Charting the Type Space

The Case of Linear Public Good Experiments

Christoph Engel, Carina I. Hausladen & Marcel H. Schubert

2021-09-07

Behavior in economic games is not only noisy. One has reason to believe that heterogeneity is patterned. A prominent application is the linear public good. It is widely accepted that choices result from participants holding discernible types. Proposed types, like freeriders or conditional cooperators, are intuitive. But the composition of the type space is neither theoretically nor empirically settled. In this paper we leverage machine learning methods to chart the type space. We use simulation to understand what can be achieved with machine learning. We rely on these insights to find clusters in a large (N = 12,414) set of experimental data points from public good games. We discuss ways in which these clusters could be rationalized. Finally, we offer two outlooks, one traditional economic approach and another rooted in supervised learning, on how to go forward with our results.

JEL: C71, H41

1 Introduction

Standard theory predicts the tragedy of the commons. Everybody maximises individual profit, and exploits socially minded choices of others. If members of the community interact repeatedly, but it is known when interaction will stop, the gloomy prediction still holds. A robust experimental literature shows that, in the aggregate, results look different. In a standard symmetric linear public good, average contributions typically start considerably above zero but tend to decline over time (J. O. Ledyard, 1995; Zelmer, 2003; Chaudhuri, 2011). A substantial theoretical literature rationalises these results, usually by introducing some form of social preference into the utility function (for an excellent overview see Fehr and Schmidt, 2002). While such extensions of motives can generate a starting point above zero, it is more difficult for them to also explain the downward trend. For this, one needs a reactive element. It has been prominently introduced into the literature with the concept of conditional cooperation (Fischbacher, Gachter, and Fehr, 2001). A conditional cooperator is willing to act unselfishly provided she expects or knows that others will do so as well. In principle, the downward trend could result from the fact that conditional cooperation is imperfect. While participants would not be outright selfish, they would still try to outperform their peers, albeit only slightly (Fischbacher and Gachter, 2010). Engel and Rockenbach (2020) has shown that this explanation is not supported by the data. Rather the downward trend results from bad experiences. If participants, in the previous period, have been overly optimistic about the contributions of their peers, they adjust their beliefs and, in turn, their contributions, in the subsequent period. Critically they overreact to negative experiences.

This is where the present project starts. If the population were homogeneous, and completely consisted of conditional cooperators, there could not be a downward trend. The source of the trend, and hence the need for at least some form of institutional intervention to sustain cooperation, must be heterogeneity. Even if many individuals are in principle good-natured and happy to cooperate in good times, their willingness to do so is fragile. If they experience exploitation, they react. While the claim is intuitive that populations are heterogeneous, understanding the character of this heterogeneity is inherently difficult. One needs estimates about the utility functions of group members: is an individual outright selfish? Is she so strongly motivated by the common good that she does not care about the choices of others? Or does she react? If so what does she react to? And how strongly? There could also be mixed types: individuals freeride or cooperate for that matter, unconditionally as long as a certain threshold is not crossed. Reaction functions might have an exploratory component: while an individual is in principle of a certain type, she occasionally tests the waters, by contributing more or less than suggested by her ordinary reaction function. Reaction functions could be non-linear. Conditional cooperators might for instance be happy to tolerate an occasional bad experience (maybe attributing it to others having made a mistake), but they might lose faith, and react very strongly, if bad experiences repeat. There might be individuals who try to educate their groups, by showing them what could happen if others do not stop misbehaving. For that purpose, they might once contribute nothing, and go back to high contributions in the following period. Reaction functions may also depend on the effects of occasional exploitation. In the standard setting (group size

4, marginal per capita rate .4) 3 loyal members still make a small profit if they continue to cooperate (and accept that the free rider gains a windfall profit).

All these behavioural programs resonate with data from public good experiments. But these are only ex post rationalisations. Moreover, not every dataset could be reasonably explained with all of these behavioural programs. Before the field can move forward, and better targeted interventions can be designed, one needs a much deeper understanding of behavioural heterogeneity. Ultimately it would be highly desirable to formally define, and experimentally test, these reaction functions. But a necessary first step is exploratory: which reaction functions exist, and how prevalent are they? Charting the type space is the aim of the present project. We start from the assumption that the theoretical possibilities for the composition of the type space are at best partly understood. We further note that reactions may not only differ in kind, but also in degree, which is why parameters must be estimated. This is why we revert to machine learning. We use a reasonably large dataset of earlier linear public good games to find types, and discuss reaction functions that would rationalize the reaction patterns.

In principle, choice data is well suited for our endeavour. The choices of others in previous periods are the only information to which participants can react in an anonymous linear public good. For each individual, we can check whether, and if so in which ways, they have reacted to past choices of the remaining members of their group. We can represent the development of their choices over time as a timeseries. We can use the rich set of methods developed in the machine learning community for clustering the timeseries of choices, giving the algorithm the possibility to use the average choices of the remaining group members in the previous period as an input. From these clusters we can extract what machine learners call a prototype.

This approach, however, presupposes that reaction functions can indeed be inferred from choices. Arguably this will depend on at least two features of the data: the precision with which an individual participant has reacted to experiences, and the character of these experiences. The former depends on the noise rate. Potentially individuals have a certain reaction function, but they do not act upon it at all times. The latter depends on group composition, and on initial choices. To illustrate: in a group of three straightforward free riders, a conditional cooperator can be expected to quickly make choices that are indistinguishable from the choices of native free riders. Discriminating between the choices of conditional cooperators and of free riders will be the more difficult the lower the initial contribution of a conditional cooperator. It should be equally difficult to discriminate between conditional cooperators and genuinely cooperative participants if a single conditional cooperator is surrounded by a group of native cooperators.

Before using machine learning for clustering participants in real data, we ,therefore, investigate with simulated data the framework conditions under which potentially powerful algorithms can find types. In simulations, we can systematically vary the composition of the type space, the definition of individual types, and the noise rate. This first step yields one important insight: machine learning methods find patterns. If the choice program of an individual is reactive, one and the same choice pattern may result from different reaction

functions, depending on the choices the remaining group members have made in the previous period. Consequently, there is no one to one mapping between patterns and types. This must be reflected in the design of the clustering algorithm. We show that interpretation becomes much easier if one estimates a number of patterns that is considerably bigger than the expected number of types and hence reaction functions.

Simulation also helps us with two further tasks. We can estimate the richness of the data that is required for making the exercise meaningful. And we can check in which ways fine-tuning the algorithm improves estimation.

As explained above, we do not take it for granted that the type space has already been understood completely. A major motive for our project is the possibility that there are further types that have not been theorised. Yet for our simulations we need to build in types that have already been conceptualised. In the simulations, we work with groups consisting of different fractions of the following five types: altruists, whom we define as participants who do not react to experiences, and who start with relatively high contributions. Such participants may exhibit variance, the more so the higher the noise rate. But they show no trend. The corresponding type at the lower end is total freeriders. They in principle do not make contributions to the public project, but may occasionally deviate from this program. Pure conditional cooperators start with relatively high contributions but adjust them to experiences. Following Fischbacher, Gachter, and Fehr (2001), we allow for hump shaped contributions: up to a value near half the endowment they increase contributions in reaction to good experiences, but they exhibit a perverse reaction to even better experiences. Following Engel and Rockenbach (2020), we finally implement farsighted freeriders. For some initial periods "they feed the cow" by making substantial contributions, but then start "milking" it by reducing their contributions below average contributions in the previous period.

The remainder of this paper is organized as follows: In Section 2, we situate our endeavour in the literature. A small number of types have already been theorized. We use these types to simulate data, Section 3. We use this dataset for two purposes: In Section 4 we show why a naive approach cannot work: once one allows for choice functions (of at least some types) to be reactive, there is no one-to-one mapping between types and what one can observe in the data, i.e. choice patterns and their corresponding experience patterns. In Section 5 we use theory and an extended grid search to find the best algorithmic configuration for clustering this kind of data. This prepares the main Section 6, where we apply the method to a sizeable set of experimental data. It turns out that empirical choice/experience patterns are much richer, and quite different from the patterns resulting if one exclusively assumes types that have already been theorized. Section 8 concludes with discussion.

2 Literature

It has often been noted that choices in public good experiments are not homogeneous (see only Fischbacher, Gachter, and Fehr, 2001; Fischbacher and Gachter, 2010). But the literature has only relatively recently begun to define the type space more precisely. Amin,

Abouelela, and Soliman (2018) use theory derived from Fischbacher, Gachter, and Fehr (2001) to classify 72 participants from a new experiment into 7 types, and then use simulation to find out which fraction of which type is required to sustain cooperation in a linear public good. Lucas, Oliveira, and Banuri (2012) show with simulation that cooperation is hard to sustain in a linear public good if the group consists of heterogeneous types (which they take from Fischbacher and Gachter (2010)). Arifovic and J. Ledyard (2012) develop a model that combines social preferences with learning. In the framework of this model, conditional cooperation is not a type but develops endogenously. They use data from, among others, Isaac and Walker (1988) and Andreoni (1995) to calibrate their model, and argue that it has a good fit. We have a different goal. On the one hand, we do not expect individual choices to be merely noisy. We consider the possibility that heterogeneity is patterned. On other hand, we do not assume that the behavioral forces that drive this heterogeneity are already fully understood. We, to the contrary, want to find patterns that are hard to reconcile with extant theoretical concepts. The purpose of our exercise is hypothesis generation. Testing these hypotheses would require a series of new experiments. That is beyond the scope of the present paper.

Engel (2020) also uses machine learning to organize the type space for experimental data, and demonstrates the approach with data from Fischbacher and Gachter (2010). Yet he has a different research question. He wants to compare the performance of a finite mixture model (that estimates the type space and choices conditional on type simultaneously) with a two-step approach (that first estimates the type space from the data, and then choices conditional on type in a mixed effects model that interacts the types estimated in the first step with the effect of experimental manipulations). He also uses a different approach for estimating types, using the coefficients of local (per participant) regressions as inputs for a classification and regression tree.

A third group of contributions is more remote. Game theory usually starts with a complete definition of the game which includes the strategies available. Yet when they are exposed to one of the games of life, individuals often do not know that much. They must learn what game they are playing and what strategies are available. This task is even harder if they cannot exclude that the population with whom they play is heterogeneous. However, games can be too complex for solving them analytically. Then solutions must be found computationally. Ficici, Parkes, and Pfeffer (2012) make the game tractable by first compressing a large number of agents into a manageable number of clusters, and then solve the simplified game analytically.

Closest in spirit are Bapna et al. (2004) and Lu et al. (2016). Both papers aim at classifying bidding strategies in online auctions (Bapna et al., 2004) and in flower auctions (Lu et al., 2016), using machine learning methods. Vorobeychik, Wellman, and Singh (2007) use machine learning methods to find the strategy space of infinite games . Mao et al. (2017) use experimental data from a prisoner's dilemma to specify a classic learning model helping them to divide players of a prisoners' dilemma game into two distinct behavioural types. The main difference to us is that we look at a different, more compley game (a dilemma) to which prior results are not easily transferred. Moreover, we use experimental data, and exploit the power of algorithms for the classification of time series data.

