Fear in Mediation: Exploiting the "Windfall of Malice"

J. Díaz * D. Mitsche * N. Rustagi † J. Saia †

Abstract

We consider a problem at the intersection of distributed computing and game theory, namely: Is it possible to achieve the "windfall of malice" even without the actual presence of malicious players? Our answer to this question is "Yes and No". Our positive result is that for the virus inoculation game, it is possible to achieve the windfall of malice by use of a mediator. Our negative result is that for congestion games that are known to have a windfall of malice, it is not possible to design a mediator that achieves this windfall. In proving these two results, we develop novel techniques for mediator design that we believe will be helpful for creating non-trivial mediators to improve social welfare in a large class of games.

1 Introduction

Recent results show that malicious players in a game may, counter-intuitively, improve social welfare [16, 8, 13, 18, 14]. For example, Moscibroda, Schmidt and Wattenhofer show that for a virus inoculation game, the existence of malicious players, who may lie about the action they perform, will actually lead to better social welfare for the remaining players than if such malicious players are absent [16] . This improvement in the social welfare with malicious players has been referred to as the "windfall of malice" [8]. The existence of the windfall of malice for some games leads to an intriguing question: Can we achieve the windfall of malice even without the actual presence of malicious players?

In this paper, we show that the answer to this question is sometimes "Yes". How do we achieve the beneficial impact of malicious players without their actual presence? Our approach is to use a mediator. Informally, a mediator is a trusted third party that suggests actions to each player. The players retain free will and can ignore the mediator's suggestions. The mediator proposes actions privately to each player, but the algorithm the mediator uses to decide what to propose is public knowledge. The contributions of this paper are threefold.

- We introduce a general technique for designing mediators that is inspired by careful study of the "windfall of malice" effect. In our approach, the mediator

*Llenguatges i Sistemes Informàtics, UPC, 08034 Barcelona; email: {diaz,dmitsche}@lsi.upc.edu

†Department of Computer Science, University of New Mexico; email: {rustagi,saia}@cs.unm.edu. This research was partially supported by NSF CAREER Award 0644058, NSF CCR-0313160, and an AFOSR MURI grant.

makes a random choice of one of two possible configurations, where a configuration is just a set of proposed actions for each player. The first configuration is optimal: the mediator proposes a set of actions that achieves the social optimum (or very close to it). The second configuration is "fear inducing": the mediator proposes a set of actions that leads to catastrophic failure for those players who do not heed the mediators advice. The purpose of the second configuration is to ensure that the players follow the advice of the mediator when the optimal configuration is chosen. Thus, the random choice of which configuration is chosen must be hidden from the players.

- We show the applicability of our technique by using it to design mediators for two games. First, we design a mediator for the virus inoculation game from [16], that achieves a social welfare that is asymptotically optimal. Second, we design a mediator for a variant of the El Farol game [3, 12, 10, 15] that improves the social welfare over the best Nash equilibria. Surprisingly, our technique works for the El Farol game, even though this game does not have a windfall of malice.

- We show the limits of our technique by proving an impossibility result that shows that for a large class of games, no mediator will improve the social welfare over the best Nash equilibria. In particular, this impossibility result holds for the congestion games that Babaioff, Kleinberg and Papadimitriou show have a windfall of malice [8]. Thus, we show that some games with a windfall of malice effect can not be improved by the use of a mediator.

1.1 Related Work

The concept of a mediator is closely related to that of a correlated equilibrium, which was introduced by Aumann in [7]. In particular, if a mediator proposes actions to the players such that it is in the best interest of each player to follow the mediators proposal, then the mediator is said to implement a correlated equilibrium. There are several recent results on correlated equilibria and mediators. Papadimitriou and Roughgarden [17] give polynomial time algorithms that can optimize over correlated equilbria, via a linear programming approach, for a large class of multiplayer games that are "succinctly representable" in the sense that the set of possible strategy vectors over all players is polynomial. Christodoulou and Koutsoupias [11] study the price of anarchy and stability in congestion games where each edge has a linear cost function with positive coefficients. They show that in such a setting, the price of anarchy for pure equilibria is almost the same as the price of anarchy of correlated equilibria: a difference of no more than 1.4%. Balcan et al. [9], describe techniques for moving from a high cost Nash equilibrium to a low cost Nash equilibrium via a "public service advertising campaign". They show that in many games, even if not all players follow instructions, it is possible to ensure such a move . While their result does not explicitly consider mediators, it is similar in flavor to ours in the sense that an outside third party is acting to improve social welfare.

A major motivation of our use of a mediator is recent work by Abraham et al. [1, 2]. Their work shows that it is possible to implement mediators just by having the players

talk amongst themselves ("cheap talk"). In other words, there exists a distributed algorithm for talking among the players that enables the simulation of a mediator. Moreover, Abraham et al. show it is possible to achieve this in a robust manner, even with up to linear size coalitions and up to a constant fraction of adversarial players.

Several recent results study the use of mediators that may act on behalf of a player [4, 19, 21, 20]. In other words, these results consider the situation where if a player decides to use the mediator, it first communicates any relevant information to the mediator and then the mediator acts for the player, without the player having the opportunity to change the mediators action.

Paper Organization: Section 1.2, below, gives basic definitions. Next, Section 2 describes an asymptotically optimal mediator for the virus inoculation game. Section 3 states and proves our impossibility result. Section 4 describes our result for the El Farol game, and shows how this game is similar to virus inoculation. Finally, Section 5 concludes and gives open problems.

1.2 Basic definitions and notation

A *correlated equilibrium* is a probability distribution over strategy vectors that ensures that no player has incentive to deviate. We define a *configuration* for a given game to be a vector of pure strategies for that game, one for each player. We define a *mediator* for a game to be a probability distribution $\mathcal{D}(\mathcal{C})$ over a finite set of different configurations $\mathcal{C}$. The set of configurations $\mathcal{C}$ and the distribution $\mathcal{D}(\mathcal{C})$ are known to all players. However, the actual configuration chosen is unknown, and the advice the mediator gives to a particular player based on the chosen configuration is known only to that player. We say that a mediator is *valid* if all players are incentivized to follow its advice. In this case, the mediator implements a correlated equilibrium. From a distributed computing viewpoint, the major difference between a correlated equilibria and a Nash equilibria is that in a correlated equilibria, players share a global coin, but in a Nash equilibria, players only have access to private coins.

