence includes all 144 perspectives.[13]

This procedure makes random bunching of similar games unlikely because, for every block of 16 perspectives, each perspective belongs to a bin different than the last one.

The randomizing of the labeling in each sequence is less complicated: for every perspective, we randomly re-label using symmetry operations either player 1's actions or player 2's actions. We do not re-label using the players' positions since that would alter the perspective.

Finally, we randomly assigned one of the 15 treatment sequences to each subject; our assignment is balanced so that we have an equal sample size for each treatment. Previous experiments in the literature (e.g., Fudenberg and Liang, 2019) have randomized per subject - we elected to randomize using the treatments so that we can test if how we presented the games to subjects has a statistical effect.

5.2 Stage 2: Risk-elicitation

We elicited subjects' risk preferences using the risk-taking item from the Global Preferences Survey (Falk et al., 2018). After having them complete the item as done in the original Global Preferences Survey for the participant's country, we also had them complete a modified version of the same item which we re-scaled the payoff to match the payoff scale of the games played in stage 1. The positive affine transformation changed the maximum value of the lotteries in the task to the money value of receiving 4 points and the minimum value of receiving 1 point from a game. The values of the lotteries were all presented in the currency of the country from where the subject participated in the experiment.

5.3 Stage 3: Exit questionnaire

In the exit questionnaire, we asked subjects to describe in their own words how they came to make decisions in the games. We also asked subjects questions on a Likert scale about how they perceived their strategy across stage 1 and information about their demographics (e.g., age, education, gender identity, to name a few).

5.4 Implementation

The experiment was programmed using oTree (Chen, Schonger, and Wickens, 2016).

In stage 1, we paid a show-up fee of 1.70 GBP for completing the experiment and paid a rate of 2.00 GBP per the average number of points across three randomly selected games. This means subjects could earn anywhere from 2.00 GBP to 8.00 GBP. No feedback was given from round to round and subjects were randomly matched with a new opponent each round. After completing

[13]The reason we choose to have subjects play the perspectives of G^{**} instead of all the games in G is because we assume that the behavior of the games is invariant to relabeling. This is, of course, not taken for granted and we control for this by randomizing the labels of the games and balancing the random assignment of labels.

48 and 96 games respectively, there was a forced 30-second break to mitigate fatigue during the experiment.

We recruited 450 subjects from Prolific with ethical approval from CIREP-UPF. To control for regional variations, we restricted our sample only to participants in the United Kingdom. The subjects were predominantly women (295/450 = 65.56% women, 153/450 = 34.00% men) from ages 25-44 (138/450 = 30.67% age 25-34, 137/450 = 30.44% age 35-44) with some college education completed (110/450 = 24.44% some college, 130/450 = 28.89% four-year degree). The average completion time of the entire experiment was nearly 20 minutes.

6 Results

In this section, we detail the results of the experiment first by showing the estimated empirical similarity classes and then making sense of them by analyzing aggregate and individual behavior.

To estimate the empirical similarity classes, we estimated the joint frequency distribution of the actions in each game by computing the proportion of subjects that played each action across all treatments (see table 2 for the estimates). We then calculated 99% confidence intervals using the binomial test for each game.[14] The empirical rule $\hat{R}$ will thus consist of bivariate confidence intervals for every game.

6.1 Empirical similarity classes

From estimating the joint frequency distribution of each game, we get 11 similarity classes distinguished, see Figure 3, namely:

- **Games where all rules coincide.** (NE = LK(α) = NES) These are games where the main rules from the solution concepts included in our analysis predict the same outcome. It is a subset of the single-outcome games from the level-k(α) similarity classes. This includes all DD games but the Prisoner's Dilemma (game 57), most OD1 games, and one CO1 game (34 games).
- **Games where the LK(α) solution is not an equilibrium.** (NE $\neq$ LK(α)) These are games where the level-k(α) solution of the game leads to an off-equilibrium outcome, starting at the $k = 1$ step. This includes all of the OD2 games and some of the CO2 games (21 games).
- **Games with no pure strategy Nash equilibria but unique behavioral solutions** (MP - G65) These are games that have unique level-1(α) or unique near-equal split solutions. It includes all MP games except for game 65 (9 games).
- **Games with inefficient equilibria.** (PDs) These games have unique pure NE that are either Pareto inferior or have strictly lower payoff sums than another outcome. It includes the Prisoner's Dilemma (game 57) and other OD1 games (4 games).

[14] Our results are robust to estimating using bootstrap confidence intervals.

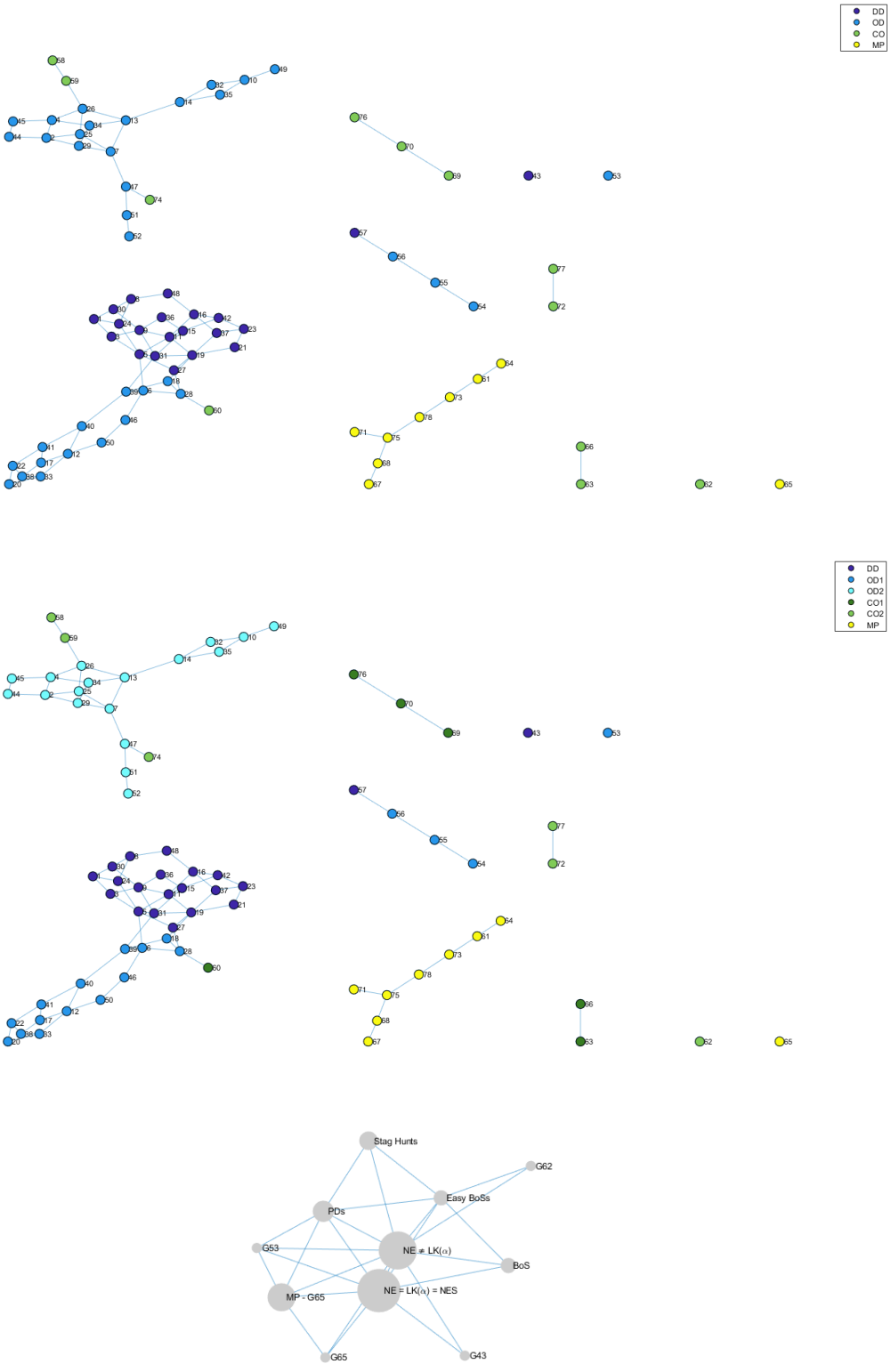

Figure 3: Graph of the empirical similarity classes in G^* from aggregate behavior, colored by standard types (top), level-k(α) types (middle), and how the classes connect in the Robinson-Goforth topology in G^* (bottom).

- **Games with a risky equilibrium and a distinct efficient one.** (Stag Hunts) These games in which one of the pure NE can achieve the near-equal split outcome whereas the other is the LK(α) solution of the game. It includes the canonical Stag Hunt (game 69) and other CO1 games (3 games).
- **Games with risky equilibria.** (BoSs) These games have two pure NE but the LK(α) solution leads players off-equilibria. It consists of the canonical Battle of the Sexes (game 77) and game 72 (2 games).
- **Games with multiple equilibria, but one is the LK(α) solution.** (Easy BoSs) These games have two pure NE but the LK(α) solution leads players to one of them in $k = 1$ step. Nonetheless, this equilibrium gives a better payoff to one player than in the other equilibrium. It consists of game 66 and game 63 (2 games).
- **Special games.** These are games that form similarity classes of their own, consisting of game 43 (DD), game 53 (OD1), game 62 (CO2), and game 65 (MP) (4 games).

It is easy to make sense of these empirical similarity classes by noticing the differences in the solution concepts across the games. But do the subjects generalize according to the solution concepts in our analysis? In the remainder of this section, we address this question by first demonstrating that examining the aggregate behavior of the subjects is not enough to rationalize these similarity classes. However, examining the individual behaviors of subject and how they are explained by the solution concepts does.

6.2 Aggregate behavior

If aggregate behavior were sufficient to explain the empirical similarity classes, one might expect that looking at the solution concepts that best fit the aggregate behavior would be enough.

One approach is to do a classification exercise similar to the one in Fudenberg and Liang (2019) wherein we compute the accuracy of each rule in predicting modal play (that is, the most frequently played action in each of the 144 perspectives).

The best-performing rule is Level-1(α) with an accuracy of 94.44% (136/144 perspectives), followed by a statistically close performance from Level-5(α) and Level-5 (see table 1 or figure 4). In the 8 perspectives, where Level-1(α) fails to predict, both sNES and PDNE predict 100.00% (8/8 perspectives) of the modal play. These results are consistent with what Fudenberg and Liang (2019) find in 3×3 games, where most of the modal play is predicted by Level-1(α), but in games where Level-1(α) is predicted to perform badly, modal play resembles PDNE.

Even though level-1(α) predicts more than 90% of the most frequently played action, the similarity classes of level-1(α) alone do not match our empirical similarity classes (in fact, figure A8 in the Appendix shows that level-1(α) does not separate any games in G).[15] A natural objection might be

[15]This same exercise can be repeated with the top five best-fitting rules and the result does not change.

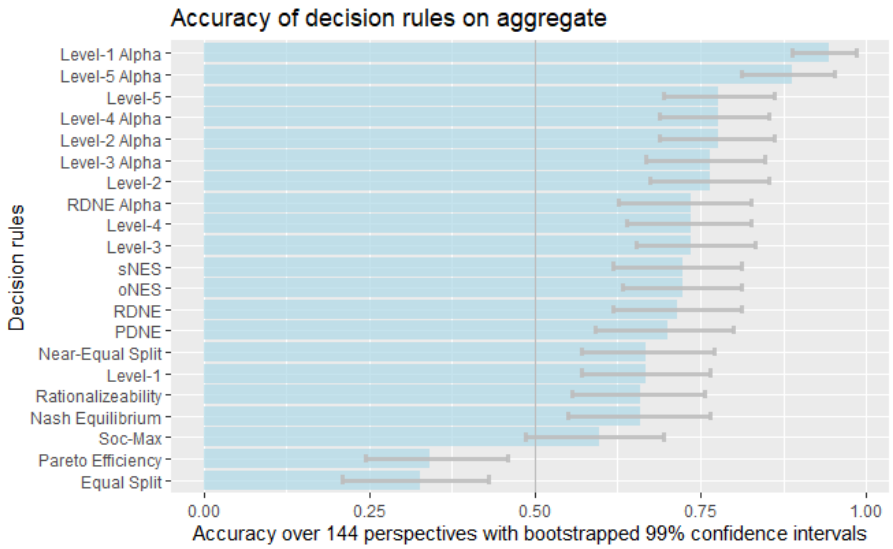

Figure 4: Accuracy of decision rules in predicting the most frequently played action.

that predicting the most frequently played action may not be appropriate since our similarity classes are generated using the overlap of the joint frequency distributions as opposed to identifying the most frequent action. Instead, it might be preferable to do a prediction task in which we compute the mean-squared error of each rule in predicting the frequency distribution of each of the 144 perspectives.