3 Data Generating Process

Linear Public Good Games While we believe our method to be applicable more generally for finding patterned heterogeneity in repeated, interactive experiments, our specific object of investigation is a linear public good. The game is defined by the following profit function

$$\pi_{it} = e - c_{it} + \mu \sum_{k=1}^{K} c_{kt} \tag{1}$$

where π is profit of individual i in period t. Every period, the individual receives an endowment e. She can keep the endowment, or make a contribution c to the public project of the group. Marginal per capita rate $0 < \mu < 1$ creates the dilemma. As $\mu < 1$, each individual is best off keeping the entire endowment for herself. Yet as $K\mu > 1$, the group is best off if all members contribute their complete endowments. Most frequently, $e = 20, \mu = .4, G = 4$ have been chosen (J. O. Ledyard, 1995; Zelmer, 2003; Chaudhuri, 2011). Then 3 loyal group members still make a small profit. This serves as a buffer against the rapid decline of contributions.

Simulated type space In their seminal paper, Fischbacher, Gachter, and Fehr (2001) argue that (in their one-shot version of this game) there are three types: free-riders, conditional co-operators, and "hump-shaped" players. In his reanalysis of Fischbacher and Gachter (2010), Engel (2020) further finds a small, but discernible fraction of altruists. In their reanalysis of Fischbacher and Gachter (2010), Engel and Rockenbach (2020) use a combination of belief and choice data to distinguish a fifth group, which they call far-sighted freeriders. In our simulations, we allow for these five types. We focus on a partner design. Groups stay together for the full duration of the game. We always allow for an individual random effect η_i and residual error $\sigma_{it} \perp \eta_i$, which we both define to be normally distributed with mean 0 and standard deviation .3 ($\sim \mathcal{N}(0, .3)$). We thus implement the type space as defined in Table 1, where $c_{-i,t-1}$ is the average contribution of the remaining group members in the previous period $p - 1$.

Table 1: Simulated Type Space

type	$p = 1$	$p > 1$
short-sighted freerider	0	0
far-sighted freerider	10	$c_{-i,t-1}$ if $t < \tau$ 0 if $t \geq \tau$
conditional cooperator	10	$c_{-i,t-1}$
hump shaped	5	$c_{-i,t-1}$ if $c_{-i,t-1} \leq 10$ $-c_{-i,t-1}$ if $c_{-i,t-1} > 10$
altruist	20	20

We have groups of size $G = 4$, and we allow for $t = 5$ types. Participants choose their contributions to the public good simultaneously, which is why their order does not matter.

We consider the possibility that types are present more than once in a group. Hence we have a problem of unordered sampling with replacement. This gives us a total type space of

$$N = \binom{t+G-1}{G} = \frac{(5+4-1)!}{(5-1)!4!} = 70 \tag{2}$$

different group combinations. In our simulations, we include each of these 70 combinations of types 4 times. As three of the five types (conditional cooperators, far-sighted freeriders, hump shaped players) are reactive, we give the classification algorithm access to the exact same experiences that participants make in this design, i.e. the mean contribution of the remaining group members in the previous period. Hence the object of clustering is a two-dimensional time series consisting of the own contributions over time as well as the lagged past experiences in terms of average contributions within the group over time. We run the simulations for different number of periods $p \in 10, 15, 20, 25, 30$ The results do not differ with the number of periods P.

4 The Naive Approach

Confusion matrix Simulation is routinely employed to test the performance of an estimator. One generates a data set where one knows ground truth and checks whether a proposed estimator reconstructs the simulated parameters reasonably well. If an alternative estimator outperforms a competing estimator, one adopts the better performing method. Simulation gives the researcher confidence in the use of an estimator with data where she does not know ground truth.

When applied to our estimation problem, the seemingly straightforward criterion for choosing an estimator would be the frequency of identifying the simulated types. Assessed with this criterion, the results reported in Table 2 are sobering.[1] Each of the 5 types is exactly 224 times present in the dataset. Yet the size of the clusters ranges from 92 to 400. All clusters except the third are fairly impure: participants from different simulated types are put into the same cluster. Even knowing ground truth, it is hard to match clusters with types. Cells are highlighted in green if, at least, the most frequent type per cluster, and the most frequent cluster per type, coincide. In the example dataset, this only holds for the two non-reactive types: altruists and short-sighted freeriders. But even the purity of these two clusters is low. In cluster 5, 33% are actually conditional cooperators. In cluster 1, only 35% are indeed short-sighted free-riders. The remaining 65% consist of 27% hump-shaped types, 22% far-sighted (and hence partly reactive) free-riders, and 17% conditional cooperators. For all reactive types, one needs secondary (hump-shaped types, yellow cell) or tertiary (far-sighted free-riders: red cell) criteria for matching clusters with types. For cluster 3, no unique type can be found (as altruists are even more prominent in cluster 5). This is why one cannot even match the highest frequency in the cluster with the highest frequency in

[1] For consistency, we use the same algorithmic configuration that we develop in Section 5, and that we later apply to the experimental data in Section 6.

a type if one no longer considers clusters and types that have already been matched in an earlier round of matching.

Table 2: Confusion Matrix

cluster	1	2	3	4	5	Total
altruist			92		132	224
conditional cooperator	68	92			64	224
far-sighted freerider	86	74		64		224
hump shaped	106	94		24		224
short-sighted freerider	140	68		16		224
Total	400	328	92	104	196	1120

Clusters are patterns, not types Figure 1 shows why the attempt fails to validate 5 clusters by comparing them with 5 simulated types. The algorithm does a reasonably good job at clustering the data. But it clusters patterns of observed contributions, combined with patterns of observed experiences in past rounds (henceforth experiences). There is no one-to-one mapping of 5 patterns to 5 types.

Cluster 3 is the only pure cluster. It is defined by contributions being high, irrespective of experiences. These altruists do even accept outright exploitation. The remaining altruists are in cluster 5. By definition, their own contributions are also at the top. But now experiences are more favourable. This is why the algorithm lumps altruists together with conditional cooperators. As they are reactive, in quite some members of this cluster, contributions drop in the middle of the timeseries. The kink of course results from the presence of far-sighted freeriders who start cashing in. Many of the far-sighted freeriders are put into cluster 4. They are together with hump shaped players and short-sighted free riders, who both make low contributions throughout the game. Apparently the decisive feature for putting a participant into this cluster is not her own contributions, but the contrast between low contributions (at least for some part of the timeseries) and considerably more favourable experiences. By the same token, clusters 1 and 2 are distinguished. In both clusters, contributions are rather low. But in cluster 1, experiences are low as well, while they are discernibly higher than contributions in cluster 2.

It is even more instructive to consider which types are put into which clusters (Figure 1b) in the Appendix. Altruists, conditional cooperators and far-sighted freeriders are split into clearly distinct subgroups. In the case of altruists, the critical feature is experiences. If experiences are good, they are in cluster 5. Otherwise they are in cluster 3. For conditional cooperators, the match between the level of their own contributions and experiences is decisive. If both are high, they end up in cluster 5. If both are low, they end up in cluster 1. If both are in the intermediate range, they are put into cluster 2. The same logic applies to far-sghted freeriders. They are in cluster 2 with intermediate and in cluster 1 with low contributions and experiences. Yet if both contributions and experiences start high, they

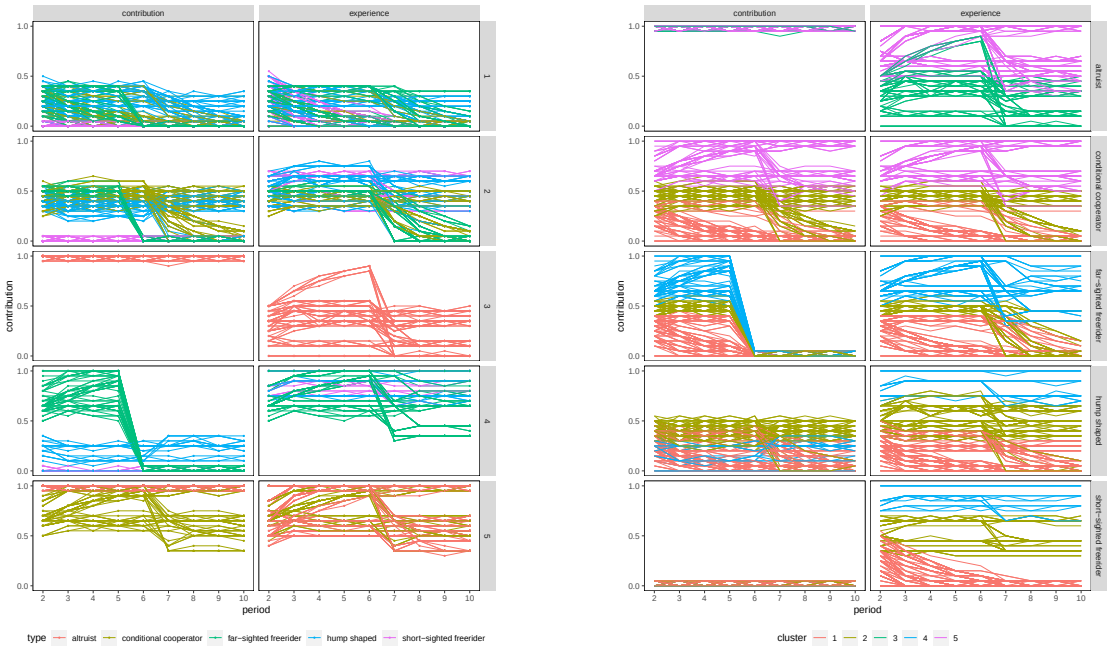
are not put in cluster 5, but in cluster 4. In principle, this is also the logic for hump-shaped players. If contributions and experiences are low, they are assigned to cluster 1. If contributions and experiences are intermediate, they are assigned to cluster 2. The only difference results from perverse reactions if experiences are too good, so that participants react with reducing their own contributions. These participants are put into cluster 4. Finally, by design short-sighted players cannot be distinguished by their own contributions. The same as hump-shaped players, they are distributed across clusters 4, 2 and 1, depending on the level of contributions by the remaining group members.

Hence upon closer scrutiny, there is not a problem with the performance of the algorithm. It just does not do what one might have naïvely expected. The object of classification is not types, but time series. Three of the types that we have simulated are reactive themselves. Unless the environment exclusively consists of short-sighted free-riders or altruists (which only holds for 2 of 70 simulated group compositions), individuals with a consistent reaction function respond to a variety of environments. If we impose 5 clusters, the algorithm must distribute pairs of experiences and choices across these clusters as best it can.

If one allows for types to be reactive, one cannot directly infer reaction functions from the data. Precisely because types are allowed to be reactive, one and the same reaction function may lead to distinctly different choice patterns. Actually just considering choice patterns would be misleading as well. One would miss the possibility that, in certain environments, multiple types exhibit very similar behaviour. In Figure 1a, the point is most forcefully illustrated by the biggest cluster, cluster 1. Since overall cooperativeness is low in these groups, the choices of conditional cooperators, hump-shaped players, far-sighted free-riders and short-sighted free-riders look very similar.

One needs an indirect strategy if one wants to infer potentially reactive types from the data. The proximate object of discovery cannot be types. It must be two-dimensional patterns, i.e. combinations of the development of experiences over time with the development of choices over time. The data can only inform the researcher about the distinct characteristics of these patterns. As the next step in the research process, she must attempt to rationalise these patterns.

In Section 5 we discuss alternative approaches for this task, and define our preferred algorithmic configuration. In Section 6 we apply this approach to the experimental data.



