Quantitative Verification in Rational Environments

Anshul Gupta
Department of Computer Science
University of Liverpool
Liverpool, United Kingdom
anshul@liverpool.ac.uk

Sven Schewe
Department of Computer Science
University of Liverpool
Liverpool, United Kingdom
sven.schewe@liverpool.ac.uk

***Abstract*—We study optimal equilibria in turn based multi-player mean-payoff games. Nash equilibria are a standard way to define rational behaviour of different players in multi-player games. These equilibria treat all players equally. We study settings where a leader has additional power over the game: she has the power to assign strategies to all participating players, including herself. We argue that a leader who assign the strategies, may not want to comply with the common restrictions imposed by Nash equilibria. This setting provides the basis for the quantitative analysis of the distributed systems, where the leader can take the role of a controller or an adversary, while the other players form a rational environment. We show that the leader always has an optimal strategy in this setting, and that no Nash equilibrium can be superior to it. Finding this equilibrium is NP-complete and, for a fixed number of players, there is a polynomial time reduction to solving two player mean-payoff games.**

I. Introduction

There has been a recent trend to replace traditional model checking by quantitative model checking (see [11] for a recent survey). In traditional model checking, a qualitative property such as 'a system is always eventually granted access to a resource', ($\square\lozenge\mathsf{access}$), is checked. This property can be reflected by the deterministic Büchi automata (DBA) from Figures 1 and 2. The first automaton, $\mathcal{A}$, shown in Figure 1 is in an accepting state whenever the process requests (and is granted) access to a resource, while the second automaton $\mathcal{B}$, shown in Figure 2 is in an accepting state whenever the player is granted access.

Traditionally, a DBA would accept an infinite play if the accepting state is visited infinitely many times, and all of these paths would be of equal quality. In quantitative model checking, this qualitative measure is replaced by a quantitative measure, where the quality of a path would be measured by the limit average share of accepting states occurring in a run of the DBA: it defines a mean-payoff condition. In this setting, $\mathcal{B}$ refers to the limit average share of the time that a system's critical resource is used, while $\mathcal{A}$ refers to the limit average frequency with which a process asks for (and receives) access on an infinite path.

This inspired us to consider a setting, where different selfish players follow different objectives defined by such DBAs. In our example from Figures 3 and 4, the environment consists of two selfish players who want to maximise the frequency in which they are granted access to a critical resource (using $\mathcal{A}$ for their respective objective), while the control objective of the system is to maximise the utilisation of the system (using $\mathcal{B}$ for this objective).

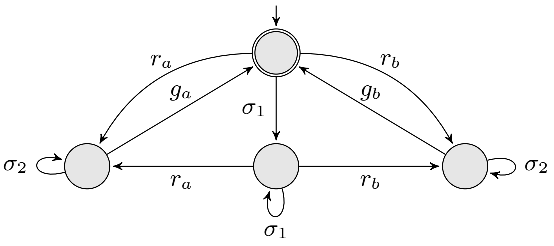

Fig. 1. $\sigma_1 = \neg(r_a \vee r_b)$, $\sigma_2 = \neg(r_a \vee r_b \vee g_a \vee g_b)$

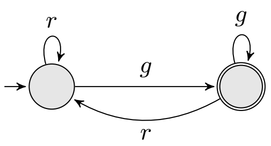

Fig. 2. $g = g_a \vee g_b \vee g'_a \vee g'_b$, $r = r'_a \vee {r_b}' \vee r_a \vee r_b \vee \varepsilon$
Two property automata: Figure 1 and Figure 2

In some states of this model the players have choices. In our example, the two players have the choice to make two different kinds of requests, r_a (resp. r'_a for the second player), which shall trigger an access for just one time unit, represented by g_a (resp. g'_a for the second player), or a request r_b (resp. r'_b), which shall trigger an access for three time units, represented by three g_b (resp. g'_b). They can also use a local ε move.

In order to keep the model simple, we focus on models where, on each state, there is one player who resolves the choice. The states in which a player resolves the choice are depicted as squares. Slightly more general, we consider mean-payoff games (MPGs) [25], [12], [9], [2], [20], [3], [6]. MPGs are finite turn-based games of infinite duration. They are played on a game arena – a directed graph, whose vertices are owned by different players. An MPG is played by placing a token on a vertex, and allowing the player who owns the vertex to push the token forward along an edge in the arena. Thus, the players successively create an infinite play. The edges of the game hold rewards for each player, and the objective of each player is to maximise her limit average reward. Figure 5 depicts the multi-player mean-payoff game defined by players from Figures 3 and 4 with their respective properties.

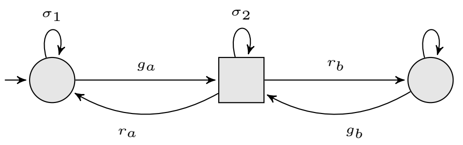

Fig. 3. $\sigma_1 = r'_a \vee r'_b \vee g'_a \vee g'_b \vee \varepsilon$, $\sigma_2 = \sigma_1 \vee g_a \vee g_b$

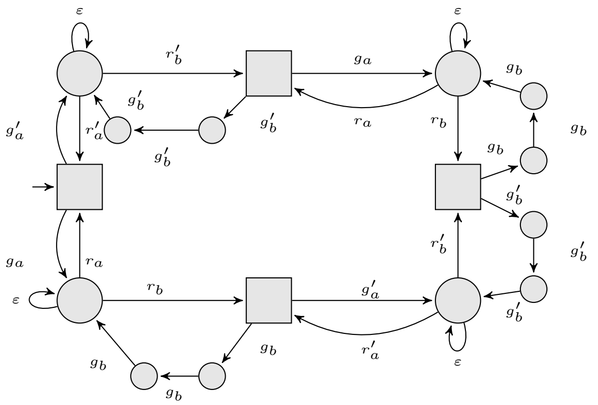

Fig. 4. The rational environments (Figure 3) and the system (Figure 4), shown as automata that coordinate on joint actions.

