Lower Bounds on Implementing Robust and Resilient Mediators

Ittai Abraham
School of Computer Science and Engineering
The Hebrew University of Jerusalem
Jerusalem, Israel
ittaia@cs.huji.ac.il

Danny Dolev*
School of Computer Science and Engineering
The Hebrew University of Jerusalem
Jerusalem, Israel
dolev@cs.huji.ac.il

Joseph Y. Halpern†
Cornell University
Ithaca, NY 14850
halpern@cs.cornell.edu

August 11, 2008

Abstract

We provide new and tight lower bounds on the ability of players to implement equilibria using cheap talk, that is, just allowing communication among the players. One of our main results is that, in general, it is impossible to implement three-player Nash equilibria in a bounded number of rounds. We also give the first rigorous connection between Byzantine agreement lower bounds and lower bounds on implementation. To this end we consider a number of variants of Byzantine agreement and introduce reduction arguments. We also give lower bounds on the running time of two player implementations. All our results extended to lower bounds on (k, t)-*robust* equilibria, a solution concept that tolerates deviations by coalitions of size up to k and deviations by up to t players with unknown utilities (who may be malicious).

*Part of the work was done while the author visited Cornell university. The work was funded in part by ISF, ISOC, NSF, CCR, and AFOSR.

†Supported in part by NSF under grants CCR-0208535, ITR-0325453, and IIS-0534064, by ONR under grant N00014-01-10-511, by the DoD Multidisciplinary University Research Initiative (MURI) program administered by the ONR under grants N00014-01-1-0795 and N00014-04-1-0725, and by AFOSR under grant FA9550-05-1-0055.

1 Introduction

The question of whether a problem in a multiagent system that can be solved with a trusted mediator can be solved by just the agents in the system, without the mediator, has attracted a great deal of attention in both computer science (particularly in the cryptography community) and game theory. In cryptography, the focus on the problem has been on *secure multiparty computation*. Here it is assumed that each agent i has some private information x_i. Fix functions $f_1, \ldots, f_n$. The goal is have agent i learn $f_i(x_1, \ldots, x_n)$ without learning anything about x_j for $j \neq i$ beyond what is revealed by the value of $f_i(x_1, \ldots, x_n)$. With a trusted mediator, this is trivial: each agent i just gives the mediator its private value x_i; the mediator then sends each agent i the value $f_i(x_1, \ldots, x_n)$. Work on multiparty computation (see [Gol04] for a survey) provides conditions under which this can be done. In game theory, the focus has been on whether an equilibrium in a game with a mediator can be implemented using what is called *cheap talk*—that is, just by players communicating among themselves (see [Mye97] for a survey).

There is a great deal of overlap between the problems studied in computer science and game theory. But there are some significant differences. Perhaps the most significant difference is that, in the computer science literature, the interest has been in doing multiparty computation in the presence of possibly malicious adversaries, who do everything they can to subvert the computation. On the other hand, in the game theory literature, the assumption is that players have preference and seek to maximize their utility; thus, they will subvert the computation iff it is in their best interests to do so. Following [ADGH06], we consider here both rational adversaries, who try to maximize their utility, and possibly malicious adversaries (who can also be considered rational adversaries whose utilities we do not understand).

1.1 Our Results

In this paper we provide new and optimal lower bounds on the ability to implement mediators with cheap talk. Recall that a *Nash equilibrium* σ is a tuple of strategies such that given that all other players play their corresponding part of σ then the best response is also to play σ. Given a Nash equilibrium σ we say that a strategy profile ρ is a *k-punishment strategy for σ* if, when all but k players play their component of ρ, then no matter what the remaining k players do, their payoff is strictly less than what it is with σ. We now describe some highlights of our results in the two simplest settings: (1) where rational players cannot form coalitions and there are no malicious players (this gives us the solution concept of Nash equilibrium) and (2) where there is at most one malicious player. We describe our results in a more general setting in Section 1.2.

No bounded implementations: In [ADGH06] it was shown that any Nash equilibrium with a mediator for three-player games with a 1-punishment strategy can be implemented using cheap talk. The expected running time of the implementation is constant. It is natural to ask if implementations with a bounded number of rounds exist for all three-player games. Theorem 2 shows this is not the case, implementations must have infinite executions and cannot be bounded for all three-player games. This lower bound highlights the importance of using randomization. An earlier attempt to provide a three-player cheap talk implementation [Ben03] uses a bounded implementation, and hence cannot work in general. The key insight of the lower bound is that when the implementation is bounded, then at some point the punishment strategy must become ineffective. The details turn out to be quite subtle. The only other lower bound that we are aware of that has the same flavor is the celebrated FLP result [FLP85] for reaching agreement in asynchronous systems, which also shows that no bounded implementation exists. However, we use quite different proof techniques than FLP.

Byzantine Agreement and Game Theory: We give the first rigorous connection between Byzantine agreement lower bounds and lower bounds on implementation. To get the lower bounds, we need to consider a number of variants of Byzantine agreement, some novel. The novel variants require new impossibility results. We have four results of this flavor:

1. Barany [Bar92] gives an example to show that, in general, to implement an equilibrium with a mediator in a three-player game, it is necessary to have a 1-punishment strategy. Using the power of randomized Byzantine agreement lower bounds we strengthen his result and show in Theorem 4 that we cannot even get an ϵ-implementation in this setting.

2. Using the techniques of [BGW88] or [For90], it is easy to show that any four-player game Nash equilibrium with a mediator can be implemented using cheap talk even if no 1-punishment strategy exists. Moreover, these implementations are *universal*; they do not depend on the players' utilities. In Theorem 3 we prove that universal implementations do not exist in general for three-player games. Our proof uses a nontrivial reduction to the weak Byzantine agreement (WBA) problem [Lam83]. To obtain our lower bound, we need to prove a new impossibility result for WBA, namely, that no protocol with a finite expected running time can solve WBA.

3. In [ADGH06] we show that for six-player games with a 2-punishment strategy, any Nash equilibrium can be implemented even in the presence of at most one malicious player. In Theorem 5 we show that for five players even ϵ–-implementation is impossible. The proof uses a variant of Byzantine agreement; this is related to the problem of *broadcast with extended consistency* introduced by Fitzi et al. [FHHW03]. Our reduction maps the rational player to a Byzantine process that is afraid of being detected and the malicious player to a standard Byzantine process.

4. In Theorem 8, we show that for four-player games with at most one malicious player, to implement the mediator, we must have a PKI setup in place, even if the players are all computationally bounded and even if we are willing to settle for ϵ–implementations. Our lower bound is based on a reduction to a novel relaxation of the Byzantine agreement problem.

Bounds on running time: We provide bounds on the number of rounds needed to implement two-player games. In Theorem 9(a) we prove that the expected running time of any implementation of a two-player mediator equilibrium must depend on the utilities of the game, even if there is a 1-punishment strategy. This is in contrast to the three-player case, where the expected running time is constant. In Theorem 9(b) we prove that the expected running time of any ϵ–implementation of a two-player mediator equilibrium for which there is no 1-punishment strategy must depend on ϵ. Both results are obtained using a new two-player variant of the secret-sharing game. The only result that we are aware of that has a similar spirit is that of Boneh and Naor [BN00], where it is shown that two-party protocols with "bounded unfairness" of ϵ must have running time that depends on the value of ϵ. The implementations given by Urbano and Vila [UV02, UV04] in the two-player case are independent of the utilities; the above results show that their implementation cannot be correct in general.

1.2 Our results for implementing robust and resistant mediators

In [ADGH06] (ADGH from now on), we argued that it is important to consider deviations by both rational players, who have preferences and try to maximize them, and players that can be viewed as malicious, although it is perhaps better to think of them as rational players whose utilities are not known by the other players or mechanism designer. We considered equilibria that are (k, t)-*robust*; roughly speaking, this means that the equilibrium tolerates deviations by up to k rational players, whose utilities are presumed known, and up to t players with unknown utilities (i.e., possibly malicious players). We showed how (k, t)-robust equilibria with mediators could be implemented using cheap talk, by first showing that, under appropriate assumptions, we could implement secret sharing in a (k, t)-robust way using cheap talk. These assumptions involve standard considerations in the game theory and distributed systems literature, specifically, (a) the relationship between k, t and n, the total number of players in the system; (b) whether players know the exact utilities of other players; (c) whether there are broadcast channels or just point-to-point channels; (d) whether cryptography is available; and (e) whether the game has a $(k + t)$-*punishment strategy*; that is, a strategy that, if used by all but at most $k + t$ players, guarantees that every player gets a worse outcome than they do with the equilibrium strategy. Here we provide a complete picture of when implementation is possible, providing lower bounds that match the known upper bounds (or improvements of them that we have obtained). The following is a high-level picture of the results. (The results discussed in Section 1.1 are special cases of the results stated below. Note that all the upper bounds mentioned here are either in ADGH, slight improvements of results in ADGH, or are known in the literature; see Section 3 for the details. The new results claimed in the current submission are the matching lower bounds.)

- If $on > 3k + 3t$, then mediators can be implemented using cheap talk; no punishment strategy is required, no knowledge of other agents' utilities is required, and the cheap-talk strategy has bounded running time that does not depend on the utilities (Theorem 1(a) in Section 3).
- If $n \leq 3k+3t$, then we cannot, in general, implement a mediator using cheap talk without knowledge of other agents' utilities (Theorem 3). Moreover, even if other agents' utilities are known, we cannot, in general, implement a mediator without having a punishment strategy (Theorem 4) nor with bounded running time (Theorem 2).
- If $n > 2k + 3t$, then mediators can be implemented using cheap talk if there is a punishment strategy (and utilities are known) in finite expected running time that does not depend on the utilities (Theorem 1(b) in Section 3).
- If $n \leq 2k+3t$, then we cannot, in general, ϵ-implement a mediator using cheap talk, even if there is a punishment strategy and utilities are known (Theorem 5).
- If $n > 2k + 2t$ and we can simulate broadcast then, for all ϵ, we can ϵ-implement a mediator using cheap talk, with bounded expected running time that does not depend on the utilities in the game or on ϵ (Theorem 1(c) in Section 3). (Intuitively, an ϵ-implementation is an implementation where a player can gain at most ϵ by deviating.)
- If $n \leq 2k + 2t$, we cannot, in general, ϵ-implement a mediator using cheap talk even if we have broadcast channels (Theorem 7). Moreover, even if we assume cryptography and broadcast channels, we cannot, in general, ϵ-implement a mediator using cheap talk with expected running time that does not depend on ϵ (Theorem 9(b)); even if there is a punishment strategy, then we still cannot, in general, ϵ-implement a mediator using cheap talk with expected running time independent of the utilities in the game (Theorem 9(a)).
- If $n > k + 3t$ then, assuming cryptography, we can ϵ-implement a mediator using cheap talk;

moreover, if there is a punishment strategy, the expected running time does not depend on ϵ (Theorem 1(e) in Section 3).