Throughout this paper, we will only consider mediators that treat all players equally, i.e., once having decided (by a random experiment according to $\mathcal{D}(\mathcal{C})$) which is the configuration the mediator is choosing from, all players have the same probability to be proposed a particular strategy. Also, throughout the paper we assume that the number of strategic players, n, is very large (tending to infinity). Finally, we will use the notation $a(n) \sim b(n)$ if $a(n) = b(n)(1 \pm o(1))$. We also use the notation $[n] = \{1, \ldots, n\}$.

2 Virus Inoculation Game

We now describe the *virus inoculation game* from [16, 6]. There are n players, each corresponding to a node in a square grid G. Each player has two choices: either to inoculate itself (at a cost of 1) or to do nothing and risk infection (which costs L). After the decision of the nodes to inoculate or not, one node selected uniformly at random is infected with a virus. A node v that chooses not to inoculate gets infected by the virus if either the virus starts at v or the virus starts at another node v' and there is a path of not inoculated nodes connecting v and v'.

We define the *attack graph* G_a to be the graph induced on G by the set of all nodes that do not inoculate. Aspnes et al. [5] proved that in a pure Nash equilibrium every component of the attack graph has size n/L. The social welfare achieved in such an equilibria is thus $\Theta(n)$. However, Moscibroda et al. [16] proved that the minimum social cost is $\Theta(n^{2/3}L^{1/3})$ for the grid, which occurs when the components in G_a are of size $(n/L)^{2/3}$. Moreover, they show that the existence of enough Byzantine players, who can never be trusted to inoculate, ensures that the social welfare of any Nash equilibria is slightly better than $\Theta(n)$.

Based on the result from [16], we observe that the main problem in this game is that the individual players do not have enough fear of being infected. In particular, they are unable to achieve the optimal social welfare because they form connected components in G_a that are too large. Thus, we design a mediator that randomly chooses between two configurations (see Figure 1). The first configuration is optimal: all components in G_a are of size $(n/L)^{2/3}$. The second configuration is "fear inducing": any node that does not inoculate in this configuration has probability about $1/2$ of being infected. The only purpose of the second configuration is to ensure that the selfish players follow the advice of the mediator when the optimal configuration is chosen.

Clearly, we only want to choose the fear inducing configuration with very small probability. The critical fact that enables us to do this is the fact that for a given player, when that player is advised to inoculate, the posterior probability that the mediator is in the second configuration increases significantly over the prior probability. This is the case because so many more nodes are told to inoculate in the second configuration. Thus, players that are told to inoculate are more likely to be infected. Finally, we also note that nodes that are told not to inoculate are more likely to be in the first configuration and thus not to be attacked.

We now formally describe the mediator for this game.[1] The mediator will choose randomly between one of the following two configurations C_1 and C_2.

Configuration C_1: The mediator proposes a pattern of inoculation such that 1) all nodes that do not inoculate are in one giant component in G_a; 2) each node has equal probability of being chosen to inoculate; and 3) the probability that a fixed node inoculates is $\frac{1}{2} - \frac{1}{2\sqrt{n}}$. The mediator accomplishes this in the following manner:

1. The mediator flips a coin. If it comes up heads, it proposes that all nodes in even columns do not inoculate. If it comes up tails, it proposes that all nodes in odd columns do not inoculate.

2. The mediator chooses a random integer, x, uniformly between 1 and $\sqrt{n}$. For each of the columns that have not already been told not to inoculate, the mediator proposes that each node in that column inoculate except for the x-th node in that column.

[1] For ease of analysis, we assume that both $\sqrt{n}$ and $(\frac{n}{L})^{1/3}$ are integers. Also, $\sqrt{n}$ should be an integer multiple of $(\frac{n}{L})^{1/3}$ (this assumption can be removed easily without effecting our asymptotic results)

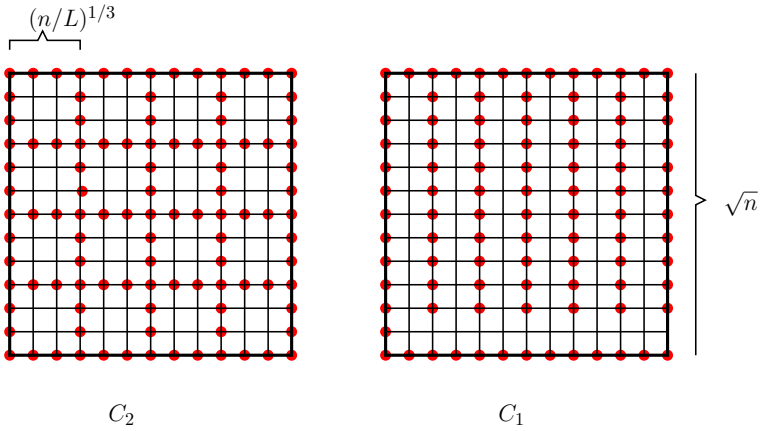

Figure 1: The $\sqrt{n} \times \sqrt{n}$ grid with two configurations C_1, C_2 for the virus inoculation game

Configuration C_2: The mediator proposes a pattern of inoculation that ensures that 1) each component in G_a is of size no more than $(\frac{n}{L})^{2/3}$; 2) each node is chosen to inoculate with equal probability; and 3) the probability that a fixed node inoculates is at most $2(L/n)^{1/3}$. It does this as follows.

1. The mediator chooses integer x uniformly at random in the range 1 to $(n/L)^{1/3}$.
2. For every node v in row r and column c, if one of the following two conditions hold, the mediator proposes v to inoculate: 1) $r \equiv x \mod (n/L)^{1/3}$; or 2) $c \equiv x \mod (n/L)^{1/3}$. Otherwise the mediator tells v not to inoculate.

For these two configurations C_1 and C_2 we now define the probability distribution $\mathcal{D}(\{C_1, C_2\})$ with $p_1 = cL^{-2/3}n^{-1/3}$ and $p_2 = (1 - cL^{-2/3}n^{-1/3})$, where $c > 0$ can be chosen to be any small constant satisfying $c > 2L/(L-1)$ (in particular $c = 4$ always suffices).