The way each player plays can be captured by a strategy. A set of strategies, one for each player, is called a strategy profile. A strategy profile is in a Nash equilibrium if no player has an incentive for unilateral deviation, i.e., if all other players adhere to their strategy, a player cannot increase her payoff by changing her strategy. Nash equilibria [18], [16], [23], [19], [8] are a common way to describe stable strategies with the intuition that only if no player gains from changing her strategy unilaterally, the strategy will be maintained. Qualitative Nash equilibria have been used to refine a worst case analysis. In [11], for example, the distributed development of a system is considered, where teams develop components that try to establish individual specifications. In this symmetric setting, a component can safely be assumed not to be malicious to the extent that, for harming others, it would sacrifice compliance with its main objective. Within this constraint, however, it is conservatively considered to be adversarial.

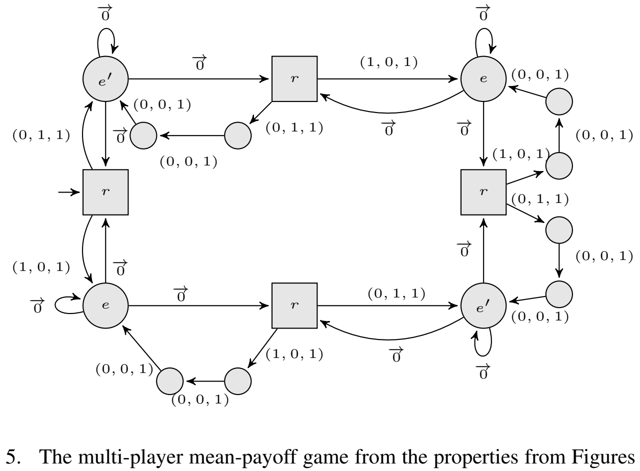

Fig. 5. The multi-player mean-payoff game from the properties from Figures 1,2 and 3,4. The nodes are labelled by the players who own them, e and e' for the rational environment players, and r for the system player. The payoff is shown in the order payoff for e, e', r.

In this paper, we raise the question of how an interested party, henceforth called a *leader*, can capitalise on setting an equilibrium strategy. The question we set out to answer can therefore be phrased as:

How should a reflective leader control a system?

When we view the leader as adversarial, we can use the same techniques to determine how she can coordinate an attack on the system to minimise the payoff.

If we allow the leader to select the complete strategy profile, the question whether achieving a Nash equilibrium is the right target for her begs to be asked. The constraints on the strategies of the other players are clearly a pre-requisite for a stable strategy, but how about hers? If the leader can select strategies that she can improve over, she has more leeway when selecting a strategy profile. We therefore argue that she should be allowed to 'discriminate' against herself. As we will show in Section III-A that the leader may suffer from restricting her strategy in the Nash sense. Note that this effect does not occur in the qualitative setting, where such an equilibrium would also be Nash.

In mean-payoff games based on quantitative specifications, studying a leader of this type has numerous natural justifications. For example, we might seek optimal control of a system that is used by the external players, who are not under our control, but to whom we can communicate the rules of its use. It is natural to assume that the rules will only be complied with, if the external players have no incentive to deviate, while the controller can take a higher perspective and take the indirect effect (in form of non-compliance by the external players) of her deviation into account when setting the rules. Similarly, an adversarial leader has to take the rationality of the external players involved into account, but she can herself resolve the remaining nondeterminism in the system in any way that complies with this constraint.

In our example, the leader will seek to maximise (in the controller model) resp. minimise (in the attacker model) the time any process is using the critical resource, while the individual processes attempt to maximise their own access to it.

In software engineering, the environment is often regarded as antagonistic. This relates to a two player zero-sum game, where the environment forms a monolithic block whose gains and losses are the losses and gain, respectively, of the system, represented by the leader in our setting.

A. Results

A first result is the introduction of the concept of *leader equilibria* and *leader strategy profiles*. A leader strategy profile is a relaxation of a Nash equilibrium, where all players except for the leader do not benefit from unilaterally changing their respective strategy. They are the strategy profiles the leader can enforce: in all other strategy profiles, rational players would refuse to follow her, as they would have preferable options.

We call optimal leader strategy profiles *leader equilibria*. While they are equilibria in the sense that no participating player benefits from deviation, they are equilibria on a different level: while the effect of unilateral deviation of the remaining players is considered in a Nash sense, the leader takes the effect on the stability of the strategy profile into account.

We show that leader equilibria always exist and that no Nash equilibrium can be superior over them, where the latter is a simple corollary from the fact that each Nash equilibrium is in particular a leader strategy profile. We establish the NP completeness of the related decision problem 'is there a leader equilibrium with payoff greater or equal to a threshold' (which equals the bound for Nash equilibria [24]), and sharpen this bound by showing that they cannot be approximated. We also show that the NP hardness depends on the number of players: for a bounded number of players, we give a polynomial time reduction to solving two player mean-payoff games (2MPGs).

As the complexity of solving two player MPGs is wide open, we cannot hope for determining the precise complexity for solving MPGs with a bounded number of players without solving their complexity first. However, we get simple corollaries for the complexity of finding leader and Nash equilibria for a bounded number of players: it can be done in pseudo polynomial time [6], it can be done in smoothed polynomial time [3], there are fast randomised [2] and deterministic [20] strategy improvement algorithms, and the decision problem is in UP∩CoUP [12], [25].

B. Related Work

The existence of Nash equilibria in multi-player mean-payoff games has been established in [26]. Ummels and Wojtczak [24] studied the complexity of determining the existence of Nash equilibria, where each reward falls into a given closed interval in multi-player mean-payoff games. Both sides of the NP completeness proofs are closely related to ours. A key ingredient of using reward and punish strategy profiles for mean-payoff games are inspired by [24], [5] and similar strategies in stateless games [27].

In [22], Ummels has studied the concept of subgame perfect equilibrium for the case of infinite games. He has given simple examples to show that subgame perfect equilibrium, where choice of strategy should be such that it is optimal for initial history of the game and not for just initial vertex, exists in the case of infinite games.

There are quite a few works on optimal equilibria, in particular on equilibria that are 'best for society', which is usually defined as the optimal sum. This definition is, for example, used in the definition of the Price of Anarchy [15] for network and internet related games, or in traffic routing games [1], [10]. In [17], the authors study a virus inoculation game on social networks, in which players think of their neighbour's welfare. In [7], the authors have modelled a society game and shown how the equilibria are affected if players think of society rather than thinking of themselves.