In this case, the best-performing rule is unambiguously Level-1 (see figure 5) and most other rules perform almost the same against our data. For example, level-1(α) and Nash equilibrium have a statistically indistinguishable gap. Unlike level-1(α), level-1 does separate games into multiple similarity classes (see figure A9 in the Appendix). However, its performance by itself in producing similar partitions to the empirical similarity classes is not better than level-kα: the adjusted Rand index (ARI) between level-1 and the empirical similarity classes is 0.3603 (99% bootstrap confidence interval: 0.2356, 0.5661, non-adjusted Rand index: 0.7409) whereas the ARI between level-k(α) and the empirical similarity classes is 0.6382 (C.I.: 0.4427, 0.7158, n.a. RI: 0.8362).[16] Furthermore, neither level-1, level-1(α), nor level-k(α) similarity classes are able to distinguish the most well-known games as being distinct behaviorally (e.g., PDs are not separated from other DD games).

Since one rule (or family of rules) is not able to fully explain the similarity classes, subjects may adopt different rules and thus the empirical similarity classes may reflect the heterogeneity in the subjects' behavior.

[16]The adjusted Rand index is a measure of comparison between two partitions that adjusts for partitions being the same by chance.

Table 1: Accuracy of decision rules in predicting the most frequently played action.

Decision Rule	Games correctly predicted	99% CI lower bound	% of 144	99% CI upper bound	% of games if L1(α) incorrect
L1(α)	136	0.89	0.94	0.99	0.00
L5(α)	128	0.81	0.89	0.95	0.62
L2(α)	112	0.69	0.78	0.86	0.75
L4(α)	112	0.71	0.78	0.85	0.75
L5(α)	112	0.71	0.78	0.85	0.75
L2	110	0.69	0.76	0.83	0.75
L3(α)	110	0.70	0.76	0.83	0.62
L3	106	0.65	0.74	0.83	0.75
L4	106	0.64	0.74	0.80	0.75
oNES	104	0.63	0.72	0.81	1.00
sNES	104	0.62	0.72	0.81	1.00
RDNE	103	0.62	0.72	0.81	0.75
PDNE	101	0.59	0.70	0.80	1.00
NES	96	0.55	0.67	0.76	1.00
L1	96	0.56	0.67	0.74	0.00
NE	95	0.55	0.66	0.76	0.62
RAT	95	0.56	0.66	0.74	0.62
Soc-Max	86	0.49	0.60	0.69	1.00
PE	49	0.24	0.34	0.46	0.88
ES	47	0.21	0.33	0.40	0.88

6.3 Individual behavior

Since every subject plays 144 games, we can find regularities in how subjects play that allow us to distinguish which solution concepts each subject finds most salient.

Nonetheless, given that there are only two possible actions per perspective, there may be natural objections to trying to infer what rules are used by each subject. For instance, it may well be that the best-fitting decision rule for a subject turns out to be level-1(α) but that, were the subject to have changed one or two choices, we may have classified that subject instead as level-2(α) or level-3(α). In other words, since decision rules could overlap frequently in their predictions, the gap between the model that best explains a subject's behavior and the second- or third-best may in principle be small, which could give rise to instability in identifying what solution concepts subjects use.

To examine this concern, we analyze the predictions of all rules and their overlap with each other (see figure 6). While it is true that the level-k rules and their variants for $k > 1$ only distinguish themselves from RAT and NE in games where both actions are rationalizable (the green area of the RAT rule in the figure), this still leaves 36 perspectives in which subjects can play differently. In those 36 perspectives, the predictions made by the different level-k rules are indeed different, except for level-1(α) and level-5(α), which coincide. However, there are still 18 perspectives where level-1(α) and level-5(α) do not coincide. It is easy to see that the near-equal split rules are markedly different to the level-k and best-reply-based rules. Within the near-equal split rules, sNES and oNES differ in the predictions over 16 perspectives (predominantly on games where the

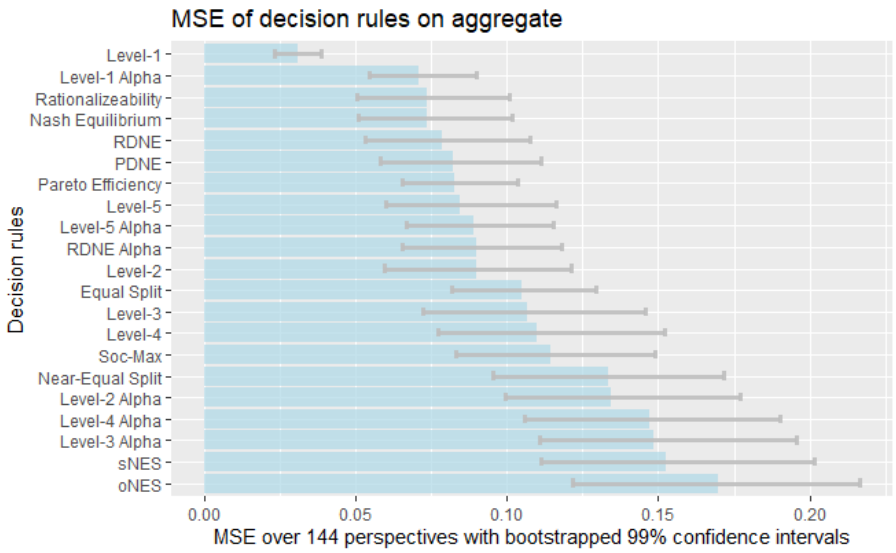

Figure 5: MSE of decision rules in predicting the frequency distribution of perspectives.

other rules all predict the same action). Finally, the prediction of Soc-Max differs with that of Near-Equal Split in 24 perspectives.

Given the differences between the rules' predictions and the fact that we have subjects play all 144 perspectives to span the entire set of 1-2 2×2 games, we estimate the best-fitting decision rule of each subject for all 144 perspectives that they play. Since some subjects' best-fitting decision rule could inevitably have low predictive accuracy, we also estimate the distribution of how well a randomly generated rule could predict a subject's choices – whenever a subject's best-fitting decision rule fails to reject having an accuracy as good as the randomly generated rule at a 5% significance level, we classify the subject as Level-0 since their behavior is better predicted by randomness than by any of the solution concepts in our analysis. Otherwise, we classify the subjects according to their best-fitting rules.

The distribution of the subjects' best-fitting decision rule is summarized by figure 7. Level-1(α) again appears as a clear favorite (now among the individual subjects as opposed to modal play), followed by the two near-equal split decision rules, followed by level-5(α). Furthermore, PDNE does not appear as a best-fitting rule for any subject. The degree to which subjects consistently play according to their best-fitting decision rule is especially high for subjects best fit by level-1(α). When level-1(α) subjects do deviate though, they tend to deviate to playing near-equal split (see figure 8). This also works the other way around, suggesting that the presence of level-1(α) and near-equal split in the subject distribution is not merely the result of subjects belonging to one of "two types" but instead landing on a spectrum of using a combination of these two rules. In fact, for most subjects best fit by level-1(α) (near-equal split), more than 90% of their choices become explained by adding near-equal split (level-1(α)). More broadly, the added accuracy from joining level-k(α) rules with near-equal split together to predict individual behavior is statistically significant and not easily matched by other rule combinations (see figure A10 in the Appendix for

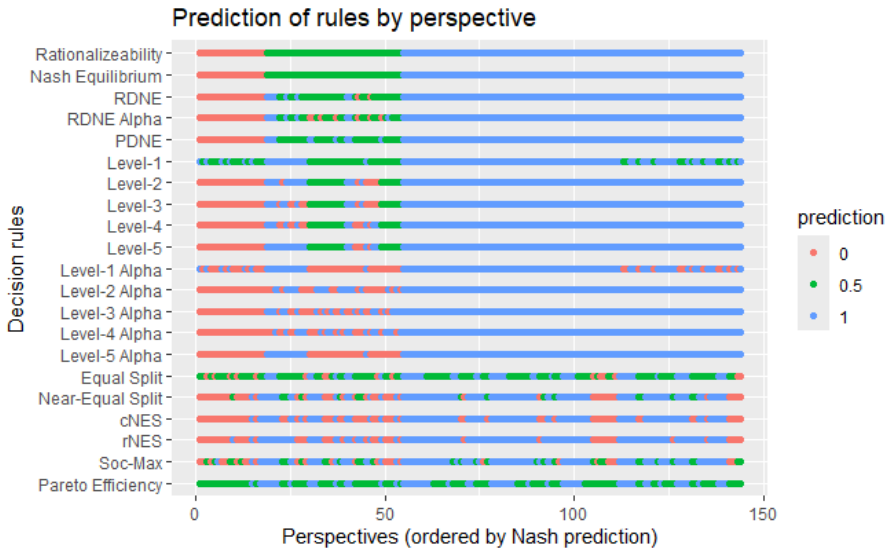

Figure 6: The prediction of each decision rule across the 144 perspectives. Note: This figure reports sNES and oNES as rNES and cNES respectively.

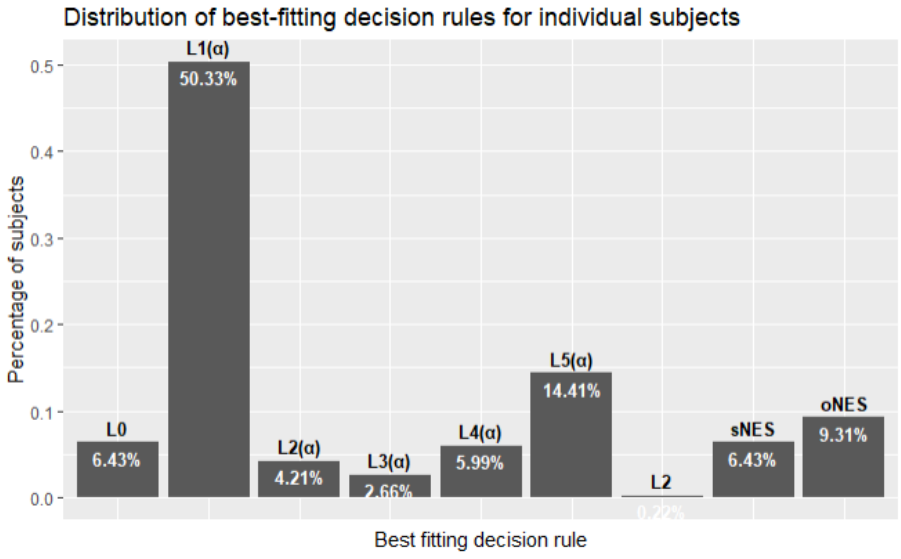

Figure 7: The distribution of the individual subjects' best-fitting decision rules. Note: This figure reports sNES and oNES as rNES and cNES respectively.

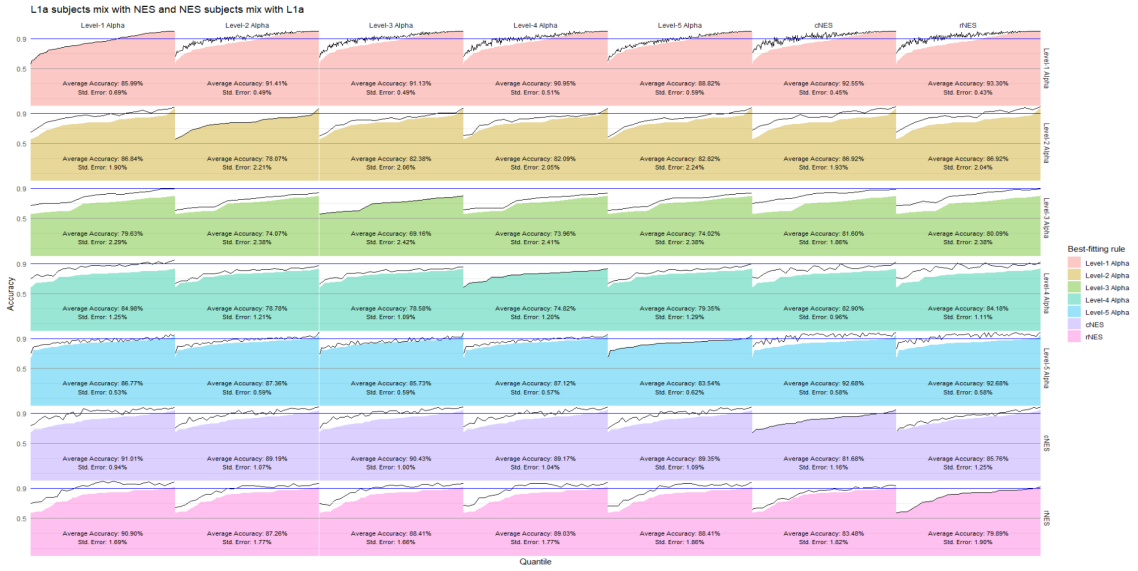

Figure 8: The quantile function of the accuracy of each subject's best-fitting rule (area in color) plus the added accuracy of other rules in games where the best-fitting rule predicts incorrectly. The rows correspond to the groups of subjects according to their best-fitting rule. The columns correspond to the rule used to predict games where the best-fitting rule predicts incorrectly. This figure reports sNES and oNES as rNES and cNES respectively.

a complete version of figure 8).