- If $n \leq k + 3t$, then even assuming cryptography, we cannot, in general, ϵ-implement a mediator using cheap talk (Theorem 8).
- If $n > k + t$, then assuming cryptography and that a PKI (Public Key Infrastructure) is in place,[1] we can ϵ-implement a mediator (Theorem 1(d) in Section 3); moreover, if there is a punishment strategy, the expected running time does not depend on ϵ (Theorem 1(e) in Section 3).

The lower bounds are existential results; they show that if certain conditions do not hold, then there exists an equilibrium that can be implemented by a mediator that cannot be implemented using cheap talk. There are other games where these conditions do not hold but we can nevertheless implement a mediator.

1.3 Related work

There has been a great deal of work on implementing mediators, both in computer science and game theory. The results above generalize a number of results that appear in the literature. We briefly discuss the most relevant work on implementing mediators here. Other work related to this paper is discussed where it is relevant.

In game theory, the study of implementing mediators using cheap talk goes back to Crawford and Sobel [CS82]. Barany [Bar92] shows that if $n \geq 4$, $k = 1$, and $t = 0$ (i.e., the setting for Nash equilibrium), a mediator can be implemented in a game where players do not have private information. Forges [For90] provides what she calls a *universal mechanism* for implementing mediators; essentially, when combining her results with those of Barany, we get the special case of Theorem 1(a) where $k = 1$ and $t = 0$. Ben-Porath [Ben03] considers implementing a mediator with cheap talk in the case that $k = 1$ if $n \geq 3$ and there is a 1-punishment strategy. He seems to have been the first to consider punishment strategies (although his notion is different from ours: he requires that there be an equilibrium that is dominated by the equilibrium that we are trying to implement). Heller [Hel05] extends Ben-Porath's result to allow arbitrary k. Theorem 1(b) generalizes Ben-Porath and Heller's results. Although Theorem 1(b) shows that the statement of Ben-Porath's result is correct, Ben-Porath's implementation takes a bounded number of rounds; Theorem 2 shows it cannot be correct.[2] Heller proves a matching lower bound; Theorem 5 generalizes Heller's lower bound to the case that $t > 0$. (This turns out to require a much more complicated game than that considered by Heller.) Urbano and Vila [UV02, UV04] use cryptography to deal with the case that $n = 2$ and $k = 1$;[3] Theorem 1(e)) generalizes their result to arbitrary k and t. However, just as with Ben-Porath, Urbano and Vila's implementation takes a bounded number of rounds; As we said in Section 1.1, Theorem 9(a) shows that it cannot be correct.

In the cryptography community, results on implementing mediators go back to 1982 (although this terminology was not used), in the context of *(secure) multiparty computation*. Since there are no utilities in this problem, the focus has been on essentially what we call here *t-immunity*: no group of t players can prevent the remaining players from learning the function value, nor can they learn the other

[1] We can replace the assumption of a PKI here and elsewhere by the assumption that there is a trusted preprocessing phase where players may broadcast.

[2] Although Heller's implementation does not take a bounded number of rounds, it suffers from problems similar to those of Ben-Porath.

[3] However, they make somewhat vague and nonstandard assumptions about the cryptographic tools they use.

players' private values. Results of Yao [Yao82] can be viewed as showing that if $n = 2$ and appropriate computational hardness assumptions are made, then, for all ϵ, we can obtain 1-immunity with probability greater than $1 - \epsilon$ if appropriate computational hardness assumptions hold. Goldreich, Micali, and Wigderson [GMW87] extend Yao's result to the case that $t > 0$ and $n > t$. Ben-Or, Goldwasser, and Wigderson [BGW88] and Chaum, Crépeau, and Damgard [CCD88] show that, without computational hardness assumptions, we can get t-immunity if $n > 3t$; moreover, the protocol of Ben-Or, Goldwasser, and Wigderson does not need an ϵ "error" term. Although they did not consider utilities, their protocol actually gives a (k, t)-robust implementation of a mediator using cheap talk if $n > 3k + 3t$; that is, they essentially prove Theorem 1(a). (Thus, although these results predate those of Barany and Forges, they are actually stronger.) Rabin and Ben-Or [RB89] provide a t-immune implementation of a mediator with "error" ϵ if broadcast can be simulated. Again, when we add utilities, their protocol actually gives an ϵ–(k, t)-robust implementation. Thus, they essentially prove Theorem 1(c). Dodis, Halevi, and Rabin [DHR00] seem to have been the first to apply cryptographic techniques to game-theoretic solution concepts; they consider the case that $n = 2$ and $k = 1$ and there is no private information (in which case the equilibrium in the mediator game is a *correlated equilibrium* [Aum87]); their result is essentially that of Urbano and Vila [UV04] (although their protocol does not suffer form the problems of that of Urbano and Vila).

Halpern and Teague [HT04] were perhaps the first to consider the general problem of multiparty computation with rational players. In this setting, they essentially prove Theorem 1(d) for the case that $t = 0$ and $n \geq 3$. However, their focus is on the solution concept of *iterated deletion*. They show that there is no Nash equilibrium for rational multiparty computation with rational agents that survives iterated deletion and give a protocol with finite expected running time that does survive iterated deletion. If $n \leq 3(k + t)$, it follows easily from Theorem 2: that there is no multiparty computation protocol that is a Nash equilibrium, we do not have to require that the protocol survive iterated deletion to get the result if $n \leq 3(k + t)$. Various generalizations of the Halpern and Teague results have been proved. We have already mentioned the work of ADGH. Lysanskaya and Triandopoulos [LT06] independently proved the special case of Theorem 1(c) where $k = 1$ and $t + 1 < n/2$ (they also consider survival of iterated deletion); Gordon and Katz [GK06] independently proved a special case of Theorem 1(d) where $k = 1$, $t = 0$, and $n \geq 2$.

In this paper we are interested in implementing equilibrium by using standard communication channels. An alternate option is to consider the possibility of simulating equilibrium by using much stronger primitives. Izmalkov, Micali, and Lepinski [IML05] show that, if there is a punishment strategy and we have available strong primitives that they call *envelopes* and *ballot boxes*, we can implement arbitrary mediators perfectly (without an ϵ error) in the case that $k = 1$, in the sense that every equilibrium of the game with the mediator corresponds to an equilibrium of the cheap-talk game, and vice versa. In [LMPS04, LMS05], these primitives are also used to obtain implementation that is perfectly collusion proof in the model where, in the game with the mediator, coalitions cannot communicate. (By way of contrast, we allow coalitions to communicate.) Unfortunately, envelopes and ballot boxes cannot be implemented under standard computational and systems assumptions [LMS05].

It is reasonable to ask at this point whether mediators are of practical interest. After all, if three companies negotiate, they can just hire an arguably trusted mediator, say an auditing firm. The disadvantage of this approach in a setting like the internet, with constantly shifting alliances, there are always different groups that want to collaborate; a group may not have the time and flexibility of hiring a mediator, even assuming they can find one they trust. Another concern is that our results simply shift the role of what has to be trust elsewhere. It is certainly true that our results assume point-to-point communication that cannot be intercepted. If $n \leq k + 3t$, then we must also assume the existence of a public-key infrastructure. Thus, we have essentially shifted from trusting the mediator to trusting the PKI. In practice, individuals who want to collaborate may find point-to-point communication and a PKI more trustworthy than an intermediary, and easier to work with.

The rest of this paper is organized as follows. In Section 2, we review the relevant definitions. In Section 3, we briefly discuss the upper bounds, and compare them to the results of ADGH. In Section 4, we prove the lower bounds. The missing proofs appear in the appendix.

2 Definitions

In this section, we give detailed definitions of the main notions needed for our results. Sometimes there are subtle differences between our differences and those used in the game-theory literature. We discuss these differences carefully.

2.1 Mediators and cheap talk

We are interested in implementing mediators. Formally, this means we need to consider three games: an *underlying game* Γ, an extension Γ_d of Γ with a mediator, and a cheap-talk extension Γ_{CT} of Γ. Our underlying games are *(normal-form) Bayesian games*. These are games of incomplete information, where players make only one move, and these moves are made simultaneously. The "incomplete information" is captured by assuming that nature makes the first move and chooses for each player i a *type* in some set $\mathcal{T}_i$, according to some distribution that is commonly known. Formally, a Bayesian game Γ is defined by a tuple $(N, \mathcal{T}, A, u, \mu)$, where N is the set of players, $\mathcal{T} = \times_{i \in N} \mathcal{T}_i$ is the set of possible types, μ is the distribution on types, $A = \times_{i \in N} A_i$ is the set of action profiles, and $u_i : \mathcal{T} \times A$ is the utility of player i as a function of the types prescribed by nature and the actions taken by all players.

A *strategy* for player i in a Bayesian game Γ is a function from i's type to an action in A_i; in a game with a mediator, a strategy is a function from i's type and message history to an action. We allow behavior strategies (i.e., randomized strategies); such a strategy gets an extra argument, which is a sequence of coin flips (intuitively, what a player does can depend on its type, the messages it has sent and received if we are considering games with mediators, and the outcome of some coin flips). We use lower-case Greek letters such as σ, τ, and ρ to denote a strategy profile; σ_i denotes the strategy of player i in strategy profile σ; if $K \subseteq N$, then σ_K denotes the strategies of the players in K and σ_{-K} denotes the strategies of the players not in K. Given a strategy profile σ a player $i \in N$ and a type $t_i \in T_i$ let $u_i(t_i, \sigma)$ be the expected utility of player i given that his type is t_i and each player $j \in N$ is playing the strategy σ_j.