II. Preliminaries

A *multi-player mean-payoff game* (MMPG) is a tuple $\langle P, V, \{V_p \mid p \in P\}, v_0, E, \{r_p : E \to \mathbb{Q} \mid p \in P\}\rangle$, where

- P is a set of players,
- V is a set of vertices with a designated initial vertex $v_0 \in V$,
- $\{V_p \mid p \in P\}$ is a partition of the vertices V,
- $E \subseteq V \times V$ is a set of edges, such that each vertex has a successor ($\forall v \in V\ \exists v' \in V,\ (v, v') \in E$), and
- $\{r_p \mid p \in P\}$ is a family of reward functions $r_p : E \to \mathbb{Q}$, that assign, for each player $p \in P$, a reward to each transition to p.

An MMPG is intuitively played by placing a token on the initial vertex. Each time the token is on the vertex of a player p, player p chooses an outgoing transition and moves the token along this transition. This way, the players jointly construct an infinite *play* $\pi \in V^\omega$. For each player p, a play $\pi = v_0, v_1 \ldots$ is evaluated to

$$r_p(\pi) = \liminf_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} r_p\big((v_i, v_{i+1})\big).$$

If the reward functions sum up to 0, i.e., if $\sum_{p \in P} r_p(e) = 0$ holds for all edges $e \in E$, then we call the MMPG a zero-sum game.

The way that the respective players choose the successor vertex is a function $\sigma_p : V^* V_p \to V$ from an initial sequence of a play that ends in some vertex $v \in V_p$ of player p to a vertex v', such that $(v, v') \in E$. A family of strategies $\sigma = \{\sigma_p \mid p \in P\}$ is called a strategy profile. A strategy profile σ defines a unique play π_σ, and therefore a reward $r_p(\sigma) = r_p(\pi_\sigma)$ for each player p.

A strategy profile is a Nash equilibrium if no player has an incentive to change her strategy, provided that all other players keep theirs. That is, if, for all players $p \in P$ and for all $\sigma' = \{\sigma'_q \mid q \in P\}$ with $\sigma_q = \sigma'_q$ for all $q \neq p$, $r_p(\sigma) \geq r_p(\sigma')$ holds.

For a designated leader $l \in P$, a strategy profile is a *leader strategy profile* if no *other* player has an incentive to deviate from her strategy. That is, if, for all players $p \in P \smallsetminus \{l\}$ and for all $\sigma' = \{\sigma'_q \mid q \in P\}$ with $\sigma_q = \sigma'_q$ for all $q \neq p$, $r_p(\sigma) \geq r_p(\sigma')$ holds. Thus, a leader strategy profile allows for solutions, where the leader could improve upon her reward by changing her strategy. While this may on first glance not be in the interest of a leader, we will see that she can obtain better results with leader strategy profiles than with Nash equilibria.

We use two player zero-sum mean-payoff games (2MPGs) to determine the outcome of MMPGs when, from some point onwards, one player, say p, is playing against all others, where the objective of p is inherited from the multi-player MPG, while the objective of the remaining players is to minimise her reward. As the objective of the remaining players is defined by the objective of p, we use only r_p to describe the objective of the game. The 2MPG *for* p of an MMPG $\mathcal{M} = \langle P, V, \{V_p \mid p \in P\}, v_0, E, \{r_p : E \to \mathbb{Q} \mid p \in P\}\rangle$, denoted $\mathsf{2mpg}(\mathcal{M}, p)$, is therefore the game $\langle P, V, \{V_p, V \setminus V_p\}, v_0, E, r_p\rangle$. 2MPGs have optimal memoryless strategies for both players, and the outcome is, when starting in any vertex $v \in V$, determined [25]. By abuse of notation, we denote this value for a vertex v by $r_p(v)$.

III. Leader equilibria

A leader strategy profile that provides the maximal reward for the leader among all leader strategy profiles is called a *leader equilibrium*. In the remainder, we show that

1) leader equilibria are generally better (for the leader) than Nash equilibria (Theorem 3.2),
2) determining if there is a strategy profile σ with $r_l(\sigma) = 1$, such that the strategy profile σ is a Nash equilibrium or leader strategy profile is NP hard even for zero-sum MMPGs with reward functions whose domain is $\{-1, 1\}$ (Theorem 3.3), and the optimal reward of the leader cannot be approximated efficiently (Corollary 3.4), and
3) leader equilibria (and optimal Nash equilibria) always exist, and, for a fixed set of players, finding them in MMPGs is polynomial time reducible to solving two player MPGs (Corollary 3.16).

For social optima, it suffices to add a social reward to the reward function, e.g., the sum of the individual rewards, without letting the respective player own any vertex. The technique introduced in this paper can then be used to optimise the social payoff.

We start with a trivial inference of the existence of leader strategy profiles.

Lemma 3.1: Leader strategy profiles exist for all multi-player mean-payoff games.

Proof: It is shown in [5] that Nash equilibria for multi-player mean-payoff games exist. By definition, any strategy profile in Nash equilibrium, is also a leader strategy profile.

What remains to be shown is that optimal leader strategy profiles (i.e., leader equilibria) exist, but that the optimum can be taken is implied by the construction from Section III-C (cf. Corollary 3.11).

A. Superiority of leader equilibria

In this subsection, we show that leader equilibria are superior over Nash equilibria: a benign leader who assigns strategies in such a way that she only makes sure that no *other* player has an incentive to deviate, while allowing for the use of 'modest' strategies that she can improve upon when the other players stick to their strategies, is more successful than a leader who follows the short sighted egoistic approach to chose only among strategies she cannot improve upon herself.

On first glance, it may not seem to be in the interest of the leader to be benign. To the contrary, it would seem that the leader could improve upon such strategy profiles by simply adjusting her strategy. A second look, however, reveals that she has to comply with less constraints and can, consequently, choose from a larger pool of strategy profiles. In particular, this implies that the optimum cannot be worse.