From the complementarity between level-k(α) and near-equal split, we can make sense of why the empirical similarity classes appear to further separate the similarity classes implied by level-k(α). For instance, the Stag Hunts class is separated from the Easy BoSs class because the fact an equilibrium is near-equal split makes it focal for not only the 15% of subjects that are best fit by near-equal split rules, but also the other subjects that habitually deviate from their rules to play near-equal split. Likewise, the PDs are separated from the rest of the OD1 games because regardless of whether subjects know that there is a unique pure NE, the fact that there is a different near-equal split solution to the game makes the frequency distribution different from that of OD1 games.

6.4 Conclusion

This paper analyzes human behavior across a comprehensive set of one-shot 2×2 games, using a new approach that groups games into similarity classes. These classes reflect both the predicted outcomes of game-theoretic solution concepts and observed behavior in the lab. Our results show that empirical similarity classes, derived from aggregate data, correspond closely with theoretical similarity classes predicted by Nash equilibrium and level-k reasoning. This suggests that these solution concepts can explain behavior across a wider range of games than traditionally demonstrated, supporting their use in broader contexts.

While we find that level-k reasoning and Nash equilibrium are strong predictors of aggregate behavior, no single solution concept is sufficient to capture all observed patterns across game

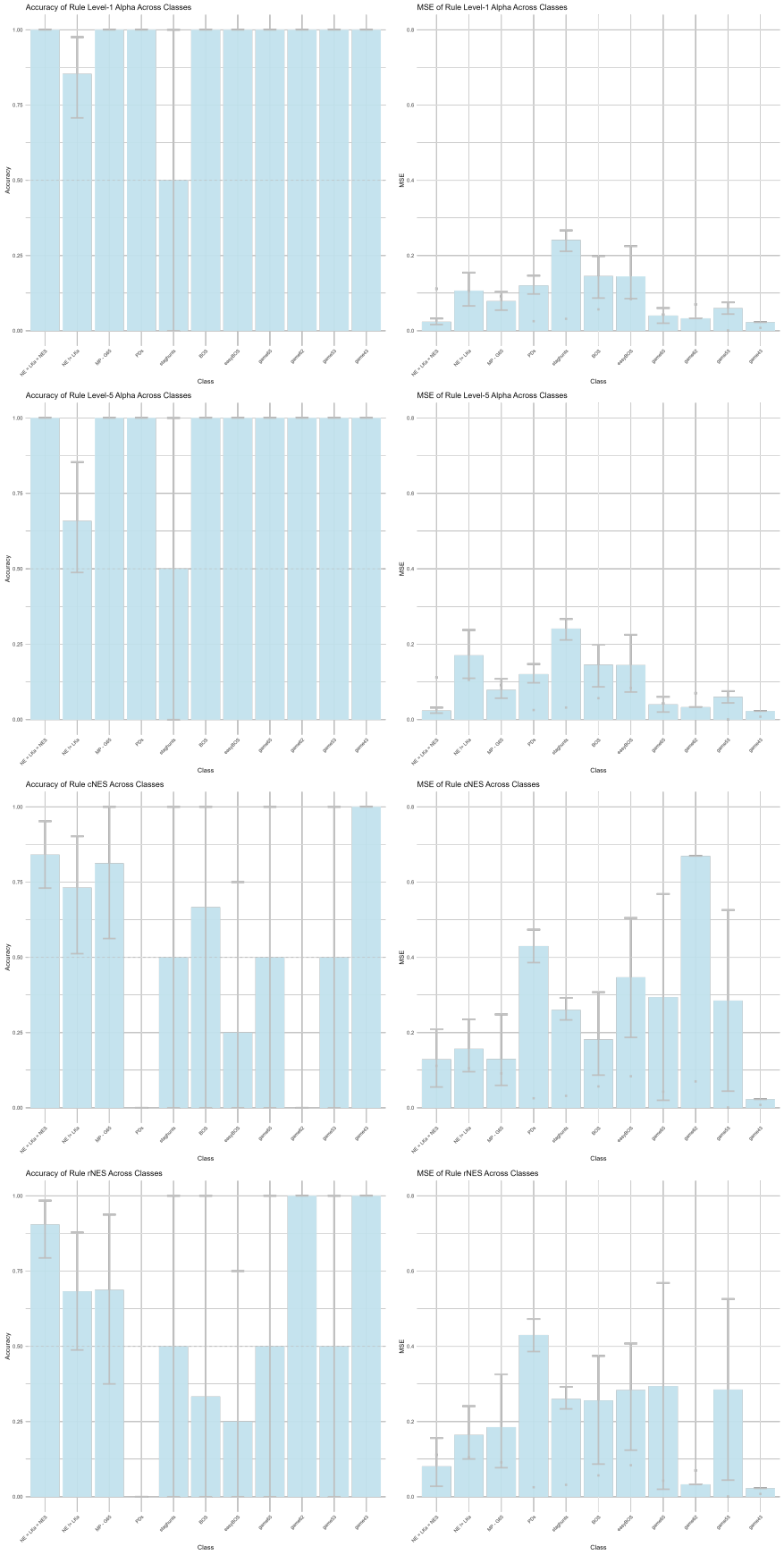

Figure 9: Accuracy and mean squared error (MSE) of decision rules across different empirical similarity classes. The gray dotted line corresponds to the predicted accuracy for a random rule. The gray squares are the MSE of predicting that any action is equally likely for every game (i.e., 50% A and 50% B if A and B are the actions) within a similarity class.

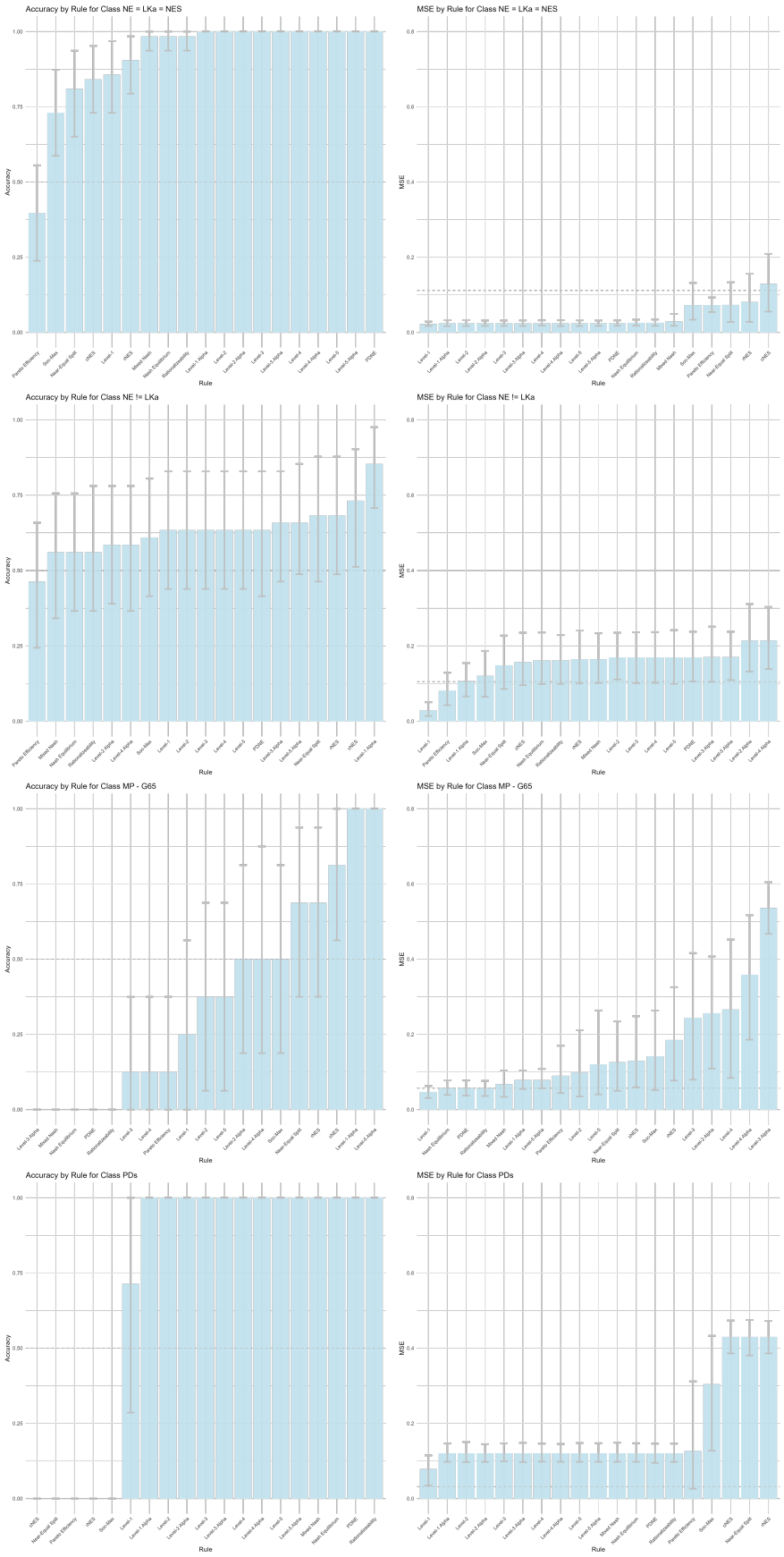

Figure 10: Accuracy and mean squared error (MSE) distributions of decision rules across different empirical similarity classes. The gray dotted line corresponds to the predicted accuracy for a random rule. The gray squares are the MSE of predicting that any action is equally likely for every game (i.e., 50% A and 50% B if A and B are the actions) within a similarity class.

types. This highlights the importance of accounting for variation in individual decision-making and suggests that models combining multiple solution concepts could be useful for more complex or mixed-strategy settings.

The next steps in this work will extend the analysis in several directions. First, we will assess how these solution concepts perform across different game contexts to confirm whether our findings hold with additional players, payoff structures, or repeated interactions. Second, we plan to study the overlap among solution concepts by identifying which games prompt similar predictions across different rules. Finally, we will investigate the properties of "special games" within the empirical similarity classes to better understand the features of games that players find consistently salient. These steps aim to deepen our understanding of strategic behavior and help refine models that predict human decision-making across various settings.

References

[1] **Aliprantis, Charalambos D. and Kim C. Border.** 2006. "Infinite Dimensional Analysis: A Hitchhiker's Guide." *Springer.*

[2] **Camilo Moore, Amil and Rosemarie Nagel.** 2023. "Not uncovered by the machine learning algorithm: equal split as a (best) predictor in initial play?" *Universitat Pompeu Fabra.*

[3] **Biggar, Oliver, and Iman Shames.** 2023. "The graph structure of two-player games." *Scientific Reports*, 13(1): 1833.

[4] **Blume, Lawrence E. and William R. Zame.** 1994. "The Algebraic Geometry of Perfect and Sequential Equilibria." *Econometrica.*

[5] **Bolton, Gary E. and Alex Ockenfels.** 2000. "ERC: A Theory of Equity, Reciprocity, and Competition." *American Economic Review.*

[6] **Bruns, Bryan.** 2010. "Navigating the Topology of 2×2 Games: An Introductory Note on Payoff Families, Normalization, and Natural Order." *arXiv.*

[7] **Camerer, Colin and Teck Hua Ho.** 2003. "Experience-weighted Attraction Learning in Normal-Form Games." *Econometrica.*

[8] **Camerer, Colin, Teck Hua Hong, and Juin-Kuan Chong.** 2004. "A Cognitive Hierarchy Model of Games." *Quarterly Journal of Economics.*

[9] **Carlsson, H., and E. Van Damme.** 1993. Global games and equilibrium selection. Econometrica 61 (5), 989–1018.

[10] **Chen, Daniel L., Martin Schonger, and Chris Wickens.** 2016. "oTree — An open source platform for laboratory, online, and field experiments." *Journal of Behavioral and Experimental Finance.*

[11] **Costa-Gomes, Miguel, Vincent P. Crawford, and Bruno Broseta.** 2001. "Cognition and Behavior in Normal-Form Games: An Experimental Study." *Econometrica.*

[12] **Crawford, V.P., Costa-Gomes, M.A., and N. Iriberri.** 2013. Structural models of nonequilibrium strategic thinking: theory, evidence, and applications. Journal of Economic Literature 51(1): 5–62

[13] **Erev, Ido and Alvin E. Roth.** 1995. "Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria." *American Economic Review.*

[14] **Falk, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde.** 2018. "Global Evidence on Economic Preferences." *Quarterly Journal of Economics.*

[15] **Fehr E., and K.M. Schmidt.** 1999. A Theory of Fairness, Competition, and Cooperation. The quarterly journal of economics 114(3): 817-868.

