Learning Equilibria in Repeated Congestion Games

Moshe Tennenholtz
Microsoft Israel R&D Center, Herzlia, Israel
and Faculty of Industrial Eng. and Management
Technion–Israel Institute of Technology
Haifa, Israel
moshet@microsoft.com

Aviv Zohar
School of Engineering and Computer Science
The Hebrew University of Jerusalem
Jerusalem, Israel and
Microsoft Israel R&D Center, Herzlia, Israel
avivz@cs.huji.ac.il

ABSTRACT

While the class of congestion games has been thoroughly studied in the multi-agent systems literature, settings with incomplete information have received relatively little attention. In this paper we consider a setting in which the cost functions of resources in the congestion game are initially unknown. The agents gather information about these cost functions through repeated interaction, and observations of costs they incur. In this context we consider the following requirement: the agents' algorithms should *themselves* be in equilibrium, *regardless* of the actual cost functions and should lead to an efficient outcome. We prove that this requirement is achievable for a broad class of games: repeated symmetric congestion games. Our results are applicable even when agents are somewhat limited in their capacity to monitor the actions of their counterparts, or when they are unable to determine the exact cost they incur from every resource. On the other hand, we show that there exist asymmetric congestion games for which no such equilibrium can be found, not even an inefficient one. Finally we consider equilibria with resistance to the deviation of more than one player and show that these do not exist even in repeated resource selection games.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*Multiagent Systems*; J.4 [**Social and Behavioral Sciences**]: Economics

General Terms

Theory, Economics.

Keywords

Learning Equilibrium, Congestion Games, Repeated Games.

1. INTRODUCTION

The general class of congestion games is known to model many real-world systems quite well. In congestion games, agents use resources which they are allowed to pick from a given set. The cost that is associated with each resource depends on the number of agents that use it. For example, in a transportation setting, roads can be thought of as resources that are being used by the drivers. The cost (travel time) of using a road is increased if other drivers have chosen to use it as well. Another example is advertisement. Advertisers can choose to place ads with different agencies or publishers. The effectiveness of these ads decreases and their price increases as more agents attempt to advertise in the same place.

An appealing property of congestion games is that if the costs of resources are common knowledge, the game is guaranteed to have a pure Nash equilibrium. However, in many scenarios the cost functions are initially unknown to the participating agents. One approach to deal with this uncertainty is by incorporating probabilistic assumptions about the cost functions. The agents are then often assumed to have common knowledge about the governing probability distributions – an assumption that is sometimes unrealistic. An alternative model makes no assumptions about the information possessed by agents at the beginning of the interaction. Instead, agents gather information by learning from past observations and then adjust their behavior accordingly.

The above learning process is carried out through repeated interactions. Indeed, interactions in multi-agent systems are often repeated. For example, drivers travel to work every weekday and can slowly accumulate information about congestion in different routes. As they obtain this information they may start to behave differently to minimize their travel time.

Game theory, among its many other goals, aspires to suggest a "reasonable" behavior for agents that are interacting in a strategic environment. Ideally we want a game-theoretic solution to have two main properties. The first property is optimality – if the agents follow the prescribed behavior the outcome should be efficient. The second property is stability. Stability guarantees that agents will indeed follow the prescribed behavior, as any agent who deviates from it can only lose. *Learning Equilibrium* [5] has been suggested as an equilibrium between learning algorithms employed by the players in a repeated setting. The equilibrium is achieved in an *ex-post* manner: a single player will not change its behavior even if it knows all the missing information.

In contrast to the full information setting, in which the Nash equilibrium is guaranteed to exist, it is uncommon to have an ex-post equilibrium in partial information settings, and a Learning Equilibrium is thus much more rare.

Cite as: Learning Equilibria in Repeated Congestion Games, Moshe Tennenholtz and Aviv Zohar, *Proc. of 8th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, Decker, Sichman, Sierra and Castelfranchi (eds.), May, 10–15, 2009, Budapest, Hungary, pp. XXX-XXX.

Our main result in this paper shows that a pure Learning Equilibrium exists in a relatively large class of games: the class of repeated symmetric congestion games. The equilibrium we demonstrate is efficient as it maximizes social welfare for the set of players, and uses no randomization (it is in pure strategies). With this result, we generalize a previous one that has shown the existence of (mixed strategy) Learning Equilibria in symmetric and monotonic resource selection games [3] and greatly extend the set of games for which Learning Equilibrium is known to exist[1].

The equilibrium relies on the fact that agents are able to see each other's actions and are able to observe the cost they themselves incur from selecting a specific resource within some bundle. We go on to show that even these assumptions can be relaxed: it is sufficient that agents see only the actions of players that have selected resources they themselves have selected. It is also enough if players observe only the total cost they endure from their selected bundle of resources, without any detail on which resource is responsible for any part of the cost. In the latter case, we demonstrate the existence of a mixed strategy Learning Equilibrium (in contrast to the pure strategy equilibrium we show in our other results).

We proceed to show that the case of asymmetric games is not as favorable in general. In contrast to the symmetric case, there are asymmetric repeated congestion games in which no Learning Equilibrium exists (not even an inefficient one). We also demonstrate that in some games it is impossible to reach an efficient solution that resists deviation by more than one player, even in the highly restricted setting of resource selection games.

Our work should be contrasted with the line of research that deals with convergence to a Nash equilibrium but without ensuring the stability of the convergence behavior itself (in our case it is the behavior of players that is in equilibrium and not necessarily the action profile to which they converge).