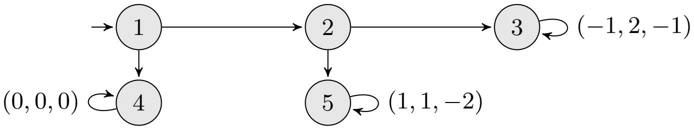

Fig. 6. An MMPG, where the leader equilibrium is strictly better than all Nash equilibria. The rewards are depicted in the order first player, leader, passive player. The rewards on the edges $(1, 2)$, $(1, 4)$, $(2, 3)$, and $(2, 5)$ is not shown, because these edges can only be taken once in a play. Their rewards therefore have no impact on the payoff for any player.

The MMPG from Figure 6 exemplifies why this may lead to an increased payoff. It shows a simple MMPG with five vertices, 1 through 5, where the *leader* owns vertex 2, and a player *first* owns the initial vertex, vertex 1. The other vertices have exactly one successor (themselves), such that it does not matter who owns them. We assume that they are assigned to a third player, player *passive*.

Initially, player *first* can either play to vertex 4, or to vertex 2. When playing to vertex 4, every player will receive a payoff of 0. When she plays to vertex 2, the *leader* can either move on to vertex 3, securing herself a payoff of 2, to the cost of the *first* and the *passive* player, who both receive a payoff of -1. Alternatively, she could play to vertex 5, where both the *leader* and the *first* player receive a payoff of 1, to the cost of the *passive* player, whose payoff is -2. In a Nash equilibrium, the *leader* will never move to vertex 5, as she can improve over such strategies by simply choosing to go to vertex 3. Consequently, the *first* player will not move to vertex 2 in any Nash equilibrium, as this would result in a payoff of -1 for her, such that moving to vertex 4 is preferable. Thus, the only play produced by a Nash equilibrium is the play $1 \cdot 4^{\omega}$.

But the leader has a better leader equilibrium: she can benignly waive her option to move to vertex 3, and instead move to vertex 5. Then, it becomes preferable for the *first* player to move to vertex 2. This results in an improved reward for the leader.

Theorem 3.2: Nash equilibria cannot be superior to leader equilibria, and leader equilibria can be strictly better than all Nash equilibria.

B. NP hardness

In order to establish NP hardness, we reduce the satisfiability of a 3SAT formula over n atomic propositions with m conjuncts to solving a zero-sum MMPG with $2n + 1$ players and $4m + 2n + 1$ vertices that uses only payoffs 0 and 1. We consider the reduction for the example of the 3SAT formula $C_1 \wedge C_2 \wedge C_3$ with $C_1 = p \vee \neg q \vee \neg r$, $C_2 = \neg p \vee q \vee \neg r$, and $C_3 = \neg p \vee \neg q \vee \neg r$.

The $2n + 1$ players consist of $2n$ players for the $2n$ literals corresponding to the n variables, and the *leader* who intuitively tries to validate the formula. The game is played in two phases, an initial *formula phase*, and a subsequent *payoff* phase.

In the formula phase, the leader cycles through the m disjuncts, where we denote her vertices by C_1 through C_m. In each of these vertices, she selects a literal, e.g., (C_1, p) from

C_1 in the example. Such a vertex is owned by the player of this literal, e.g., (C_1, p) is owned by p. This literal player can either move to an absorbing vertex, or move to the next literal (when in conjunct 1 through $m-1$) or to a designated vertex in the evaluation phase (when in conjunct m).

In the evaluation phase, all nodes are owned by the leader. The leader cycles through the atomic propositions, either taking an edge to p, where player $\neg p$ receives a payoff of 0 while all other players receive a payoff of 1, or to $\neg p$, where player p receives a payoff of 0 while all other players receive a payoff of 1. The absorbing sink has only one self-loop with payoff 1 for all literal players and 0 for the leader.

For the example formula, the leader has a leader equilibrium which constructs the path $C_1, (C_1, p), C_2, (C_2, q), C_3, (C_3, \neg r), (p, q, \neg r)^\omega$.

This path provides a payoff of 1 to the leader and the players p, q, and $\neg r$, and a payoff of $\frac{2}{3}$ for the remaining players. No player has therefore an incentive to deviate unilaterally, and it is clear that such a play exists whenever the formula is satisfiable. Thus, for satisfiable formulas, there is a leader equilibrium with payoff 1 for the leader.

If the formula is not satisfiable, any run that ends up in the evaluation phase would have to path by (C_i, p) and $(C_j, \neg p)$ for some atomic proposition p. But then, the sum of the payoffs of these players is at most $2 - \frac{1}{n}$, and hence at least one of these players receives a payoff < 1. This player has an incentive to deviate by going to the sink instead. Thus, all leader stable strategy profiles end in the sink vertex for unsatisfiable formulas, and a leader equilibrium has a payoff of 0 for the leader.

a) **Zero-sum games:** To progress from here to zero sum games, we can simply replace the rewards of 0 to -1 and add $(n-1)$ additional players who own no vertex and always have a reward of -1. The result that these games cannot be approximated clearly carries over.

Theorem 3.3: The decision problem of whether or not a leader strategy profile or Nash equilibrium σ with reward $r_l(\sigma) = 1$ of the leader exists in games with rewards in $\{0, 1\}$ resp. zero sum games with rewards in $\{-1, 1\}$, such that the reward of the leader is always in $\{0, 1\}$ resp. $\{-1, 1\}$ is NP complete.
The proof is closely related to the NP hardness proof from [24].

Corollary 3.4: Unless P=NP, no tractable algorithm can approximate the optimal reward of the leader closer than 0.5.

C. Reward and punish strategy profiles for leader equilibria

Let us consider a leader strategy profile σ. We first show that we can obtain a leader strategy profile with a similar payoff for the leader by applying a punishment to the first player who deviates from σ. The power to define the equilibrium allows the leader to use the power of all remaining players to punish this deviator.

That is, we use a strategy profile where all players cooperate to produce π_σ. Note that the *leader* solicits cooperation from every player who owns some vertex in the game. Further, the strategy profile σ offers the reward $r_p(\pi_\sigma)$ to a player p, which is at least as good as the reward that player p would have received in the two player game $2\mathsf{mpg}(\mathcal{M}, p)$ from any vertex in $\mathsf{ver}(\pi_\sigma)$. But, if a player deviates from σ, all other players co-operate to harm this player, throwing their own interests to the wind.