(a) 5 Clusters for 5 Types

(b) 5 Types in 5 Clusters

Figure 1: Clusters versus Types

5 Method

Clustering time series data Repeated experiments produce time series data. It is meaningful to relate the choices of an individual at a given point in time to the choices this individual has made at an earlier point in time, and that she will make at a later point in time. From the development of choices over time one can infer the program this individual has followed. In principle, one could capture the dependence of choices over time with the help of parameters of an appropriate transformation, and then cluster individuals with classic algorithms for static data (Liao, 2005); this is how Engel (2020) proceeds, using the coefficients of linear local (per participant) regressions as input for the classifier. This straightforward approach may well be sufficient for many practical applications. Yet the approach requires that the local regressions adequately capture the characteristics of an individual's choice program. As in this project we want to find the best way to characterize these programs, we prefer a classifier that remains open to unexpected features of the individual timeseries. This is why we work with the raw time series, and use algorithms that have been specifically developed for time series data (for overviews see Liao, 2005; Sardá-Espinosa, 2017).

Multivariate clustering Actually, many standard experiments are not only repeated. They are also interactive and produce panel data. In an interactive experiment, the program of an individual participant may react to the experiences she has made with the choices of others.

This may hold for a cognitive reason: the individual learns from others; or for a motivational reason: the individual wants to react to the choices of others. In principle, the reactive component of the individual choice program could be captured by regressing individual choices on the experiences resulting from the choices made by other group members. Yet this approach assumes that the reaction to experiences stays consistent over time. We are, of course, open to this possibility, but do not want to impose it by the design of our estimation. This is why, instead, we provide the algorithm with the exact information that participants receive in the experiment. It consists of the average choice of the remaining group members in the previous period. The algorithm thus simultaneously receives two times series: the development of the choices over time that each participant has made; and the corresponding development of the average choices made by the remaining group members in the respective previous period.

Choice of the clustering algorithm Multiple methods have been developed for clustering (raw) multivariate time series, and they all come with multiple degrees of freedom (for overviews again see Liao, 2005; Sardá-Espinosa, 2017). For our purposes, we need a clustering algorithm that is able to deal with multivariate data. In principle, this algorithm could be either hierarchical or partitional. In general, hierarchical approaches are preferable if one has reason to believe that the type space exhibits a discernible structure. This is not the case with our data, which is why we use a partitional algorithm (Hastie, Tibshirani, and Friedman, 2009, chapter 13).

Cluster evaluation The number of clusters k is a free parameter. In order to select the best k, one has to use cluster validation indices. As our clustering problem is unsupervised, we have to rely on internal cluster validation indices. The following validation indices are well-established in the literature:

- Silhouette index (`Sil`)[2]
- Dunn index (`D`)
- COP index (`COP`)
- Davies-Bouldin index (`DB`)
- modified Davies-Bouldin index (`DBstar`)
- Calinski-Harabasz index (`CH`)
- Score Function (`SF`)

These CVIs differ by the emphasis they put on cluster cohesion over cluster separation; whether they combine parameters by way of summation or division; whether or not they rely on normalization (for detail see Arbelaitz et al., 2013). As we have no strong conceptual reasons to prefer one CVI over the other, we employ all methods and aggregate over the outcomes.[3]

[2] Letters refer to the code in `R` package `dtwclust`.

[3] As the clustering algorithm has a random starting point, we repeat the comparison with 15 different starting points and use the mean index per CVI. Three of these indices (COP, DB, and DBstar) are to be minimized.

For simplicity, in the literature one picks either on or two CVIS without any specific criteria as to why or the choice is often made by majority vote. The former would seem arbitrary, yet for several of the choices that we have to make, the majority vote is inconclusive. We therefore proceed the following way: for each choice parameter in question, we rank the scores of each CVI. For each outcome, we calculate the sum over all 7 ranks. We choose the parameter that receives the highest sum of ranks.

Selection of the optimal range for k Section 4 makes it clear that we have to expect more patterns than types, and hence should estimate a number of clusters that is larger than 5. But which is the optimal number? As we know the data generating process, for the simulated data we can derive the maximum from theory. In the dataset, we have 5 types who interact in groups of 4. From (2) we know that this leads to 70 distinct group compositions. One might think that the number of environments that a player may face is smaller, as there are only 3 others in the group. Yet others are potentially themselves reactive. Then the choices the individual in question has made in the past have shaped the experiences others have made in previous periods, to which they have reacted in turn. Hence theoretically, there are 5 types $\cdot$ 70 environments $=$ 350 different patterns. Imposing that many clusters would almost surely lead to overfitting. To strike a balance between overfitting and underfitting, we proceed in two steps: In the first step, for a given dataset, we only consider any k for which the within-cluster variation ssw is $25\% \leq ssw \leq 10\%$ of the respective maximum and minimum. In Figure 2, we apply this method to the simulated data. As one sees, the range for k is within a sensible margin: considerably greater than 5, but much smaller than the upper bound of 350.

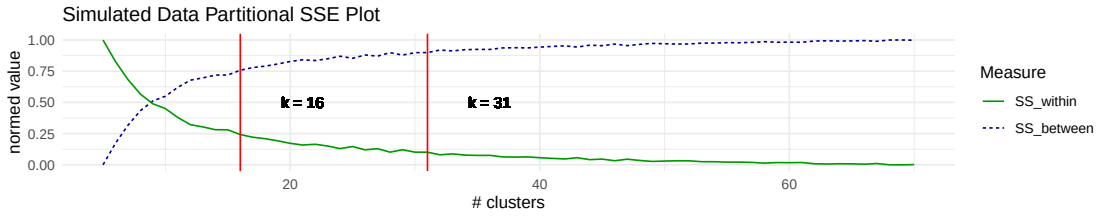


Figure 2: Simulated Data: Acceptable Range for k

Within the range thus defined, we then select the optimal k, using the cvi-ranking method introduced above.

Distance measure We further have to define the distance measure. For the clustering of time series, traditional measures like euclidean distance are inappropriate, as they would overly depend on local differences. The most popular alternative is Dynamic Time Warping DTW. It can capture similarity even if one timeseries is slightly shifted, or has a slightly different shape (Berndt and Clifford, 1994). Yet DTW is computationally costly. The procedure may occasionally even lead to pathological matches. Both concerns motivate

For comparability we invert the scores of these CVIs.

the imposition of constraints. They limit the area that can be reached by the algorithm. We consider the GAK and the "soft DTW" (sDTW) constraints (Cuturi, 2011; Cuturi and Blondel, 2017).

Centroid We finally consider two methods for defining the centroid of the respective cluster. With Partition Around Medoids (PAM) the centroid always is an existing time-series, while DTW barycenter averaging (DBA) constructs a synthetic centroid, which makes the method more robust (Petitjean, Ketterlin, and Gançarski, 2011).

Parameter search For finding the best specifications in distance measure, smoothing parameter γ when distance is sDTW as well as centroid function, we generate a grid of 4608 simulated datasets. We allow for individual specific error $\eta \in \{0.6, 0.7, 0.8, 0.9\}$, residual error $\sigma \in \{0.6, 0.7, 0.8, 0.9\}$, sample size $N \in \{1, 2, 3, 4, 5, 6\} \cdot 70$, i.e. the full set of possible type combinations, and the window within which dynamic time warping is executed $w \in \{1, 2, 3\}$. For each point in the grid, we run the clustering algorithm one time for each possible variation of parameters, i.e. 9 times.[4] The best-performing variation for each grid point is then selected according to the rank-sum method outlined in the previous section.

We have the following results: sDTW is always the preferred distance measure. Which smoothing parameter γ is optimal depends on the number of clusters k as well as the size of the dataset N. For larger numbers of clusters, i.e. $k \geq 35$ and a larger dataset, lower smoothing in the range $0.007 \leq \gamma \leq 0.085$ is preferred. For $k < 35$ and a smaller dataset, a smoothing of $\gamma = 0.01$ is preferred. For the majority of all cases, the preferred centroid function is DBA, irrespective of remaining parameters. Consequently, we use these parameters for the paritional algorithm.

6 Experimental Data

Section 4 has demonstrated in which ways, in a linear public good, a pair of two timeseries is related to the reaction function of a participant. The development of choices over time must be seen in the light of the development of experiences this participant has made. As we have explained, there is no one-to-one mapping between this two-dimensional times series and the reaction function, and hence the participant's type. Yet we have shown in which indirect ways the type can be inferred. As we expect the type space to be limited, we use clustering (of two-dimensional time series data) to organise the evidence. This gives us a methodology for the ultimate purpose of writing this paper: we want to infer from clustering real, experimental data whether the true type space differs from, or is richer than, the five types that have already been established and theorized.

Data Table 3 defines the dataset. We only use data from linear public good games without any experimental intervention, i.e. data from voluntary contribution mechanisms. We

[4] For a detailed description of all variations, please see Appendix A1

Table 3: Information on Experimental Studies Included

study	periods	endowment	group size	MPCR	subjects
Diederich, Goeschl, and Waichman (2016)	7	40	10	0.3	360
Diederich, Goeschl, and Waichman (2016)	7	40	40	0.3	200
Diederich, Goeschl, and Waichman (2016)	7	40	100	0.3	500
Diederich, Goeschl, and Waichman (2016)	7	1,000	10	0.3	50
Engel, Kube, and Kurschilgen (2020)	10	20	4	0.4	96
Nikiforakis and Normann (2008)	10	20	4	0.5	24
Engel and Rockenbach (2020)	20	20	3	0.4	30
Kosfeld, Okada, and Riedl (2009)	20	20	4	0.4	40
Kosfeld, Okada, and Riedl (2009)	20	20	4	0.6	176

have a total of 12,414 observations from 1,476 participants. Figure A.2 visually represents the dataset. On average, all experiments featured in the dataset exhibit the characteristic negative time trend. Yet there is considerable variance. The level of cooperativeness is differently high. The decay in cooperation is differently steep. In one experiment, contributions are even almost stable over time. We see this variance as an advantage. It gives us more scope for finding unknown reaction functions, in particular due to variance in the experiences participants have made.

Table 3 shows that the experimental studies exhibit unique characteristics. Simultaneously clustering the complete dataset would obscure these differences. The most critical parameters seem to be the number of rounds played, and the size of the group. Technically the difference in the number of rounds could be normalized by way of linear interpolation. Yet as we show in Figure A.3a and Figure A.3b, interpolation introduces artificial noise into time series that, otherwise, appear quite regular. To avoid such artefacts, we separately cluster the data for subsets defined by the length of the interaction and the size of the group.

Results Figure 3, Figure 4, and Figure 5 display the resulting clusters by subset. Each cluster's prototype is highlighted in bold. The individual time series in the respective cluster are represented by thin grey lines. This also informs how many pairs of time series are in the respective cluster. Comparing Figure 3, Figure 4, and Figure 5, a first result is patent: the prototypes differ profoundly between the three subsets. The typespace is not only richer than extant behavioral theory; it is also conditional on the context defined by the respective experimental protocol. The most striking difference likely results from the size of the group. While groups had size 3 or 4 in the remaining experiments, in the experiments with $t = 7$, groups had size $N = 10$, 40 or 100. Regression to the mean is the likely reason why experiences in all clusters with length 7 are nearly flat, and close to the middle of the range. By contrast, with $t = 10$ and $t = 20$, experiences exhibit much greater variance, both within and across clusters.