1.1 Structure of the Paper

The remainder of the paper is organized as follows: In Section 2 we briefly survey the related work. In Section 3 we formally define congestion games, the repeated games setting, and the Learning Equilibrium. We then go on to prove in Section 4 that symmetric congestion games have a pure strategy Learning Equilibrium. In Section 5 we extend this result to a case where agents have a more limited view of the actions taken by other players. Then we consider, in Section 6, the case in which agents can only observe the total cost they incur instead of the cost per resource, and show that an equilibrium exists there too. We then exhibit in Section 7 a result that shows that asymmetric congestion games may have no Learning Equilibrium. Finally, in Section 8 we show that there exist simple repeated congestion games that do not have an equilibrium that is resistant to deviations of more than one player.

2. RELATED WORK

Congestion games [20, 18] are central to work in CS/AI, Game Theory, Operations Research, and Economics. In particular, congestion games have been extensively discussed in the price of anarchy literature, e.g., [15]. Most of the work on congestion games assumes that all parameters of the game are commonly known, or at least that there is commonly known Bayesian information regarding the unknown parameters (see [11, 12]). However, in many situations, the game, and in particular the resource cost functions are unknown. When the game under discussion is played only once, one has to analyze it using solution concepts for games with incomplete information without probabilistic information (known also as pre-Bayesian games)[2]. Alternatively, if the game is played repeatedly the players may learn about the resource cost functions by observing the feedback for actions they performed in the past. This brings us to the study of reinforcement learning in (initially unknown) repeated congestion games [3].

Learning in the context of multi-agent interactions has attracted the attention of researchers in psychology, economics, artificial intelligence, and related fields for quite some time ([17, 14, 7, 4, 8, 13, 9, 10]). Much of this work uses repeated games as a model for such interactions. There are various definitions of what would define a satisfactory learning process. In this paper we adopt a most desirable and highly demanding requirement: we wish the players' learning algorithm to conform a *Learning Equilibrium* [5, 6, 2, 19] that leads to an economically efficient outcome. Other works also explore ex-post equilibria in settings that are somewhat different than a repeated game (but still with repeated interactions) [16, 21]. As illustrated in the above mentioned work, it is a highly attractive and non-trivial challenge to characterize general sets of games for which Learning Equilibria exist. One such result has been obtained for resource selection games [3]. This paper extends this study to the much broader and central class of symmetric congestion games.

3. PRELIMINARIES

We begin by defining a congestion game G.

DEFINITION 3.1. *Let $\mathcal{N} = \{1, \ldots, n\}$ be a set of players. Let $\mathcal{R}$ be a set of resources, and let $\Sigma_i \subseteq 2^{\mathcal{R}}$ be the set of allowed resource bundles that player i may select. Each resource $r \in \mathcal{R}$ has a cost function associated with it: $c_r : \mathbb{N} \to \mathbb{R}$ that describes the cost of the resource, as it depends on the number of players that use it. Each player i selects a subset of resources $\sigma_i \in \Sigma_i$ and then endures a cost that is the sum of costs from all the resources in its selected bundles:*

$$Cost_i = \sum_{r \in \sigma_i} c_r(k_r(\sigma)) \quad (1)$$

where we denote by $k_r(\sigma)$ the number of players that chose resource r in profile σ. To simplify notation, we will denote the cost attributed to resource r in a strategy profile σ by $c_r(\sigma)$ while keeping in mind that this still only depends on the number of agents that use resource r. We assume that the cost of each resource is bounded by below by some value

[1]Resource selection games are congestion games in which players are only allowed to select bundles of resources that consist of a single resource. Here we consider all congestion games and do not require monotonicity

[2]Such an analysis was done in [1] in a resource selection game, in which the number of participants is unknown.

[3]This study should be distinguished from that of best- and better-response dynamics that are known to converge to equilibrium in congestion games with complete information (see [18]).

L: $\forall r, k \quad L \leq c_r(k)$. *The congestion game G is then defined as the tuple* $G = (\mathcal{N}, \mathcal{R}, \{\Sigma_i\}_{i\in\mathcal{N}}, \{c_r\}_{r\in\mathcal{R}})$.

Symmetric congestion games are then defined as follows:

DEFINITION 3.2. *A congestion game is symmetric if all players have the same set of allowed bundles Σ. I.e.,*

$$\forall i \in \mathcal{N} \quad \Sigma_i = \Sigma$$

the game is then defined by the tuple $G = (\mathcal{N}, \mathcal{R}, \Sigma, \{c_r\}_{r\in\mathcal{R}})$.

Notice that much of the work in computer science deals with symmetric congestion games. In particular the class of resource selection games is a very restricted instance of symmetric congestion games in which resource bundles always contain a single resource.

We adopt the standard notation in game theory for a strategy profile: $\sigma \in (\Sigma_1 \times \ldots \times \Sigma_n)$. We denote by σ_{-i} a strategy profile of all players but the i'th player: $\sigma_{-i} \in (\Sigma_1 \times \ldots \times \Sigma_{i-1} \times \Sigma_{i+1} \times \ldots \times \Sigma_n)$ so that $\sigma = (\sigma_i, \sigma_{-i})$. We also extend this notation in a similar manner so that $\sigma_{-(i,j)}$ is a strategy profile of all players except players i and j.

3.1 Repeated Games and Learning Equilibria

Since we are interested in scenarios in which agents can learn about their environment, we will be looking at a repeated games setting. The players will be interacting for a predetermined number of rounds T. The setting is that of partial information: there is a set of possible states of the world $\mathcal{S}$ from which a specific state $S \in \mathcal{S}$ is selected. This state is unknown to the players, and in our case consists of the costs of each resource. At every round of the repeated game, players play the same congestion game, and have costs as we have defined above (given the state of the world). At the end of the repeated game, their total cost is the average obtained during all rounds of play[4].