Thus, not complying with the requirement to produce π_σ will lead to a payoff of the deviating player, which equals the payoff of this player in a two player game that starts at the point of her deviation, i.e., at the vertex owned by her, where she is supposed to play in accordance with σ. We call any such strategy profile a *reward and punish strategy profile* and define it as a σ that offers reward $r_p(\pi_\sigma)$ to a player p and any deviation from σ by a player p will eventually lead her to get a (not necessarily strictly) lower payoff than $r_p(\pi_\sigma)$.

Note that, for reward and punish strategy profile, π_σ essentially defines σ.

Lemma 3.5: If σ is a leader strategy profile, then there is a reward and punish strategy profile σ', which is also a leader strategy profile and defines the same play $\pi_\sigma = \pi_{\sigma'}$. If σ is a Nash equilibrium, so is σ'.

Proof: We first observe that π_σ alone defines the reward of all players for the strategy profile σ and thus, due to $\pi_\sigma = \pi_{\sigma'}$, of σ'.

Let us assume for contradiction that a player $p \in P$ for Nash equilibria resp. $p \in P \smallsetminus \{l\}$ for leader strategy profiles has an incentive to deviate from her strategy in σ'. Then her payoff in σ' will be determined by the result of the two player zero-sum MPG 'her against the rest' as defined by the reward and punish strategy profiles. Note that the initial play up to this point has no impact on the limit reward.

But she can deviate from her strategy in σ at the same position with at least the same reward, by simply assuming that she plays against all other players in the same game. Consequently, she has an incentive to deviate in the strategy profile σ, too, which contradicts the assumption that σ is a Nash equilibrium resp. leader strategy profile.

This observation allows us to concentrate on reward and punish strategy profiles only. Let $\mathsf{ver}(\pi)$ be the set of vertices that occur in a play, and let $\mathsf{own}(S) = \{p \in P \mid S \cap V_p \neq \emptyset\}$ be the set of players that own some vertex in S. With these terms, it is simple to characterise reward and punish strategy profiles.

Lemma 3.6: For an MMPG $\mathcal{M}$, a play π_σ is the outcome of a reward and punish strategy profile σ, which is a Nash equilibrium resp. leader strategy profile, if, and only if, for all vertices $v \in \mathsf{ver}(\pi)$ and all players $p \in \mathsf{own}(\mathsf{ver}(\pi_\sigma))$ resp. $p \in \mathsf{own}(\mathsf{ver}(\pi_\sigma)) \smallsetminus \{l\}$ that control a vertex that occurs in the play, it holds that $r_p(\pi_\sigma) \geq r_p(v)$.

Proof: To show the 'if' direction, we assume for contradiction that $r_p(\pi_\sigma) < r_p(v)$ holds for some vertex $v \in \mathsf{ver}(\pi)$, which is owned by p (resp. owned by $p \neq l$). Then player p can improve on her strategy by following her strategy until v is reached, and henceforth follow the strategy from $2\mathsf{mpg}(\mathcal{M}, p)$. As the initial play does not influence the limit inferior, her payoff would be at least $r_p(v)$, which is strictly greater than

$r_p(\pi_\sigma)$. $\lightning$

To show the ‘only if’ direction, we assume for contradiction that $r_p(\pi_\sigma) \geq r_p(v)$ holds, but no reward and punish strategy profile defines π_σ. Assume that player p, deviates in vertex v from π_σ. Then the other players will join to diminish her payoff henceforth. Taking into account that the initial sequence up to this point has no influence on the the payoffs, they can follow the optimal strategy of the opponents of p from $\mathsf{2mpg}(\mathcal{M}, p)$, restricting the payoff of player p to $r_p(v)$. $\lightning$

In the next step, we show that we can determine the existence of a *well behaved* reward and punish strategy profile that satisfies such a constraint system. A strategy profile is *well behaved* if the ratio in which every edge occurs has a limit, that is, if, for all edges $(s,t) \in E$, there is a $r_{(s,t)} = \lim\limits_{n\to\infty} \frac{\#_n^{(s,t)}(\pi_\sigma)}{n}$, where $\#_n^{(s,t)}(v_0, v_1, v_2 \ldots) = \big|\{i < n \mid (v_i, v_{i+1}) = (s,t)\}\big|$ is the number of edges (s,t) among the first n edges that occur in a play $v_0, v_1, v_2 \ldots$. (This limit does not necessarily exist for general strategy profiles.)

b) **Linear programs for well behaved reward and punish strategy profiles:** The first central observation is that, if we already know

- the set of vertices Q visited in π_σ and
- a (strongly connected) set S of vertices such that $S \subseteq Q$ contains all vertices that are visited infinitely often (and is therefore strongly connected),

then we can infer a constraint system by Lemma 3.6, which is necessary and sufficient for a well behaved reward and punish strategy profile. The constraint system consists of two parts. One part is the ratios, where we use the $p_{(s,t)}$ from above for edges $(s,t) \in E \cap S \times S$, and similarly p_v for the limit ratio of each vertex v in S. (Obviously, the limit ratio of each vertex not in S and of each edge not in $S \times S$ must be 0.)

This provides a first part of a constraint system, namely

- the ratio of vertices and edges that are not in S resp. $S \times S$ is 0,
  - $p_v = 0$ for all $v \in V \smallsetminus S$ *and*
  - $p_e = 0$ for all $e \in E \smallsetminus S \times S$,
- the ratio of vertices and edges that are in S resp. $S \times S$ is ≥ 0,
  - $p_v \geq 0$ for all $v \in S$ *and*
  - $p_e \geq 0$ for all $e \in E \cap S \times S$,
- the sum of the ratio of vertices is 1, $\sum\limits_{v \in V} p_v = 1$,
- the ratio of a vertex is the sum of the ratios of its incoming edges and
the ratio of a vertex is the sum of the ratios of its outgoing edges,
  - $p_s = \sum\limits_{t.(s,t)\in E} p_{(s,t)}$ for all $s \in S$ *and*
  - $p_t = \sum\limits_{s.(s,t)\in E} p_{(s,t)}$ for all $t \in S$.