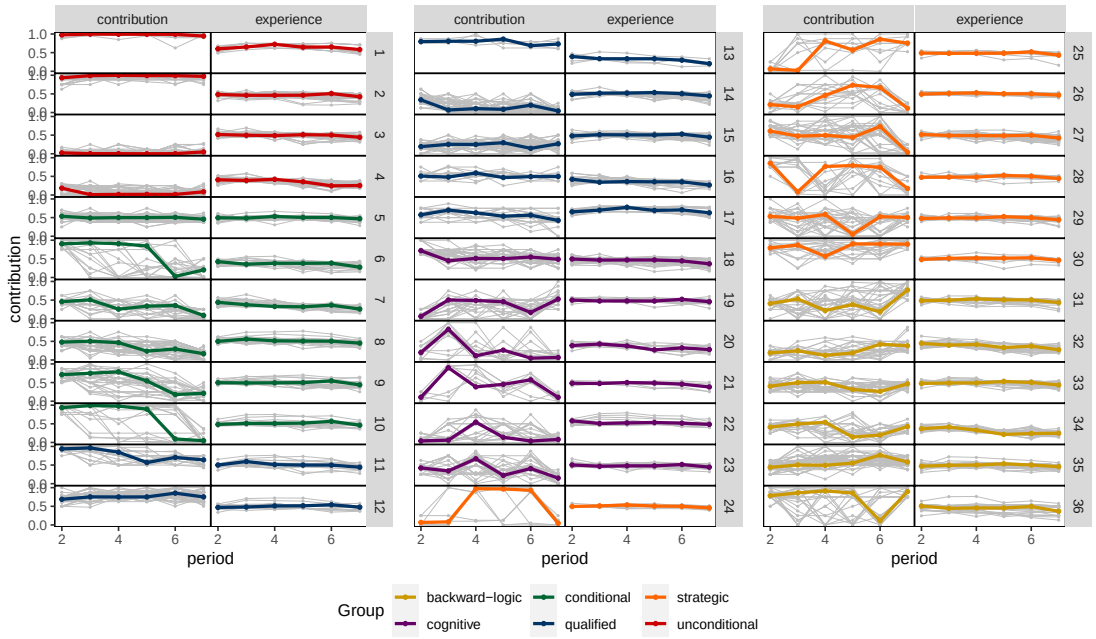


Figure 3: Cluster by Experimental Subsets, $t = 7$

Short Panel, Large Group We start interpretation with Figure 3, $t = 7$. For our purposes, the high degree of homogeneity in experiences is fortunate. We effectively see reaction patterns while holding constant that participants experience fairly homogeneous mean contributions of others. These experiences are mildly favorable: the level of contributions is in the middle of the range; they do not change much over time.

We only find few clusters with approximately unconditional choices. In cluster 1 and 2, choices are close to the top, while experiences are much lower. We can explain this pattern with unconditional altruism. In cluster 3 and 4 we find unconditional, short-sighted freeriders. Participants of this type always contribute 0, or (in cluster 4) close to it.

In contrast, strict conditional cooperators should precisely match the experiences they have made (provided they expect others to behave the next period the same way as in the present period). Cluster 5 can be rationalized this way. As experiences are so consistently close to the midpoint, there is no room for the behavior theorized as "hump shaped". The only other previously theorized type that can be traced in the data is farsighted free-riders. This type invests in cooperation in early periods, and exploits others in later periods. This holds for clusters 6 to 10.

The remaining 26 clusters are hard or even impossible to rationalise with the behavioural programs hitherto discussed in the literature. Clusters 11-17 could be interpreted as qualified versions of known types. In clusters 11–13, contributions are substantially above experiences, but not at the top. This could be the behaviour of an altruist who is, however, not willing to be completely blind to the choices of others. Clusters 14 and 15 are the mirror image

at the low end. Contributions are not immediately and not completely at zero, but always below experiences. Finally clusters 16 and 17 are imperfect cases of conditional cooperation. In cluster 16, contributions are consistently slightly above experiences, while they are consistently slightly below experiences in cluster 17.

Another potential behavioral program is of a cognitive nature. Participants are surprised by experiences and adjust their choices to the behavioural environment. This explanation is most intuitive in early periods, and if the participant aligns her own choices with experiences. Clusters 18 and 19 closely fit this explanation. The participant had been either overly optimistic or overly pessimistic about the level of contributions. The remaining clusters with pronounced changes in the initial periods (clusters 19, 20, 21, 22, 23) require a more involved behavioural program. A consistent interpretation would be exploration. Exploration is reasonable if the participant in question does not only herself have a reactive choice program, but considers the possibility that others have reactive programs as well. In that case she needs to test the waters and find out what is going to happen if she changes her own moves. The participant deliberately risks falling below the attainable period income as an investment into more profitable moves in the future. This explanation is particularly plausible for changes in early periods.

A participant who engages in (potentially) costly exploration can be said to act strategically. But in this interpretation the strategy is confined to making a better informed decision herself in future periods. The choice patterns in clusters 24–30 suggest a more encompassing strategic motive. The participant in question does not only aim at optimising her own future choices. She intends to induce other group members to behave in a way she considers more appropriate. In cluster 24, the participant seems to try leading by example. In clusters 25–27, the participant also, at least in some periods, contributes more than the group average. This could be motivated by the aim of signalling good intentions and the possibility of a brighter future to the group. In cluster 28, the participant might want to combine a warning what could happen if others don't follow suit with a positive signal later on. In clusters 29 and 30 only negative signals can be found.

In the final group of six clusters, participants increase contributions in the final or the penultimate period. We thus find an inverse endgame effect. As the game has a defined end, such choices cannot be motivated strategically. Participants must have deontological motives. If they had contributed less than average in earlier periods, a consistent interpretation is repent, leading to (at least partial) compensation. In cluster 36 the opposite interpretation as punishment invites itself. In the remaining clusters 33–35, contributions had been at or even slightly above the group average. At some point contributions go down, but go up again. This pattern would be consistent with an expression of discontent.

Short Panel, Small Group 10 periods are just three more than 7, yet the patterns in the clusters for $t = 10$ – displayed in Figure 4 – look very different from the ones for $t = 7$.[5] The

[5] The total number of observations with $t = 10$ is much smaller than with $t = 7$. If we keep the upper limit of k at the level derived from the theoretical number of type combinations, we get too many clusters with very few observations. We therefore adjust the upper limit to $N/4 = 30$. With the help of the ranksum

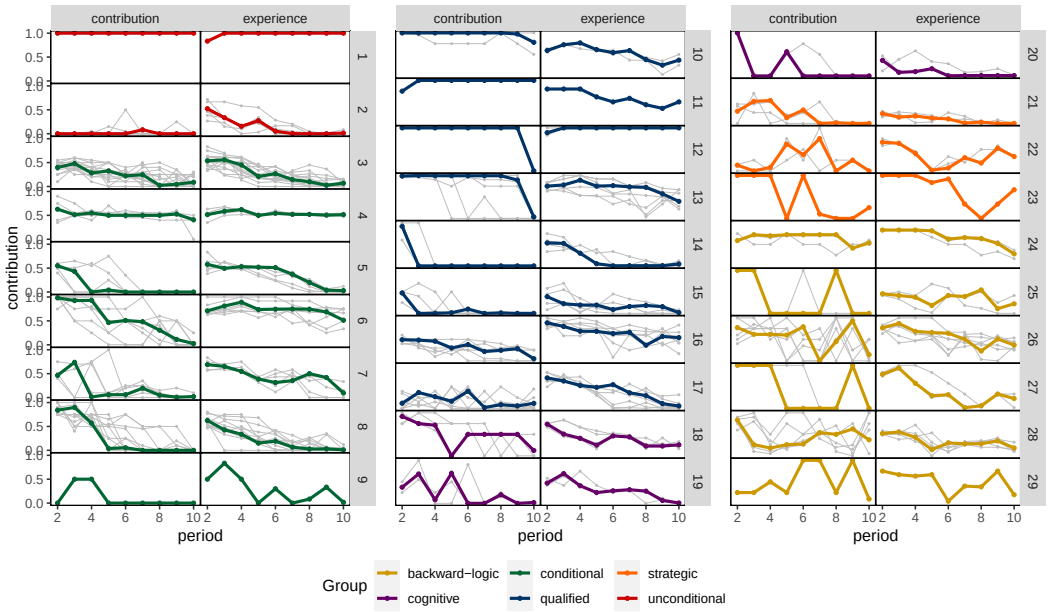


Figure 4: Cluster by Experimental Subsets, $t = 10$

obvious source of the difference is in the experiences participants are making. While in bigger groups, regression to the mean conceals variance in the types of other group members, this variance may play itself out in the groups of 4 from which these data are taken. In many clusters we also observe the downward trend that has often been reported in the literature on public goods (see e.g, clusters 3, 7, 8, and 17).

In this set of experiments, we find choice patterns consistent with unconditional altruism (cluster 1) and unconditional free riding (cluster 2). But there are only very few observations in these two clusters. Again, not many clusters are such that contributions track experiences (clusters 3 and 4), which is what would be expected from a textbook conditional cooperator. A somewhat larger number of clusters is consistent with far-sighted free riding, i.e. making reasonably high contributions in early periods, in the interest of cashing in at a later point (clusters 5, 6, 7, 8, and 9).

A further set of clusters are at best qualified versions of the previously theorised types. In clusters 10–13, in most periods contributions are at the top. But unconditional altruists would neither need a period to go to the top (had the participant initially been concerned about the degree of exploitation?), nor go down in the final period. Likewise, in clusters 14 and 15, contributions are low or even zero in most periods, but not in the beginning. Have these been far-sighted free-riders who do not consider investment in the corporation spirit worth the while, given what they experience in the first period? Finally in clusters 16 and 17 choices grosso modo track experiences, but at a lower level. Are these conditional cooperators intending to at least slightly outperform the group?

over all internal CVIs, we end up with the meaningful number of 29 clusters, depicted in Figure 4.

In clusters 18–20, we observe stark changes in early periods, either downwards (cluster 18), upwards (cluster 20), or both (cluster 19). These choice patterns could be motivated by exploration.

While exploration is also a possible interpretation in clusters 21–23, these patterns could also be motivated by the intention to educate the group, and thereby improve the outcome for all.

Finally in clusters 24–29, we see upward moves in the final or the penultimate periods, i.e. an inverse endgame effect. As explained with $t = 7$, one needs deontological motives to rationalise such choice patterns. In all clusters at least for some periods the participant had contributed less than the average in the previous period, and might feel morally urged to at least partly give back to the group.