We assume that players start the game without knowledge about the state – i.e., they do not know the cost functions of each resource. They only have information about the allowed bundles Σ and on the number of other players. The strategy of a player at a given round t will depend on its past observation history H_i^t. We denote by $\mathcal{H}$ the set of all possible histories. Thus a strategy s in the repeated game is a function from the set of all possible histories, to the set of resource bundles a player may choose from $s : \mathcal{H} \to \Sigma$. We will overload notation and denote by $Cost_i(s)$ the average cost for player i when the strategy profile s is played in the corresponding repeated game; in the case where mixed strategies are considered $Cost_i(s)$ will refer to the expected average cost for the player.

The exact history that is available to the players differs according to the exact scenario. In different cases players may be able to observe different things about the game.

DEFINITION 3.3. *We say that a repeated congestion game has* perfect monitoring *if each player is able to view the actions of all other players, and the cost he himself endured per resource. We shall say that a repeated congestion game has* imperfect monitoring *if a player can only observe his cost on each resource and the identity of other players who have selected resources that he uses and is unable to see the actions of other players on resources that he does not use at the time.*

[4]There are several possible alternatives to this formulation of the repeated game. For example, an infinite game can also be considered, with or without a discount factor on the payments at each round. These formulations lead to analogous results to those that we show in this paper.

We will mostly be interested in games with perfect monitoring. In the next section we show that these games have a pure $\epsilon-$Learning Equilibrium. We will then extend our results to games with imperfect monitoring, and later briefly discuss other possible limitations on the level of monitoring.

DEFINITION 3.4. *A strategy profile $s = (s_1, \ldots, s_n)$ of the players is considered an ϵ-Learning Equilibrium if a deviating player will not gain more than ϵ utility from deviating no matter what state of the world has been selected. That is,*

$$\forall i \in \mathcal{N} \quad \forall S \in \mathcal{S} \quad \forall s_i' \quad Cost_i(s_i, s_{-i}) < Cost_i(s_i', s_{-i}) + \epsilon$$

4. A LEARNING EQUILIBRIUM IN SYMMETRIC CONGESTION GAMES

In this section we shall describe an equilibrium strategy profile for players in a repeated symmetric congestion game. Notice that our result applies to general congestion games, in which the cost of each resource may increase or decrease as more players use it, and not only to monotonic games.

While the setting we examine is not cooperative, it is useful to observe the best cooperative solution that can be played. We denote by OPT the best aggregate social utility achievable if all players cooperate in the single shot congestion game. $OPT = \min_\sigma \sum_{i=1}^n Cost_i(\sigma)$

We now show that if the game is allowed to continue long enough, then we have an ϵ-equilibrium for any $\epsilon > 0$, and that this equilibrium can be as close as we want to the optimal social welfare (i.e., we are able to get close to the cooperative solution even in a non-cooperative partial information setting).

THEOREM 4.1. *Let G be a symmetric congestion game. For any $\epsilon \in \mathbb{R}$, $\epsilon > 0$ there exists $T \in \mathbb{N}$ such that a repeated game on G with perfect monitoring that lasts $t > T$ rounds has an ϵ-Learning Equilibrium* in pure strategies*, where the cost of each player is at most $\frac{OPT}{n} + \epsilon$.*

Before we prove the Theorem, we describe the equilibrium strategy itself. It consists of three phases:

1. Cooperative learning: In this phase players explore the costs of resources under different congestion conditions. A deterministic schedule in which every player experiences every resource under every possible congestion setting is selected and players perform their part in this schedule. If no player deviates, then by the end of this phase all of the values $c_r(k)$ are known by all players and the playing optimally phase begins. If any player deviates from the schedule at any point, then the learn-or-punish phase is initiated immediately (other players can detect this because they are able to observe each other's actions during play).

2. Playing optimally: In this phase, each player computes the strategy profile that yields the optimal social utility (ties between different profiles are broken according to a predetermined order). The players play this profile, while cycling through the different roles in it (i.e., they take turns using the different bundles this profile dictates). This guarantees each of them an equal share of the optimal payment (when

the game goes on long enough). This phase goes on until the game ends, or until some player deviates from the planned schedule, at which point the learn-or-punish phase begins.

3. Learn-or-punish: This phase is reached if any of the players has deviated from the planned sequence of actions in any of the previous phases[5]. It goes on indefinitely. Note, that there may be values of $c_r(k)$ that are still unknown to some or all of the $n-1$ honest players. The actions taken at this phase guarantee that one of the honest players either learns a previously unknown cost of a resource, or that the deviating player is punished. To do so, the $n-1$ honest players optimistically estimate the cost of every resource with an unknown cost $c_r(k)$. We define the optimistic estimate $\hat{c}_r(k)$ as follows. for $k = 1, 2, \ldots$

$$\hat{c}_r(k) = \begin{cases} c_r(k) & \text{if the value is known,} \\ L & \text{(the lower bound on costs) otherwise.} \end{cases} \quad (2)$$

The players then play a Nash equilibrium in the congestion game with only $n-1$ players (they ignore the existence of the n'th player). If one of the honest players observes a previously unknown value, in the next rounds it will signal the value it learned to the other players. This signaling is done through the bundles that this player selects in the following rounds (which is something that the other players can observe). After this signaling is complete, all players have shared knowledge regarding this new value and they resume playing the Nash equilibrium for $n-1$ players with newly calculated values of $\hat{c}_r(k)$. If no new information is learned, they continue to play the same Nash equilibrium indefinitely. We will show below that in this case, the n'th player is being punished.

Clearly, if all players play according to the proposed strategy and no one deviates, they all learn the costs associated with various resources and receive their share (up to some ϵ that is associated with the costs they endure during the learning phase) of the OPT/n payment. All that remains is to show that if one of the players deviates, he will have a higher cost.