We can now prove the main theorem of this section which shows that $\mathcal{D}(\{C_1, C_2\})$ is asymptotically optimal.

Theorem 2.1. *$\mathcal{D}(\{C_1, C_2\})$ is a mediator with social welfare $\Theta(n^{2/3}L^{1/3})$.*

Proof. To prove the statement, we need a few definitions. Define by $\mathcal{E}_I^j$ the event that the mediator advises player j to inoculate and define by $\mathcal{E}_{\bar{I}}^j$ the event that the mediator advises player j not to inoculate. Since all players are to be treated equally by the mediator, we will omit the index j. Define also by $\mathcal{E}_A$ the event that a not inoculated node gets infected by the virus, and denote by $\mathcal{C}_A$ the infection cost of a not inoculated node. We also use the notation $\mathcal{C}_I$ to denote the cost of inoculation (clearly $\mathcal{C}_I = 1$). We first need to show that $\mathcal{D}(\{C_1, C_2\})$ indeed yields a mediator. That is, we have to verify the following conditions of a correlated Nash equilibrium:

$$\begin{aligned} \mathbf{E}\left[\mathcal{C}_A|\mathcal{E}_I\right] &\geq \mathbf{E}\left[\mathcal{C}_I|\mathcal{E}_I\right] = 1 \\ \mathbf{E}\left[\mathcal{C}_A|\mathcal{E}_{\bar{I}}\right] &\leq \mathbf{E}\left[\mathcal{C}_I|\mathcal{E}_{\bar{I}}\right] = 1, \end{aligned}$$

which is equivalent to showing that

$$\begin{aligned} \mathbf{Pr}\left(\mathcal{E}_A|\mathcal{E}_I\right) &\geq 1/L \quad (1)\\ \mathbf{Pr}\left(\mathcal{E}_A|\mathcal{E}_{\bar{I}}\right) &\leq 1/L, \quad (2)\end{aligned}$$

since for any event $\mathcal{E}$ with $\mathbf{Pr}\left(\mathcal{E}\right) > 0$, we have that $\mathbf{E}\left[\mathcal{C}_A|\mathcal{E}\right] = L\mathbf{Pr}\left(\mathcal{E}_A|\mathcal{E}\right)$. We denote furthermore by $\mathcal{E}_i$, $i = 1, 2$, the event that configuration C_i, $i = 1, 2$ is chosen. Note that $\mathbf{Pr}\left(\mathcal{E}_A|\mathcal{E}_1\right) = 1$. To prove (1), first observe that

$$\begin{aligned} \mathbf{Pr}\left(\mathcal{E}_1|\mathcal{E}_I\right) &= \mathbf{Pr}\left(\mathcal{E}_1, \mathcal{E}_I\right)/\mathbf{Pr}\left(\mathcal{E}_I\right)\\ &\sim \frac{p_1(1/2 - 1/(2\sqrt{n}))}{p_1(1/2 - 1/(2\sqrt{n})) + 2p_2(L/n)^{1/3}},\end{aligned}$$

and similarly for $\mathbf{Pr}\left(\mathcal{E}_2|\mathcal{E}_I\right)$. Now, plugging in the values of p_1, p_2 and using that $L \in o(n)$ we get [2]

$$\begin{aligned} \mathbf{Pr}\left(\mathcal{E}_A|\mathcal{E}_I\right) &= \mathbf{Pr}\left(\mathcal{E}_A, \mathcal{E}_1|\mathcal{E}_I\right) + \mathbf{Pr}\left(\mathcal{E}_A, \mathcal{E}_2|\mathcal{E}_I\right)\\ &= \mathbf{Pr}\left(\mathcal{E}_A|\mathcal{E}_1, \mathcal{E}_I\right)\mathbf{Pr}\left(\mathcal{E}_1|\mathcal{E}_I\right) + \mathbf{Pr}\left(\mathcal{E}_A|\mathcal{E}_2, \mathcal{E}_I\right)\mathbf{Pr}\left(\mathcal{E}_2|\mathcal{E}_I\right)\\ &\geq \mathbf{Pr}\left(\mathcal{E}_1|\mathcal{E}_I\right) + \frac{1}{L^{2/3}n^{1/3}}\mathbf{Pr}\left(\mathcal{E}_2|\mathcal{E}_I\right)\\ &\sim \frac{p_1(1/2 - 1/(2\sqrt{n}))}{p_1(1/2 - 1/(2\sqrt{n})) + 2p_2(L/n)^{1/3}}\\ &+ (L^{-2/3}n^{-1/3})\frac{2p_2(L/n)^{1/3}}{p_1(1/2 - 1/(2\sqrt{n})) + 2p_2(L/n)^{1/3}}\\ &\sim \frac{(c/2)L^{-2/3}n^{-1/3} + 2L^{-1/3}n^{-2/3}}{(c/2)L^{-2/3}n^{-1/3} + 2(L/n)^{1/3}}\\ &= \frac{2cL^{2/3}n^{2/3} + 4Ln^{1/3}}{2cL^{2/3}n^{2/3} + 4L^{5/3}n^{2/3}}\\ &\sim \frac{c}{c + 2L},\end{aligned}$$

which is greater than $1/L$ for $c > (2L)/(L - 1)$. Similarly, to prove (2), note that

$$\begin{aligned} \mathbf{Pr}\left(\mathcal{E}_1|\mathcal{E}_{\bar{I}}\right) &= \mathbf{Pr}\left(\mathcal{E}_1, \mathcal{E}_{\bar{I}}\right)/\mathbf{Pr}\left(\mathcal{E}_{\bar{I}}\right)\\ &\sim \frac{p_1(1/2 + 1/(2\sqrt{n}))}{p_1(1/2 + 1/(2\sqrt{n})) + p_2(1 - 2(L/n)^{1/3})},\end{aligned}$$