[16] **Fudenberg, Drew and Annie Liang.** 2019. "Predicting and Understanding Initial Play." *American Economic Review*, 109(12): 4112-4141.

[17] **Germano, Fabrizio.** 2006. "On some geometry and equivalence classes of normal form games." *International Journal of Game Theory.*

[18] **McKelvey, Richard D. and Thomas Thomas R. Palfrey.** 1995. "Quantal Response Equilibria for Normal Form Games." *Games and Economic Behavior.*

[19] **Nagel, Rosemarie.** 1995. "Unraveling in Guessing Games: An Experimental Study." *American Economic Review.*

[20] **Omidshafiei, Shayegan, et al.** 2020. "Navigating the landscape of multiplayer games." *Nature communications*, 11(1): 5603.

[21] **Rapoport, Anatol, Melvin J. Guyer, and David G. Gordon.** 1976. *The* 2×2 *Game.* University of Michigan Press.

[22] **Robinson, David and David Goforth.** 2005. *The Topology of* 2×2 *Games: A New Periodic Table.* Routledge.

[23] **Schnauel, Stephen, Leo K. Simon, and William R. Zame.** 1991. "The Algebraic Geometry of Games and the Tracing Procedure." *Game Equilibrium Models II.*

Appendix

A Proofs

Proof of Theorem 1

Proof. Suppose $g \sim_R g'$. Then there exists a path from g to $\psi g'$ for some ψ such that the rule is constant. Therefore, $\psi g'$ is in the same component of $(G, E(A_\Psi))$ as g'. Since $E(A_\Psi) \subset E(A_\Psi + (1 - A_\Psi) \odot A_N \odot A_R)$, this implies that there is a path from g to g' in $(G, E(A_\Psi + (1-A_\Psi) \odot A_N \odot A_R))$. Thus, g and g' are in the same component of $(G, E(A_\Psi + (1 - A_\Psi) \odot A_N \odot A_R)$.

Suppose now that g and g' are in the same component of $(G, E(A_\Psi + (1 - A_\Psi) \odot A_N \odot A_R))$. This implies the existence of a path (g_t) such that $g_1 = g$ and $g_n = g'$ through $(G, E(A_\Psi + (1 - A_\Psi) \odot A_N \odot A_R))$. Without loss of generality, we can rewrite (g_t) as

$$(g_t) = (g_{1,1}, \ldots, g_{1,l_1}, g_{2,1}, \ldots, g_{2,l_2}, \ldots, g_{k,1}, \ldots, g_{k,l_k}),$$

where k is the number of times (g_t) passes through non-equivalent through symmetry operation games and $g_{i,j} = \psi g_{i,h}$ for any $j \neq h$ for some $\psi \in \Psi$. This rewriting splits the sequence into distinct subsequences of equivalent through symmetry operation games. Let ψ_i be such that $g_{i,l_i} = \psi_i g_{i,1}$. Consider now the following sequence:

$$(g_t^*) = \left(\Big(\prod_{s=0}^{k-1} \psi_{k-s} \Big)^{-1} \prod_{s=0}^{k-t} \psi_{k-s} g_{t,1} \right)_{t=1}^{k}$$

We now prove that (g_t^*) is a path from g to $\psi g'$ for some $\psi \in \Psi$ such that for any consecutive games along the path, $g_t^* \in N(g_{t+1}^*)$ and $R(g_t^*) = R(g_{t+1}^*)$, which shows, by construction, and $g \sim_R g'$.

To show $g_1^* = g$, it's enough to see that $g_1^* = g_{1,1} = g_1 = g$. To show that $g_k^* = \psi g'$ for some $\psi \in \Psi$, likewise one can see that $g_k^* = (\psi_k \times \cdots \times \psi_1)^{-1} \psi_k \times g_{k,1} = (\psi_k \times \cdots \times \psi_1)^{-1} g_{k,l_k} = (\psi_k \times \cdots \times \psi_1)^{-1} g_n = (\psi_k \times \cdots \times \psi_1)^{-1} g'$.

Finally, to prove that $g_t^* \in N(g_{t+1}^*)$ and $R(g_t^*) = R(g_{t+1}^*)$, note that if $g_t^* \in N(g_{t+1}^*)$, then $\psi g_t^* \in N(\psi g_{t+1}^*)$ for any $\psi \in \Psi$, and that by relabeling invariance, if $R(g_t^*) = R(g_{t+1}^*)$, then $R(\psi g_t^*) = R(\psi g_{t+1}^*)$ for any $\psi \in \Psi$.[17] Thus, comparing these conditions between g_t^* and g_{t+1}^* is the same as comparing them between $\psi_t g_{t,1}$ and $g_{t+1,1}$. Since $\psi_t g_{t,1} = g_{t,l_t}$, $(g_{t,l_t}, g_{t+1,1}) \in (g_t)$, and g_{t,l_t} is not equivalent by symmetry operations to $g_{t+1,1}$ (by construction), it follows that $g_{t,l_t} \in N(g_{t+1,1})$ and $R(g_{t,l_t}) = R(g_{t+1,1})$ because (g_t) is a path in $(G, E(A_\Psi + (1 - A_\Psi) \odot A_N \odot A_R))$. Therefore, $g_t^* \in N(g_{t+1}^*)$ and $R(g_t^*) = R(g_{t+1}^*)$.

Since we've shown that (g_t^*) is a path from g to $\psi g'$ for some $\psi \in \Psi$ such that for any consecutive

[17] The invariance of the topology to relabeling is a trivial property for the Robinson-Goforth case but can be extended to other topologies that have the same property.

games along the path, $g_t^* \in N(g_{t+1}^*)$ and $R(g_t^*) = R(g_{t+1}^*)$, the result immediately arrives that $g \sim_R g'$.

□

Proof of the claim in footnote 3

Proof. The formal statement is as follows: let G be a set of games, $G_t \subset G_{t+1} \subseteq G$, $R_t(g) = R_{t+1}(g)\ \forall g \in G_t$ but $R_t(g') \neq R_{t+1}(g')$ for some $g' \in G_{t+1}$, where R_t, R_{t+1} are correspondences with finite ranges that satisfy relabeling invariance (meaning, their values can only be finite sets and for any $\psi \in \Psi$, if g and g' produce the same value, then ψg and $\psi g'$ should also produce the same). If for some $\gamma \in G_t, \gamma' \in G_{t+1} \backslash G_t$, $\gamma \sim_{R_t} \gamma'$ but $R_t(\gamma') \neq R_{t+1}(\gamma')$, then $\gamma \not\sim_{R_{t+1}} \gamma'$ (in other words, γ and γ' must belong to different similarity classes according to R_{t+1}).

The proof is by contradiction. Let $\gamma \sim_{R_t} \gamma'$ but $R_{t+1}(\gamma') \neq R_t(\gamma')$ for $\gamma' \in G_{t+1} \backslash G_t, \gamma \in G_t$. Then, without loss of generality, that implies the existence of a path of neighboring games $p_0 = (g_1^{p_0}, \ldots, g_{n_0}^{p_0})$ such that $g_1^{p_0} = \gamma, g_{n_0}^{p_0} = \gamma'$, $R_t(\gamma) = R_{t+1}(\gamma) = R_t(\gamma') \neq R_{t+1}(\gamma')$, and R_t is continuous over p_0. Importantly, p_0 has continuity in R_t but not in R_{t+1} because the endpoint values of R_t and R_{t+1} differ. Suppose that $\gamma \sim_{R_{t+1}} \gamma'$. This implies the existence of a path of neighboring games $p_1 = (g_1^{p_1}, \ldots, g_{n_1}^{p_1})$ such that $g_1^{p_1} = \gamma, g_{n_1}^{p_1} = \psi\gamma'$ for some $\psi \in \Psi$ and R_{t+1} is continuous over p_1. Let $p_2 = (\psi g_1^{p_0}, \ldots, \psi g_{n_0}^{p_0})$ be a path which effectively relabels the games on the path in p_0 so that it has the same endpoint as p_1. Note that if p_2 is just a relabeling of p_0, R_{t+1} should have the same discontinuities over p_2 as it does over p_0, which must mean that if R_t and R_{t+1} differed at the different ends of p_0, they also should at the different ends of p_2. But if R_t and R_{t+1} differ on γ, this is a contradiction to $R_t(\gamma) = R_{t+1}(\gamma)$ for $\gamma \in G_t$. □

B Additional figures

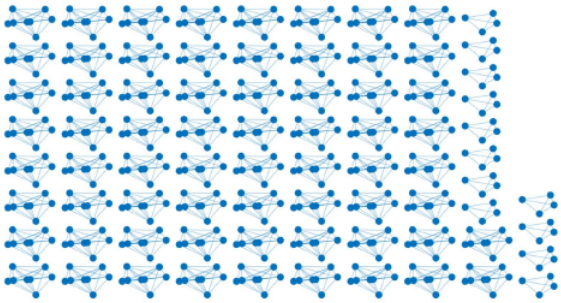

Figure A1: The graph on G, where edges represent equivalence between two games. Each of the 78 components consists of 4 or 8 equivalent games in G, depending on whether the component is one of symmetric or asymmetric games. Formally, this graph is denoted as $(G, E(A_\Psi))$.

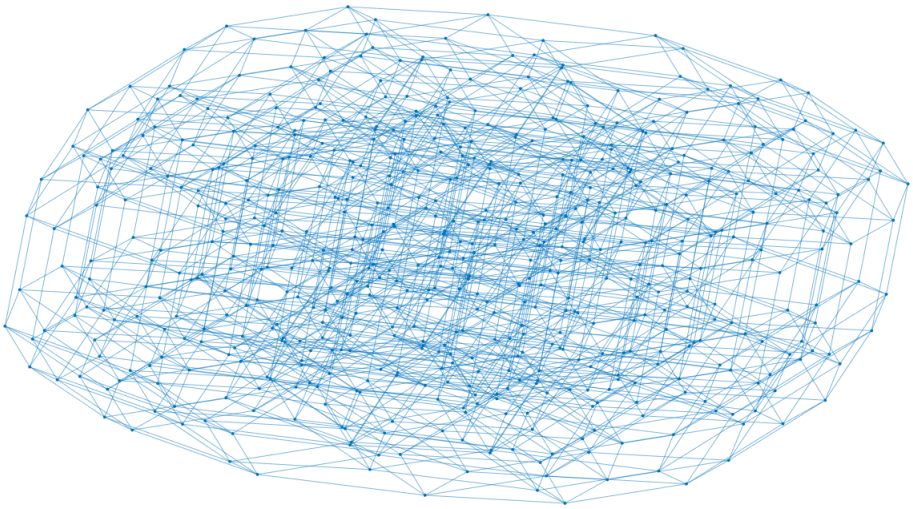

Figure A2: The graph on G, where edges represent whether two games are neighbors in the Robinson-Goforth topology. Formally, this graph is denoted as $(G, E(A_N))$.

C Additional similarity classes

Example 3. (Near-Equal Split similarity classes) When considered on their own, both rules based on self-favoring or other-favoring Near-Equal Split (sNES, oNES), since they always select a unique outcome, always yield a unique similarity class, comprising all games in G (and similarly for G^*). However, taking the two rules together leads to nine similarity classes; see Figure A11. The same nine classes obtain if one takes the Near-Equal Split (NES) rule alone.

What are the distinctions between the different classes? The largest class (dark blue nodes, 65 nodes) contains all games with a unique NES which is selected by both Near-Equal-Split players. In all other games, there are two NES cells which either are in the same row, column or (off)-diagonal. The second largest similarity class (lighter blue nodes, 4 nodes) contains OD and DD games with the NES rule spanning two outcomes (with 4-3 or 3-4 payoffs) in the same row. The next largest class (olive green nodes, 3 nodes) has spans two outcomes with the 4-3 and 3-4 cells on the main diagonal and thus two actions for both players. The remaining games have two 2-3

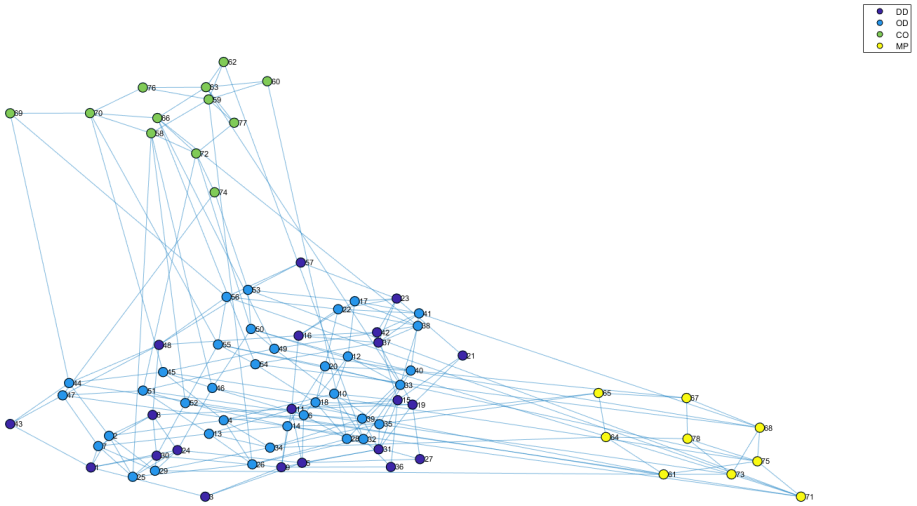

Figure A3: Graph of the Robinson-Goforth topology in G^* colored by types of games.