We allow players to signal the values of $c_r(k)$ that they detect to each other, so that if one of the honest players learn a value, he can communicate it to the others. This can be done either through communication channels that they share (cheap talk) or through the actions that they select that are visible to the other players.

Formally, during the learn-or-punish phase, all honest players play a Nash equilibrium in the game

$$\hat{G} = (\mathcal{N} \setminus \{n\}, \mathcal{R}, \Sigma, \{\hat{c}_r\}_{r \in \mathcal{R}})$$

that has $n-1$ players and optimistic resource costs $\hat{c}_r$. We denote by $\hat{C}_i()$ the costs of players in the game $\hat{G}$. The following lemma demonstrates the idea that is at the heart of the learn-or-punish behavior:

LEMMA 4.2. *Assuming that the honest players play in the game G a Nash equilibrium that was computed according to the parameters of the game $\hat{G}$, then if the n'th player does not receive a lower payment than all other players, some players learns the value of a previously unknown resource cost function.*

[5]Since we are only interested in proving resilience to the deviation of one player, we do not describe the actions of players when more than a single player has deviated.

The intuition behind the proof of the lemma is that if the deviating player fairs better than some other player i in the game G, then this is because the deviator selected a bundle that has cheaper resources. Player i may be using some of these resources as well, and is paying a similar cost for this subset, therefore the difference must come from resources that the deviator and i did not both choose. Because the game is symmetric, player i could have chosen the bundle the deviator picked which would get him a lower cost even in the game $\hat{G}$. Since the bundle was not picked some of the items in i's current bundle are under-estimated. This only occurs if their exact value is unknown, and so player i learns something new. A more formal proof follows below:

PROOF OF LEMMA 4.2. Let σ be the strategy profile that is played in the congestion game. Without loss of generality, we assume that the deviating player is player n. The other $n-1$ players are following the prescribed behavior and are playing a Nash equilibrium σ_{-n} of $\hat{G}$ (only they play it in the game G in reality). I.e.,

$$\forall i \in \mathcal{N} \setminus \{n\} \quad \forall \sigma_i' \in \Sigma \quad \hat{C}_i(\sigma_i, \sigma_{-(n,i)}) \leq \hat{C}_i(\sigma_i', \sigma_{-(n,i)}) \quad (3)$$

Let us also assume that the n'th player has a strategy that costs him less than the cost attained by some other player i in the game G. I.e.,

$$\sum_{r \in \sigma_i} c_r(\sigma) > \sum_{r \in \sigma_n} c_r(\sigma) \quad (4)$$

For resources that both players n and i use, the cost is equal, and so the inequality must come from the resources that are not shared by both:

$$\sum_{r \in \sigma_i \setminus \sigma_n} c_r(\sigma) > \sum_{r \in \sigma_n \setminus \sigma_i} c_r(\sigma) \quad (5)$$

Furthermore, the unshared resources of players i, n have the same cost if the other player is removed from the game:

$$\forall r \in \sigma_i \setminus \sigma_n \qquad c_r(\sigma) = c_r(\sigma_i, \sigma_{-(n,i)}) \quad (6)$$

$$\forall r \in \sigma_n \setminus \sigma_i \qquad c_r(\sigma) = c_r(\sigma_n, \sigma_{-(n,i)}) \quad (7)$$

If we assume contrary to the lemma that player i learns nothing in this round of the game, then he must know all values $c_r(\sigma)$ for resources $r \in \sigma_i$. We therefore have:

$$\sum_{r \in \sigma_i \setminus \sigma_n} \hat{c}_r(\sigma_i, \sigma_{-(n,i)}) = \sum_{r \in \sigma_i \setminus \sigma_n} c_r(\sigma_i, \sigma_{-(n,i)}) > \\ > \sum_{r \in \sigma_n \setminus \sigma_i} c_r(\sigma_n, \sigma_{-(n,i)}) \geq \sum_{r \in \sigma_n \setminus \sigma_i} \hat{c}_r(\sigma_n, \sigma_{-(n,i)}) \quad (8)$$

For the shared resources between players i and n we also know:

$$\sum_{r \in \sigma_i \cap \sigma_n} \hat{c}_r(\sigma_i, \sigma_{-(n,i)}) = \sum_{r \in \sigma_i \cap \sigma_n} \hat{c}_r(\sigma_n, \sigma_{-(n,i)}) \quad (9)$$

Now, if we combine Equations 8 and 9 we get:

$$\sum_{r \in \sigma_i} \hat{c}_r(\sigma_i, \sigma_{-(n,i)}) > \sum_{r \in \sigma_n} \hat{c}_r(\sigma_n, \sigma_{-(n,i)}) \quad (10)$$

This contradicts the fact that σ_{-n} is a Nash equilibrium in $\hat{G}$, as the i'th player gains by switching to strategy σ_n. □

Now that we are armed with Lemma 4.2, we can proceed with the proof of Theorem 4.1:

PROOF SKETCH FOR THEOREM 4.1. If all players follow the equilibrium strategy, they have a cost of OPT/n (on average) once they start the playing optimally phase. This is preceded by the cooperative learning phase in which they suffer a higher cost. Note however that this more costly phase is of finite length and so we can select the length of the game T to be large enough so that their average cost is no larger than $OPT/n + \epsilon/2$.