[2] if $L = \theta(n)$, then any pure Nash equilibria is trivially asymptoticaly optimal

and analogously for $\mathbf{Pr}(\mathcal{E}_2|\mathcal{E}_{\bar{I}})$. Hence,

$$\begin{aligned}
\mathbf{Pr}(\mathcal{E}_A|\mathcal{E}_{\bar{I}}) &= \mathbf{Pr}(\mathcal{E}_A,\mathcal{E}_1|\mathcal{E}_{\bar{I}}) + \mathbf{Pr}(\mathcal{E}_A,\mathcal{E}_2|\mathcal{E}_{\bar{I}}) \\
&= \mathbf{Pr}(\mathcal{E}_A|\mathcal{E}_1,\mathcal{E}_{\bar{I}})\mathbf{Pr}(\mathcal{E}_1|\mathcal{E}_{\bar{I}}) + \mathbf{Pr}(\mathcal{E}_A|\mathcal{E}_2,\mathcal{E}_{\bar{I}})\mathbf{Pr}(\mathcal{E}_2|\mathcal{E}_{\bar{I}}) \\
&\leq \mathbf{Pr}(\mathcal{E}_1|\mathcal{E}_{\bar{I}}) + \frac{1}{L^{2/3}n^{1/3}}\mathbf{Pr}(\mathcal{E}_2|\mathcal{E}_{\bar{I}}) \\
&\sim \frac{p_1(1/2+1/(2\sqrt{n}))}{p_1(1/2+1/(2\sqrt{n}))+p_2(1-2(L/n)^{1/3})} \\
&+ (L^{-2/3}n^{-1/3})\frac{p_2(1-2(L/n)^{1/3})}{p_1(1/2+1/(2\sqrt{n}))+p_2(1-2(L/n)^{1/3})} \\
&\sim \frac{(c/2)L^{-2/3}n^{-1/3}+L^{-2/3}n^{-1/3}}{(c/2)L^{-2/3}n^{-1/3}+1} \\
&\sim \frac{c+2}{2L^{2/3}n^{1/3}},
\end{aligned}$$

which is smaller than $1/L$ since $L \in o(n)$. Thus, we have shown that $\mathcal{D}(\{C_1, C_2\})$ indeed is a valid mediator in that players will follow its advice. We next compute the social cost for this mediator. Let $\mathcal{I}_1$ ($\bar{\mathcal{I}}_1$) be the set of nodes that inoculate (respectively do not inoculate) in C_1, and let $\mathcal{I}_2$ ($\bar{\mathcal{I}}_2$) be the set of nodes that inoculate (respectively do not inoculate) in C_2. Then the social cost for the mediator can be written as

$$\begin{aligned}
& p_1(|\mathcal{I}_1| + \sum_{v\in\bar{\mathcal{I}}_1} L\mathbf{Pr}(\mathcal{E}_A|\mathcal{E}_1,\mathcal{E}_{\bar{I}})) + p_2(|\mathcal{I}_2| + \sum_{v\in\bar{\mathcal{I}}_2} L\mathbf{Pr}(\mathcal{E}_A|\mathcal{E}_2,\mathcal{E}_{\bar{I}})) \\
\sim\ & \frac{c}{L^{2/3}n^{1/3}}(n/2+(n/2)L) + (2n^{2/3}L^{1/3} + nL\frac{1}{L^{2/3}n^{1/3}}) \\
=\ & (3+(c/2))n^{2/3}L^{1/3} + (c/2)(n/L)^{2/3} = \Theta(n^{2/3}L^{1/3}).
\end{aligned}$$

□

3 Impossibility Result

In light of the results in the previous section, a natural question is: Is it possible to design a mediator that will always improve the social welfare in any game for which there is a windfall of malice? Unfortunately, the answer to this question is "No", as we show in this section. In particular, we show that the congestion games which Babaioff, Kleinberg and Papadimitriou have proven have a windfall of malice effect [8] do not admit a mediator that is able to improve the social welfare. In fact, we prove a stronger impossibility result, showing that for any non-atomic, symmetric congestion game where the cost of a path never decreases as a function of the flow through that path (of which class of games, the examples in [8] are special instances), no mediator can improve the social optimum. In the rest of this section, we first define the congestion games we consider and then prove our impossibility result for these games.

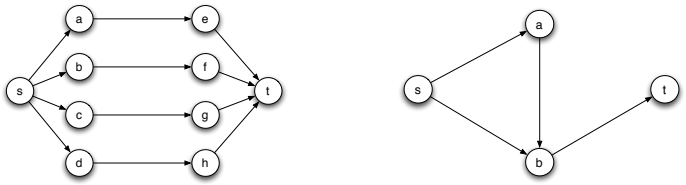

Figure 2: Examples where Theorem 3.1 applies

A non-atomic, symmetric *congestion game* (henceforth, simply a congestion game) is a specified by a set of $n \to \infty$ players; a set of E facilities (or edges); $A \subset 2^E$ actions (or paths); and finally, for each facility e a cost function f_e associated with that facility. A pure strategy profile $\mathcal{A} = (A_1, \ldots, A_n)$ is a vector of actions, one for each player. The cost of player i for action profile $\mathcal{A}$ is given by $F_i(\mathcal{A}) = \sum_{e \in A_i} f_e(x_e(\mathcal{A}))$ where $x_e(\mathcal{A})$ is the fraction of players using e in $\mathcal{A}$. As in [8], we assume that the game is *non-atomic*: since $n \to \infty$ the contribution of a single player to the flow over a facility is negligible; and *symmetric*: all players have the same cost functions.

For an action a and a flow $x \in [0, 1]$, let $\mathcal{F}_h(a, x)$ be the maximum possible cost of following action a when the total fraction of players following this action is x, where the maximum is taken over all ways that the remaining flow of $1-x$ can be distributed over other actions. Similarly, let $\mathcal{F}_\ell(a, x)$ be the *minimum* cost of following action a when the total fraction of players following this action is x.

We prove the following theorem for congestion games where the cost function of every action is always non-decreasing in the fraction of players performing that action. The theorem says that for such games, coordination between the agents in order to establish a correlated equilibrium will not decrease the social cost.