The second part of the constraint system stems from Lemma 3.6. For a well behaved strategy profile σ, $r_p(\pi_\sigma) = \sum_{e\in E} p_e r_p(e)$ is simply the weighted sum of the rewards of the individual edges. This provides us with a constraint

$$\sum_{e\in E} p_e r_p(e) \geq \max_{v\in Q}(r_p(v))$$

for all $p \in \mathsf{own}(Q)$ for Nash equilibria, and for all $p \in \mathsf{own}(Q) \smallsetminus \{l\}$ for a leader strategy profile.

Before we define the objective function, we state a simple corollary from Lemma 3.6.

Corollary 3.7: Every well behaved reward and punish strategy profile satisfies these constraints, and every well behaved strategy profile σ, whose play π_σ satisfies these constraints, defines a reward and punish strategy profile.

The objective of the leader is obviously to maximise $r_l(\pi_\sigma) = \sum_{e\in E} p_e r_l(e)$. Once we have this linear programming problem, it is simple to determine a solution in polynomial time [13], [14].

The relevant points are first to establish that a well behaved reward and punish strategy profile exists for each such solution, and second, to show that non-well behaved reward and punish strategy profiles cannot be preferable for the leader.

c) **From Q, S, and a solution to the linear programs to a well behaved reward and punish strategy profile:** We start with the simple case that the vertices and edges with non-0 ratio are strongly connected.

We design π_σ as follows. We first go from the initial vertex v_0 through states in Q to some state in S. (Note that this initial path has no bearing on the limit inferior that defines the payoff of the individual players.)

Once we have reached S, we intuitively keep a list for each vertex in S. In this list, we keep the number of times each outgoing edge with non-0 ratio has been taken. We also apply an arbitrary (but fixed) order on the outgoing edges. Each time we are in this vertex, we choose the first edge (according to this order) that has been taken less often (from this vertex) than $\frac{p_e}{p_v}$, the ratio p_e of the edge divided by the ratio p_v of this vertex, suggests. If no such edge exists, we take the first edge.

Lemma 3.8: An implementation of such a list is finite: let r_e be the ratio of an outgoing edge e of a vertex v divided by the ratio of the vertex v it emerges from, and let d be the least common denominator of these ratios for a vertex v. Then we can re-set the counters for the outgoing edges to 0 after d steps.

The result is obviously a well behaved strategy profile and the first part of the constraint system is clearly satisfied. It therefore suffices to convince ourselves that the second part is satisfied as well.

Now assume for contradiction that this is not the case. Let q_v and q_e be the real ratio of the vertices and edges, respectively. Note that our simple rule for the selection of vertices implies that $\frac{p_e}{p_v}$ is correct for all edges $e = (v, v') \in E \cap S \times S$. Then there must be a vertex $v \in S$, which has the highest factor $\frac{q_v}{p_v}$. As it is the highest factor, none of its predecessors in $E \cap S \times S$ can have a higher ratio; consequently, they must have the same ratio. By a simple inductive argument, this expands to the complete strongly connected set of non-0 vertices. As $\sum_{v \in S} p_v = 1 = \sum_{v \in S} q_v$ holds, this implies $p_v = q_v$ for all $v \in S$.

To extend this argument to the general case, we first observe that the non-0 vertices and edges form islands of (maximal) SC parts C_1, through C_k. We use this observation to compose a play as follows.

We start with an initial part, a transfer from v_0 to C_1 as in the simple case. We then continue by playing a C_1^1 part, a transfer, a C_2^1 part, a transfer, …, a C_k^1 part, transfer C_1^2, and so forth. To achieve a well behaved strategy profile we do the following.

1) We fix the ratio $\sum_i C_1^i : \sum_i C_2^i : \ldots : \sum_i C_k^i$ according to the the sum of the p_v for vertices v in the respective component. This ratio never changes, and it is given by natural numbers $c_1, c_2, \ldots, c_k$, such that $c_1 : c_2 : \ldots : c_k$ satisfies this ratio.
2) We let C_j^i grow slowly with i. We can, for example, use $i \cdot c_j$.
   Note that the transfer part has constant length, bounded by $|S|$. Thus the limit ratio of transfer is 0.
3) We let the transfer to C_{j+1}^i go to the vertex, in which C_j^i was left. Note that the transfer may contain vertices of various components, but as the overall ratio of the transfer is 0, this does not affect the limit probability.
   Consequently, we can use the controller from the simple case of one SCC for the sequence $C_i^1, C_i^2, C_i^3 \ldots$, which only focuses on the relevant part of the i^{th} component.

In effect, we have simple controllers for the individual components, and a single counting controller that manages the transfer between the components.

It is easy to see that the resulting controller inherits the correct ratios from the simple individual controllers. Together with Corollary 3.7 we get:

Theorem 3.9: If the linear program from above for sets Q of reachable states and S of states visited infinitely often has a solution, then there is a well behaved reward and punish strategy profile that meets this solution.

Finally, we show that non-well behaved reward and punish strategy profiles cannot provide a better solution than the one provided by the previous theorem.

Theorem 3.10: For given sets Q and S, non-well behaved reward and punish strategy profiles cannot provide better rewards for the leader than the reward r_l for the leader obtained by the well behaved reward and punish strategy profiles described above.

Proof: We have shown in Lemma 3.6 that there exists a well defined constraint system obeyed by all reward and punish strategy profiles with set Q of reachable states and all $p \in \mathsf{own}(Q)$ for Nash, and for all $p \in \mathsf{own}(Q) \smallsetminus \{l\}$ for the leader strategy profile. Let us assume for contradiction that there is a reward and punish strategy profile σ that defines a play π_σ with a strictly better reward $r_l(\pi_\sigma) = r_l + \varepsilon$ for some $\varepsilon > 0$.

Let k be some position in π_σ such that, for all $i \geq k$, only positions in the infinity set S of π_σ occur. Let π be the tail $v_k v_{k+1} v_{k+2} \ldots$ of π_σ that starts in position k. Obviously $r_p(\pi) = r_p(\pi_\sigma)$ holds for all players $p \in P$.

We observe that, for all $\delta > 0$, there is an $l \in N$ such that, for all $m \geq l$, $\frac{1}{m} \sum_{i=0}^{m-1} r_p\big((v_i, v_{i+1})\big) > r_p(\pi) - \delta$ holds for all $p \in P$, as otherwise the limit inferior property would be violated.