Now, if a player deviates from the prescribed behavior, the other players immediately switch to the learn-or-punish behavior. From this point on, the deviating player will receive a payment that is no better than any other player. Note that the worst player always has a cost of at least OPT/n (Because OPT is the lowest social utility achievable). This happens in all rounds with the exception of a finite number of rounds in which the other players learn the values of previously unknown resources. I.e., the deviator has a finite number of rounds with a low cost, and all remaining rounds have a cost that is at least OPT/n. We can therefore set the number of rounds T to be large enough so that the deviating player suffers an average cost of at least $OPT/n - \epsilon/2$. This implies that the deviator does not gain more than ϵ in utility from the deviation. □

5. AN EQUILIBRIUM WITH IMPERFECT MONITORING

It is sometimes unreasonable to assume that a player is able to view the actions of all other players. For example, if our players are processes that are using resources such as CPUs, one could expect that each player could see who is using the same resources that he is using, but will be unaware of other actions. We show that there is a Learning Equilibrium even with imperfect monitoring.

THEOREM 5.1. *Let G be a symmetric congestion game. For any $\epsilon \in \mathbb{R}$, $\epsilon > 0$ there exists $T \in \mathbb{N}$ such that a repeated game on G with imperfect monitoring that lasts $t > T$ rounds has an ϵ-Learning Equilibrium* in pure strategies*, where the cost of each player is at most $\frac{OPT}{n} + \epsilon$.*

The main difficulty in proving this theorem is identifying which player has deviated, and then punishing him successfully. Our proof will use a strategy that ensures us that a deviating player will be identified by the others, or will otherwise be among a pair of suspect players and will still be punished.

PROOF SKETCH. The equilibrium strategy is very similar to that used in Theorem 4.1. Once again, we have several phases:

1. Cooperative Learning: Similarly to Theorem 4.1, players act according to a predetermined schedule that allows each player to choose every resource with every possible combination of loads. If some player notices a deviation by another (not all players always notice at the same time because of the limited monitoring), it moves to the blaming phase (that is described below). Otherwise, players move on to the playing optimally phase after everyone has learned every needed value.

2. Playing-Optimally: This phase is also similar to that in Theorem 4.1, and again, any player that notices a deviation moves to the blaming phase. Otherwise, this phase goes on indefinitely.

3. Blaming: In this phase players initially cause other players to notice a deviation (by selecting resources in a manner that will conflict with the scheduled tasks of others). Once all players are aware that a deviation by someone has occured, the players go on to signal to each other[6] which player they have seen deviating first (this deviator may be the original one, or just a player that has previously observed a deviation and signalled them), and at what time this deviation originally occured. Once each player has signalled to the other players who has deviated and when, they begin the learn-or-punish phase.

4. Learn-Or-Punish: This phase is reached after a blaming phase has been completed. At this point, all players have shared knowledge of the claims of players regarding deviations. Let us denote by i the player that has been reported (by another player) as the earliest deviator. If more than one player reports i as the deviator then he clearly must be guilty, and the remaining players play an equilibrium of $n-1$ players in the game just as in Theorem 4.1. Otherwise, only one player j has reported that i deviated. We consider both i and j as suspects. The $n-2$ players who are not suspects will play their predetermined roles in the Nash equilibrium for $n-1$ players. Players i and j will both be required to play the same role (of player $n-1$) in this equilibrium. This goes on until one of the players learns some new value, in which case he signals it to the other players. Notice that signalling to the other players about new values is a bit difficult, but a player that has something to signal, can notify others that he has something to signal to them by deviating in a manner that they can observe, and then signalling to them. We discuss more details about exactly how to signal below.

Notice, that if indeed we have only one deviator, and he is identified by one of the players, then he is always one of the players i, j. Either he is the earliest deviating player that caused the chain of deviations and triggered the blaming phase, or he tries to escape this by assigning blame to some other player that he claims has deviated earlier. Either way, one of the players i, j is the guilty party, so we can trust the $n-2$ other players to do their part in the learn-or-punish phase. Then, at least one of the players i, j has been falsely accused, and can be trusted to play the role required to complete the Nash equilibrium of $n-1$ players in the game. Therefore, as we have shown in Lemma 4.2, the deviating player will be punished, or one of the other players will learn a new and previously unknown value.

If the player who learns this new value is one of the $n-2$ trustworthy players, then that player can signal this to the others, otherwise, one of the trustworthy players switches roles with the role of the suspect player in the Nash equilibrium, and he is guaranteed to learn this new value, or he will be able to recognize the deviating player among the two suspects. He can then signal his findings to the others.

Since all signalling, and learning rounds are of a finite and bounded number, a sufficiently large number of rounds can be selected to guarantee that the deviator does not gain more than ϵ utility, for any positive ϵ. □

[6]The players can signal to each other by selecting the same bundle together, and then taking turns in communicating bits (conveyed as selecting the same bundle as the others or some other bundle)

6. NON-DETAILED MONITORING

At times, the exact contribution of a specific component in the system to the congestion is unknown, and only the total cost that is payed for a certain bundle can be observed. We define the following:

DEFINITION 6.1. *A repeated congestion game has* non-detailed monitoring *if players are able to observe the actions of their counterparts, but can only observe the sum of costs they themselves incur over different resources they chose and cannot observe in detail the exact cost of every resource.*

Is there a Learning Equilibrium in this restricted case of monitoring? Since players only observe the total cost of the bundle, it is not as simple to under-estimate the cost functions as we have done in the detailed-monitoring case. The following example shows that a pure strategy equilibrium is more complicated than in the detailed-monitoring case. If such an equilibrium exists, then at the very least, players will have to rotate between strategy profiles in order to punish.