Given an underlying Bayesian game Γ as above, a game Γ_d with a mediator d that extends Γ is, informally, a game where players can communicate with the mediator and then perform an action from Γ. The utility function of a player i in Γ_d is the same as that in Γ; thus, the utility of a player i in Γ_d depends just on the types of all players and the actions taken by all players. Formally, we view both the mediator and the players as interacting Turing machines with access to an unbiased coin (which thus allows them to choose uniformly at random from a finite set of any size). The mediator and players interact for some (possibly unbounded) number of stages. A mediator is characterized by a function

$\mathcal{P}$ that maps the inputs it has received up to a stage (and possible some random bits) to an output for each player. Given an underlying Bayesian game Γ where player i's actions come from the set A_i and a mediator d, the interaction with the mediator in Γ_d proceeds in stages, where each stage consists of three phases. In the first phase of a stage, each player i sends an input to d (player i can send the empty input, i.e., no input at all); in the second phase, d sends each player i an output according to $\mathcal{P}$, (again, the mediator can send the empty output); and in the third phase, each player i chooses an action in A_i or no action at all. A player can play at most one action from A_i in each execution (play) of Γ_d. Player i's utility function in Γ_d is the same as that in the underlying game Γ, and depends only on the action profile in A played by the players and the types. To make this precise, we need to define what move an action in A_i is played by player i in executions of Γ_d where i in fact never plays an action in A_i. For ease of exposition, we assume that for each player i, some default action $a_i^* \in A$ is chosen. There are other ways of dealing with this issue (see, for example, [AH03] for an alternative approach). Our results do not depend on the choice, since in our upper bounds, with probability 1, all players do play an action in equilibrium, and our impossibility results are independent of the action chosen if players do not choose an action. (We remark that the question of what happens after an infinite execution of the cheap-talk game becomes much more significant in asynchronous systems; see [ADH].)

Although we think of a *cheap-talk* game as a game where players can communicate with each other (using point-to-point communication and possibly broadcast), formally, it is a game with a special kind of mediator: player i send the mediator whatever messages it wants to send other players in the first phase of a round; the mediator forwards these messages to the intended recipients in the second phase. We can model broadcast messages by just having the mediator tag a message as a broadcast (and sending the same message to all the intended recipients, of course).

We assume that cheap talk games are always unbounded; players are allowed to talk forever. The *running time* of an execution of a joint strategy σ is the number of steps taken until the last player makes a move in the underlying game. The running time may be infinite.

It is standard in the game theory literature to view cheap talk as *pre-play* communication. Thus, although different plays of Γ_{CT} may have different running times (possibly depending on random factors), it is assumed that it is commonly known when the cheap-talk phase ends; then all players make their decisions simultaneously. It is not possible for some players to continue communicating after some other players have decided (see, for example, [Hel05], where this assumption is explicit). For their possibility results, ADGH define games and give recommended strategies for these games such that, as long as players use the recommended strategy, all players make a move at the same time in each play of the cheap-talk game. However, it is not assumed that this is the case off the equilibrium path (that is, if players do not follow the recommended strategy). The assumption that all players stop communicating at the same time seems to us very strong, and not implementable in practice, so we do not make it here. (Dropping this assumption sometimes makes our impossibility results harder to prove; see the proof of Theorem 7 in Appendix A.5 for an example.) Thus, there is essentially only one cheap-talk game extending an underlying game Γ,; Γ_{CT} denotes the cheap-talk extension of Γ.

When we consider a deviation by a coalition K, we want to allow the players in K to communicate with each other. If Γ' is an extension of an underlying game Γ (including Γ itself) and $K \subseteq N$, let $\Gamma' + CT(K)$ be the extension of Γ where the mediator provides private cheap-talk channels for the players in K in addition to whatever communication there is in Γ'. Note that $\Gamma_{\mathrm{CT}} + CT(K)$ is just Γ_{CT}; players in K can already talk to each other in Γ_{CT}.

2.2 Implementation

Note that a strategy profile—whether it is in the underlying game, or in a game with a mediator extending the underlying game (including a cheap-talk game)—induces a mapping from type profiles to distributions over action profiles. If Γ_1 and Γ_2 are extension of some underlying game Γ, then strategy σ_1 in Γ_1 *implements* a strategy σ_2 in Γ_2 if both σ and σ' induce the same function from types to distributions over actions. Note that although our informal discussion in the introduction talked about *implementing mediators*, the formal definitions (and our theorems) talk about implementing strategies. Our upper bounds show that, under appropriate assumptions, for *every* (k, t)-robust equilibrium σ in a game Γ_1 with a mediator, there exists an equilibrium σ' in the cheap-talk game Γ_2 corresponding to Γ_1 that implements σ; the lower bounds in this paper show that, if these conditions are not met, there exists a game with a mediator and an equilibrium in that game that cannot be implemented in the cheap-talk game. Since our definition of games with a mediator also allow arbitrary communication among the agents, it can also be shown that every equilibrium in a cheap-talk game can be implemented in the mediator game: the players simply ignore the mediator and communicate with each other.

2.3 Solution concepts

We can now define the solution concepts relevant for this paper. In particular, we consider a number of variants of (k, t) robustness, and the motivation behind them.

In defining these solution concepts, we need to consider the expected utility of a strategy profile conditional on players having certain types. We abuse notation and continue to use u_i for this, writing for example, $u_i(t_K, \sigma)$ to denote the expected utility to player i if the strategy profile σ is used, conditional on the players in K having the types t_K. Since the strategy σ here can come from the underlying game or some extension of it, the function u_i is rather badly overloaded. We sometimes include the relevant game as an argument to u_i to emphasize which game the strategy profile σ is taken from, writing, for example, $u_i(t_K, \Gamma', \sigma)$.

k-resilient equilibrium: A strategy profile is a Nash equilibrium if no player can gain any advantage by using a different strategy, given that all the other players do not change their strategies. We want to define a notion of *k-resilient equilibrium* that generalizes Nash equilibrium, but allows a coalition of up to k players to change their strategies. One way of capturing this, which goes back to Aumann [Aum59], is to require that there be no deviations that result in everyone in a group of size at most k doing better.

To make this intuition precise, we need some notation. Given a type space $\mathcal{T}$, a set K of players, and $t \in \mathcal{T}$, let $\mathcal{T}(t_K) = \{t' : t'_K = t_K\}$. If Γ is a game over type space $\mathcal{T}$, σ is a strategy profile in Γ, and $\Pr$ is the probability on the type space $\mathcal{T}$, let

$$u_i(t_K, \sigma) = \sum_{t' \in \mathcal{T}(t_K)} \Pr(t' \mid \mathcal{T}(t_K)) u_i(t', \sigma).$$

Thus, $u_i(t_K, \sigma)$ is i's expected payoff if everyone uses strategy σ and types are restricted to $\mathcal{T}(t_K)$.

Definition 1. *σ is a k-*resilient$'$ equilibrium *if, for all $K \subseteq N$ and all types $t \in \mathcal{T}$, it is not the case that there exists a strategy τ such that $u_i(t_K, \tau_K, \sigma_{-K}) > u_i(t_K, \sigma)$ for all $i \in K$.*

Thus, σ is k-resilient$'$ if no subset K of at most k players can all do better by deviating, even if they share their type information (so that if the true type is t, the players in K know t_K). This is essentially Aumann's [Aum59] notion of resilience to coalitions, except that we place a bound on the size of coalitions.

As the prime suggests, this will not be exactly the definition we focus on. ADGH consider a stronger notion, which requires that there be no deviation where even one player does better.

Definition 2. *σ is a* strongly k-resilient$'$ equilibrium *if, for all $K \subseteq N$ with $|K| \leq k$ and all types $t \in \mathcal{T}$, it is not the case that there exists a strategy τ such that $u_i(t_K, \tau_K, \sigma_{-K}) > u_i(t_K, \sigma)$ for some $i \in K$.*

Both of these definitions have a weakness: they implicitly assume that the coalition members cannot communicate with each other beyond agreeing on what strategy to use. While, in general, there are equilibria in the cheap-talk game that are not available in the underlying game (so having more communication can increase the number of possible equilibria), perhaps surprisingly, allowing communication between coalition members can also *prevent* certain equilibria, as the following example shows.

Example 1. Consider a game with four players. Players 1 and 2 have a type in $\{0, 1\}$; the type of players 3 and 4 is 0. All tuples of types are equally likely. Players 3 and 4 can each choose an action in the set $\{0, 1, \text{PUNISH}, \text{PASS}\}$; players 1 and 2 each choose an action in $\{\text{PUNISH}, \text{PASS}\}$. If anyone plays PUNISH, then everyone gets a payoff of -1. If no one plays PUNISH, the payoffs are as follows: If player 3 plays PASS, then 3 gets a payoff of 1; similarly, if player 4 plays PASS, then 4 gets a payoff of 1. If player 3 plays 0 or 1 and this is 1's type, then 3 gets 5; if not, then 3 gets -5; similarly for player 4. Finally, if player 2's type is 0, then player 1 and 2's payoffs are the same as player 3's payoffs; similarly, if player 2 has a 1, then 1 and 2's payoffs are the same as 4's payoffs. It is easy to see that everyone playing PASS is a 3-resilient$'$ equilibrium in the underlying game and it is also an equilibrium in the game with a mediator, if the players cannot communicate. However, if players can communicate for one round, then players 1, 2, and 3 can do better if player 1 sends player 3 his type, and player 3 plays it. This guarantees player 3 a payoff of 5, while players 1 and 2 get an expected payoff of 2.5.