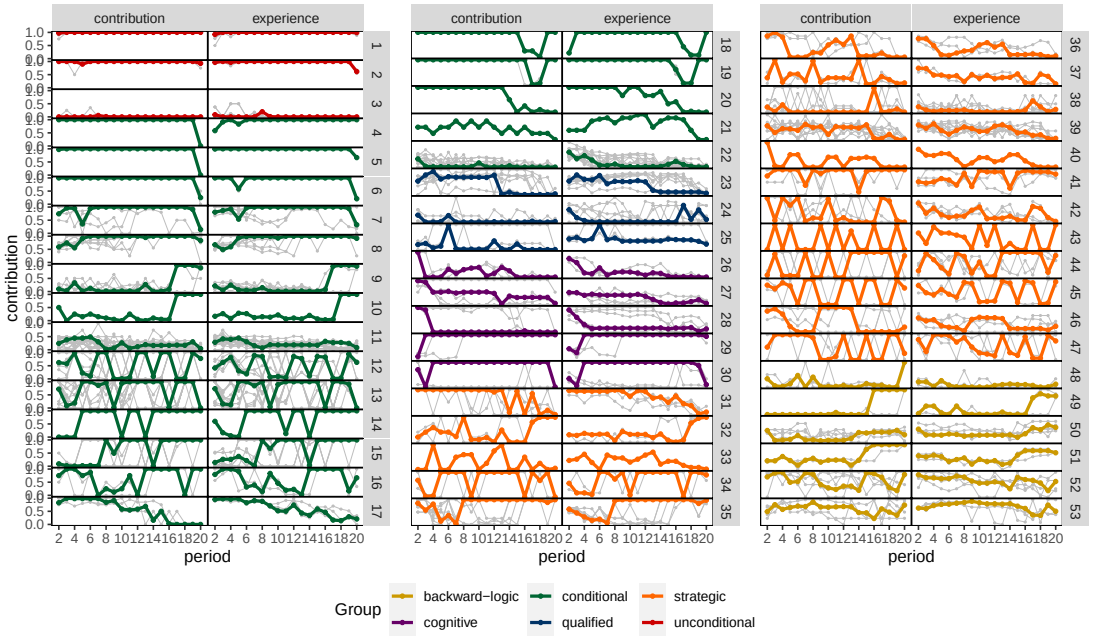


Figure 5: Cluster by Experimental Subsets, $t = 20$

Long Panel, Small Group The final set of observations is displayed by Figure 5 and comes from experiments with $t = 20$, and $N = 3$ or $N = 4$. Hence in these experiments, participants stay together for a long time and therefore have many opportunities for learning from each other, and for attempting to influence each other. Both learning and influencing is meaningful as there are only few interaction partners. The average contribution is a reasonable proxy for the composition of the type space. As Figure 5 shows, the different institutional setting again matters profoundly. In this context, experiences exhibit considerable variance. Choices quite often track experiences fairly closely.

Using the same set of cognitive and motivational effects as with $t = 7$ and $t = 10$, we can

organise the type space. Yet it is composed very differently. We do find a very small number of arguably unconditional altruists (clusters 1 and 2) and unconditional free riders (cluster 3). We also find clusters in which participants track the generally high (clusters 4–8) or low contributions of the remaining group members (clusters 9–11). The fact that there is at least some variance in experiences mirrored in variance in choices demonstrates that these participants are not qualified unconditional types, but react to what they observe.

The most characteristic difference between $t = 7$ and also $t = 10$ on the one hand, and $t = 20$ on the other hand is the zig-zaggy pattern of both experiences and choices in many of the clusters. Apparently the longer time horizon has not helped, but hurt. Multiple opportunities for observing each other have made it difficult for many groups to get in sync. They have oscillated between low and high contributions, although there is no point in taking turns in a linear public good, which is a cooperation, not a coordination problem.

In clusters 12–17 choices and experiences match so closely that behaviour fits classic conditional cooperation. By contrast, in clusters 18–22 it is the participant who triggers the download trend of the group, which suggests that these participants are farsighted free-riders.

In clusters 23–25, choices slightly deviate from experiences, but no straightforward alternative interpretation invites itself. This is why we have classified these clusters as instances of qualified conditional cooperation.

With $t = 20$, we do not find many clusters where choices can be rationalised by a merely cognitive effect. In clusters 26–28, participants might have been too optimistic initially. In cluster 29 they might initially have been too pessimistic. Cluster 30 is the only cluster that suggests exploration in the first two periods.

In the bulk of not only clusters, but also data, participants seem to aim at inducing a higher contribution level in their groups. They either periodically go up, if not to the top. Or they signal that others should not take their benevolence for granted, by temporarily reducing contributions, frequently to 0 (clusters 31–47).

Finally we again observe five clusters with an inverse endgame effect (clusters 48–53). In cluster 50, contributions had clearly been below experiences for many periods. In this cluster, compensating the other group members for anti-social behaviour is a consistent explanation. This explanation might also matter in clusters 52–53. By contrast choices in clusters 48 and 49 more look like an expressive act, showing others how much more favourable outcomes might have been, had they been less selfish.

7 Rationalisation

Section 6 has shown pronounced heterogeneity. It is beyond the scope of the present paper to theorize all the many patterns that we observe. The fact that we find discernible clusters gives us confidence that these patterns are meaningful. In the previous section we have offered plausible interpretations. But these interpretations are, of course, only hypotheses. New experiments will be needed to isolate potential cognitive and motivational channels. In this section, by way of illustration, we zero in on one striking difference: with short panels

and large groups (Figure 3) and with slightly longer panels and small groups (Figure 4), individual and group patterns are much smoother than with longer panels and small groups (Figure 5): in the data from games repeated for 20 announced periods, we find a lot more zigzagging. Obviously the longer shadow of the future has a strong behavioral effect. In this section we discuss an explanation for the observed sudden, drastic changes in behavior.

In a linear public good, payoff is given by (1). In the stage game, a participant maximizes profit by complete freeriding, i.e. by setting $c_i = 0$. As the linear public good is a prisoner's dilemma, $c_i > 0$ is dominated. If a selfish player assumes that all other group members are selfish as well, i.e. when assuming common knowledge of rationality, $c_{i,t} = 0, \forall t$ is the best response. This is the well known unraveling prediction. As participant i expects all other participants j to choose $c_{j,t=T} = 0$ in the final period T, there is no scope for investing in cooperation in early rounds.

In their seminal paper, Kreps et al. (1982) have shown that this result breaks down when allowing for behavioral uncertainty. They show this for the belief that the counterpart (in a 2x2 prisoner's dilemma with discrete {cooperate, defect} action space) might either play tit for tat, or might be a conditional cooperator. Yet in their model, a longer shadow of the future is unequivocally beneficial. If T is large enough, in early rounds a rational player never defects. She would reveal that she is actually selfish. At best, her counterpart plays tit for tat, and punishes her for her deviation from the cooperative path of the group. If her counterpart is selfish as well, they end up in the the {defect, defect} equilibrium for all future rounds. We, by contrast, find the opposite. With a longer shadow of the future, there are more deviations. We also find upward, not only downward deviations. We thus need a different model to rationalize this observation.

In the initial step, we only allow for genuinely cooperative players. For this type $u_i(c_i < \bar{c}_j) < u_i(c_i = \bar{c}_j)$, where c_i is the contribution of player i to the public good, and $\bar{c}_j$ is the average contribution of the remaining group members to the public good. We work with this average as, in the experiments from which we have data, participants did not get feedback about individual contributions of the remaining members of their groups.

In principle, such a group is able to sustain cooperation. Yet no player wants to be the sucker: $u_i(c_i = \bar{c}_j) > u_i(c_i > \bar{c}_j)$. When choosing how much to contribute in period t, participants do not know how much the remaining group members are going to contribute. They must work with the expectation $E(\bar{c}_{j,t})$. In later periods, they have a signal: $E(\bar{c}_{j,t}) \approx \bar{c}_{j,t-1}$. But in the initial period, they must work with their home-grown beliefs. Cooperation may fail. Not because other group members are genuinely selfish, but because at least some of them have been too sceptical initially.

Against this backdrop, it can be rationalized that a player sets $c_{i,t} > \bar{c}_{j,t-1}$: she uses this to signal her type. The signal is credible, as a player who is not willing to sustain cooperation has no reason to do that. The signal can be interpreted as an investment. The participant accepts to be temporarily exploited, in the interest of lifting the entire group to a higher contribution level. Such an investment is the more profitable the longer the shadow of the future. This explains why such choices are more frequent in games with a larger number of periods.

The fact that $\bar{c}_{j,t=1} < e$, where e is the endowment, and hence the maximum contribution, may have more than one reason. Either all other group members were indeed conditionally cooperative, but too sceptical; or at least one of them was actually (short-sightedly) selfish. If the participant has invested into signalling her cooperative type in period $t = 2$, she must wait another period to learn. If $\bar{c}_{j,t=3} = e$, the group has coordinated at the maximum. In this logic, participant i not only sets $c_{i,t=2} = e$, but also $c_{i,t=3} = e$. She thus gives the other group members a chance to adapt to the cooperation signal she has sent in $t = 2$, and only reverts to $c_i < e$ in period $t = 4$ if $\bar{c}_{j,t=3}$ has proven that her strategy did not work out as the group is actually not cooperative.

While this strategy is consistent it puts a high burden on the group member who attempts to trigger the virtuous cycle. If she was too optimistic about the composition of the type space, she has to accept exploitation for two consecutive periods. Now her strategy is motivated by the possibility that an all-cooperative group is stuck in a bad equilibrium. The fact that she signals her cooperative type in $t = 2$ can be interpreted as the contribution to a second order public good Yamagishi (1986) Heckathorn (1989): one of the cooperative players must accept temporary exploitation for signalling her type. This interpretation provides scope for an alternative strategy. The group member who has made the initial move, in $t = 2$, expects other group members to follow suit in $t = 3$. Hence her strategy would be

- $c_{i,t=2} = e$
- $c_{i,t=3} = \bar{c}_{j,t=1}$
- $c_{it>3} = e|\bar{c}_{j,t=3} = e$; $\bar{c}_{j,t=1}$ otherwise

Hence this group member expects to be compensated in $t = 3$ by the remaining group members for her initial cooperative move, by them tolerating that she reduces her contribution to the original contribution level in the group, or even to 0. If cooperative participants use this strategy, we should see one of two patterns. In groups that are actually non-cooperative, we should see one period jumps to the maximum that have no consequences at the group level. In groups that are actually cooperative, we should see one period jumps to the maximum, followed by a jump downwards, followed by coordination at a high level. Hence we should observe zigzagging.

The same strategy does also work if the cooperative group member who takes the initiative is less optimistic about the composition of the type space. She may be open to the possibility that one or more of the remaining group members are actually selfish, but willing to sustain cooperation, as they expect the long-term profit from this strategy to be higher than early defection. Such players mimic genuinely cooperative players.

The player who takes the initiative may also be willing to tolerate partial defection. This is particularly plausible in the canonical design of the game, with 4 group members and MPCR = .4. Then 3 group members who cooperate fully still have a slightly higher payoff than from defection (24, rather than 20). Tolerating partial defection is also easier in the design of the game investigated in this paper as participants only get feedback at the group level. They can therefore not see whether $\bar{c}_j < e$ is due to one player defecting, or all other players contributing less than the maximum, but the same amount.

Yet if either possibility is taken into account, observing $\bar{c}_{j,t} > \bar{c}_{j,t-1}$ is less informative. Per se, genuinely cooperative players have no reason to revert to lower contributions in later rounds. By contrast a player who only mimics a cooperative type will start defecting once the expected payoff from defecting before others outweighs the gains from cooperation for future rounds. If other genuinely cooperative players are concerned about this possibility, they may themselves reduce contributions, as they loose faith in the willingness of others to cooperate. This concern looms even larger if cooperative gains are below maximum, as then gains from continuing to cooperate are smaller.