We now fix, for all $a \in \mathbb{N}$, a sequence $\pi_a = v_k v_{k+1} v_{k+2} \ldots v_{k+m_a}$, such that $v_{k+m_a+1} = v_k$ and $\frac{1}{m} \sum_{i=0}^{m_a-1} r_p\big((v_i, v_{i+1})\big) > r_p(\pi) - \frac{1}{a}$ holds for all $p \in P$.

Let $\pi_0 = v_0 v_1 \ldots v_{k-1}$. We now select $\pi' = \pi_0 {\pi_1}^{b_1} {\pi_2}^{b_2} {\pi_3}^{b_3} \ldots$, where the b_i are natural numbers big enough to guarantee that $\frac{b_i \cdot |\pi_i|}{|\pi_{i+1}| + |\pi_0| + \sum_{j=1}^{i} b_j \cdot |\pi_j|} \geq 1 - \frac{1}{i}$ holds.

Letting b_i grow this fast ensures that the payoff, which is at least $r_p(\pi) - \frac{1}{i}$ for all players $p \in P$, dominates till the end of the first iteration[1] of $|\pi_{i+1}|$.

The resulting play belongs to a well behaved (as the limit exists) strategy profile, and can thus be obtained by a well behaved reward and punish strategy profile by Lemma 3.6. It thus provides a solution to the linear program from above, which contradicts our assumption.

In particular, Theorems 3.9 and 3.10 imply together with Lemma 3.1 the existence of a leader equilibrium.

Corollary 3.11: Leader equilibria exist for all multi-player mean-payoff games.

*d) **Decision & optimisation procedures:*** The *decision problem* related to the construction of optimal equilibria asks whether or not, for a given threshold r_{thld}, there exists a strategy profile σ, which is a Nash equilibrium resp. leader strategy profile and provides a reward $r_l(\pi_\sigma) \geq r_{\mathsf{thld}}$ for the leader.

In Lemma 3.6 and Theorem 3.10 we have established that it is enough to consider well behaved reward and punish strategy profiles. The relevant behaviour of these strategy profiles is captured by the set of reachable vertices, the set of infinite vertices S, and the ratio of the edges in $E \cap S \times S$.

We use this observation in various algorithms, starting with a nondeterministic one.

Theorem 3.12: For an MMPG $\mathcal{M}$ and a threshold r_{thld}, the respective decision problem for leader strategy profiles and Nash equilibria is NP complete, both in the general case and when restricted to zero-sum games with payoffs in $\{-1, 1\}$.

Proof: We use nondeterminism to first guess a set Q of visited vertices, a set S of vertices visited infinitely often and then the linear program defined by them and a solution thereof. Note that the linear program is polynomial in $\mathcal{M}$ and, consequently, has a polynomial solution.

After having a closer look at the sets Q and S, we can check that there is a possible path from the initial vertex to S, that S is strongly connected, that Q and S define the guessed linear program, its constraint system is satisfied by the solution and the reward of the leader is at least the threshold r_{thld} given.

[1] Including the first iteration of π_{i+1} is a technical necessity, as a complete iteration of π_{i+i} provides better guarantees, but without the inclusion of this guarantee, the π_j's might grow too fast, preventing the existence of a limes.

All of these tests can obviously be performed in polynomial time.

The respective hardness results have been established in Theorem 3.12.

Although there is no perfectly fitting lemma or theorem for citing in, the inclusion in NP could have been inferred from [24], and the techniques used there are quite similar to ours. We re-proved it as we need the intermediate results below.

The hardness result uses a polynomial number of players. This raises a question if the complexity is better for a bounded number of up to k players.

We first assume that we are already provided with solutions to the 2MPGs to $\mathcal{M}$. To devise a decision procedure, we start with a simple observation:

Lemma 3.13: For a given MMPG $\mathcal{M}$ with k players and n vertices, there are at most $(n+1)^k$ many different thresholds in the related linear programs.

Proof: For each player p, there is either the threshold $r_p(v)$ for some vertex v of $\mathcal{M}$, or no restriction on the threshold at all in Part II of the constraint system of a linear program.

Consequently, we only have to consider the most liberal constraint systems.

Lemma 3.14: For a given MMPG $\mathcal{M}$ with k players and n vertices and a threshold as referred to in the proof of Lemma 3.13, it suffices to refer to up to n first parts of the constraint system of the linear programming problem.

Proof: For each Part II of the constraint system as referred to in the proof of Lemma 3.13, it is easy to determine the maximal set Q of nodes that can be visited. For this maximal Q, we can determine the strongly connected components S_1, $S_2, \ldots$ of $(V,E) \cap Q$ that are reachable from the initial vertex v_0. Obviously, there are at most n of them.

It is now easy to see that, for all Q', S' that define Part II of the constraint system, Q' is contained in Q and S' is contained in one SCC S_i from above. Now Q and S_i define a more liberal Part I of a constraint system than Q' and S'. Thus, every solution for Q' and S' is a solution for Q and S_i, too.

Thus, for a given k, there are only polynomially many linear programming problems to consider, and they are easy to construct. Solving linear programming problems requires only polynomial time [14], [13]. We thus obtain the following theorem.

Theorem 3.15: If we are provided with the solutions to the 2MPGs defined by an MMPG with a fixed number k of players, then we can determine an optimal solution in polynomial time.

Corollary 3.16: MMPGs with a fixed number of players can be solved in polynomial time by a machine with an oracle for solving two player zero-sum MPGs. If 2MPGs are solvable in polynomial time, so are MMPGs with a fixed number of players.

D. Reduction to two player mean-payoff games

Thus, finding optimal strategy profiles in MMPGs with a fixed number of players can be derived from solutions to 2MPGs. Various works have been published on solving 2MPGs. In [6], the authors give an improved pseudopolynomial procedure to solve two player mean-payoff games. [2] provides a randomised strongly subexponential and pseudopolynomial algorithm, and [12], [25] contain an UP∩CoUP algorithm for the respective decision problems. There are wilder reductions like one to symbolic linear programming [21] and a smoothed polynomial time complexity [3].

Corollary 3.16 therefore provides the following:

Corollary 3.17: MMPGs with a fixed number of players can be solved in UP∩coUP, in pseudo polynomial time, in smoothed polynomial time, and in randomised subexponential time.