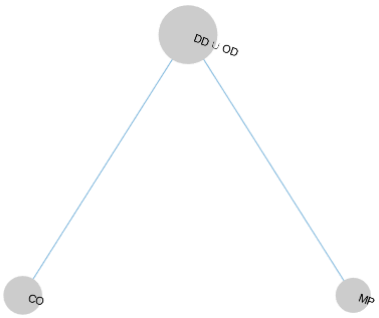

Figure A4: Graph of the Robinson-Goforth topology in G^*, with nodes representing the similarity classes according to Nash equilibrium and edges how they connect.

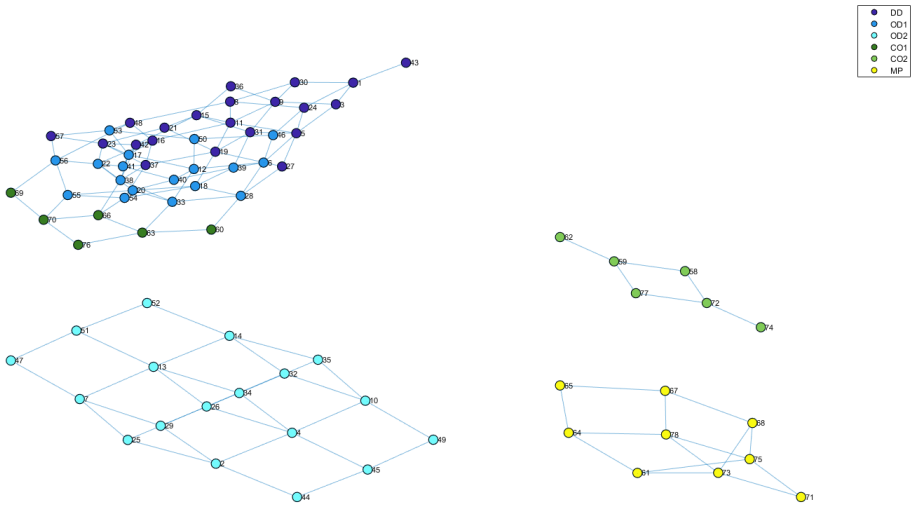

Figure A5: Graph of the similarity classes in G^* implied by Level-$k(\alpha)$ rules, $k = 1, \ldots, 5$.

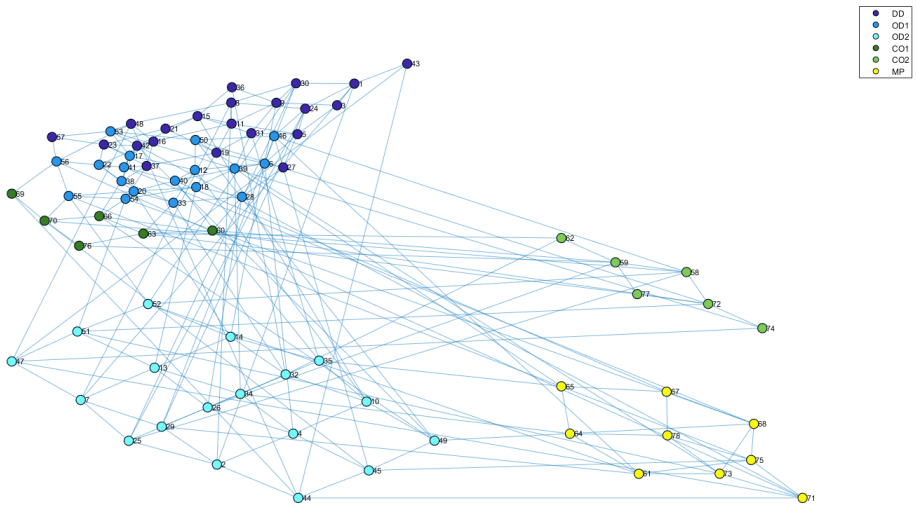

Figure A6: Graph of the Robinson-Goforth topology in G^* colored by similarity classes implied by Level-$k(\alpha)$ rules, $k = 1, \ldots, 5$.

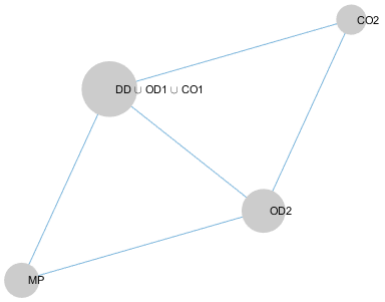

Figure A7: Graph of the Robinson-Goforth topology in G^*, with nodes representing the similarity classes according to Level-$k(\alpha)$ rules, $k = 1, \ldots, 5$, and edges how they connect.

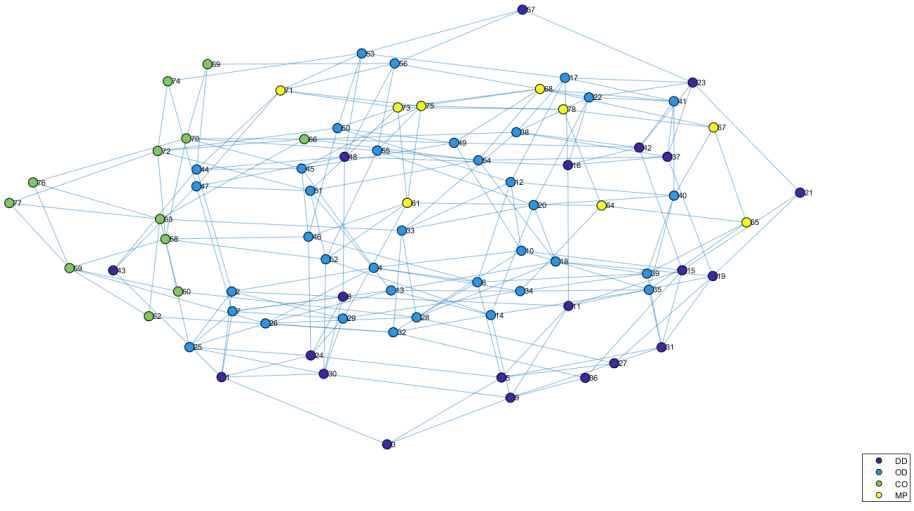

Figure A8: Graph of the similarity classes in G^* implied by the Level-1(α) rule. It does not separate any games.

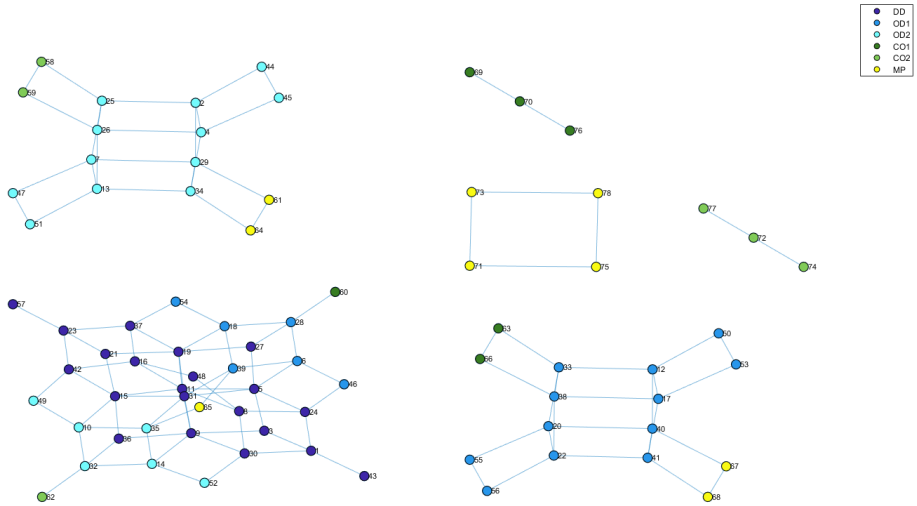

Figure A9: Graph of the similarity classes in G^* implied by the Level-1 rule.

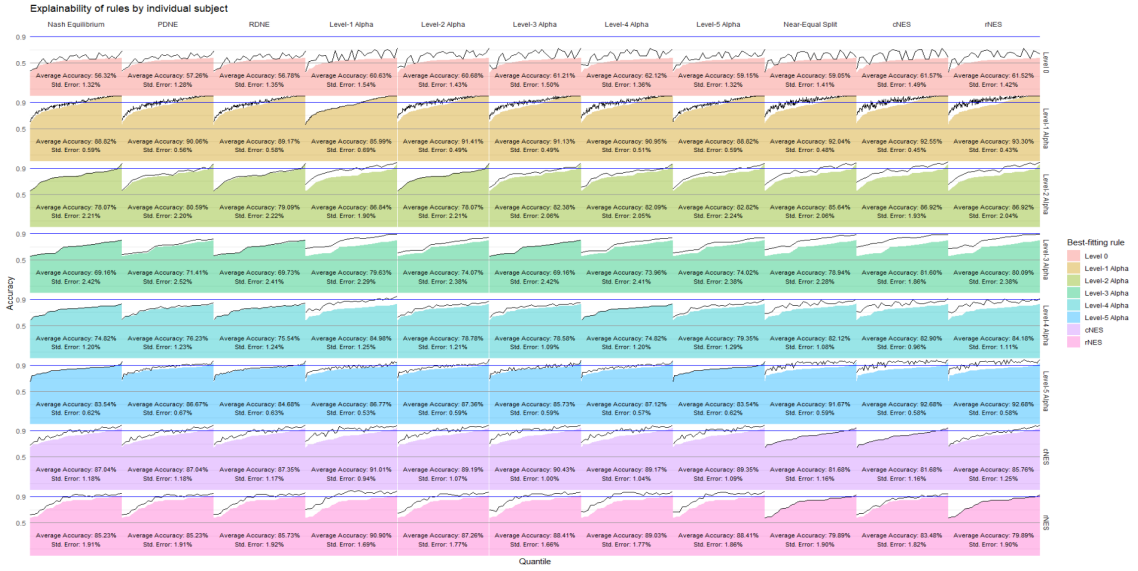

Figure A10: The quantile function of the accuracy of each subject's best-fitting rule (area in color) plus the added accuracy of other rules in games where the best-fitting rule predicts incorrectly. The rows correspond to the groups of subjects according to their best-fitting rule. The columns correspond to the rule used to predict games where the best-fitting rule predicts incorrectly. This figure reports sNES and oNES as rNES and cNES respectively.

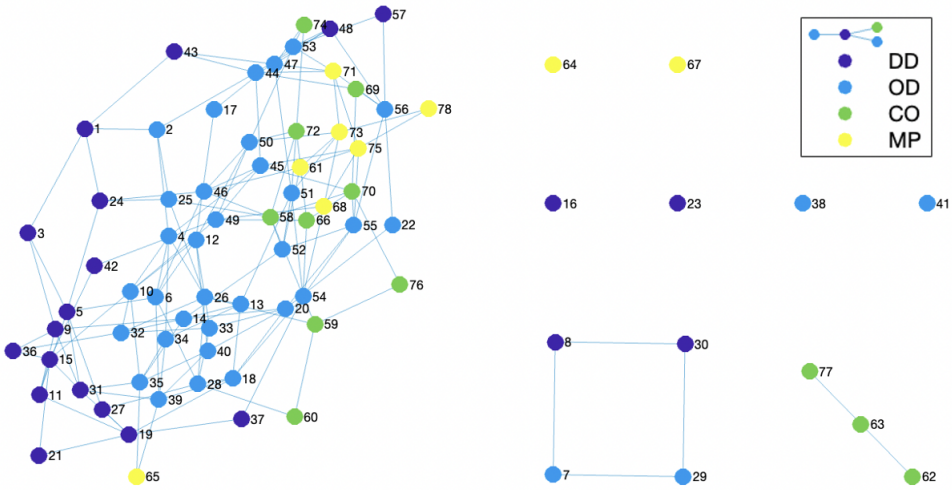

Figure A11: Graph of the similarity classes in G^* implied by self-favoring and other-favoring Near-Equal Split.

and 3-2 or 4-2 and 2-4 outcomes in different constellations. ■

D Table of games in G^*

In the 4-page table below, in the columns freq1 and freq2, we report the population frequencies of player 1 and player 2 respectively, where the entries in the first row represent the data without the random players, and the entries in the second row represent the raw data; similarly n1 and n2 are the number of subjects playing the corresponding game as player 1 and player 2. Moreover and importantly, the single number freq1 or freq2 represents the frequency of the player's action associated with the own payoff of 4. The column denoted by "type" distinguishes four broad types of games: games with one-sided strict dominance (OD), games with two-sided strict dominance (DD), matching pennies type games (MP), and coordination type games (CO). However, it is useful to further refine OD and CO depending on whether the Level-$k(\alpha)$ rules select just one or more than one outcome. Hence we have two types of games in OD: ones where the Level-$k(\alpha)$ rules select a single outcome (OD1), and ones where they select two outcomes (OD2). Similarly, there are two types of coordination games: ones where Level-$k(\alpha)$ rules select a single NE (CO1), and ones ones where Level-$k(\alpha)$ rules span all outcomes and hence both NE (CO2).