Taking these possibilities into account, we can also rationalize upward jumps in later rounds. By the same logic as in $t = 2$, a (genuinely or strategically) cooperative type wants to stabilize cooperation, by sending this cooperative signal. Again the longer the time horizon, the more this strategy is profitable. And again, such a cooperative player may expect to be compensated in the subsequent period, by others tolerating her one-period defection. So again zigzagging can be rationalized.

Once we allow for the belief that groups are heterogeneous, in the defined sense, we can also rationalize temporaneous downward jumps. This is straightforward if a player only mimics a cooperative type: she tests the waters. If others react she knows that they are vigilant, so that (early) defection does not pay. Yet a temporaneous downward jump can also be rational for a genuinely cooperative player who is sceptical about the motives of other cooperators. If they react by reducing their contributions, she knows that their cooperation is not genuine, and she can react by reverting to low contributions herself.

8 Discussion

The linear public good is one of the workhorses of behavioural economics. Hundreds of experiments have been run with this paradigm. The design is appealing as, in a stylised way, it captures what arguably is the essence of many conflicts of life, running from the degradation of the environment over the instability of a cartel to the precarious nature of any constraining institutional framework. The design implements a multi-person, multi-period prisoners dilemma with a known end. If one assumes that actors exclusively maximise individual profit, the repeated game has a unique solution. In the final period, all group members will contribute nothing to the common project. Through unraveling, this is also the prediction for any earlier period.

The first experiments undertaken with this design have already refuted this prediction. On average, contributions start at some higher level but decay over time. Per se, social preferences can rationalise positive contributions, but they do not predict the decay. Interestingly, per se the prominent concept of conditional cooperation cannot predict the decay either. If all group members are perfect conditional cooperators, and expect all others to follow the same behavioural program, any level of cooperation can be sustained, depending on initial beliefs. Fischbacher and Gachter (2010) propose a consistent explanation: the decay could result from conditional cooperation being imperfect. Participants would be willing to let themselves be guided by the level of cooperativeness in their group. But they would always

try to slightly undercut. Yet in their reanalysis of Fischbacher's and Gächter's data, Engel and Rockenbach (2020) have shown that true conditional cooperation is actually near-perfect. The decay results from heterogeneity. By the combination of choice data with belief data, they show that the decay results from the presence of short-and far-sighted free-riders. This is where the present project starts. It uses machine learning methods to cast light on this heterogeneity, and chart the type space.

The paper makes a methodological and a substantive contribution. On the methodology side, it shows in which ways clustering can be used to infer the composition of the type space. On the substantive side it shows that extant theories about behavioural types can only explain a very narrow fraction of the data.

Repeated experiments generate time series data. In principle, the large family of algorithms for clustering time series data are therefore appropriate. Yet contributions could not exhibit a downward trend unless at least some participants hold a choice program that is reactive. If we were to deprive the algorithm of the experiences participants make, it would lump together choice patterns that are generated by completely different behavioural programs. This is why we use multivariate clustering and feed the algorithm with pairs of experiences and choices.

One might naïvely think that the algorithm will find as many clusters as there are distinct behavioural programs. With simulation, we show why this approach must fail. We simulate all combinations of five behavioral programs that have been theorized in the literature: altruists, conditional cooperators, far-sighted freeriders, hump-shaped contributors, and short-sighted freeriders. For investigating these five behavioural programs, we need many more clusters. Yet we also show that we do not need the theoretical maximum of 350 clusters; this would make the approach next to unusable for real data, as one would need a huge amount of data for that many clusters to be credible. We use internal cluster validation indices to find the appropriate trade-off between underusing and overusing the evidence.

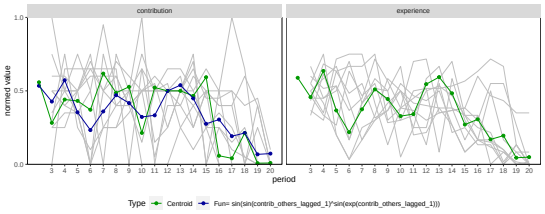
We apply this methodology to a large dataset consisting of 12,414 observations. Results clearly show that the true type space is much richer than thus far assumed by the literature. Only a few of the clusters that we find in the experimental data can be rationalised with any of the five theoretical behavioural programs that we have used to simulate data. Obviously, the type space is considerably richer than typically assumed in the behavioral literature.

The main limitation of our approach is its exploratory nature. We alert the research community that the behavioural programs participants employ in an experiment seemingly as simple as the linear public good are very likely much richer, and much more heterogeneous, than typically assumed when designing and analysing these experiments. This can obviously only be a first step. In the next step, frequent choice programs must be rationalised. In conclusion, we sketch two approaches that could productively prove complementary.

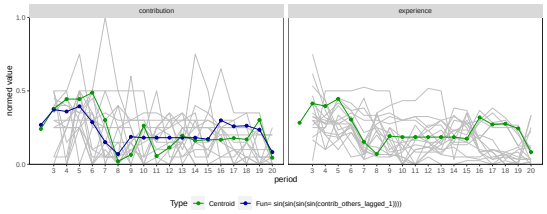
The first approach capitalizes on methods originating in physics. In physics (Udrescu and Tegmark, 2020), same as in industry (Francone et al., 1999; Castillo et al., 2002), problems are often too complex to start the analysis from first principles. Rather one begins with the data and searches for succinct ways of rationalising them. The approach is known as symbolic regression. Its main advantage is flexibility. Different from maximum likelihood, one

need not impose functional form. Given (weaker) constraints on depth, a set of pre-defined functions from which the procedure may choose, as well as operators, symbolic regression aims to find the best combination of these inputs and the data, resulting in a functional form with the least error to the original values. Technically, symbolic regression is a supervised technique.[6]

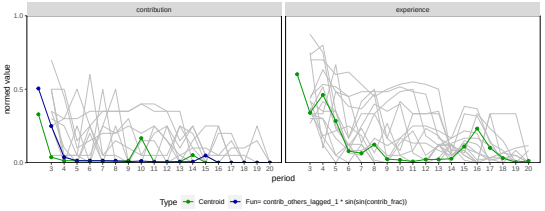
In the spirit of a proof of concept, in Figure 6 we show the resulting rationalisations for three selected clusters (from experiments with $t = 20$). As can be seen, the approximation works well. But the functional form that happens to fit the data not only differs widely across clusters; this could be necessary for defining the character of the heterogeneity. The functional form does also not easily lend itself to interpretation. If the collective endeavour of charting the type space progresses, one might be able to add meaningful explanations. But it might also turn out that symbolic regression is only good at prediction, not at explanation.



(a) Cluster 11



(b) Cluster 17



(c) Cluster 42

Figure 6: Exemplary Results Symbolic Regression

The alternative approach is more in the spirit of behavioral economics. In this perspective, one would read the evidence presented in this paper as a challenge for the development, and subsequent experimental testing, of more appropriate behavioral theory. One would aim at replacing the "as if" models from symbolic regression with process models. The interpretations proposed in section 6 offer building blocks for this enterprise. We conclude by explicating the ones that we deem most promising.

Unless a participant has an unconditional behavioural program, in the initial period she must work with beliefs about the behavioural programs that other group members are implementing. Participants might conceptually be allowed to condition their own program on the direction and the amount by which these beliefs turn out to be false. Additionally, participants might conceptually be allowed to invest in exploring the behavioural programs of

[6]We make use of the R-package gramEvol (Noorian, de Silva, and Leong, 2016) which uses genetic programming (Kotanchek, Smits, and Kordon, 2003) to efficiently traverse the space of possible solutions.

the rest of their groups.

On the motivational side, one may enrich the concept of conditional cooperation. Participants would not blindly copy the experiences they are making. They would rather consider the possibility to influence the choices of other group members in future periods by their own choices in the present period. This could be theorized as costly signaling of type, including the intention to enforce a normative conviction in the future.

Besides such a forward looking, strategic perspective, there might also be backward looking motives. Participants might want to express their discontent and frustration, even if this reduces their own prospect for a higher profit. Or, conversely, they might want to reward others for unselfish acts. They might also, relatedly, repent their own past behaviour, and aim at partial reparation.

Sometimes, the next step forward in a line of research is not the final answer, but the right question. With the present project, we aim at demonstrating that heterogeneity in dynamic games, and in the linear public good, in particular, is a promising frontier. This investigation is urgent if one hopes to learn from experimental data about the behavioural determinants of social dilemmas, in the interest of designing more powerful interventions.

References

Amin, Engi, Mohamed Abouelela, and Amal Soliman (2018). "The Role of Heterogeneity and the Dynamics of Voluntary Contributions to Public Goods: An Experimental and Agent-Based Simulation Analysis". In: *Journal of Artificial Societies and Social Simulation* 21.1.

Andreoni, James (1995). "Cooperation in Public-Goods Experiments: Kindness or Confusion?" In: *American Economic Review*, pp. 891–904.

Arbelaitz, Olatz et al. (2013). "An Extensive Comparative Study of Cluster Validity Indices". In: *Pattern Recognition* 46.1, pp. 243–256.

Arifovic, Jasmina and John Ledyard (2012). "Individual Evolutionary Learning, Other-Regarding Preferences, and the Voluntary Contributions Mechanism". In: *Journal of Public Economics* 96.9-10, pp. 808–823.

Bapna, Ravi et al. (2004). "User Heterogeneity and its Impact on Electronic Auction Market Design: An Empirical Exploration". In: *MIS Quarterly*, pp. 21–43.

Berndt, Donald J. and James Clifford (1994). "Using Dynamic Time Warping to Find Patterns in Time Series." In: *KDD workshop*. Vol. 10. 16. Seattle, WA, USA: pp. 359–370.

Castillo, Flor A et al. (2002). "Symbolic regression in design of experiments: A case study with linearizing transformations". In: *Proceedings of the 4th Annual Conference on Genetic and Evolutionary Computation*, pp. 1043–1047.

Chaudhuri, Ananish (2011). "Sustaining Cooperation in Laboratory Public Goods Experiments: a Selective Survey of the Literature". In: *Experimental Economics* 14.1, pp. 47–83.

Cuturi, Marco (2011). "Fast Global Alignment kernels". In: *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 929–936.

Cuturi, Marco and Mathieu Blondel (2017). "Soft-DTW: a Differentiable Loss Function for Time-Series". In: *arXiv preprint arXiv:1703.01541*.

Diederich, Johannes, Timo Goeschl, and Israel Waichman (2016). "Group Size and the (In)Efficiency of Pure Public Good Provision". In: *European Economic Review* 85, pp. 272–287.

Engel, Christoph (2020). "Estimating Heterogeneous Reactions to Experimental Treatments". Working paper.

Engel, Christoph, Sebastian Kube, and Michael Kurschilgen (2020). "Managing Expectations: How Selective Information Affects Cooperation". Working paper.

Engel, Christoph and Bettina Rockenbach (2020). "What Makes Cooperation Precarious?" Working paper.

Fehr, Ernst and Klaus Schmidt (2002). "Theories of Fairness and Reciprocity. Evidence and Economic Applications". In: *Advances in Economics and Econometrics. 8th World Congress*. Ed. by Mathias Dewatripont and Stephen J. Turnovsky. Cambridge: Cambridge University Press, pp. 208–257.

Ficici, Sevan G., David C. Parkes, and Avi Pfeffer (2012). "Learning and Solving Many-Player Games Through a Cluster-Based Representation". In: *arXiv preprint arXiv:1206.3253*.