Theorem 3.1. *Consider a non-atomic, anonymous congestion game. If for all $a \in A$ and $0 \leq x \leq x' \leq 1$, $\mathcal{F}_h(a, x) \leq \mathcal{F}_\ell(a, x')$ then the smallest social cost achieved by a correlated equilibrium is no less than the smallest social cost achieved by a Nash equilibrium.*

Figure 2 gives examples of congestion games for which Theorem 3.1 applies. In these graphs, if the costs of all edges are non-decreasing in flow, then the smallest social cost achieved by a correlated equilibria is no better than the smallest social cost achieved by a Nash equilibria. In both examples, all players must travel from the source node s to the sink node t, so the set of allowable actions are just the set of all paths from s to t. The graph on the left is a specific example of a more general class of graphs for which all paths are disjoint and edge costs are non-decreasing, for which Theorem 3.1 applies. The graph on the right is a generalization of the congestion game from [8], which they show has a positive windfall of malice for certain non-decreasing cost functions. In the next section and in Figure 3 described therein, examples of congestion games for which Theorem 3.1 does *not* hold are given.

We next give a high level sketch of how we prove this theorem. We will fix a non-atomic, anonymous congestion game G with q actions, $a_1, \ldots, a_q$, and n players. We define a *configuration*, C, for such a game to be a partitioning of the set of

players across the q actions. We note that the number of possible configurations is finite; in particular, q^n. We next fix a mediator, M, for this game. We assume the mediator uses ℓ different configurations $C_1, \ldots, C_\ell$; that $0 \leq x_{i,j} \leq 1$ is the fraction of the players in configuration C_j assigned to action a_i; and that $c_{i,j} \in \mathbb{R}$ is the cost in configuration C_j for action a_i. We further assume that for all $j \in [\ell]$, p_j is the probability with which the mediator M chooses C_j.

For any two actions a, a' we define the *a posteriori cost* of a given a' as the expected cost for a player of performing action a when action a' is suggested by the mediator M; formally, $\mathrm{POST}(a, a') = \mathbf{E}[\mathcal{C}_a | \mathcal{E}_{a'}]$, where $\mathcal{C}_a$ is a random variable (over the configuration chosen by the mediator) and $\mathcal{E}_{a'}$ is the event that action a' is recommended by the mediator. We define the *a priori cost* of action a as the cost of a player completely ignoring what the mediator suggests and always performing action a; formally, $\mathrm{PRI}(a) := \sum_{j=1}^{\ell} p_j c_{i,j}$.

The sketch behind our proof for this theorem is as follows. First, we show in Lemma 3.2 that for all actions a, if the cost of a is non-decreasing in the flow through a, then $\mathrm{POST}(a, a) \geq \mathrm{PRI}(a)$. We show this by repeated decompositions of terms in summations for the a priori and posterior costs. The proof is straightforward but technical and so is included in the appendix. Next, let Y be the cost of a player listening and following the advice of the mediator, and let X be the cost of the player if she just ignores the advice of the mediator and always chooses the action a that minimized $\mathrm{PRI}(a)$. In Lemma 3.3 we show that it must be that $E(Y) \leq E(X)$. This lemma is shown by summing up inequality constraints on the mediator. Finally, we use these two lemmas to show the main theorem by showing that if Lemma 3.2 holds, then $E(Y) > E(X)$. The main technical challenge is the fact that we must show that $E(Y) > E(X)$ even though Lemma 3.2 does not necessarily give a strict inequality. We address this problem by a subtle case analysis in the proof of the main theorem, and by augmenting Lemma 3.2 to show that in some cases, the inequality it implies is strict.

We now present the detailed proof of the theorem. Observe that the condition for all $a \in A$ and $0 \leq x \leq x' \leq 1$, $\mathcal{F}_h(a, x) \leq \mathcal{F}_\ell(a, x')$ implies that for all $i \in [m]$, $\forall j, k \in [\ell]$ we have that $x_{ij} \leq x_{ik}$ implies $c_{ij} \leq c_{ik}$, and so the conditions of the following lemma are satisfied. We begin with Lemma 3.2, whose proof is in Appendix A.

Lemma 3.2. *Given $\ell \geq 2$ configurations $C_1, \ldots, C_\ell$, with corresponding probabilities $p_r > 0$, $r \in [\ell]$. If for $i \in [m]$, $\forall j, k \in [\ell]$ we have that $x_{ij} \leq x_{ik}$ implies $c_{ij} \leq c_{ik}$, then $POST(a_i, a_i) \geq PRI(a_i)$. Moreover, if for any $i \in [q]$, not all c_{ij}, $j \in [\ell]$ are the same, then $POST(a_i, a_i) > PRI(a_i)$.*

Define by $a_{pri} := \mathrm{argmin}_a \mathrm{PRI}(a)$. Given a mediator over a fixed set of configurations, let X be the random variable denoting the cost of an arbitrary player when he decides to use action a_{pri}, i.e., $\mathbf{E}[X] = \sum_{j=1}^{\ell} p_j c_{a_{pri}j}$. Denote also by Y the random variable of the cost when following the advice of the mediator, i.e., $\mathbf{E}[Y] = \sum_{i=1}^{m} \mathrm{POST}(a_i, a_i) \mathbf{Pr}(\mathcal{E}_i) = \sum_{i=1}^{m} \sum_{j=1}^{\ell} p_j x_{ij} c_{ij}$. We have the following relationship between Y and X.

Lemma 3.3. *For any mediator we have $\mathbf{E}[Y] \leq \mathbf{E}[X]$.*

Proof. Assume without loss of generality that action a_1 is the action with a_{pri}. The constraints for a correlated Nash equilibrium are that for all actions a_i and a_j, $\mathbf{E}\left[\mathcal{C}_{a_i}|\mathcal{E}_{a_i}\right] \leq \mathbf{E}\left[\mathcal{C}_{a_j}|\mathcal{E}_{a_i}\right]$. These constraints imply that

$$\forall_{i:2\leq i\leq q} : \sum_{j=1}^{\ell} p_j x_{ij} c_{ij} \leq \sum_{j=1}^{\ell} p_j x_{ij} c_{1j}.$$

Summing all of these $q-1$ inequalities together gives the single inequality, which we can rearrange as follows to show our result:

$$\begin{aligned}
\sum_{i=2}^{q}\sum_{j=1}^{\ell} p_j x_{ij} c_{ij} &\leq \sum_{i=2}^{q}\sum_{j=1}^{\ell} p_j x_{ij} c_{1j} \iff \\
\sum_{j=1}^{\ell}\sum_{i=2}^{m} p_j x_{ij} c_{ij} &\leq \sum_{j=1}^{\ell} p_j c_{1j} \sum_{i=2}^{q} x_{ij} \iff \\
\sum_{j=1}^{\ell}\sum_{i=2}^{q} p_j x_{ij} c_{ij} &\leq \sum_{j=1}^{\ell} p_j c_{1j}(1 - x_{1j}) \iff \\
\sum_{j=1}^{\ell}\sum_{i=1}^{q} p_j x_{ij} c_{ij} &\leq \sum_{j=1}^{\ell} p_j c_{1j} \iff \\
\mathbf{E}[Y] &\leq \mathbf{E}[X].
\end{aligned}$$

□

We are now ready to prove the main theorem.