EXAMPLE 6.1. *Let us define a repeated congestion game with non-detailed monitoring for two players with three resource bundles A, B, C. Each bundle contains resources as depicted in Figure 1(a) (the names of the resources match the bundles they appear in). Assume that there exists a pure-strategy Learning Equilibrium in this setting, and that player* 1 *is following it (player* 2 *will be our deviator). Now, since player* 1 *is playing a pure strategy σ_1, we can construct the deviation strategy as follows:*

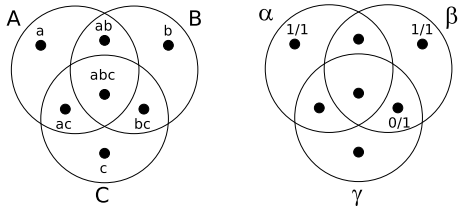
$$\textit{If } \sigma_1 \textit{ will pick bundle } \begin{cases} A, & \textit{pick } C; \\ B, & \textit{pick } A; \\ C, & \textit{pick } B. \end{cases}$$

I.e., in Figure 1(a), the deviator will always select the bundle that is counter clockwise from the one picked by player 1*. Now, assume that in this case, player* 1 *always observes a cost of* 1*. He can assign costs to resources in many ways.*

One possible assignment of costs is depicted in Figure 1(b) (where resources that do not have costs written next to them are assumed to have a cost of 0 when used by one or two players, and a cost denoted by x/y means a cost of x units for a single player and y units for two players).

Notice that the cost of the deviator may be $0, 1,$ *or* 2 *depending on which pair of bundles are picked. Therefore, if player 1 chooses correctly (resource bundle γ or β), he will guarantee that the deviator never gains. However, player 1 cannot distinguish which resource bundle among A, B, C matches α, β, γ. This is because all his observations are symmetric – both the game structure, and the costs he has seen.*

If player 1 *keeps rotating between bundles A, B, C equally, it can be shown that he will either observe a previously unseen value, or he will guarantee player* 2 *an average cost of* 1*, so a more complex punishment strategy may yet exist.*



(a) The structure of the game.

(b) A possible assignment of costs.

Figure 1: The game in Example 6.1.

At this point, we are not sure if games with non-detailed monitoring have a pure strategy Learning Equilibrium (but we conjecture that they do have one). However, we exhibit the following theorem, that guarantees the existence of a Learning Equilibrium in mixed strategies:

THEOREM 6.1. *Let G be a symmetric congestion game. For any $\epsilon \in \mathbb{R},\ \epsilon > 0$ there exists $T \in \mathbb{N}$ such that a repeated game on G with non-detailed monitoring that lasts $t > T$ rounds has an ϵ-Learning Equilibrium* in mixed strategies*, where the cost of each player is at most $\frac{OPT}{n} + \epsilon$.*

The proof relies on the technique presented in [6], where all symmetric 2 player games are shown to have a Learning Equilibrium (the authors extend the proof to more players, but require a correlation device to coordinate the actions of the players when they punish the deviator — we show that this is not needed in congestion games).

PROOF SKETCH. The equilibrium strategy is similar to that in Theorem 4.1, with the exception of the learn-or-punish phase. If all players cooperate, they can observe all costs for all possible combinations of actions, and then play optimally. If a deviating player is discovered at any point, the learn-or-punish phase that is described below is initiated. We assume without loss of generality, that the n'th player is the deviator. To describe the learn-or-punish phase we require the following definition:

DEFINITION 6.2. *Let Σ be the set of available resource bundles. We say that a bundle $\sigma_n \in \Sigma$ that belongs to the deviating player is* fully known *when the $n-1$ honest players know their costs for all possible action profiles:*

$$\forall i \neq n \quad \forall \sigma_{-n} \in \Sigma^{n-1} \quad Cost_i(\sigma_n, \sigma_{-n}) \textit{ is known.} \qquad (11)$$

Let $\mathcal{K}$ denote the set of fully known bundles at a given moment in time. Notice that the players are aware of all possible payments they can get if the deviator selects a bundle from $\mathcal{K}$, and more specifically, the costs in the sub-game in which only bundles from $\mathcal{K}$ are selected by any of the players are also known. This subgame is in fact a symmetric congestion game that is equivalent to: $G' = (\mathcal{N}, \mathcal{R}, \mathcal{K}, \{c_r\}_{r \in \mathcal{R}})$. In the learn-or-punish phase players take one of two actions:

1. An exploratory action: The player randomly selects a resource bundle from Σ according to the uniform distribution and plays that selection.

2. A punishment action: The player plays (in G) his role in the Nash equilibrium of $n-1$ players of the game $G' = (\mathcal{N}, \mathcal{R}, \mathcal{K}, \{c_r\}_{r \in \mathcal{R}})$.

The learn or punish phase proceeds as follows: The honest players begin by doing exploratory actions. They proceed until some bundle of the deviator becomes *fully known* (as in Definition 6.2). Once a bundle is fully known, each player performs an exploratory action with some small probability p, and performs a punishment action with probability $1-p$.

From Lemma 4.2, we know that if all the honest players play a Nash equilibrium of $n-1$ players in G', and the deviator also plays a strategy from this subgame, then the cost of

the deviator is no better than that of any other player. The probability p is set to be low enough to make sure that if the deviator keeps playing only bundles from the fully known set of bundles, it will be punished with a high enough probability. On the other hand, if the deviator plays bundles that are outside of the fully known set $\mathcal{K}$, then there is a small chance that the players will all perform exploratory actions and will learn the costs associated with a previously unknown action profile — this eventually leads to a larger set of fully known resource bundles. The length of the repeated game can be set to be long enough to ensure that players have enough time to learn (in expectation) all the unknown strategy profiles in the game, and then to punish the deviator long enough to ensure he does not gain more than ϵ from the deviation. □