Now suppose that we consider a variant of this game, where the actions are the same and the payoffs for players 3 and 4 are the same, but the payoffs for players 1 and 2 are modified as follows. If player 2's type is 0, then if no one plays PUNISH, player 1 and 2's payoffs are the same as player 3's payoffs if player 4 plays PASS; if player 4 plays 0 or 1, then their payoff is -5. Similarly, if player 2's type is 1, then player 1 and 2's payoffs are the same player 4's payoffs if player 3 plays PASS; if player 3 plays 0 or 1, then their payoff is -5. It is easy to show that everyone playing PASS is still 3-resilient$'$ if we allow only one round of communication. But with two rounds of communication, everyone playing PASS is no longer 3-resilient$'$: players 1, 2, and 3 can do better if player 2 sends player 1 his type, player 1 sends player 3 his type if player 2's type is 0 (and sends nothing otherwise), and player 3 plays player 1's type if player 1 sends it. □

Since it seems reasonable to assume that coalition members will communicate, it seems unreasonable to call everyone playing PASS 3-resilient if some communication among coalition members can destroy that equilibrium. More generally, we clearly cannot hope to implement a k-resilient equilibrium in the mediator game using cheap talk if the equilibrium does not survive once we allow communication among the coalition members. This motivates the following definition.

Definition 3. *σ is a* (strongly) k-resilient equilibrium *in a game Γ' if, for all $K \subseteq N$ with $|K| \leq k$ and all types $t \in \mathcal{T}$, it is not the case that there exists a strategy τ such that $u_i(t_K, \Gamma' + CT(K), \tau_K, \sigma_{-K}) > u_i(t_K, \Gamma', \sigma)$ for some $i \in K$.*

This definition makes precise the intuition that players in the coalition are allowed arbitrary communication among themselves.

Note that Nash equilibrium is equivalent to both 1-resilience and strong 1-resilience; however, the notions differ for $k > 1$. It seems reasonable in many applications to bound the size of coalitions; it is hard to coordinate a large coalition! Of course, the appropriate bound on the size of the coalition may well depend on the utilities involved. Our interest in strong k-resilience was motivated by wanting to allow situations where one player effectively controls the coalition. This can happen in practice in a network if one player can "hijack" a number of nodes in the network. It could also happen if one player can threaten others, or does so much better as a result of the deviation that he persuades other players to go along, perhaps by the promise of side payments. While it can be argued that, if there are side payments or threats, they should be modeled in the utilities of the game, it is sometimes more convenient to work directly with strong resilience. In this paper we consider both resilience and strong resilience, since the results on implementation obtained using the different notions are incomparable. Just because a strongly resilient strategy in a game with a mediator can be implemented by a strongly resilient strategy in a cheap-talk game, it does not follow that a resilient strategy with a mediator can be implemented by a resilient strategy in a cheap-talk game, or vice versa. However, as we show, we get the same lower bounds for both resilience and strong resilience: in our lower bounds, we give games with mediators with a strongly k-resilient equilibrium σ and show that there does not exist a cheap-talk game and a strategy that σ' that implements σ and is k-resilient. Similarly, we can show that we get the same upper bounds with both k-resilience and strong k-resilience.

Other notions of resilience to coalitions have been defined in the literature. For example, Bernheim, Peleg, and Whinston [BPW89] define a notion of *coalition-proof Nash equilibrium* that, roughly speaking, attempts to capture the intuition that σ is a coalition-proof equilibrium if there is no deviation that allows all players to do better. However, they argue that this is too strong a requirement, in that some deviations are not *viable*: they are not immune from further deviations. Thus, they give a rather complicated definition that tries to capture the intuition of a deviation that is immune from further deviations. This work is extended by Moreno and Wooders [MW96] to allow correlated strategies. Although it is beyond the scope of this paper to go through the definitions, it is easy to see that our impossibility results apply to them, because of the particular structure of the games we consider.

For some of our results we will be interested in strategies that give "almost" k-resilience, in the sense that no player in a coalition can do more than ϵ better by deviating, for some small ϵ.

Definition 4. *If $\epsilon \geq 0$, then σ is an* ϵ–k-resilient equilibrium *in a game Γ' if, for all $K \subseteq N$ with $|K| \leq k$ and all types $t \in \mathcal{T}$, it is not the case that there exists a strategy τ such that $u_i(t_K, \Gamma' + CT(K), \tau_K, \sigma_{-K}) > u_i(t_K, \Gamma', \sigma) + \epsilon$ for all $i \in K$.*

Clearly if $\epsilon = 0$, then an ϵ–k-resilient equilibrium is a k-resilient equilibrium.

(k, t)-robust equilibrium: We now define the main solution concept used in this paper: (k, t)-robust equilibrium. The k indicates the size of coalition we are willing to tolerate, and the t indicates the number of players with unknown utilities. These t players are analogues of faulty players or adversaries in the distributed computing literature, but we can think of them as being perfectly rational. Since we do not know what actions these t players will perform, nor do we know their identities, we are interested in strategies for which the payoffs of the remaining players are immune to what the t players do.

Definition 5. *A strategy profile σ in a game Γ is* t-immune *if, for all $T \subseteq N$ with $|T| \leq t$, all strategy profiles τ, all $i \notin T$, and all types $t_i \in \mathcal{T}_i$ that occur with positive probability, we have $u_i(t_i, \Gamma + CT(T), \sigma_{-T}, \tau_T) \geq u_i(t_i, \Gamma, \sigma)$.*

Intuitively, σ is t-immune if there is nothing that players in a set T of size at most t can do to give the remaining players a worse payoff, even if the players in T can communicate.

Our notion of (k, t)-robustness requires both t-immunity and k-resilience. In fact, it requires k-resilience no matter what up to t players do. That is, we require that no matter what t players do, no subset of size at most k can all do better by deviating, even with the help of the t players, and even if all $k + t$ players share their type information.

Definition 6. *Given $\epsilon \geq 0$, σ is an* ϵ–(k, t)-robust equilibrium *in game Γ if σ is t-immune and, for all $K, T \subseteq N$ such that $|K| \leq k$, $|T| \leq t$, and $K \cap T = \emptyset$, and all types $t_{K\cup T} \in \mathcal{T}_{K\cup T}$ that occur with positive probability, it is not the case that there exists a strategy profile τ such that*

$$u_i(t_{K\cup T}, \Gamma + CT(K \cup T), \tau_{K\cup T}, \sigma_{-(K\cup T)}) > u_i(t_i, \Gamma + CT(T), \tau_T, \sigma_{-T}) + \epsilon$$

for all $i \in K$. A (k, t)-robust equilibrium is just a 0–(k, t)-robust equilibrium.

We can define a *strongly (k, t)-robust equilibrium* by analogy to the definition strongly k-resilient equilibrium: we simply change "for all $i \in K$" in the definition of (k, t)-robust equilibrium to "for some $i \in K$". Thus, in a strongly (k, t)-robust equilibrium, not even a single agent in K can do better if all the players in K deviate, even with the help of the players in T.

Note that a $(1, 0)$-robust equilibrium is just a Nash equilibrium, and an ϵ–$(1, 0)$-robust equilibrium is what has been called an ϵ-Nash equilibrium in the literature. A (strongly) $(k, 0)$-robust equilibrium is just a (strongly) k-resilient equilibrium. The notion $(0, t)$-robustness is somewhat in the spirit of Eliaz's [Eli02] notion of t fault-tolerant implementation. Both our notion of $(0, t)$-robustness and Eliaz's notion of t-fault tolerance require that what the players not in T do is a best response to whatever the players in T do (given that all the players not in T follow the recommended strategy); however, Eliaz does not require an analogue of t-immunity.

In this paper, we are interested in the question of when a (k, t)-robust equilibrium σ in a game Γ_d with a mediator extending an underlying game Γ can be implemented by an ϵ–(k, t)-robust equilibrium σ' in the cheap-talk extension Γ_{CT} of Γ. If this is the case, we say that σ' is an ϵ–(k, t)-*robust* implementation of σ. (We sometimes say that $(\Gamma_{\mathrm{CT}}, \sigma')$ is an ϵ–(k, t)-*robust* implementation of (Γ_d, σ) if we wish to emphasize the games.)

3 The Possibility Results

All of our possibility results have the flavor "if there is a (k, t)-robust equilibrium in a game with a mediator, then (under the appropriate assumptions) we can implement this equilibrium using cheap talk." To state the results carefully, we must define the notions of a punishment strategy and a utility variant.

Definition 7. *If Γ_d is an extension of an underlying game Γ with a mediator d, a strategy profile ρ in Γ is a k-*punishment strategy* with respect to a strategy profile σ in Γ_d if for all subsets $K \subseteq N$ with $|K| \leq k$, all strategies ϕ in $\Gamma + CT(K)$, all types $t_K \in T_K$, and all players $i \in K$:*

$$u_i(t_K, \Gamma_d, \sigma) > u_i(t_K, \Gamma + CT(K), \phi_K, \rho_{-K}).$$

If the inequality holds with $\geq$ replacing $>$, ρ is a weak *k-punishment strategy with respect to σ.*

Intuitively, ρ is k-punishment strategy with respect to σ if, for any coalition K of at most k players, even if the players in K share their type information, as long as all players not in K use the punishment strategy in the underlying game, there is nothing that the players in K can do in the underlying game that will give them a better expected payoff than playing σ in Γ_d.

Notice that if $k + t < n \leq 2k + t$, Γ_d is a mediator game extending Γ, and σ is a (k, t)-robust equilibrium in Γ_d, then there cannot be a $(k + t)$-punishment strategy with respect to σ. For if σ is a $(k + t)$-punishment strategy, consider the strategy in the mediator game where a set T with $|T| = t \geq n - (k + t)$ players do not send a message to the mediator, and just play ρ. If σ is t-immune, $u_i(\sigma_{N-T}, \rho_T) \geq u_i(\sigma)$. But in the underlying game, if the players in $N - T$ share their types, they can compute what the mediator would have said, and thus can play σ_{N-T}, contradicting the assumption that σ is a punishment strategy.

The notion of utility variant is used to make precise that certain results do not depend on knowing the players' utilities; they hold independently of players' utilities in the game. A game Γ' is a *utility variant* of a game Γ if Γ' and Γ have the same game tree, but the utilities of the players may be different in Γ and Γ'. Note that if Γ' is a utility variant of Γ, then Γ and Γ' have the same set of strategies. We use the notation $\Gamma(u)$ if we want to emphasize that u is the utility function in game Γ. We then take $\Gamma(u')$ to be the utility variant of Γ with utility functions u'.