Proof. Denote by $a_{post} := \text{argmin}_s \text{POST}(s, s)$ the action with minimum a posteriori cost. We will consider two cases.
Case 1: Not all actions have the same a posteriori cost. Then, we have:

$$\begin{aligned}
\mathbf{E}[Y] &> \text{POST}(a_{post}, a_{post}) \\
&\geq \text{PRI}(a_{post}) \text{ by Lemma 3.2} \\
&\geq \text{PRI}(a_{pri}) = \mathbf{E}[X].
\end{aligned}$$

Case 2: All action have the same a posteriori cost. In this case, we make use of the fact that there always must be some action that does not have equal costs in each configuration. Assume not. Then the cost of each action is the same in every configuration, and so any particular configuration must be a Nash equilibrium that achieves social cost equal to the social cost of the correlated equilibrium. Thus, we let a_x be some action that does not have the same cost in all configurations. Then we have:

$$\begin{aligned}
\mathbf{E}[Y] &= \text{POST}(a_x, a_x) \\
&> \text{PRI}(a_x) \text{ by Lemma 3.2} \\
&\geq \text{PRI}(a_{pri}) = \mathbf{E}[X].
\end{aligned}$$

In both cases we have $\mathbf{E}[Y] > \mathbf{E}[X]$. This however contradicts Lemma 3.3, hence there can not exist a correlated equilibrium achieving social cost less than the optimal Nash equilibrium. □

4 El Farol

We end this paper on a positive note, by describing a simple congestion game where we can show that a mediator will improve the social optimum. This simple game gives additional insight into why our mediator for the virus inoculation game works.

The game we consider is a variant of the *El Farol* game [3, 12, 10, 15]. El Farol is a[3] tapas bar in Santa Fe. Every Thursday night, a population of people decide whether or not to go to the bar. If too many people go, they will have a worse time than if they stayed home, since the bar will be too crowded. In our variant of the problem, we also assume that if too few people go, they will have a worse time than if they stayed home, because the bar will be too boring. We can model this as a non-atomic, symmetric congestion game as follows. There are two facilities e_1 and e_2, and two actions $a_1 = \{e_1\}$ and $a_2 = \{e_2\}$. For all $0 \leq x \leq 1$, $f_{e_1}(x) = 1/2$ and $f_{e_2}(x) = |1 - 2x|$.

We observe that the social cost in our game is minimized when the flow over both edges is $1/2$, in which case, the social cost is $1/4$. This configuration, however, is not a Nash equilibrium. Pure Nash equilibria occur when the top flow is $1/4$ or the top flow is $3/4$, for a social cost of $1/2$. We now describe a mediator that achieves the social optimum for this game.

Configuration C_1: The mediator advises all players to perform action a_1.

Configuration C_2: The mediator advises half of the players to perform action a_1, and advises the other half to perform action a_2.

For these two configurations C_1 and C_2 consider now the probability distribution $\mathcal{D}(\{C_1, C_2\})$ with $p_1 = 1/3$ and $p_2 = 2/3$. The proof of the following observation is straightforward but is included in the appendix for completeness.

Observation 4.1. *$\mathcal{D}(\{C_1, C_2\})$ is a mediator with social welfare $1/3$. Moreover, $1/3$ is the optimal value that can be obtained by a mediator.*

Figure 3 illustrates the two games we have described for which mediation helps. The left subfigure portrays our variant of the El Farol game, where the cost of the top path a_1 is always $1/2$ and the cost of the bottom path varies as shown in the plot below the graph. The values of $\mathcal{F}_\ell(a_2, x)$ and $\mathcal{F}_h(a_2, x)$ are equal, since in this game, when the flow through the top path is known, the cost of the bottom path is exactly determined. The two x's on the plot show the configurations used by the mediator. As implied by Theorem 3.1, for mediation to be effective, one of these x's must be below and to the right of the other on the plot. The right subfigure in Figure 3 portrays virus inoculation as a congestion game. The cost of the top path a_1 for this game is always 1. The cost of the bottom path, a_2, is any point in the polygon shown in the plot. We now have a polygon, rather than a line, because for a fixed number of nodes that do not inoculate, the cost of not inoculating varies depends on

[3] very tasty

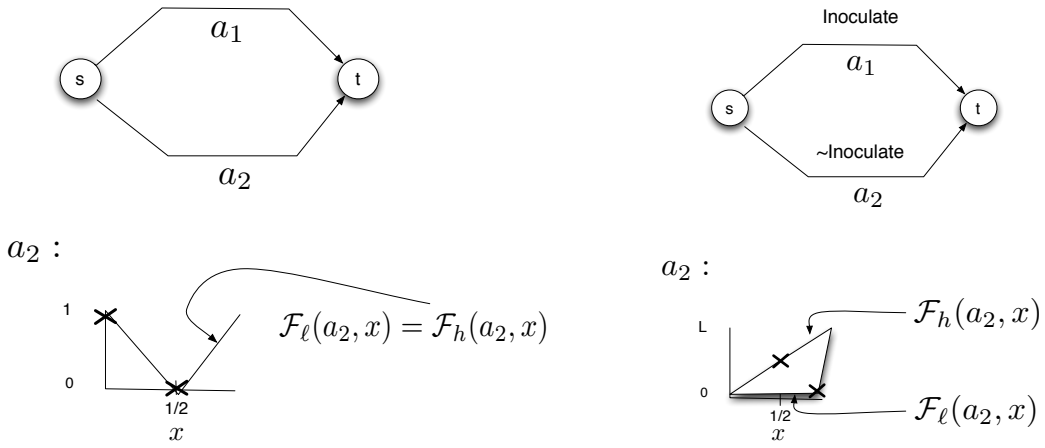

Figure 3: Congestion Games where mediation helps

how the inoculated nodes are positioned on the grid. $\mathcal{F}_\ell(a_2, x)$ is the bottom border of this polygon and $\mathcal{F}_h(a_2, x)$ is the top border. Again the two x's on the plot show the configurations used by the mediator, and again it is critical that one of these x's be below and to the right of the other. For the virus inoculation problem, we needed a clever arrangement of the inoculated nodes in one of the configurations to achieve this.