7. ASYMMETRIC CONGESTION GAMES

As we have seen, symmetric congestion games posses a Learning Equilibrium. We relied heavily on the ability of players to chose the same bundle of resources as the deviating player and thus learn its true cost, or even punish the deviator by adding congestion to that bundle. What happens in asymmetric congestion games, when players have access to different bundles? We exhibit the following result:

THEOREM 7.1. *There exists a repeated asymmetric congestion game that has no Learning Equilibrium (not even in mixed strategies), even with perfect monitoring.*

PROOF. We shall show a simple congestion game with only 3 resources and 2 players in which an equilibrium does not exist. Let $\mathcal{R} = \{A, B, C\}$ and let the allowed bundles for the players be $\Sigma_1 = \{\{A\}, \{B\}\}$ and $\Sigma_2 = \{\{A\}, \{C\}\}$. We define the costs of the resources as follows:

$$\begin{aligned} c_A(1) = 0 \quad &; \quad c_A(2) = 1 \\ c_B(1) \in \{0.5, \alpha\} \quad &; \quad c_C(1) \in \{0.5, \alpha\} \end{aligned} \tag{12}$$

where resources B, C each have two possible costs for the case a single player chooses them, and $\alpha >> 1$ is some large cost. Notice, that since these resources are each accessible only by a single player, only this player can learn their cost. Thus, the cost of every player's privately accessible resource is in fact private. We will show that there is no possible Learning Equilibrium in a repeated game of this form. A Learning Equilibrium (if one existed) would have to provide us with an equilibrium strategy for each state in an ex-post fashion. That is, no matter which costs are selected for resources B and C, no player will deviate from the proposed strategy, even if he is aware of the exact state of nature.

We will describe the states of nature in this example using tuples of the form $(c_B(1), c_C(1))$. For example the state of nature $(0.5, \alpha)$ describes the case in which resource B costs 0.5, and resource C costs α.

We assume by contradiction that an equilibrium strategy profile σ exists. The following facts then apply:

CLAIM 7.2. *In the state $(\alpha, 0.5)$, if both players play an equilibrium strategy, then player 1 uses resource A during at least $1 - 1/\alpha$ of the time.*

PROOF OF CLAIM. Observe that player 1's cost from using resource A is at most $c_A(2) = 1$, while its cost when using resource B is exactly $c_B(1) = \alpha$. In equilibrium, player 1 cannot pay an average cost per round that is higher than 1 unit, because otherwise it would benefit him to deviate from this strategy and select resource A constantly, thereby paying less. If the average cost cannot exceed 1, then player 1 cannot choose resource B too often. Let ρ denote the fraction of the time that player 1 chooses resource B. If we optimally assume that player 1's cost whenever he uses resource A is 0, his cost is then:

$$1 \geq Cost_1 \geq \rho \cdot \alpha + (1 - \rho) \cdot 0 \tag{13}$$

Which implies $\rho \leq \frac{1}{\alpha}$. □

CLAIM 7.3. *In the state $(\alpha, 0.5)$, when both players follow the equilibrium strategy, it cannot be that player 2 selects resource A more than $2/\alpha$ of the time.*

PROOF OF CLAIM. Player 2 has a fall-back strategy that will allow him to receive a payment of 0.5 every round, regardless of the actions of the other player. His equilibrium strategy must therefore yield a payment that is no smaller. According to Claim 7.2, We know that player 1 does not access resource A at most $\frac{1}{\alpha}$ of the time. We denote by γ the fraction of the time in which player 2 accesses resource A together with player 1. If we assume that player 2 selects resource A whenever player 1 does not, we can bound the cost of player 2 as follows:

$$\frac{1}{\alpha} \cdot 0 + \gamma \cdot 1 + (1 - \gamma - \frac{1}{\alpha}) \cdot 0.5 \leq Cost_2 \leq 0.5 \tag{14}$$

which implies $\gamma \leq \frac{1}{\alpha}$. Therefore, for the remainder of the time, player 1 accesses resource A alone. I.e., for a period of at least $1 - \frac{2}{\alpha}$ he suffers 0 cost. □

Now, to conclude our proof and reach a contradiction, we shall examine the behavior of the players in equilibrium in state $(0.5, 0.5)$. Notice, that at most one player can occupy resource A alone at any given round, and so at least one player access resource A alone less than half of the time. Without loss of generality, we assume that this player is player 1. If we optimistically assume that player 1 managed to avoid selecting resource A at the same time as player 2 we obtain the following bound on his cost:

$$Cost'_1 \geq 0.5 \cdot 0 + 0.5 \cdot 0.5 = 0.25 \tag{15}$$

However, if that player deviates, it can gain a better payoff. All it has to do is play *as if* the state of the world is $(\alpha, 0.5)$. In that case he has exclusive access to resource A for the majority of the time (according to claims 7.2 and 7.3), and his cost is thus less than $\frac{2}{\alpha} \cdot 1$. This is a contradiction to our assumption that an equilibrium strategy profile exists.

8. STRONG EQUILIBRIA IN REPEATED CONGESTION GAMES

It is sometimes possible to find an equilibrium strategy that resists deviation by more than one player. We show here that this does not hold for the general class of repeated congestion games. In fact, there exist very simple repeated resource selection games (with complete information) where there always exists a coalition of players that can deviate to a profile that is strictly better for *all* players in the coalition.