We say that *broadcast can be simulated*, if for all $\delta > 0$, broadcast channels can be implemented with probability $1 - \delta$. Broadcast can be simulated if, for example, there are broadcast channels; or if there is a trusted preprocessing phase where players may broadcast and assuming cryptography; or if unconditional pseudo-signatures are established [PW96].

In the theorem, we take "assuming cryptography" to be a shorthand for the assumption that *oblivious transfer* [Rab, EGL85] can be implemented with probability $1-\epsilon$ for any desired $\epsilon > 0$. It is known that this assumption holds if *enhanced trapdoor permutations* exist, players are computationally bounded, and the mediator can be described by a polynomial-size circuit [Gol04].

Theorem 1. *Suppose that Γ is Bayesian game with n players and utilities u, d is a mediator that can be described by a circuit of depth c, and σ is a (k, t)-robust equilibrium of a game Γ_d with a mediator d.*

(a) If $3(k + t) < n$, then there exists a strategy σ_{CT} in $\Gamma_{\mathrm{CT}}(u)$ such that for all utility variants $\Gamma(u')$, if σ is a (k, t)-robust equilibrium of $\Gamma_d(u')$, then $(\Gamma_{\mathrm{CT}}(u'), \sigma_{\mathrm{CT}})$ implements $(\Gamma_d(u'), \sigma)$. The running time of σ_{CT} is $O(c)$.

(b) If $2k + 3t < n$ and there exists a $(k + t)$-punishment strategy with respect to σ, then there exists a strategy σ_{CT} in Γ_{CT} such that σ_{CT} implements σ. The expected running time of σ_{CT} is $O(c)$.

(c) If $2(k + t) < n$ and broadcast channels can be simulated, then, for all $\epsilon > 0$, there exists a strategy $\sigma^{\epsilon}_{\mathrm{CT}}$ in Γ_{CT} such that $\sigma^{\epsilon}_{\mathrm{CT}}$ ϵ-implements σ. The running time of $\sigma^{\epsilon}_{\mathrm{CT}}$ is $O(c)$.

(d) If $k + t < n$ then, assuming cryptography and that a PKI is in place, there exists a strategy $\sigma^{\epsilon}_{\mathrm{CT}}$ in Γ_{CT} such that $\sigma^{\epsilon}_{\mathrm{CT}}$ ϵ-implements σ. The expected running time of $\sigma^{\epsilon}_{\mathrm{CT}}$ is $O(c) \cdot f(u) \cdot O(1/\epsilon)$ where $f(u)$ is a function of the utilities.

(e) *If $k + 3t < n$ or if $k + t < n$ and a trusted PKI is in place, and there exists a $(k + t)$-punishment strategy with respect to σ, then, assuming cryptography, there exists a strategy $\sigma^{\epsilon}_{\mathrm{CT}}$ in Γ_{CT} such that $\sigma^{\epsilon}_{\mathrm{CT}}$ ϵ-implementers σ. The expected running time of $\sigma^{\epsilon}_{\mathrm{CT}}$ is $O(c) \cdot f(u)$ where $f(u)$ is a function of the utilities but is independent of ϵ.*

Note that in part (a) we say "the running time", while in the other parts we say "the expected running time". Although all the strategies used are behavior strategies (and thus use randomization), in part (a), the running time is bounded, independent of the randomization. In the remaining parts, we cannot put an *a priori* bound on the running time; it depends on the random choices. As our lower bounds show, this must be the case.

We briefly comment on the differences between Theorem 1 and the corresponding Theorem 4 of ADGH. In ADGH, we were interested in finding strategies that were not only (k, t)-robust, but also survived iterated deletion of weakly dominated strategies. Here, to simplify the exposition, we just focus on (k, t)-robust equilibria. For part (a), in ADGH, a behavioral strategy was used that had no upper bound on running time. This was done in order to obtain a strategy that survived iterated deletion. However, it is observed in ADGH that, without this concern, a strategy with a known upper bound can be used. As we observed in the introduction, part (a), as stated, actually follows from known results in multiparty computation [BGW88, CCD88]. Part (b) here is the same as in ADGH. In part (c), we assume here the ability to simulate broadcast; ADGH assumes cryptography. As we have observed, in the presence of cryptography, we can simulate broadcast, so the assumption here is weaker. In any case, as observed in the introduction, part (c) follows from known results [RB89]. Parts (d) and (e) are new, and will be proved in [ADGH07]. The proof uses ideas from [GMW87] on multiparty computation. For part (d), where there is no punishment strategy, ideas from [EGL85] on getting ϵ-*fair* protocols are also required. (An ϵ-fair protocols is one where if one player knows the mediator's value with probability p, then other players know it with probability at least $p - \epsilon$.) Our proof of part (e) shows that if $n > k + 3t$, then we can essentially set up a PKI on the fly. These results strengthen Theorem 4(d) in ADGH, where punishment was required and n was required to be greater than $k + 2t$.

4 The Impossibility Results

No bounded implementations

We prove that it is impossible to get an implementation with bounded running time in general if $2k + 3t < n \leq 3k + 3t$. This is true even if there is a punishment strategy. This result is optimal. If $3k + 3t < n$, then there does exist a bounded implementation; if $2k + 3t < n \leq 3k + 3t$ there exists an unbounded implementation that has constant *expected* running time.

Theorem 2. *If $2k + 3t < n \leq 3k + 3t$, there is a game Γ and a strong (k, t)-robust equilibrium σ of a game Γ_d with a mediator d that extends Γ such that there exists a $(k + t)$-punishment strategy with respect to σ for which there do not exist a natural number c and a strategy σ_{CT} in the cheap talk game extending Γ such that the running time of σ_{CT} on the equilibrium path is at most c and σ_{CT} is a (k, t)-robust implementation of σ.*

Proof. We first assume that $n = 3$, $k = 1$, and $t = 0$. We consider a family of 3-player games $\Gamma_3^{n,k+t}$, where $2k + 3t < n \leq 3k + 3t$, defined as follows. Partition $\{1, \ldots, n\}$ into three sets B_1, B_2, and B_3, such that B_1 consists of the first $\lfloor n/3 \rfloor$ elements in $\{1, \dots, n\}$, B_3 consists of the last $\lceil n/3 \rceil$ elements, and B_2 consists of the remaining elements.

Let p be a prime such that $p > n$. Nature chooses a polynomial f of degree $k + t$ over the p-element field GF(p) uniformly at random. For $i \in \{1, 2, 3\}$, player i's type consists of the set of pairs $\{(h, f(h)) \mid h \in B_i\}$. Each player wants to learn $f(0)$ (the secret), but would prefer that other players do not learn the secret. Formally, each player must play either 0 or 1. The utilities are defined as follows:

- if all players output $f(0)$ then all players get 1;
- if player i does not output $f(0)$ then he gets -3;
- otherwise players i gets 2.

Consider the mediator game where each player is supposed to tell the mediator his type. The mediator records all the pairs (h, v_h) it receives. If at least $n - t$ pairs are received and there exists a unique degree $k + t$ polynomial that agrees with at least $n - t$ of the pairs then the mediator interpolates this unique polynomial f' and sends $f'(0)$ to each player; otherwise, the mediator sends 0 to each player.

Let σ_i be the strategy where player i truthfully tells the mediator his type and follows the mediator's recommendation. It is easy to see that σ is a $(1, 0)$-robust equilibrium (i.e., a Nash equilibrium). If a player i deviates by misrepresenting or not telling the mediator up to t of his shares, then everyone still learns; if the player misrepresents or does not tell the mediator about more of his shares, then the mediator sends the default value 0. In this case i is worse off. For if 0 is indeed the secret, which it is with probability 1/2, i gets 1 if he plays 0, and -3 if he plays 1. On the other hand, if 1 is the secret, then i gets 2 if he plays 1 and -3 otherwise. Thus, no matter what i does, his expected utility is at most $-1/2$. This argument also shows that if ρ_i is the strategy where i decides 0 no matter what, then ρ is a 1-punishment strategy with respect to σ.

Suppose, by way of contradiction, that there is a cheap-talk strategy σ' in the game Γ_{CT} that implements σ such that any execution of σ' takes at most c rounds. We say that a player i *learns the secret by round b of σ'* if, for all executions (i.e., plays) r and r' of σ' such that i has the same type and the same message history up to round b, the secret is the same in r and r'. Since we have assumed that all plays of σ' terminate in at most c rounds, it must be the case that all players learn the secret by round c of σ'. For if not, there are two executions r and r' of σ' that i cannot distinguish by round c, where the secret is different in r and r'. Since i must play the same move in r and r', in one case he is not playing the secret, contradicting the assumption that σ' implements σ. Thus, there must exist a round $b \le c$ such that all three players learn the secret at round b of σ' and, with nonzero probability, some player, which we can assume without loss of generality is player 1, does not learn the secret at round $b - 1$ of σ'. This means that there exists a type t_1 and message history h_1 for player 1 of length $b - 1$ that occurs with positive probability when player 1 has type t_1 such that, after $b - 1$ rounds, if player 1 has type t_1 and history h_1, player 1 considers it possible that the secret could be either 0 or 1. Thus, there must exist type profiles t and t' that correspond to polynomials f and f' such that $t_1 = t_1'$, $f(0) \neq f'(0)$ and, with positive probability, player 1 can have history h_1 with both t and t', given that all three players play σ'.