Fischbacher, Urs and Simon Gachter (2010). "Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Goods Experiments". In: *American Economic Review* 100.1, pp. 541–56.

Fischbacher, Urs, Simon Gachter, and Ernst Fehr (2001). "Are People Conditionally Cooperative? Evidence from a Public Goods Experiment". In: *Economics Letters* 71.3, pp. 397–404.

Francone, Frank D et al. (1999). "Homologous Crossover in Genetic Programming." In: *GECCO*. Citeseer, pp. 1021–1026.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

Heckathorn, Douglas D (1989). "Collective action and the second-order free-rider problem". In: *Rationality and Society* 1.1, pp. 78–100.

Isaac, R. Mark and James M. Walker (1988). "Group Size Effects in Public Goods Provision: The Voluntary Contributions Mechanism". In: *Quarterly Journal of Economics* 103.1, pp. 179–199.

Kosfeld, Michael, Akira Okada, and Arno Riedl (2009). "Institution Formation in Public Goods Games". In: *American Economic Review* 99.4, pp. 1335–55.

Kotanchek, Mark, Guido Smits, and Arthur Kordon (2003). "Industrial strength genetic programming". In: *Genetic programming theory and practice*. Springer, pp. 239–255.

Kreps, David M et al. (1982). "Rational cooperation in the finitely repeated prisoners' dilemma". In: *Journal of Economic Theory* 27.2, pp. 245–252.

Ledyard, John O (1995). "Public Goods: A Survey of Experimental Research". In: *Handbook of Experimental Economics*. Ed. by John Kagel and Al Roth. Princeton: Princeton University Press, pp. 111–194.

Liao, T. Warren (2005). "Clustering of Time Series Data – a Survey". In: *Pattern Recognition* 38.11, pp. 1857–1874.

Lu, Yixin et al. (2016). "Exploring Bidder Heterogeneity in Multichannel Sequential B2B Auctions". In: *MIS Quarterly* 40.3, pp. 645–662.

Lucas, Pablo, Angela de Oliveira, and Sheheryar Banuri (2012). "The Effects of Group Composition and Social Preference Heterogeneity in a Public Goods Game: An Agent-Based Simulation". In: *Journal of Artificial Societies and Social Simulation* 17.3, pp. 148–174.

Mao, Andrew et al. (2017). "Resilient cooperators stabilize long-run cooperation in the finitely repeated Prisoner's Dilemma". In: *Nature communications* 8.1, pp. 1–10.

Nikiforakis, Nikos and Hans-Theo Normann (2008). "A Comparative Statics Analysis of Punishment in Public-Good Experiments". In: *Experimental Economics* 11.4, pp. 358–369.

Noorian, Farzad, Anthony M. de Silva, and Philip H. W. Leong (2016). "gramEvol: Grammatical Evolution in R". In: *Journal of Statistical Software* 71.1, pp. 1–26.

Petitjean, François, Alain Ketterlin, and Pierre Gançarski (2011). "A Global Averaging Method for Dynamic Time Warping, with Applications to Clustering". In: *Pattern Recognition* 44.3, pp. 678–693.

Sardá-Espinosa, Alexis (2017). "Comparing Time-Series Clustering Algorithms in R using the dtwclust Package". In: *R Package Vignette* 12, p. 41.

Udrescu, Silviu-Marian and Max Tegmark (2020). "AI Feynman: A physics-inspired method for symbolic regression". In: *Science Advances* 6.16, eaay2631.

Vorobeychik, Yevgeniy, Michael P. Wellman, and Satinder Singh (2007). "Learning Payoff Functions in Infinite Games". In: *Machine Learning* 67.1-2, pp. 145–168.

Yamagishi, Toshio (1986). "The provision of a sanctioning system as a public good". In: *Journal of Personality and Social Psychology* 51.1, pp. 110–116.

Zelmer, Jennifer (2003). "Linear Public Goods Experiments: A Meta-Analysis". In: *Experimental Economics* 6.3, pp. 299–310.

Appendix

A1 Data Generating Process, Simulated Data, and Grid Search

For our grid-search, we assumed the data-generating process of the standard public good game as laid out in the main section of the paper.

Simulated Type Space In their seminal paper, Fischbacher, Gachter, and Fehr (2001) argue that (in their one-shot version of this game) there are three types: free-riders, conditional cooperators, and "hump-shaped" players. In his reanalysis of Fischbacher and Gachter (2010), Engel (2020) further finds a small, but discernible fraction of altruists. In their reanalysis of Fischbacher and Gachter (2010), Engel and Rockenbach (2020) use a combination of belief and choice data to distinguish a fifth group, which they call far-sighted freeriders. In our simulations, we allow for these five types. We focus on a partner design. Groups stay together for the full duration of the game. We always allow for an individual random effect η_i and residual error $\epsilon_{it} \perp \eta_i$, which we both define to be normally distributed with mean 0 and standard deviation .3 ($\sim \mathcal{N}(0, .3)$). We thus implement the type space as defined in Table 1, where $c_{-i,t-1}$ is the average contribution of the remaining group members in the previous period.

We have two types that exhibit variance (between participants due to η_i, and within participants due to ϵ_{it}), but do not react to experiences: short-sighted freeriders and altruists. The contributions of these types do also not have a trend. They are random walks, albeit with diametrically opposed starting points. By contrast, the remaining three types are reactive, which may, depending on the choices of the remaining group members $c_{-i,t-1}$, lead to a trend. We have (true) conditional cooperators start in the middle of the action space. In early periods ($t < \tau = 5$), far-sighted freeriders mimic conditional cooperators, but from period τ on, they freeride. Such participants "feed the cow" for a while, to then "start milking" it. Finally, we simulate hump-shaped participants such that they start rather low, at 5, and have them behave like conditional cooperators as long as the remaining group members, in the previous period, have on average not contributed more than half of the endowment. If $c_{-i,t-1} > 10$, they exhibit a perverse reaction. The more others have contributed, the less they contribute themselves.

We have groups of size $K = 4$, and we allow for $n = 5$ types. Participants choose their contributions to the public good simultaneously, which is why their order does not matter. We consider the possibility that types are present more than once in a group. Hence we have a problem of unordered sampling with replacement. This gives us a total type space of

different group combinations. In our simulations, we include each of these 70 combinations of types N times. As three of the five types (conditional cooperators, far-sighted freeriders, hump shaped players) are reactive, we give the classification algorithm access to the exact same experiences that participants make in this design, i.e. the mean contribution of the remaining group members in the previous period. Hence the object of clustering is a two-dimensional time series.

- Profit is given by (1), with $e = 20, K = 4, t = 10$. We do, however, only use data from periods 2 - 10 for analysis, as participants have not made any experiences in the first period.
- Each of the five types is represented in equal proportions in the population from which choices are drawn.

The following elements of the the data generating process as well as of the clustering process are kept fixed for each point in the grid:

- Individual specific error $\eta \in \{0.6, 0.7, 0.8, 0.9\}$ and residual error $\sigma \in \{0.6, 0.7, 0.8, 0.9\}$.
- Sample size $n = 70 * N | N \in \{1, 2, 3, 4, 5, 6\}$
- The number of clusters $k \in \{5, 8, 10, 15, 20, 25, 30, 35, 40, 42, 44, 46, 48, 50, 52, 54\}$
- The size of the window within which dynamic time warping is executed, running from $w \in \{1, 2, 3\}$.

The grid search thus runs over $4\eta \cdot 4\sigma \cdot 6N \cdot 3w \cdot 3\gamma \cdot 16k = 4608$ variations. The following parameters of the algorithm are varied:

- The distance used, i.e., DTW, sDTW, GAK
- The smoothing parameter $\gamma \in \{0.001, 0.01, 0.1\}$.
- The centroid function, i.e., either PAM or DBA

Each point in the dataset point in the grid is thus clustered with ten variations in the configuration of the clustering algorithm. In the simulated data set, we have 1120 pairs of time-series for experiences and individual choices. If it were necessary to estimate 350 clusters, there would be little more than 3 participants per cluster, on average. The simulated dataset would be too small for the purpose. More disturbingly, it would be very difficult to compile a set of experimental data that is big enough for the ultimate goal of this study: to find out whether there are untheorized types.

In the interest of finding the appropriate number of clusters, and of better understanding the relationship between the degree of smoothing and the number of clusters that best organize the data, we evaluate the simulated dataset.

35 clusters for 5 types A comparison between Figure 1b and Figure A.1 shows how important it is to increase the number of clusters. The algorithm can still not perfectly discriminate between the types that have generated experiences and choices. This in particular holds for clusters 25—28. In these clusters, contributions and experiences are similar. They start at a differently high level and gradually decay. This pattern is generated by conditional cooperators, hump-shaped players and (in cluster 25) also farsighted free-riders interacting with each other. If they are in a group that, otherwise, is very cooperative, hump-shaped players and short-sighted freeriders generate a similar pattern (cluster 23). Conditional cooperators and hump-shaped players look the same if they are in a group dominated by far-sighted freeriders (cluster 13) or by short-sighted freeriders (cluster 24). Altruists and conditional cooperators are lumped together if the group quickly converges to full cooperation (cluster

9). Yet even in all these clusters, while types are not perfectly separated, patterns are very cleanly characterised. The algorithm visibly does a very good job. Types are not distinguished because different reaction functions generate choice patterns that are very similar, provided a participant makes the experiences defined in the respective right panel.

For the remaining clusters, even types are identified (sometimes perfectly, sometimes nearly). Yet this degree of cleanliness is only achieved because the algorithm is allowed to split one and the same type by the experiences they make. The need for a larger number of clusters is evident with altruists. They have been simulated as non-reactive, cooperative, but noisy. This is why choices look very similar in clusters 1—8. The difference results from the experiences an altruist makes. In cluster 2, they are together with a majority of conditional cooperators, but at least one far-sighted freerider. As all group members are, at least initially, conditionally cooperative, experiences improve in early periods. Yet once a freerider starts cashing in, the conditional cooperators follow suit, which explains the kink in the second part of the timeseries. In cluster 3, experiences are more extreme, as far-sighted freeriders have a bigger impact. In cluster 4, experiences never reach the top. This pattern results if hump-shaped players or short-sighted freeriders draw down the contribution level. In clusters 5—8, experiences are flat as the influence of far-sighted freeriders is absent. The composition of the remaining types determines the (nearly or perfectly) constant level of the contributions made by the remaining group members.

At the lower end, clusters 31—34 are also pure. The contributions of short-sighted freeriders are at or near to 0 throughout. But the algorithm needs multiple clusters as experiences differ. Yet as 3 of the 5 types are themselves reactive, the experience patterns look very different from the experiences that altruists are making. In the most favourable cluster 31, the remaining members are sufficiently cooperative themselves to tolerate exploitation by a single freerider. By contrast, in clusters 32 and 33 the presence of the freerider induces cooperative types to gradually reduce their contributions. In cluster 34, a small fraction of conditionally cooperative types is quickly deterred by the prevailing degree of exploitation.

Interestingly, there are also pure clusters of reactive types. In clusters 10 and 11, all players are conditional cooperators. In cluster 10, experiences are initially positive, but deteriorate in the middle of the timeseries, due to the presence of shortsighted freeriders. In cluster 11, there are no altruists, which is why experiences and contributions never reach the top. But there are also no freeriders, which is why cooperation is stable at an intermediate level.