Table 2: Table of games in G^*

game id	**payoff matrix**	**bruns id**	**type**	**raw data**			**clean data**		
				freq1	freq2	p_bar	freq1	freq2	p_bar
1	$\begin{bmatrix} 4,4 & 3,3 \\ 2,2 & 1,1 \end{bmatrix}$	11,1 12,1	DD	0.94	0.92	0.92	0.98	0.96	0.96
2	$\begin{bmatrix} 4,4 & 3,3 \\ 2,1 & 1,2 \end{bmatrix}$	7,2 11,6	OD	0.93	0.53	0.53	0.97	0.53	0.53
3	$\begin{bmatrix} 4,4 & 3,2 \\ 2,3 & 1,1 \end{bmatrix}$	11,2	DD	0.94	0.94	0.94	0.97	0.97	0.97
4	$\begin{bmatrix} 4,4 & 3,2 \\ 2,1 & 1,3 \end{bmatrix}$	8,2 11,5	OD	0.94	0.55	0.55	0.97	0.97	0.55
5	$\begin{bmatrix} 4,4 & 3,1 \\ 2,3 & 1,2 \end{bmatrix}$	10,2 11,3	DD	0.90	0.93	0.90	0.95	0.98	0.95
6	$\begin{bmatrix} 4,4 & 3,1 \\ 2,2 & 1,3 \end{bmatrix}$	9,2 11,4	OD	0.92	0.87	0.87	0.95	0.92	0.92
7	$\begin{bmatrix} 4,3 & 3,4 \\ 2,2 & 1,1 \end{bmatrix}$	1,2 11,12	OD	0.85	0.38	0.49	0.89	0.38	0.46
8	$\begin{bmatrix} 4,3 & 3,4 \\ 2,1 & 1,2 \end{bmatrix}$	6,2 11,7	DD	0.92	0.82	0.82	0.97	0.84	0.84
9	$\begin{bmatrix} 4,3 & 3,2 \\ 2,4 & 1,1 \end{bmatrix}$	5,3 10,8	DD	0.88	0.94	0.88	0.92	0.98	0.93
10	$\begin{bmatrix} 4,3 & 3,2 \\ 2,1 & 1,4 \end{bmatrix}$	3,8 5,1	OD	0.91	0.59	0.58	0.95	0.61	0.49
11	$\begin{bmatrix} 4,3 & 3,1 \\ 2,4 & 1,2 \end{bmatrix}$	5,2 11,8	DD	0.92	0.93	0.92	0.95	0.98	0.95

Table 2: Table of games in G^* (cont.)

game id	**payoff matrix**	**bruns id**	**type**	**raw data**			**clean data**		
				freq1	freq2	pbar	freq1	freq2	pbar
12	$\begin{bmatrix} 4,3 & 3,1 \\ 2,2 & 1,4 \end{bmatrix}$	2,8 5,1	OD	0.88	0.23	0.77	0.92	0.20	0.80
13	$\begin{bmatrix} 4,2 & 3,4 \\ 2,3 & 1,1 \end{bmatrix}$	2,2 11,11	OD	0.91	0.50	0.50	0.96	0.49	0.50
14	$\begin{bmatrix} 4,2 & 3,3 \\ 2,4 & 1,1 \end{bmatrix}$	5,4 9,8	OD	0.92	0.61	0.54	0.96	0.62	0.47
15	$\begin{bmatrix} 4,2 & 3,3 \\ 2,1 & 1,4 \end{bmatrix}$	4,8 5,9	DD	0.94	0.89	0.89	0.97	0.92	0.92
16	$\begin{bmatrix} 4,2 & 3,1 \\ 2,4 & 1,3 \end{bmatrix}$	5,1 12,8	DD	0.84	0.89	0.84	0.87	0.92	0.87
17	$\begin{bmatrix} 4,2 & 3,1 \\ 2,3 & 1,4 \end{bmatrix}$	1,8 5,12	OD	0.85	0.24	0.76	0.87	0.21	0.79
18	$\begin{bmatrix} 4,1 & 3,4 \\ 2,3 & 1,2 \end{bmatrix}$	3,2 11,10	OD	0.90	0.84	0.84	0.94	0.87	0.87
19	$\begin{bmatrix} 4,1 & 3,4 \\ 2,2 & 1,3 \end{bmatrix}$	4,2 11,9	DD	0.90	0.87	0.87	0.95	0.90	0.90
20	$\begin{bmatrix} 4,1 & 3,3 \\ 2,4 & 1,2 \end{bmatrix}$	5,5 8,8	OD	0.87	0.31	0.69	0.91	0.29	0.71
21	$\begin{bmatrix} 4,1 & 3,3 \\ 2,2 & 1,4 \end{bmatrix}$	5,8	DD	0.86	0.31	0.69	0.91	0.91	0.91
22	$\begin{bmatrix} 4,1 & 3,2 \\ 2,4 & 1,3 \end{bmatrix}$	5,6 7,8	OD	0.84	0.33	0.67	0.86	0.31	0.69

Table 2: Table of games in G^* (cont.)

game id	payoff matrix	bruns id	type	raw data			clean data		
				freq1	freq2	pbar	freq1	freq2	pbar
23	$\begin{bmatrix} 4,1 & 3,2 \\ 2,3 & 1,4 \end{bmatrix}$	5,7 6,8	DD	0.86	0.82	0.82	0.88	0.84	0.84
24	$\begin{bmatrix} 4,4 & 3,3 \\ 2,1 & 1,2 \end{bmatrix}$	10,1 12,3	DD	0.94	0.89	0.89	0.98	0.92	0.92
25	$\begin{bmatrix} 4,4 & 3,3 \\ 2,1 & 1,2 \end{bmatrix}$	7,3 10,6	OD	0.92	0.50	0.50	0.97	0.50	0.50
26	$\begin{bmatrix} 4,4 & 3,2 \\ 2,1 & 1,3 \end{bmatrix}$	8,3 10,5	OD	0.91	0.51	0.51	0.95	0.51	0.51
27	$\begin{bmatrix} 4,4 & 3,1 \\ 2,3 & 1,2 \end{bmatrix}$	10,3	DD	0.90	0.90	0.90	0.94	0.94	0.94
28	$\begin{bmatrix} 4,4 & 3,1 \\ 2,2 & 1,3 \end{bmatrix}$	9,3 10,4	OD	0.90	0.84	0.84	0.95	0.87	0.87
29	$\begin{bmatrix} 4,3 & 3,4 \\ 2,2 & 1,1 \end{bmatrix}$	1,3 10,12	OD	0.91	0.46	0.50	0.95	0.47	0.50
30	$\begin{bmatrix} 4,3 & 3,4 \\ 2,1 & 1,2 \end{bmatrix}$	6,3 10,7	DD	0.88	0.86	0.86	0.91	0.89	0.89
31	$\begin{bmatrix} 4,3 & 3,2 \\ 2,4 & 1,1 \end{bmatrix}$	4,3 10,9	DD	0.91	0.88	0.88	0.94	0.93	0.93
32	$\begin{bmatrix} 4,3 & 3,2 \\ 2,1 & 1,4 \end{bmatrix}$	3,9 4,10	OD	0.94	0.68	0.46	0.97	0.71	0.36
33	$\begin{bmatrix} 4,3 & 3,1 \\ 2,2 & 1,4 \end{bmatrix}$	2,9 4,11	OD	0.89	0.28	0.72	0.92	0.26	0.74

Table 2: Table of games in G^* (cont.)

game id	**payoff matrix**	**bruns id**	**type**	**raw data**			**clean data**		
				freq1	freq2	pbar	freq1	freq2	pbar
34	$\begin{bmatrix} 4,2 & 3,4 \\ 2,3 & 1,1 \end{bmatrix}$	2,3 10,11	OD	0.92	0.51	0.51	0.95	0.51	0.51
35	$\begin{bmatrix} 4,2 & 3,3 \\ 2,4 & 1,1 \end{bmatrix}$	4,4 9,9	OD	0.93	0.64	0.51	0.96	0.64	0.43
36	$\begin{bmatrix} 4,2 & 3,3 \\ 2,1 & 1,4 \end{bmatrix}$	4,9	DD	0.90	0.90	0.90	0.93	0.93	0.93
37	$\begin{bmatrix} 4,2 & 3,1 \\ 2,4 & 1,3 \end{bmatrix}$	4,1 12,9	DD	0.85	0.84	0.84	0.88	0.87	0.87
38	$\begin{bmatrix} 4,2 & 3,1 \\ 2,3 & 1,4 \end{bmatrix}$	1,9 4,12	OD	0.81	0.31	0.69	0.85	0.29	0.71
39	$\begin{bmatrix} 4,1 & 3,4 \\ 2,3 & 1,2 \end{bmatrix}$	3,3 10,10	OD	0.88	0.86	0.86	0.92	0.91	0.91
40	$\begin{bmatrix} 4,1 & 3,3 \\ 2,4 & 1,2 \end{bmatrix}$	4,5 8,9	OD	0.90	0.20	0.81	0.93	0.16	0.84
41	$\begin{bmatrix} 4,1 & 3,2 \\ 2,4 & 1,3 \end{bmatrix}$	4,6 7,9	OD	0.82	0.25	0.75	0.87	0.22	0.78
42	$\begin{bmatrix} 4,1 & 3,2 \\ 2,3 & 1,4 \end{bmatrix}$	4,7 6,9	DD	0.84	0.88	0.84	0.86	0.91	0.86
43	$\begin{bmatrix} 4,4 & 2,3 \\ 3,2 & 1,1 \end{bmatrix}$	12,1	DD	0.85	0.85	0.85	0.89	0.89	0.89
44	$\begin{bmatrix} 4,4 & 2,3 \\ 3,1 & 1,2 \end{bmatrix}$	7,1 12,6	OD	0.89	0.55	0.55	0.92	0.56	0.56

Table 2: Table of games in G^* (cont.)

game id	payoff matrix	bruns id	type	raw data			clean data		
				freq1	freq2	pbar	freq1	freq2	pbar
45	$\begin{bmatrix} 4,4 & 2,2 \\ 3,1 & 1,3 \end{bmatrix}$	8,1 12,5	OD	0.92	0.49	0.50	0.95	0.50	0.50
46	$\begin{bmatrix} 4,4 & 2,1 \\ 3,2 & 1,3 \end{bmatrix}$	9,1 12,4	OD	0.91	0.86	0.86	0.95	0.89	0.89
47	$\begin{bmatrix} 4,3 & 2,4 \\ 3,2 & 1,1 \end{bmatrix}$	1,1 12,12	OD	0.79	0.32	0.50	0.82	0.31	0.46
48	$\begin{bmatrix} 4,3 & 2,4 \\ 3,1 & 1,2 \end{bmatrix}$	6,1 12,7	DD	0.88	0.75	0.75	0.91	0.77	0.77
49	$\begin{bmatrix} 4,3 & 2,2 \\ 3,1 & 1,4 \end{bmatrix}$	3,7 6,10	OD	0.92	0.67	0.49	0.96	0.68	0.40
50	$\begin{bmatrix} 4,3 & 2,1 \\ 3,2 & 1,4 \end{bmatrix}$	2,7 6,11	OD	0.84	0.23	0.77	0.87	0.20	0.80
51	$\begin{bmatrix} 4,2 & 2,4 \\ 3,3 & 1,1 \end{bmatrix}$	2,1 12,11	OD	0.69	0.32	0.50	0.71	0.30	0.50
52	$\begin{bmatrix} 4,2 & 2,3 \\ 3,4 & 1,1 \end{bmatrix}$	6,4 9,7	OD	0.78	0.74	0.75	0.79	0.77	0.75
53	$\begin{bmatrix} 4,2 & 2,1 \\ 3,3 & 1,4 \end{bmatrix}$	1,7 6,12	OD	0.73	0.21	0.73	0.75	0.17	0.75
54	$\begin{bmatrix} 4,1 & 2,4 \\ 3,3 & 1,2 \end{bmatrix}$	3,1 12,10	OD	0.67	0.71	0.67	0.69	0.72	0.69
55	$\begin{bmatrix} 4,1 & 2,3 \\ 3,4 & 1,2 \end{bmatrix}$	6,5 8,7	OD	0.65	0.39	0.61	0.67	0.38	0.62