Let h_2 be a history for player 2 of length $b - 1$ compatible with t and h_1 (i.e., when the players play σ', with positive probability, player 1 has h_1, player 2 has h_2, and the true type profile is t); similarly, let h_3 be a history of length $b - 1$ for player 3 compatible with t' and h_1. Note that player i's action according to σ_i is completely determined by his type, his message history, and the outcome of his coin tosses. Let $\sigma_2'[t_2, h_2]$ be the strategy for player 2 according to which player 2 uses σ_2' for the first $b - 1$

rounds, and then from round b on, player 2 does what it would have done according to σ'_2 if its type had been t_2 and its message history for the first $b-1$ rounds had been h_2 (that is, player 2 modifies his actual message history by replacing the prefix of length $b-1$ by h_2, and leaving the rest of the message history unchanged). We can similarly define $\sigma'_3[t'_3, h_3]$. Consider the strategy profile $(\sigma'_1, \sigma'_2[t_2, h_2], \sigma'_3[t'_3, h_3])$. Since $\sigma'_i[t_i, h_i]$ is identical to σ'_i for the first $b-1$ steps, for $i = 2, 3$, there is a positive probability that player 1 will have history h_1 and type t_1 when this strategy profile is played. It should be clear that, conditional on this happening, the probability that player 1 plays 0 or 1 is independent of the actual types and histories of players 2 and 3. This is because players 2 and 3's messages from time b depend only on i's messages, and not on their actual type and history. Thus, for at least one of 0 and 1, it must be the case that the probability that player 1 plays this value is strictly less than 1. Suppose without loss of generality that the probability of playing $f(0)$ is less than 1.

We now claim that $\sigma'_3[t'_3, h_3]$ is a profitable deviation for player 3. Notice that player 3 receives the same messages for the first b rounds of σ' and $(\sigma'_1, \sigma'_2, \sigma'_3[t'_3, h_3])$. Thus, player 3 correctly plays the secret no matter what the type profile is, and gets payoff of at least 1. Moreover, if the type profile is t, then, by construction, with positive probability, after $b-1$ steps, player 1's history will be h_1 and player 2's history will be h_2. In this case, σ'_2 is identical to $\sigma'_2[t_2, h_2]$, so the play will be identical to $(\sigma'_1, \sigma'_2[t_2, h_2], \sigma'_3[t'_3, h_3])$. Thus, with positive probability, player 1 will not output $f(0)$, and player 3 will get payoff 2. This means player 3's expected utility is greater than 1.

For the general case, suppose that $2k + 3t < n \leq 3k + 3t$. Consider the n-player game $\Gamma^{n,k,t}$, defined as follows. Partition the players into three groups, B_0, B_1, and B_2, as above. As in the 3-player game, nature chooses a polynomial f of degree $k + t$ over the field $\{0, 1\}$ uniformly at random, but now player i's type is just the pair $(i, f(i))$. Again, the players want to learn $f(0)$, but would prefer that other players do not learn the secret, and must output a value in F. The payoffs are similar in spirit to the 3-player game:

- if at least $n - t$ players output $f(0)$ then all players that output $f(0)$ get 1;
- if player i does not output $f(0)$ then he gets -3;
- otherwise player i gets 2.

The mediator's strategy is essentially identical to that in the 3-player game (even though now it is getting one pair (h, v_h) from each player rather than a set of such pairs from a single player). Similarly, each player i's strategy in $\Gamma_d^{n,k,t}$, which we denote σ_i^n, is essentially identical to the strategy in the 3-player game with the mediator. Again, if ρ_i^n is the strategy in the n-player game where i plays 0 no matter what his type, then it is easy to check that ρ^n is a $(k + t)$-punishment strategy with respect to σ^n.

Now suppose, by way of contradiction, that there exists a strategy σ' in the cheap-talk extension $\Gamma_{\mathrm{CT}}^{n,k,t}$ of $\Gamma^{n,k,t}$ that is a (k, t)-robust implementation of σ^n such that all executions of σ' take at most c rounds. We show in Appendix A.3 that we can use σ' to get a $(1, 0)$-robust implementation in the 3-player mediator game $\Gamma_{3,d}^{n,k+t}$, contradicting the argument above. $\square$

Byzantine Agreement and Game Theory

In [ADGH06] it is shown that if $n > 3k+3t$, we can implement a mediator in a way that does not depend on utilities and does not need a punishment strategy. Using novel connections to randomized Byzantine agreement lower bounds, we show that neither of these properties hold in general if $n \leq 3k + 3t$.

We start by showing that we cannot handle all utilities variants if $n \le 3k+3t$. Our proof exposes a new connection between utility variants and the problem of *Weak Byzantine Agreement* [Lam83]. Lamport [Lam83] showed that there is no deterministic protocol with bounded running time for *weak Byzantine agreement* if $t \ge n/3$. We prove a stronger lower bound for any randomized protocol that only assumes that the running time has finite expectation.

Proposition 1. *If $\max\{2, k+t\} < n \le 3k+3t$, all 2^n input values are equally likely, and P is a (possibly randomized) protocol with finite expected running time (that is, for all protocols P'' and sets $|T| \le k+t$, the expected running time of processes P_{N-T} given (P_{N-T}, P''_T) is finite), then there exists a protocol P' and a set T of players with $|T| \le k+t$ such that an execution of (P_{N-T}, P'_T) is unsuccessful for the weak Byzantine agreement problem with nonzero probability.*

Proof. See Appendix A.1. □

The idea of our impossibility result is to construct a game that captures weak Byzantine agreement. The challenge in the proof is that, while in the Byzantine agreement problem, nature chooses which processes are faulty, in the game, the players decide whether or not to behave in a faulty way. Thus, we must set up the incentives so that players gain by choosing to be faulty iff Byzantine agreement cannot be attained, while ensuring that a (k,t)-robust cheap-talk implementation of the mediator's strategy in the game will solve Byzantine agreement.

Theorem 3. *If $2k+2t < n \le 3k+3t$, there is a game $\Gamma(u)$ and a strong (k,t)-robust equilibrium σ of a game Γ_d with a mediator d that extends Γ such that there exists a $(k+t)$-punishment strategy with respect to σ and there does not exist a strategy σ_{CT} such that for all utility variants $\Gamma(u')$ of $\Gamma(u)$, if σ is a (k,t)-robust equilibrium of $\Gamma_d(u')$, then $(\Gamma_{\mathrm{CT}}(u'), \sigma_{\mathrm{CT}})$ is a (k,t)-robust implementation of $(\Gamma_d(u'), \sigma)$.*

Proof. See Appendix A.1. □

Theorem 3 shows that we cannot, in general, get a *uniform* implementation if $n \le 3k+3t$. As shown in Theorem 1(b)–(e), we can implement mediators if $n \le 3k+3t$ by taking advantage of knowing the players' utilities.

We next prove that if $2k+3t < n \le 3k+3t$, although mediators can be implemented, they cannot be implemented without a punishment strategy. In fact we prove that they cannot even be ϵ–implemented without a punishment strategy. Barany [Bar92] proves a weaker version of a special case of this result, where $n = 3$, $k = 1$, and $t = 0$.It is not clear how to extend Barany's argument to the general case, or to ϵ–implementation. We use the power of randomized Byzantine agreement lower bounds for this result.

Theorem 4. *If $2k+2t < n \le 3k+3t$, then there exists a game Γ, an $\epsilon > 0$, and a strong (k,t)-robust equilibrium σ of a game Γ_d with a mediator d that extends Γ, for which there does not exist a strategy σ_{CT} in the CT game that extends Γ such that σ_{CT} is an ϵ–(k,t)-robust implementation of σ.*

Proof. See Appendix A.2. □

We now show that the assumption that $n > 2k+3t$ in Theorem 1 is necessary. More precisely, we show that if $n \le 2k+3t$, then there is a game with a mediator that has a (k,t)-robust equilibrium that does not have a (k,t)-robust implementation in a cheap-talk game. We actually prove a stronger result: we show that there cannot even be an ϵ–(k,t)-robust implementation, for sufficiently small ϵ.

Theorem 5. *If $k + 2t < n \leq 2k + 3t$, there exists a game Γ, a strong (k, t)-robust equilibrium σ of a game Γ_d with a mediator d that extends Γ, a $(k+t)$-punishment strategy with respect to σ, and an $\epsilon > 0$, such that there does not exist a strategy σ_{CT} in the CT extension of Γ such that σ_{CT} is an ϵ–(k, t)-robust implementation of σ.*

The proof of Theorem 5 splits into two cases: (1) $2k+2t < n \leq 2k+3t$ and $t \geq 1$ and (2) $k+2t < n \leq 2k + 2t$. For the first case, we use a reduction to a generalization of the Byzantine agreement problem called the (k, t)-Detect/Agree *problem*. This problem is closely related to the problem of *broadcast with extended consistency* introduced by Fitzi et al. [FHHW03].

Theorem 6. *If $2k + 2t < n \leq 2k + 3t$ and $t \geq 1$, there exists a game Γ, an $\epsilon > 0$, a strong (k, t)-robust equilibrium σ of a game Γ_d with a mediator d that extends Γ, and a $(k + t)$-punishment strategy with respect to σ, such that there does not exist a strategy σ_{CT} in the CT extension of Γ which is an ϵ–(k, t)-robust implementation of σ.*

Proof. See Appendix A.4. □

We then consider the second case of Theorem 5, where $k + 2t < n \leq 2k + 2t$. Since we do not assume players know when other players have decided in the underlying game, our proof is a strengthening of the lower bounds of [SRA81, Hel05].

Theorem 7. *If $k + 2t < n \leq 2k + 2t$, there exist a game Γ, an $\epsilon > 0$, a mediator game Γ_d extending Γ, a strong (k, t)-robust equilibrium σ of Γ_d, and a $(k + t)$-punishment strategy ρ with respect to σ, such that there is no strategy σ_{CT} that is an ϵ–(k, t)-robust implementation of σ in the cheap-talk extension of Γ, even with broadcast channels.*

Proof. See Appendix A.5. □

Our last lower bound using Byzantine agreement impossibility results gives a lower bound that matches the upper bound of Theorem 1(e) for the case that $n > k + 3t$. We show that a PKI cannot be set up on the fly if $n \leq k + 3t$. Our proof is based on a reduction to a lower bound for the (k, t)-*partial broadcast problem*, a novel variant of Byzantine agreement that can be viewed as capturing minimal conditions that still allow us to prove strong randomized lower bounds.