The algorithm finds even more clusters in which all participants are far-sighted freeriders themselves. There is always the kink in the middle of the series. Clusters 14—17 differ by the experiences these far-sighted freeriders are making. In clusters 15 and 17, these experiences are fairly stable, which must result from the fact that groups are mostly exclusively composed of non-reactive types. In the remaining groups, at least some group members react to the fact that the far-sighted freerider reduces contributions, by lowering contributions themselves.

In clusters 20—22, only hump-shaped players are to be found. The most characteristic pattern is cluster 20. As long as experiences are very good, hump-shaped players reduce their own contributions. But if far-sighted freeriders start exploiting, experiences fall below the threshold, and hump-shaped players begin stabilising cooperation. By contrast, in clusters

21 and 22, the contributions of hump-shaped participants stay below the more favourable level of experiences they are making.

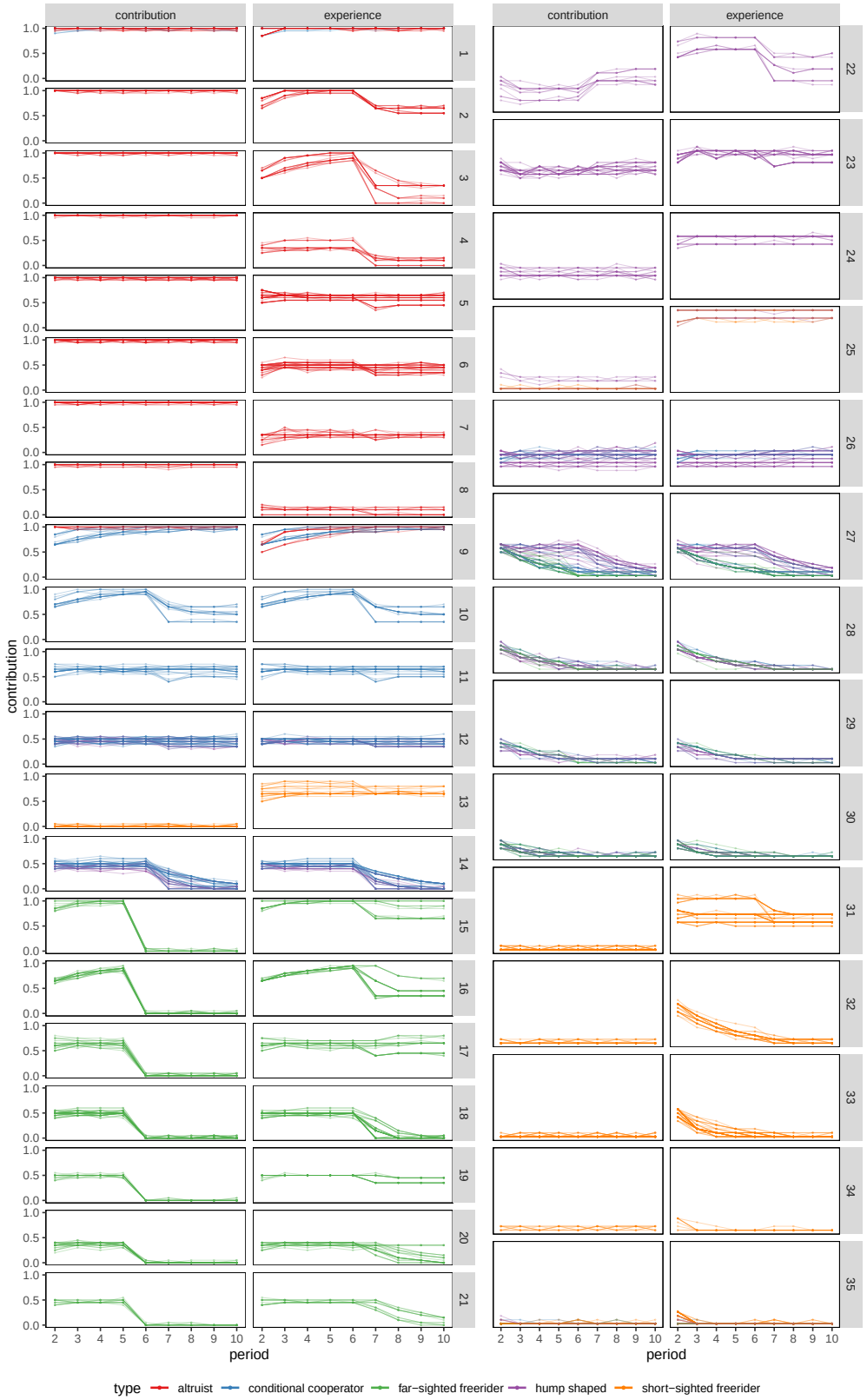


Figure A.1: Exemplary Partitioning of the Simulated Dataset into 35 Clusters for 5 Types

A2 Details on the Experimental Datasets

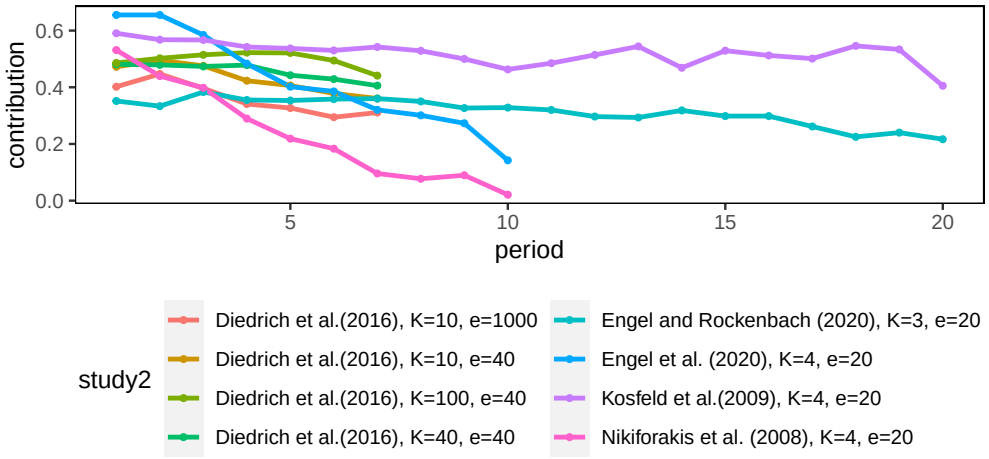


Figure A.2: Means of Participants' Contributions by Study

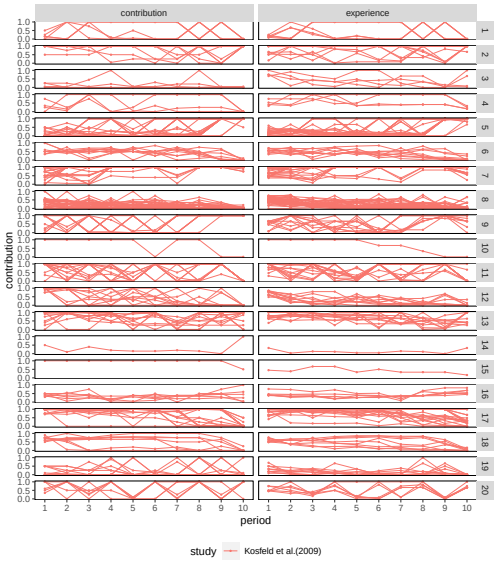
Diederich, Goeschl, and Waichman (2016) plays a standard PGG, varying the group size. We include only their treatment 1, which groups 10 individuals at a time. One specialty of their design is long-term rounds, each of exactly 72 hours. Subjects could freely decide when to participate and submit their decision via a device of their choice connected to the internet.

Engel and Rockenbach (2020) add passive bystanders who are either just present, or negative (positively) affected by the contributions active players make to a standard linear public good, to find out why conditional cooperators overreact to negative experiences. We use data from an additional baseline (not reported in the working paper) with no bystanders.

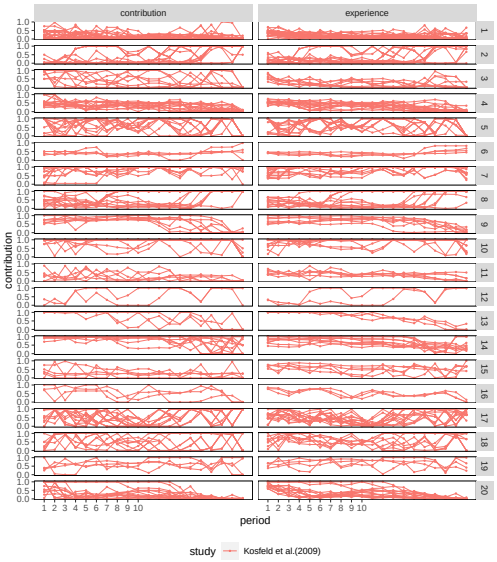
Engel, Kube, and Kurschilgen (2020) test, in a linear public good, whether selective information about the choices that unrelated third parties have made in the otherwise identical game can increase or decrease contributions. We use data from the baseline, with no manipulation of first impressions.

Kosfeld, Okada, and Riedl (2009) play an institution formation game followed by a PGG: Each player decides whether she wants to participate in an organization. Subsequently, players simultaneously determine the number of their contributions to the public good.

Nikiforakis and Normann (2008) deploys a standard PGG and adds treatments with punishment options. In our dataset, we only include their control treatment, the PGG without punishment.



(a) Interpolated, 10 periods



(b) Not Interpolated, 20 periods

Figure A.3: Separating Datasets by Periods is Crucial

Table A.1: Subsets by Period

subset	periods	groupsize	subjects
1	10	4, 3	482
2	20	4, 3	362
3	7	10, 40, 100	1210

A3 Internal Cluster Validation Indices

In Figure A.4 all indices are normalized to the unit interval. Indices to be minimized are recoded and reported as inverse.

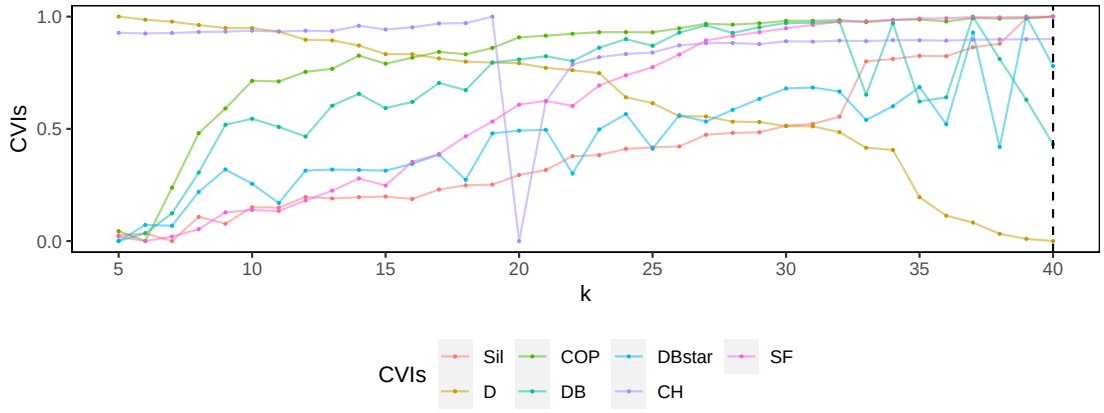


Figure A.4: Simulated Data: Internal Cluster Validation Indices

The dashed vertical line represents the pick based on the rankvote