THEOREM 8.1. *There exists a repeated resource selection game with no equilibrium that resists deviations by more than one player – even in a full information setting.*

PROOF. We give an example of such a game with 3 players and 2 resources. Let the set of players be $\mathcal{N} = \{1, 2, 3\}$ and the of resources be $\mathcal{R} = \{a, b\}$. The game is a symmetric resource selection game, that is, the allowed resources bundles for each player are $\Sigma = \{\{a\}, \{b\}\}$. The cost functions of the resources are: $c_a(1) = c_b(1) = 1$; $c_a(2) = c_b(2) = 2$; $c_a(3) = c_b(3) = 2$. The congestion game that these define has a minimal cost when two of the players select the same resource, and the third player selects a different resource—a total cost of 5. Any strategy profile s for the repeated game will have a total cost that is at least as high:

$$\sum_{i=1}^{3} Cost_i(s) \geq 5 \tag{16}$$

Now, we shall prove that in any strategy profile s, a coalition of two players can benefit from deviating. First we claim that there exists a coalition $T \in \{\{1, 2\}, \{2, 3\}, \{3, 1\}\}$ so that both players in the coalition have a cost of 1.5 or higher and at least one of the players pays strictly more. If two out of the three players have a cost below 1.5 then Equation 16 implies that the third player's cost is higher than 2 – this is impossible since the highest cost in the game is 2, and so at most one player gets a cost of 1.5 or less. It is also impossible that all three players pay a cost of 1.5 or less, since the total cost would then be only 4.5 and this contradicts Equation 16. Therefore, a player that pays more than 1.5 exists.

Next we shall show that this coalition of two players can gain by deviating to a different strategy profile in which the total expected cost of each player is lower. The strategy is as follows: At every round, The two deviating players will each choose a different resource (i.e., one will select resource a, and the other will select resource b). It is easy to see that since the third player occupies one of the resources, one of the deviators will pay a cost of 2, and the other will pay 1. I.e., their average cost will always be 1.5. (which is lower than their average cost if they do not deviate).

In order to make sure that both players in the coalition gain in expectation, they can try to distribute the costs among them in the following manner: the third player (who did not deviate) decides according to the equilibrium strategy which resource to chose. His choice may be non deterministic, so he may only assign a probability to each resource selection. The other two players can thus decide which of the two resources each of them will pick. one of them will have a higher chance of getting a cost of 2 (depending on the randomized selection of the honest player). In this manner, over many rounds they can attempt to achieve an average payment of $1.5 + \epsilon$ for one of the players, and $1.5 - \epsilon$ for the other player. If ϵ is chosen to be small enough, both deviating players gain in expectation from the deviation. □

9. REFERENCES

[1] I. Ashlagi, D. Monderer, and M. Tennenholtz. Resource Selection Games with Unknown Number of Players. In *Proc. of AAMAS06*, pages 819–825, 2006.

[2] I. Ashlagi, D. Monderer, and M. Tennenholtz. Robust learning equilibrium. In *Proc. of UAI06*, pages 34–41. AUAI Press, 2006.

[3] I. Ashlagi, D. Monderer, and M. Tennenholtz. Learning equilibrium in resource selection games. In *Proc. of AAAI07*, pages 18–23, 2007.

[4] M. Bowling and M. Veloso. Rational and convergent learning in stochastic games. In *Proc. 17th IJCAI*, pages 1021–1026, 2001.

[5] R. Brafman and M. Tennenholtz. Efficient Learning Equilibrium. *Artificial Intelligence*, 159(1-2):27–47, 2004.

[6] R. Brafman and M. Tennenholtz. Optimal Efficient Learning Equilibrium: Imperfect Monitoring. In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI) 2005*, 2005.

[7] R. I. Brafman and M. Tennenholtz. R-max – a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3:213–231, 2002.

[8] V. Conitzer and T. Sandholm. Awesome: a general multiagent learning algorithm that converges in self-play and learns best-response against stationary opponents. *Machine Learning*, 67(1-2):23–43, 2007.

[9] I. Erev and A. Roth. Predicting how people play games: Reinforcement learning in games with unique strategy equilibrium. *American Economic Review*, 88:848–881, 1998.

[10] D. Fudenberg and D. Levine. *The theory of learning in games*. MIT Press, 1998.

[11] M. Gairing, B. Monien, and K. Tiemnann. Selfish routing with incomplete information. In *Proc. of SPAA05*, pages 203–212, 2005.

[12] D. Garg and Y. Narahari. Price of Anarchy of Network Routing Games with Incomplete Information. In *Proc. of 1st Workshop on Internet and Network Economic, Springer Verlag LNCS series 3828*, pages 1066–1075, 2005.

[13] A. Greenwald, K. Hall, and R. Serrano. Correlated q-learning. In *NIPS workshop on multi-agent learning*, 2002.

[14] J. Hu and M. Wellman. Multi-agent reinforcement learning: Theoretical framework and an algorithms. In *Proc. 15th ICML*, 1998.

[15] E. Koutsoupias and C. Papadimitriou. Worst-Case Equilibria. In *Proceedings of the 16th Annual Symposium on Theoretical Aspects of Computer Science*, pages 404–413, 1999.

[16] H. Levin, M. Schapira, and A. Zohar. Interdomain routing and games. In *Proc. of STOC '08*, pages 57–66, New York, NY, USA, 2008. ACM.

[17] M. L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proc. 11th ICML*, pages 157–163, 1994.

[18] D. Monderer and L. Shapley. Potential Games. *Games and Economic Behavior*, 14:124–143, 1996.

[19] D. Monderer and M. Tennenholtz. Learning equilibrium as a generalization of learning to optimize. *Artifcial Intelligence*, 171(7):448–452, 2007.

[20] R. Rosenthal. A Class of Games Possessing Pure-Strategy Nash Equilibria. *International Journal of Game Theory*, 2:65–67, 1973.

[21] J. Shneidman and D. C. Parkes. Specification faithfulness in networks with rational nodes. In *Proc. of PODC '04*, pages 88–97, New York, NY, USA, 2004. ACM.