Theorem 8. *If $\max(2, k + t) < n \leq k + 3t$, then there is a game Γ, a strong (k, t)-robust equilibrium σ of a game Γ_d with a mediator d that extends Γ for which there does not exist a strategy σ_{CT} in the CT game that extends Γ such that σ_{CT} is an ϵ–(k, t)-robust implementation of σ even if players are computationally bounded and we assume cryptography.*

Proof. See Appendix A.6. □

Tight bounds on running time

We now turn our attention to running times. We provide tight bounds on the number of rounds needed to ϵ–implement equilibrium when $k + t < n \leq 2(k + t)$. When $2(k + t) < n$ then the expected running time is independent of the game utilities and independent of ϵ. We show that for $k + t < n \leq 2(k + t)$ this is not the case. The expected running time must depend on the utilities, and if punishment does not exist then the running time must also depend on ϵ.

Theorem 9. *If $k + t < n \leq 2(k + t)$ and $k \geq 1$, then there exists a game Γ, a mediator game Γ_d that extends Γ, a strategy σ in Γ_d, and a strategy ρ in Γ such that*

(a) for all ϵ and b, there exists a utility function $u^{b,\epsilon}$ such that σ is a (k, t)-robust equilibrium in $\Gamma_d(u^{b,\epsilon})$ for all b and ϵ, ρ is a (k, t)-punishment strategy with respect to σ in $\Gamma(u^{b,\epsilon})$ if $n > k+2t$, and there does not exist an ϵ–(k, t)-robust implementation of σ that runs in expected time b in the cheap-talk extension $\Gamma_{\mathrm{CT}}(u^{b,\epsilon})$ of $\Gamma(u^{b,\epsilon})$;

(b) there exists a utility function u such that σ is a (k, t)-robust equilibrium in $\Gamma_d(u)$ and, for all b, there exists ϵ such that there does not exist an ϵ–(k, t)-robust implementation of σ^i that runs in expected time b in the cheap-talk extension $\Gamma_{\mathrm{CT}}(u)$ of $\Gamma(u)$.

This is true even if players are computationally bounded, we assume cryptography and there are broadcast channels.

Proof. See Appendix A.7. □

Note that, in part (b), it is not assumed that there is a (k, t)-punishment strategy with respect to σ in $\Gamma(u)$. With a punishment strategy, for a fixed family of utility functions, we can implement an ϵ–(k, t)-robust strategy in the mediator game using cheap talk with running time that is independent of ϵ; with no punishment strategy, the running time depends on ϵ in general.

5 Conclusions

We have provided conditions under which a (k, t)-robust equilibrium with a mediator can be implemented using cheap talk, and proved essentially matching lower bounds. There are still a few gaps in our theorems, as well as other related issues to explore. We list some of them here.

- In Theorem 1(c), we get only an ϵ-implementation for some $\epsilon > 0$. Can we take $\epsilon = 0$ here?
- We require that the cheap-talk implementation be only a Nash equilibrium. But when we use a punishment strategy, this may require a player to do something that results in him being much worse off (although in equilibrium this will never occur, since if everyone follows the recommended strategy, there will never be a need for punishment). It may be more reasonable to require that the cheap-talk implementation be a *sequential equilibrium* [KW82] where, intuitively, a player is making a best response even off the equilibrium path. To ensure that the cheap-talk strategy is a sequential equilibrium, Ben-Porath [Ben03] requires that the punishment strategy itself be a Nash equilibrium. We believe for our results where a punishment strategy is not required, the cheap-talk strategy is in fact a sequential equilibrium and, in the cases where a punishment strategy is required, if we assume that the punishment strategy is a Nash equilibrium, then the cheap-talk strategy will be a sequential equilibrium. However, we have not checked this carefully. It would also be interesting to consider the extent to which the cheap-talk strategy satisfies other refinements of Nash equilibrium, such as *perfect equilibrium* [Sel75].
- Our focus in this paper has been the synchronous case. We are currently exploring the asynchronous case. While we had originally assumed that implementation would not be possible in the

asynchronous case, it now seems that many of the ideas of our possibility results carry over to the asynchronous case. However, a number of new issues arise. In particular, we need to be careful in dealing with uncertainty. The traditional assumption in game theory is to quantify all uncertainty probabilistically. But with asynchrony, part of the uncertainty involves, how long it will take a message to arrive and when agents will be scheduled to move. (In general, in an asynchronous setting, one player can make many moves before a second agent makes a single move.) It is far from clear what an appropriate distribution would be to characterize this uncertainty. Thus, the tradition in distributed computing has been to assume that an adversary decides message delivery time and when agents are scheduled. The results in the asynchronous case depend on how we deal with the uncertainty, which in turn affects the notion of equilibrium.

- We have assumed that in the cheap-talk game, every player can talk directly to every other player. It would be interesting to examine what happens if there is a communication network which characterizes which players a given player can talk to directly.
- In the definition of t-immunity and (k,t)-robustness, we have allowed the players in T to use arbitrary strategies. In practice, we may be interested only in restricting each player i in T to using a strategy in some predetermined set S_i'.

We hope to return to all these issues in future work.

Appendix

A Proofs

This section includes the proofs for all results stated in the main text. We repeat the statement of the results for the readers' convenience.

A.1 Proof of Theorem 3

In the weak Byzantine agreement problem, there are n processes, up to t of which may be faulty ("Byzantine"). Each process has some initial value, either 0 or 1. Some processes (chosen by nature) are faulty; their "intention" is to try to prevent agreement among the remaining processes. Each non-Byzantine process must decide 0 or 1. An execution of a protocol P is *successful for weak Byzantine agreement* if the following two conditions hold:

I. (Agreement:) All the non-Byzantine processes decide on the same value in $\{0, 1\}$.

II. (Weak Nontriviality:) If all processes are non-Byzantine and all processes have the same initial value i, then all the processes must decide i.

Proposition 1. *If $\max\{2, k + t\} < n \leq 3k + 3t$, all 2^n input values are equally likely, and P is a (possibly randomized) protocol with finite expected running time, then there exists a protocol P' and a set T of players with $|T| \leq k + t$ such that an execution of (P_{N-T}, P'_T) is unsuccessful for the weak Byzantine agreement problem with nonzero probability.*

Proof. The proof is based on the argument of [FLM86]. Partition the processes $N = \{1, \dots, n\}$ into three sets B_0, B_1, B_2 such that $|B_i| \leq k+t$. Let $r_1, \dots, r_n$ be the random tapes such that process i uses tape r_i.

Let c be an integer parameter that will be fixed later and consider the scenario consisting of $2cn$ processes arranged into $6cn$ sets $A_0, A_1, \dots, A_{6cn-1}$. The number of processes in the set A_i is $|B_{i \pmod 3}|$, and the *indexes* of processes A_i correspond to the indexes of processes in the set $B_{i \pmod 3}$. Thus, for each value j in N, there are $2c$ processes whose index is set to j. If $j \in B_i$ then there is exactly one such process in each set $A_{3\ell+i}$ for $\ell \in \{0, 1, \dots, 2c - 1\}$.

Each process whose index is $j \in N$ executes protocol P_j with random tape r_j. Messages sent by processes in A_i according to P reach the appropriate recipients in $A_{i-1 \pmod{6cn}}, A_i, A_{i+1 \pmod{6cn}}$; the processes in A_i start with 1 if $-6cn/4 \pmod{6cn} < i \leq 6cn/4 \pmod{6cn}$ and 0 otherwise.

Any two consecutive sets $A_i, A_{i+1 \pmod{6cn}}$ define a possible scenario denoted S_i where the non-faulty processes, $A_i, A_{i+1 \pmod{6cn}}$, execute P and the faulty processes simulate the execution of all the remaining sets of processes. Let e_i denote the probability that protocol P fails the weak Byzantine agreement conditions in scenario S_i.

Fix $c = 2^5 b$ and consider the scenarios S_1 and S_{3cn}. Since the expected running time is at most b then by Markov inequality, with probability at least $7/8$ the processes in S_1 require at most $8b$ rounds. So with probability at least $1/2$ both the processes in S_1 and the processes in S_{3cn} decide in at most $8b$ rounds, denote this event as $\mathcal{E}$.

Since we fixed $c = 2^5 b$ and all processes that have the same index use the same random tape then given $\mathcal{E}$, we claim that the processes in S_1 cannot distinguish their execution from an execution where all processes are non-faulty and begin with a 1 and similarly processes in S_{3cn} cannot distinguish their execution from an execution where all processes are non-faulty and begin with a 0. Therefore, given $\mathcal{E}$, the non-faulty processes in S_1 decide 1 and the non-faulty processes in S_{3cn} decide 0. Given $\mathcal{E}$, it cannot be the case that $e_i = 0$ for all i. Hence there must exist an index i such that $e_i > 0$ given $\mathcal{E}$.

Now consider the strategy P' for processes in $B_{i-1 \pmod 3}$ that guesses the random tapes of the processes in $B_i, B_{i+1 \pmod 3}$ and simulates the processes $A_0, A_1, \ldots, A_{6cn-1}$ except for the processes in $A_i, A_{i+1 \pmod{6cn}}$. With non-zero probability, the initial values of the non-faulty processes will be as the initial values of $A_i, A_{i+1 \pmod{6cn}}$. Given this, with probability $1/2$ event $\mathcal{E}$ occurs. Given this, the faulty processes may guess the tapes of the non-faulty processes with non-zero probability. Given this, with probability $e_i > 0$ the non-faulty processes will fail. $\square$

Theorem 3. *If $2k + 2t < n \leq 3k + 3t$, there is a game $\Gamma(u)$ and a strong (k,t)-robust equilibrium σ of a game Γ_d with a mediator d that extends Γ such that there exists a $(k+t)$-punishment strategy with respect to σ and there does not exist a strategy σ_{CT} such that for all utility variants $\Gamma(u')$ of $\Gamma(u)$, if σ is a (k,t)-robust equilibrium of $\Gamma_d(u')$, then $(\Gamma_{\mathrm{CT}}(u'), \sigma_{\mathrm{CT}})$ is a (k,t)-robust implementation of $(\Gamma_d(u'), \sigma)$.*

Proof. Consider the following game Γ with $2k + 2t < n \leq 3k + 3t$ players. A player's type is his initial value, which is either 0 or 1. We assume that each of the 2^n tuples of types is equally likely. Each player must choose a characteristic G (for "good") or B (for "bad"); if the player chooses G, he must also output either 0, 1, or PUNISH. The utility function u is characterized as follows:

- If there exists a set R of at least $n - (k+t)$ players that choose G such that
  - all players in R either commonly output 0 or commonly output 1; and
  - if all players in R have initial value i then the common output of R is i;

  then we call this a *good outcome*, the utility is 1 for players that choose G and output the same value as the players in R, and the utility is 0 for all other players.
- If $n - (k+t)$ or more players choose G and output PUNISH or all players choose B, then we call this a *punishment outcome*; the utility is -1 for all players.
- Otherwise we have a *bad outcome*, and the utility is $-2n$ for players that choose G and 2 for players that choose B.