5 Conclusion

We have shown that a mediator can improve the social welfare in some strategic games with a positive windfall of malice. Several open questions remain including the following. First, can we determine necessary and sufficient conditions for a game to allow a mediator that improves social welfare over the best Nash? In particular, can we find such conditions for general congestion games? What about arbitrary anonymous games? Second, for games where each player can choose among k actions, can we say how many configurations are needed by any mediator? Preliminary work in this direction shows that for 2 actions, sometimes more than 2 configurations are needed. Finally, can we use approaches similar to those in this paper for designing mediators for multi-round games? We have already made some preliminary progress in this direction for multi-round games where the number of rounds is determined by a geometric random variable.

References

[1] Ittai Abraham, Danny Dolev, Rica Gonen, and Joe Halper. Distributed computing meets game theory: robust mechanisms for rational secret sharing and multiparty computation. In *Principles of Distributed Computing(PODC)*, 2006.

[2] Ittai Abraham, Danny Dolev, Rica Gonen, and Joe Halper. Lower bounds on implementing robust and resilient mediators. In *IACR Theory of Cryptography Conference(TCC)*, 2008.

[3] Brian Arthur. Bounded rationality and inductive behavior (the el farol problem). *American Economic Review*, 84:406–411, 1994.

[4] Itai Ashlagi, Dov Monderer, and Moshe Tennenholtz. Mediators in position auctions. In *Proceedings of the ACM Conference on Electronic Commerce(EC)*, 2007.

[5] J. Aspnes. Randomized protocols for asynchronous consensus. *Distributed Computing*, 16:165–175, 2003.

[6] J. Aspnes, K. Chang, and A. Yampolskiy. Inoculation strategies for victims of viruses and the sum-of-squares partition problem. *Journal of Computer and System Science*, 72(6):1077–1093, 2006.

[7] R. Aumann. Subjectivity and correlation in randomized games. *Mathematical Economics*, 1:67–96, 1974.

[8] Moshe Babaioff, Robert Kleinberg, and Christos H. Papadimitriou. Congestion games with malicious players. In *ACM Conference on Electronic Commerce*, 2007.

[9] Balcan, Blum, and Mansour. Improved equilibria via public service advertising. In *ACM Symposium on Discrete Algorithms (SODA)*, 2009.

[10] D. Challet, M. Marsili, and Gabriele Ottino. Shedding light on el farol. *Physica A: Statistical Mechanics and Its Applications*, 332:469–482, 2003.

[11] G. Christodoulou and E. Koutsoupias. On the price of anarchy and stability of correlated equilibria of linear congetion games. In *Proceedings of the European Symposium on Algorithms(ESA)*, 2005.

[12] M. de Cara, O. Pla, and F. Guinea. Competition, efficiency and collective behavior in the "el farol" bar model. *The European Physics Journal B*, 10, 1998.

[13] Martin Gairing. Malicious bayesian congestion games. In *Proc. of the 6th Workshop on Approximation and Online Algorithms (WAOA)*, 2008.

[14] Kostka. Byzantine Caching Game, 2006. ftp://ftp.tik.ee.ethz.ch/pub/students/2006-So/SA-2006-33.pdf.

[15] H. Lus, C. Aydin, S. Keten, H. Unsal, and A. Atiligan. El farol revisited. *Physica A: Statistical Mechanics and Its Applications*, 346:651–656, 2005.

[16] Thomas Moscibroda, Stefan Schmid, and Roger Wattenhofer. When selfish meets evil: Byzantine players in a virus inoculation game. In *Principles of Distributed Computing(PODC)*, 2006.

[17] C. Papadimitriou and T. Roughgarden. Computing correlated equilibria in multi-player games. *Journal of the ACM (JACM)*, 5(14), 2008.

[18] Aaron Roth. The price of malice in linear congestion games. In *Workshop on Internet and Network Economics(WINE)*, 2008.

[19] Ola Rozenfeld and Moshe Tennenholtz. Strong and correlated strong equilibria in monotone congestion games. In *Proceedings of the Workshop on Internet and Network Economics (WINE)*, 2006.

[20] Ola Rozenfeld and Moshe Tennenholtz. Group dominant strategies. In *Proceedings of the 3rd International Workshop on Internet and Network Economics (WINE)*, 2007.

[21] Ola Rozenfeld and Moshe Tennenholtz. Routing mediators. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, 2007.

Appendix

A Proof of Lemma 3.2

Proof. Consider without loss of generality action a_1. During this proof we use the notation of x_i for x_{1i} and c_i for c_{1i}, $i \in [\ell]$. Assume also without loss of generality that the configurations are ordered in such a way that $x_1 \leq x_2 \leq \ldots \leq x_\ell$ and thus $c_1 \leq c_2 \leq \ldots \leq c_\ell$. Note that $\text{POST}(a_1, a_1) = \frac{1}{\sum_{i=1}^{\ell} p_i x_i}(\sum_{i=1}^{\ell} p_i x_i c_i)$ and $\text{PRI}(a_1) = \sum_{i=1}^{\ell} p_i c_i$. Thus we must show that:

$$\sum_{i=1}^{\ell} p_i x_i c_i \geq (\sum_{i=1}^{\ell} p_i c_i)(\sum_{i=1}^{\ell} p_i x_i).$$