IV. Discussion

The two main contributions of this paper are the introduction of leader equilibria and the concept of well behaved reward and punish strategy profiles as a technical foundation to them.

Well behaved reward and punish strategy profiles are general instruments for optimising the payoff of one player, while projecting away problems like the potential non-existence of limit average values. It is our belief that they will be useful in many related optimisation problems. The introduction of leader equilibria is a conceptual change to Nash equilibria, where an interested party overcomes the antinomy of Nash equilibria exemplified in Figure 6: the interested party (which we christened the leader) might improve her payoff by choosing a strategy, which is not stable for herself in the Nash sense of not being able to improve the payoff by unilaterally deviating from her strategy.

The concept of leader equilibria extends the set of rational control objectives. It allows, for example, for mixing environments that are rationally following their own objectives with a hostile environment. For such environments, it provides worst case *rational* results, where rational refers to the way the rational players behave: we assume that they follow a strategy that they do not have an incentive to deviate from.

The solutions one obtains can be used to create stable rules that optimise various outcomes, including social optima as well as egoistic solutions.

V. Future Work

The results of this paper can be applied to implement a tool to analyse the behaviour of distributed systems where a central controller along with other rational adversaries in a system follow their objectives of maximising the utility. This technique can be used to optimise social optima as well by adding an objective for the controller. Possible future work could be to implement the tool to represent this. It would relate to solving Multi-player mean-payoff games by reduction to the underlying two player mean-payoff games. It requires to implement the constraint system as described in Section III-C and the solution is therefore tractable.

References

[1] S. Aland, D. Dumrauf, M. Gairing, B. Monien, and F. Schoppmann. Exact price of anarchy for polynomial congestion games. *SIAM Journal of Computing*, 40(5):1211–1233, 2011.

[2] H. Björklund and S. Vorobyov. A combinatorial strongly subexponential strategy improvement algorithm for mean-payoff games. *Discrete Applied Math.*, 155(2):210–229, 2007.

[3] E. Boros, K. M. Elbassioni, M. Fouz, V. Gurvich, K. Makino, and B. Manthey. Stochastic mean-payoff games: Smoothed analysis and approximation schemes. In *Proceedings of ICALP 2011*, LNCS 6755, pages 147–158, 2011.

[4] T. Brihaye, V. Bruyère, and J. De Pril. Equilibria in quantitative reachability games. In *Proceedings of CSR 2010*, LNCS 6072, pages 72–83, 2010.

[5] T. Brihaye, J. De Pril, and S. Schewe. Multiplayer Cost Games with Simple Nash Equilibria. In *Proceedings of LFCS 2013*, LNCS 7734, pages 59–73. 2013.

[6] L. Brim, J. Chaloupka, L. Doyen, R. Gentilini, and J.-F. Raskin. Faster algorithms for mean-payoff games. *Formal Methods in System Design*, 38(2):97–118, 2011.

[7] R. Buehler, Z. Goldman, D. Liben-Nowell, Y. Pei, J. Quadri, A. Sharp, S. Taggart, T. Wexler, and K. Woods. The price of civil society. In *Proceedings of WINE*, pages 375–382, 2011.

[8] S. G. Canovas, P. Hansen, and B. Jaumard. Nash equilibria from the correlated equilibria viewpoint. *IGTR*, 1(1):33–44, 1999.

[9] K. Chatterjee, T. A. Henzinger, and M. Jurdzinski. Mean-payoff parity games. In *Proceedings of LICS 2005*, pages 178–187, 2005.

[10] P.-A. Chen and D. Kempe. Altruism, selfishness, and spite in traffic routing. In *Proceedings EC 2008*, pages 140–149, 2008.

[11] T.A. Henzinger. Quantitative reactive modeling and verification. *Computer Science - R&D*, 28(4):331–344, 2013.

[12] M. Jurdziński. Deciding the winner in parity games is in UP $\cap$ co-UP. *Information Processing Letters*, 68(3):119–124, November 1998.

[13] N. Karmarkar. A new polynomial-time algorithm for linear programming. In *Proceedings of STOC*, pages 302–311, 1984.

[14] L. G. Khachian. A polynomial algorithm in linear programming. *Dokl. Akad. Nauk SSSR*, 244:1093–1096, 1979.

[15] E. Koutsoupias and C. H. Papadimitriou. Worst-case equilibria. In *Proceedings of STACS 1999*, pages 404–413, 1999.

[16] E. Lehrer. Nash equilibria of n-player repeated games with semi-standard information. *International Journal of Game Theory*, 19(2):191–217, 1990.

[17] D. Meier, Y. A. Oswald, S. Schmid, and R. Wattenhofer. On the windfall of friendship: inoculation strategies on social networks. In *Proceedings of EC 2008*, pages 294–301, 2008.

[18] J. F. Nash. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, 36(1):48–49, 1950.

[19] M. J. Osborne and A. Rubinstein. *A course in game theory*. The MIT Press, Cambridge, USA, 1994. electronic edition.

[20] S. Schewe. An optimal strategy improvement algorithm for solving parity and payoff games. In *Proceedings of CSL 2008*, LNCS 5213, pages 368–383, 2008.

[21] S. Schewe. From parity and payoff games to linear programming. In *Proceedings of MFCS 2009*, LNCS 5734, pages 675–686, 2009.

[22] M. Ummels. Rational behaviour and strategy construction in infinite multi-player games. In *Proceedings of FSTTCS 2006*, pages 212–223, 2006.

[23] M. Ummels. The complexity of nash equilibria in infinite multi-player games. In *Proceedings of FoSSaCS 2008*, LNCS 4962, pages 20–34, 2008.

[24] M. Ummels and D. Wojtczak. The Complexity of Nash Equilibria in Limit-Average Games. In *Proceedings of CONCUR 2011*, pages 482-496, 2011.

[25] U. Zwick and M. S. Paterson. The complexity of mean-payoff games on graphs. *Theoretical Computer Science*, 158(1–2):343–359, 1996.

[26] F. Thuijsman and T. E. S. Raghavan. Perfect information stochastic games and related classes. In *Int. J. Game Theory*, pages 403-408, 1997.

[27] J. W. Friedman. A Non-cooperative Equilibrium for Supergames. In *The Review of Economic Studies*, pages 1–12, 1971.