Table 2: Table of games in G^* (cont.)

game id	**payoff matrix**	**bruns id**	**type**	**raw data**			**clean data**		
				freq1	freq2	pbar	freq1	freq2	pbar
56	$\begin{bmatrix} 4,1 & 2,2 \\ 3,4 & 1,3 \end{bmatrix}$	6,6 7,7	OD	0.67	0.40	0.60	0.69	0.40	0.60
57	$\begin{bmatrix} 4,1 & 2,2 \\ 3,3 & 1,4 \end{bmatrix}$	6,7	DD	0.68	0.68	0.68	0.71	0.71	0.71
58	$\begin{bmatrix} 4,4 & 2,3 \\ 3,1 & 1,2 \end{bmatrix}$	7,4 9,6	CO2	0.82	0.50	0.61	0.87	0.50	0.58
59	$\begin{bmatrix} 4,4 & 2,2 \\ 3,1 & 1,3 \end{bmatrix}$	8,4 9,5	CO2	0.87	0.49	0.56	0.91	0.49	0.55
60	$\begin{bmatrix} 4,4 & 2,1 \\ 3,2 & 1,3 \end{bmatrix}$	9,4	CO2	0.85	0.85	0.86	0.88	0.88	0.89
61	$\begin{bmatrix} 4,3 & 2,4 \\ 3,2 & 1,1 \end{bmatrix}$	1,4 9,12	MP	0.82	0.29	0.44	0.85	0.28	0.41
62	$\begin{bmatrix} 4,3 & 2,2 \\ 3,1 & 1,4 \end{bmatrix}$	3,10	CO1	0.82	0.82	0.73	0.85	0.85	0.58
63	$\begin{bmatrix} 4,3 & 2,1 \\ 3,2 & 1,4 \end{bmatrix}$	2,10 3,11	CO1	0.73	0.40	0.75	0.74	0.38	0.75
64	$\begin{bmatrix} 4,2 & 2,4 \\ 3,3 & 1,1 \end{bmatrix}$	2,4 9,11	MP	0.85	0.34	0.45	0.88	0.33	0.42
65	$\begin{bmatrix} 4,2 & 2,3 \\ 3,4 & 1,1 \end{bmatrix}$	3,4 9,10	MP	0.86	0.75	0.62	0.90	0.77	0.50
66	$\begin{bmatrix} 4,2 & 2,1 \\ 3,3 & 1,4 \end{bmatrix}$	1,10 3,12	CO1	0.65	0.47	0.76	0.67	0.47	0.76

Table 2: Table of games in G^* (cont.)

game id	**payoff matrix**	**bruns id**	**type**	**raw data**			**clean data**		
				freq1	freq2	pbar	freq1	freq2	pbar
67	$\begin{bmatrix} 4,1 & 2,3 \\ 3,4 & 1,2 \end{bmatrix}$	3,5 8,10	MP	0.78	0.26	0.81	0.80	0.23	0.82
68	$\begin{bmatrix} 4,1 & 2,2 \\ 3,4 & 1,3 \end{bmatrix}$	3,6 7,10	MP	0.71	0.20	0.83	0.72	0.18	0.76
69	$\begin{bmatrix} 4,4 & 1,3 \\ 3,2 & 1,2 \end{bmatrix}$	7,6	CO2	0.47	0.47	0.95	0.49	0.49	0.98
70	$\begin{bmatrix} 4,4 & 1,2 \\ 3,3 & 1,2 \end{bmatrix}$	7,5 8,6	CO2	0.52	0.52	0.97	0.52	0.52	0.96
71	$\begin{bmatrix} 4,3 & 1,4 \\ 3,2 & 1,1 \end{bmatrix}$	1,6 7,12	MP	0.41	0.24	0.60	0.38	0.22	0.56
72	$\begin{bmatrix} 4,3 & 1,1 \\ 3,2 & 1,4 \end{bmatrix}$	1,11 2,12	CO1	0.45	0.29	0.68	0.43	0.27	0.64
73	$\begin{bmatrix} 4,2 & 1,4 \\ 3,3 & 1,1 \end{bmatrix}$	2,6 7,11	MP	0.28	0.23	0.50	0.27	0.21	0.47
74	$\begin{bmatrix} 4,2 & 1,1 \\ 3,3 & 1,4 \end{bmatrix}$	1,12	CO1	0.31	0.31	0.62	0.30	0.30	0.59
75	$\begin{bmatrix} 4,1 & 1,3 \\ 3,4 & 1,2 \end{bmatrix}$	1,5 8,12	MP	0.24	0.39	0.58	0.22	0.37	0.55
76	$\begin{bmatrix} 4,4 & 1,2 \\ 3,3 & 1,2 \end{bmatrix}$	8,5	CO2	0.45	0.45	0.92	0.44	0.44	0.90
77	$\begin{bmatrix} 4,3 & 1,1 \\ 3,2 & 1,4 \end{bmatrix}$	2,11	CO1	0.39	0.39	0.78	0.38	0.38	0.76

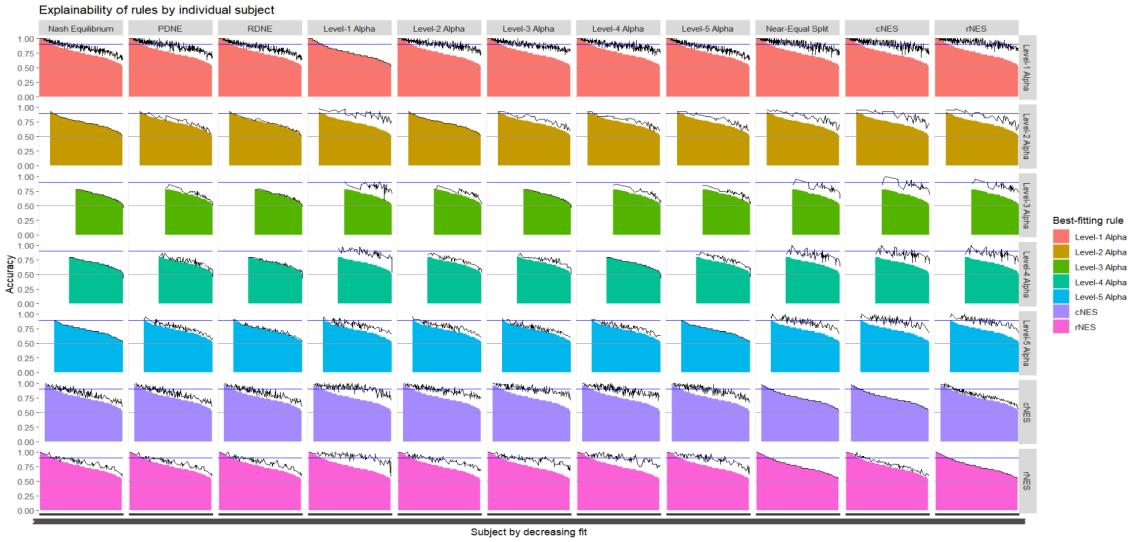

Figure A12: The accuracy of each subject's best-fitting rule (area in color) and the added accuracy of residual rules (black lines). The rows correspond to the groups of subjects that are best fit by a specific rule. The columns correspond to the residual rule. The added accuracy is equal to the best-fitting rule's accuracy plus the accuracy of the residual rule on the games where the best-fitting rule fails to predict. The gray horizontal line corresponds to 0.50, which is the expected accuracy of a random rule. The blue horizontal line corresponds to 0.90. Subjects are ordered from the best-fitting subjects to the worst (conditional on being best-fit by a specific rule).

Table 2: Table of games in G^* (cont.)

game id	**payoff matrix**	**bruns id**	**type**	**raw data**			**clean data**		
				freq1	freq2	pbar	freq1	freq2	pbar
78	$\begin{bmatrix} 4,2 & 1,4 \\ 3,3 & 1,1 \end{bmatrix}$	2,5 8,11	MP	0.36	0.27	0.59	0.34	0.26	0.57

E Double rule plots and tables

F Graphs of theoretical similarity classes of behavioral rules

F.1 Pareto efficiency

Three similarity classes, see Figure A13.

F.2 Nash equilibrium and Pareto efficiency

Eight similarity classes, see Figure A14.

A. Table of games in G^*

game id	payoff matrix				bruns id	type	freq1	freq2	n1	n2
1	4	4	3	3	11,1	DD	0,96	0,95	57	88
	2	2	1	1	12,1		0,93	0,93	60	96
2	4	4	3	3	7,2	OD1	0,93	0,60	28	743
	2	1	1	2	11,6		0,90	0,58	30	829
3	4	4	3	2	11,2	DD	1,00	1,00	56	56
	2	3	1	1			0,97	0,97	58	58
4	4	4	3	2	8,2	OD1	0,96	0,60	54	672
	2	1	1	3	11,5		0,92	0,59	60	743
5	4	4	3	1	10,2	DD	0,96	1,00	27	32
	2	3	1	2	11,3		0,90	0,95	29	37
6	4	4	3	1	9,2	OD2	0,89	0,97	27	33
	2	2	1	3	11,4		0,86	0,89	29	37
7	4	3	3	4	1,2	OD1	0,90	0,40	60	805
	2	2	1	1	11,12		0,88	0,41	64	896
8	4	3	3	4	6,2	DD	0,94	0,86	34	36
	2	1	1	2	11,7		0,89	0,85	38	39
9	4	3	3	2	5,3	DD	0,97	0,93	29	29
	2	4	1	1	10,8		0,94	0,88	33	33
10	4	3	3	2	3,8	OD1	0,95	0,55	55	55
	2	1	1	4	5,1		0,89	0,55	61	58
11	4	3	3	1	5,2	DD	0,82	1,00	28	30
	2	4	1	2	11,8		0,74	0,97	31	31
12	4	3	3	1	2,8	OD2	0,86	0,19	28	27
	2	2	1	4	5,1		0,80	0,24	30	29
13	4	2	3	4	2,2	OD1	0,96	0,47	26	691
	2	3	1	1	11,11		0,93	0,47	29	769
14	4	2	3	3	5,4	OD1	1,00	0,58	51	787
	2	4	1	1	9,8		0,96	0,56	57	871
15	4	2	3	3	4,8	DD	1,00	0,88	33	32
	2	1	1	4	5,9		0,97	0,81	37	37
16	4	2	3	1	5,1	DD	0,83	0,91	36	34
	2	4	1	3	12,8		0,82	0,82	39	38
17	4	2	3	1	1,8	OD2	0,86	0,17	29	29
	2	3	1	4	5,12		0,79	0,24	33	33
18	4	1	3	4	3,2	OD2	0,90	0,74	59	54
	2	3	1	2	11,10		0,88	0,67	60	61
19	4	1	3	4	4,2	DD	0,93	0,96	59	28
	2	2	1	3	11,9		0,89	0,94	64	31
20	4	1	3	3	5,5	OD2	0,93	0,21	27	28
	2	4	1	2	8,8		0,90	0,23	29	30

A. Table of games in G^* (cont.)

game id	payoff matrix				bruns id	type	freq1	freq2	n1	n2
21	4	1	3	3	5,8	DD	0,99	0,99	69	69
	2	2	1	4			0,93	0,93	75	75
22	4	1	3	2	5,6	OD2	0,74	0,46	31	56
	2	4	1	3	7,8		0,67	0,46	36	65
23	4	1	3	2	5,7	DD	0,94	0,77	34	31
	2	3	1	4	6,8		0,92	0,70	38	37
24	4	4	3	3	10,1	DD	0,92	0,93	36	58
	1	2	2	1	12,3		0,90	0,85	39	66
25	4	4	3	3	7,3	OD1	0,96	0,55	84	700
	1	1	2	2	10,6		0,95	0,54	95	774
26	4	4	3	2	8,3	OD1	1,00	0,53	23	816
	1	1	2	3	10,5		0,86	0,52	29	902
27	4	4	3	1	10,3	DD	0,98	0,98	83	83
	1	3	2	2			0,92	0,92	93	93
28	4	4	3	1	9,3	OD2	0,97	0,93	30	29
	1	2	2	3	10,4		0,94	0,91	32	33
29	4	3	3	4	1,3	OD1	1,00	0,49	27	726
	1	2	2	1	10,12		0,93	0,51	29	806
30	4	3	3	4	6,3	DD	0,94	0,89	88	28
	1	1	2	2	10,7		0,91	0,91	98	32
31	4	3	3	2	4,3	DD	0,93	0,97	55	30
	1	4	2	1	10,9		0,93	0,97	58	31
32	4	3	3	2	3,9	OD1	0,90	0,64	29	617
	1	1	2	4	4,10		0,82	0,63	33	682
33	4	3	3	1	2,9	OD2	0,86	0,18	56	55
	1	2	2	4	4,11		0,87	0,19	61	58
34	4	2	3	4	2,3	OD1	0,98	0,53	59	777
	1	3	2	1	10,11		0,94	0,52	66	860
35	4	2	3	3	4,4	OD1	0,91	0,56	67	618
	1	4	2	1	9,9		0,90	0,56	70	688
36	4	2	3	3	4,9	DD	0,96	0,96	84	84
	1	1	2	4			0,93	0,93	89	89
37	4	2	3	1	4,1	DD	0,88	0,78	56	23
	1	4	2	3	12,9		0,83	0,69	60	29
38	4	2	3	1	1,9	OD2	0,89	0,25	27	28
	1	3	2	4	4,12		0,83	0,23	29	30
39	4	1	3	4	3,3	OD2	0,68	0,83	25	23
	1	3	2	2	10,10		0,63	0,75	27	28
40	4	1	3	3	4,5	OD2	0,98	0,14	59	28
	1	4	2	2	8,9		0,91	0,17	66	30