If all x_i are the same, then we clearly have equality and in this case $\text{POST}(a_1, a_1) = \text{PRI}(a_1)$. Otherwise, we will show that this inequality is true by decomposing the x_i terms into x_1 and ϵ_i terms, $\epsilon_i \geq 0$ (and there exists at least one j with $\epsilon_j > 0$). For any $i \in \{2, \ldots, \ell\}$ we write $x_i = x_1 + \epsilon_1 + \ldots + \epsilon_{i-1}$. Consider only the summands in the above inequality that contain the term x_1. If $x_1 = 0$ then clearly the inequality holds for such summands. If $x_1 > 0$, we get the following chain of inequalities for the summands containing x_1:

$$\begin{aligned}
\sum_{i=1}^{\ell} p_i x_1 c_i &\geq (\sum_{i=1}^{\ell} p_i c_i)(\sum_{i=1}^{\ell} p_i x_1) \\
\sum_{i=1}^{\ell} p_i c_i &\geq (\sum_{i=1}^{\ell} p_i c_i)(\sum_{i=1}^{\ell} p_i) \\
\sum_{i=1}^{\ell} p_i c_i &\geq \sum_{i=1}^{\ell} p_i c_i,
\end{aligned}$$

so this inequality holds.

Now consider the summands in the inequality containing ϵ_j for $1 \leq j \leq \ell - 1$. We get the inequality:

$$\sum_{i=j+1}^{\ell} p_i \epsilon_j c_i \geq (\sum_{i=1}^{\ell} p_i c_i)(\sum_{i=j+1}^{\ell} p_i \epsilon_j).$$

If $\epsilon_j = 0$, the inequality holds. If $\epsilon_j > 0$, for that j showing the previous inequality is equivalent to showing

$$\sum_{i=j+1}^{\ell} p_i c_i \geq (\sum_{i=1}^{\ell} p_i c_i)(\sum_{i=j+1}^{\ell} p_i).$$

To show that this inequality is true, we decompose the c_i terms into c_1 plus δ_i terms. That is, $c_i = c_1 + \delta_1 + \ldots + \delta_{i-1}$, for $i = 1, \ldots, \ell - 1$. Consider first the c_1 term. If

$c_1 = 0$, again the inequality holds trivially. If $c_1 > 0$, we get the chain of inequalities

$$\begin{aligned}
\sum_{i=j+1}^{\ell} p_i c_1 &\geq (\sum_{i=1}^{\ell} p_i c_1)(\sum_{i=j+1}^{\ell} p_i) \\
\sum_{i=j+1}^{\ell} p_i &\geq (\sum_{i=1}^{\ell} p_i)(\sum_{i=j+1}^{\ell} p_i) \\
\sum_{i=j+1}^{\ell} p_i &\geq \sum_{i=j+1}^{\ell} p_i,
\end{aligned}$$

which holds. Next we consider the δ_k terms for $k \leq j+1$. If $\delta_k = 0$, the inequality clearly holds for summands containing this term. If $\delta_k > 0$, we get the inequality chain:

$$\begin{aligned}
\sum_{i=j+1}^{\ell} p_i \delta_k &\geq (\sum_{i=k+1}^{\ell} p_i \delta_k)(\sum_{i=j+1}^{\ell} p_i) \\
\sum_{i=j+1}^{\ell} p_i &\geq (\sum_{i=k+1}^{\ell} p_i)(\sum_{i=j+1}^{\ell} p_i)
\end{aligned}$$

which also holds. In particular, since $p_1 > 0$, we have that $(\sum_{i=j+1}^{\ell} p_i) < 1$, and so if $\delta_k > 0$, the inequality is strict. Finally, we consider the δ_k terms for $k > j+1$. If $\delta_k = 0$, the inequality holds trivially. If $\delta_k > 0$ we get the inequality chain:

$$\begin{aligned}
\sum_{i=k}^{\ell} p_i \delta_k &\geq (\sum_{i=k}^{\ell} p_i \delta_k)(\sum_{i=j+1}^{\ell} p_i) \\
\sum_{i=k}^{\ell} p_i &\geq (\sum_{i=k}^{\ell} p_i)(\sum_{i=j+1}^{\ell} p_i),
\end{aligned}$$

which also holds.

Now, we note that if not all c_i are the same for $i \in [\ell]$, it must be the case that there exists some j such that $\delta_j > 0$, and it follows that we must also have that $\epsilon_j > 0$. As shown above, in such a situation, we obtain a strict inequality over the summands containing the term δ_j, and so the entire inequality, $\mathrm{POST}\,(a_1, a_1) > \mathrm{PRI}\,(a_1)$ must be strict.

□

B Proof of Observation 4.1

Proof. Define by $\mathcal{E}_i^s$, $i = 1, 2$, $s = 1, \ldots, n$, the event that the mediator proposes to player s to go on the i'th edge and define by $\mathcal{C}_i^s$, $i = 1, 2$, $s = 1, \ldots, n$, the cost for player s of going on the i'th edge. Since the mediator treats all players equally, we

will leave out the index s. Therefore, for a mediator to implement a correlated Nash equilibrium, the following inequalities must hold:

$$\mathbf{E}\left[\mathcal{C}_2 \,|\, \mathcal{E}_1\right] \geq \mathbf{E}\left[\mathcal{C}_1 \,|\, \mathcal{E}_1\right], \tag{3}$$

$$\mathbf{E}\left[\mathcal{C}_1 \,|\, \mathcal{E}_2\right] \geq \mathbf{E}\left[\mathcal{C}_2 \,|\, \mathcal{E}_2\right]. \tag{4}$$

For the particular choice of $p_1 = 1/3$ and $p_2 = 2/3$, it is easy to see that both (3) and (4) are satisfied.

Now we show that $1/3$ is the optimal value that can be obtained by any mediator. Let x_1 be the flow on e_1 and x_2 be the flow on e_2. The argument is as follows: for (3) to be satisfied, a configuration with $x_1 \in [0, 1/4] \cup [3/4, 1]$ has to be chosen, and among all these the configuration C_1 of the previous example is the one which has minimum total cost and the same time allows for the highest probabilities for configurations outside this interval. For the remaining values of $x_1 \in [1/4, 3/4]$, C_2 minimizes the total cost. □