A. Table of games in G^* (cont.)

game id	payoff matrix				bruns id	type	freq1	freq2	n1	n2
41	4	1	3	2	4,6	OD2	0,84	0,24	56	37
	1	4	2	3	7,9		0,83	0,24	63	38
42	4	1	3	2	4,7	DD	0,77	0,87	31	31
	1	3	2	4	6,9		0,78	0,75	37	36
43	4	4	2	3	12,1	DD	0,94	0,94	102	102
	3	2	1	1			0,89	0,89	114	114
44	4	4	2	3	7,1	OD1	1,00	0,57	26	484
	3	1	1	2	12,6		0,93	0,56	29	539
45	4	4	2	2	8,1	OD1	0,89	0,61	28	754
	3	1	1	3	12,5		0,83	0,59	30	837
46	4	4	2	1	9,1	OD2	0,94	1,00	80	29
	3	2	1	3	12,4		0,89	0,91	92	33
47	4	3	2	4	1,1	OD1	0,89	0,27	28	493
	3	2	1	1	12,12		0,87	0,30	30	548
48	4	3	2	4	6,1	DD	0,93	0,73	28	30
	3	1	1	2	12,7		0,94	0,75	31	32
49	4	3	2	2	3,7	OD1	0,94	0,57	310	280
	3	1	1	4	6,10		0,90	0,58	343	319
50	4	3	2	1	2,7	OD	0,89	0,19	456	409
	3	2	1	4	6,11		0,85	0,22	505	455
51	4	2	2	4	2,1	OD1	0,69	0,27	865	865
	3	3	1	1	12,11		0,68	0,30	960	960
52	4	2	2	3	6,4	OD1	0,74	0,78	865	865
	3	4	1	1	9,7		0,71	0,77	960	960
53	4	2	2	1	1,7	OD2	0,71	0,17	325	312
	3	3	1	4	6,12		0,68	0,21	362	345
54	4	1	2	4	3,1	OD2	0,71	0,69	383	366
	3	3	1	2	12,10		0,69	0,67	424	408
55	4	1	2	3	6,5	OD2	0,66	0,45	437	428
	3	4	1	2	8,7		0,65	0,45	484	474
56	4	1	2	2	6,6	OD2	0,65	0,36	191	220
	3	4	1	3	7,7		0,65	0,40	212	250
57	4	1	2	2	6,7	DD	0,69	0,69	865	865
	3	3	1	4			0,68	0,68	960	960
58	4	4	2	3	7,4	CO2	0,91	0,53	402	377
	1	1	3	2	9,6		0,87	0,52	452	416
59	4	4	2	2	8,4	CO2	0,89	0,55	351	398
	1	1	3	3	9,5		0,87	0,54	391	437
60	4	4	2	1	9,4	CO1	0,88	0,88	57	57
	1	2	3	3			0,86	0,86	65	65

A. Table of games in G^* (cont.)

game id	payoff matrix				bruns id	type	freq1	freq2	n1	n2
61	4	3	2	4	1,4	MP	0,86	0,23	312	328
	1	2	3	1	9,12		0,83	0,26	342	367
62	4	3	2	2	3,10	CO2	0,78	0,78	747	747
	1	1	3	4			0,74	0,74	832	832
63	4	3	2	1	2,10	CO1	0,71	0,33	377	368
	1	2	3	4	3,11		0,67	0,34	421	407
64	4	2	2	4	2,4	MP	0,89	0,39	387	413
	1	3	3	1	9,11		0,85	0,42	431	458
65	4	2	2	3	3,4	MP	0,89	0,80	356	418
	1	4	3	1	9,10		0,87	0,78	397	458
66	4	2	2	1	1,10	CO1	0,61	0,39	335	375
	1	3	3	4	3,12		0,58	0,41	375	412
67	4	1	2	3	3,5	MP	0,81	0,19	865	865
	1	4	3	2	8,10		0,78	0,22	960	960
68	4	1	2	2	3,6	MP	0,64	0,17	434	431
	1	4	3	3	7,10		0,63	0,23	481	479
69	4	4	1	3	7,6	CO1	0,57	0,57	865	865
	3	1	2	2			0,56	0,56	960	960
70	4	4	1	2	7,5	CO1	0,59	0,53	375	463
	3	1	2	3	8,6		0,58	0,53	420	511
71	4	3	1	4	1,6	MP	0,44	0,19	440	425
	3	2	2	1	7,12		0,46	0,25	483	477
72	4	3	1	1	1,11	CO2	0,43	0,25	344	353
	3	2	2	4	2,12		0,44	0,28	385	392
73	4	2	1	4	2,6	MP	0,28	0,24	405	441
	3	3	2	1	7,11		0,32	0,27	453	488
74	4	2	1	1	1,12	CO2	0,26	0,26	865	865
	3	3	2	4			0,29	0,29	960	960
75	4	1	1	3	1,5	MP	0,21	0,40	409	369
	3	4	2	2	8,12		0,24	0,41	457	411
76	4	4	1	2	8,5	CO1	0,53	0,53	686	686
	2	1	3	3			0,53	0,53	770	770
77	4	3	1	1	2,11	CO2	0,31	0,31	865	865
	2	2	3	4			0,32	0,32	960	960
78	4	2	1	4	2,5	MP	0,32	0,19	404	375
	2	3	3	1	8,11		0,35	0,21	443	416

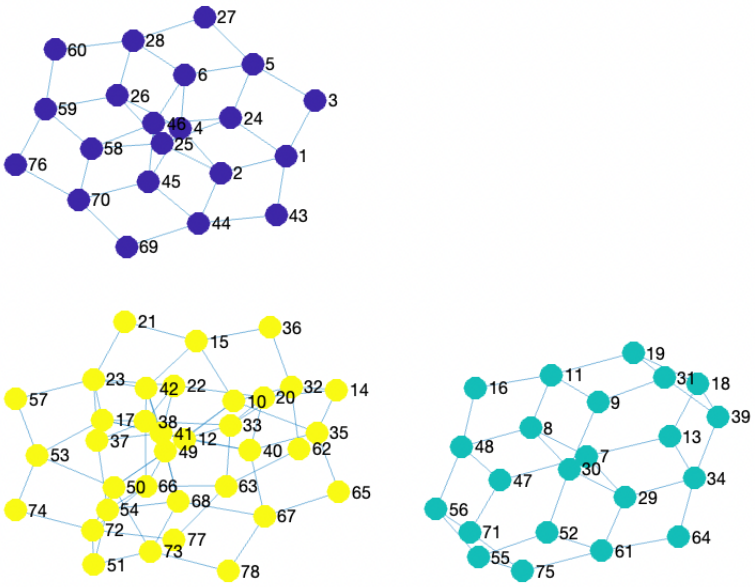

Figure A13: Graph of the theoretical similarity classes in G^* implied by Pareto efficiency.

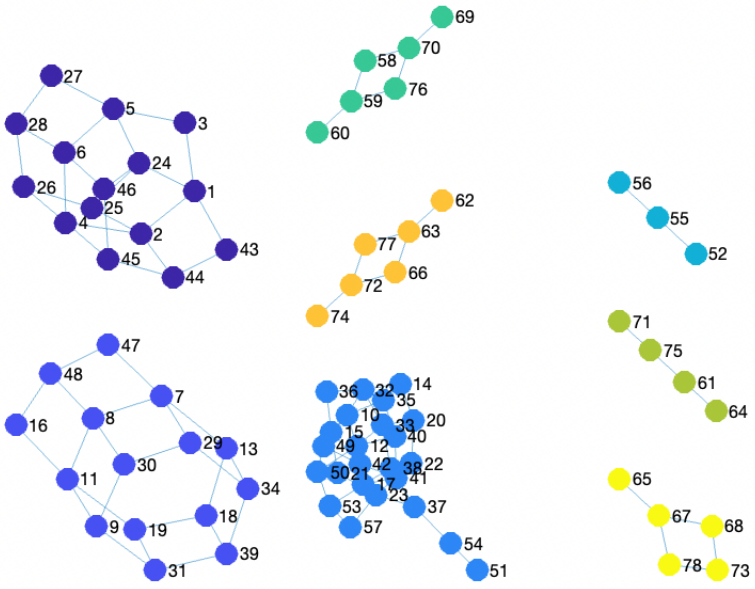

Figure A14: Graph of the theoretical similarity classes in G^* implied by Nash equilibrium and Pareto efficiency.

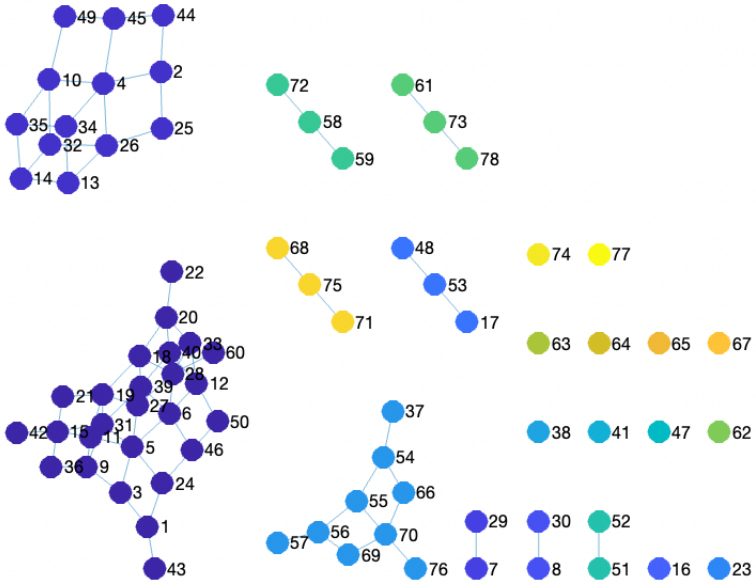

Figure A15: Graph of the theoretical similarity classes in G^* implied by all Level-$k(\alpha)$ and self-favoring and other-favoring Near-Equal Split.

F.3 Level-$k(\alpha)$ and Near-equal split

22 similarity classes, see Figure A15.

F.4 Level-k

11 similarity classes, see Figure A16.

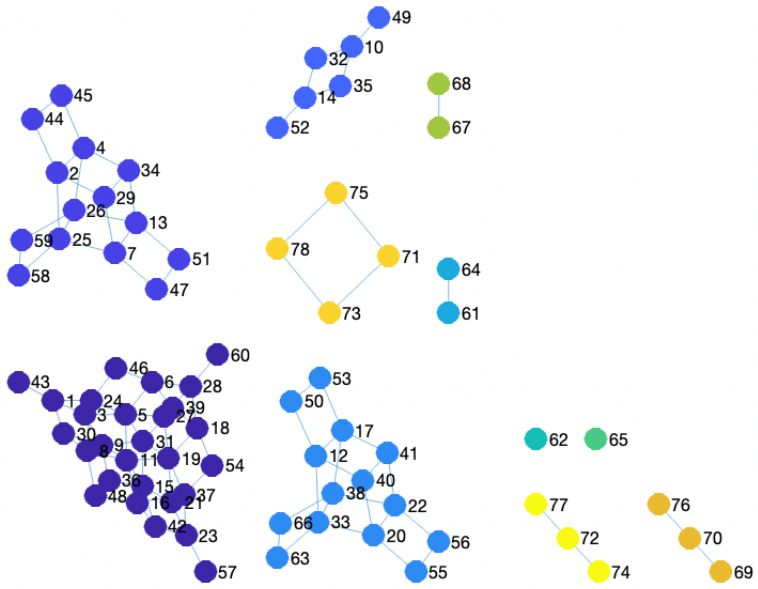

Figure A16: Graph of the theoretical similarity classes in G^* implied by all Level-k, $k = 1, \ldots, 5$.