For future reference, we take the utility function u^M to be identical to u, except that in a punishment outcome, a player that chooses B gets $2M$, while a player that chooses G gets $-2nM$ (so that $u = u^1$).

Consider the strategy σ_i for player i in the game Γ_d with a mediator based on Γ where i sends its value to the mediator, and chooses characteristic G and outputs the value the mediator sends if the value is in $\{0, 1\}$, and outputs 0 if the mediator sends PUNISH. The mediator sends PUNISH if there are less than $n - (k+t)$ values sent; otherwise it sends the majority value (in case of a tie, it sends 1). Let ρ_i be the strategy in the underlying game Γ of choosing B and outputting 0.

Lemma 2. *The strategy σ is a strong (k,t)-robust equilibrium in the utility variant game $\Gamma_d(u^M)$ for all M. Moreover, σ results in a good outcome and ρ is a $(k+t)$-punishment strategy with respect to σ.*

Proof. If $|T| \leq t$, and all the players in $N-T$ play σ, then the mediator will get at least $n-t$ values. If the majority value is i, then all the players in $N-T$ will decide i. Since $n > 2t$, there must be at least one player in $N-T$ that has type i. Moreover, $|N-T| \geq n-(k+t)$, so the players in $N-T$ constitute a set R that makes the outcome good. Thus we have t-immunity.

Now fix $K, T \subseteq N$ such that K, T are disjoint, $|K| \leq k$, and $|T| \leq t$. Clearly for any $\tau_{K\cup T} \in S_{K\cup T}$, and $i \notin K \cup T$ we have $u_i(\sigma_{-(K\cup T)}, \tau_{(K\cup T)}) = 1$. If at least $n-(k+t)$ players play σ then the outcome will be good and all the players that play σ will get a utility of 1, no matter what the other players do; moreover, any player that chooses B will get utility of 0. Here we need the fact that $n > 2k+2t$, so there cannot be two sets of size at least $n-(k+t)$ where the players output different values. It easily follows that σ is a (k,t)-robust equilibrium. Note that if any set of $n-(k+t)$ or more players play ρ in the underlying game Γ then, no matter what the remaining players do, the utility for all the players is -1, whereas, as we have seen, if $n-(k+t)$ or more players play σ in Γ_d, then these players get 1. Thus, σ is a $(k+t)$-punishment strategy with respect to σ. □

Returning to the proof of Theorem 3, by way of contradiction, suppose that there exists a strategy σ' in the CT extension Γ_{CT} of Γ such that $(\Gamma_{\mathrm{CT}}(u^M), \sigma')$ is a (k,t)-robust implementation of $(\Gamma_d(u^M), \sigma)$, for all M. Let P_i be the protocol for process i where process i sends messages according to σ'_i, taking its initial value to be its type, decides ℓ if σ'_i chooses G and outputs $\ell \in \{0,1\}$, and outputs 0 if i chooses B or chooses G and outputs PUNISH.

By Proposition 1, there exists a protocol P', sets K, T with $|K| \leq k$ and $|T| \leq T$, and an $\epsilon > 0$ such that $(P_{-(K\cup T)}, P'_{K\cup T})$ has probability ϵ of having an unsuccessful execution. Without loss of generality, we can assume that $|K| = k$ and $|T| = t$. (If not, we can just add $k-|K|$ processes to K and $t-|T|$ processes to T and have them all use protocol P.) Let σ''_j be the strategy where player j chooses B and sends messages according to P'_j. It is easy to see that $(\sigma'_{-(K\cup T)}, \sigma''_{K\cup T})$ results in a bad outcome whenever $(P_{-(K\cup T)}, P'_{K\cup T})$ results in an unsuccessful outcome. For if $(\sigma'_{-(K\cup T)}, \sigma''_{K\cup T})$ results in a punishment outcome, then all players not in $K \cup T$ output 0 with $(P_{-(K\cup T)}, P'_{K\cup T})$, so the outcome is successful. Thus, the probability of a bad outcome with $(\sigma'_{-(K\cup T)}, \sigma''_{K\cup T})$ is ϵ.

Fix $M > 2/\epsilon$. In the game $\Gamma_{\mathrm{CT}}(u^M)$, if $j \in K$, we have $u_j^M(\sigma'_{-(K\cup T)}, \sigma''_{K\cup T}) > 3$ (since j's expected utility conditional on a bad outcome is greater than $4/\epsilon$, and a bad outcome occurs with probability ϵ, while j's expected utility conditional on a good or punishment outcome is at least -1). Since σ' is a (k,t)–robust equilibrium, if $j \in K$, we must have $u_j^M(\sigma'_{-T}, \sigma''_T) \geq u_j^M(\sigma'_{-(K\cup T)}, \sigma''_{K\cup T}) > 3$. Thus, the probability of a bad outcome with $(\sigma'_{-T}, \sigma''_T)$ must be positive. Note that t-immunity guarantees that, for all $i \notin T$, $u_i^M(\sigma'_{-T}, \sigma''_T) \geq 1$. Thus, the total expected utility of the players in $N-T$ when playing $(\sigma'_{-T}, \sigma''_T)$ must be at least $n-t+2k$ (since the players in K have expected utility at least 3). However, in a good outcome, their total utility $n-(t+k)$; in a punishment outcome, their total utility is $-n+t < 0$; and in a bad outcome, their total utility is less than 0 (since even if all but one of the players in $N-(T-K)$ choose characteristic B and get utility $2M$, the player who chooses characteristic G gets utility $-2Mn$). Thus, the only way that the total expected utility of the players in $N-T$ can be greater than $n-t+2k$ is if $k=0$ and the probability of a bad outcome (or a punishment outcome) with $(\sigma'_{-(K\cup T)}, \sigma''_{K\cup T})$ is 0. This gives us the desired contradiction, and completes the proof of Theorem 3. □

A.2 Proof of Theorem 4

Theorem 4. *If $2k + 2t < n \leq 3k + 3t$, then there exists a game Γ, an $\epsilon > 0$, and a strong (k, t)-robust equilibrium σ of a game Γ_d with a mediator d that extends Γ, for which there does not exist a strategy σ_{CT} in the CT game that extends Γ such that σ_{CT} is an ϵ–(k, t)-robust implementation of σ.*

Proof. Consider a variant of the game described in the proof of Theorem 3:

Game Γ has $2k + 2t < n \leq 3k + 3t$ players. Players are partitioned into three sets B_1, B_2, B_3 such that $|B_i| \leq k + t$. Nature chooses three independent uniformly random bits $b_1, b_2, b_3 \in \{0, 1\}$ and gives each player in B_i the type b_i. Each player must choose a characteristic G or B; if the player chooses G, he must also output either 0 or 1. The utility function u is characterized as follows:

- If there exists a set R of at least $n - (k + t)$ players that choose G such that
  - all players in R either commonly output 0 or commonly output 1; and
  - if all players in R have initial value i then the common output of R is i;

  then we call this a *good outcome*, the utility is 1 for players that choose G and output the same value as the players in R, and the utility is 0 for all other players.
- Otherwise we have a *bad outcome*, and the utility is 0 for players that choose G and 16 for players that choose B.

Consider the same mediator as in Theorem 3, except that rather than sending PUNISH if fewer than $n - (k + t)$ values are sent, the mediator simply sends 0. Again, it is easy to see that the strategy σ of sending the true type and following the mediator's advice is a (k, t)-robust equilibrium in the mediator game. Note that there is no $(k + t)$-punishment strategy with respect to σ in this game.

Let σ' be any strategy in the cheap talk game Γ_{CT} such that for any set $K \cup T$ with $|K \cup T| \leq k + t$ and any protocol $\tau_{K \cup T}$ the expected running time of $(\sigma'_{N-(K \cup T)}, \tau_{K \cup T})$ is finite. Let P be the protocol for Byzantine agreement induced by σ'. Specifically, protocol P_i simulates σ'_i by giving it its initial value, sending messages according to σ'_i and finally decide on the same value that σ'_i outputs.

We use the following lower bound on randomized Byzantine agreement protocols.

Proposition 3. *If $2t < n \leq 3t$ and processes are partitioned into three sets B_1, B_2, B_3 such that $|B_i| \leq t$. Nature chooses three independent uniformly random bits $b_1, b_2, b_3 \in \{0, 1\}$ and gives each player in B_i the initial value b_i. Then there exists a function Ψ that maps protocols to protocols such that for any joint protocol P there exists a set T of processes such that $T = B_i$ for some $i \in \{1, 2, 3\}$ and the execution $(P_{N-T}, \Psi(P)_T)$ fails to reach Byzantine Agreement with probability at least 1/6. The running time of $\Psi(P)$ is polynomial in the number of players and the running time of P.*

Proof. The proof follows from [KY84]. See Proposition 5 for a self-contained proof that also handles this special case. $\square$

Let T be the set whose existence is guaranteed by Proposition 3 for protocol P. Then consider the strategy τ_T in Γ_{CT} where players choose B and play according to the protocol $\Psi(P)_T$. Since with probability at least $1/6$ the execution of $(P_{N-T}, \Psi(P)_T)$ fails to reach Byzantine Agreement, the execution of (σ'_{N-T}, τ_T) reaches a bad outcome with probability at least 1/6, and the expected utility of each member of T is > 2. Hence there does not exist a $(k + t)$-robust equilibrium in the cheap talk game that can ϵ-implement the equilibrium with a mediator for any $\epsilon < 1$. $\square$

A.3 Proof of Theorem 2

Theorem 2. *If $2k + 3t < n \leq 3k + 3t$, there is a game Γ and a strong (k,t)-robust equilibrium σ of a game Γ_d with a mediator d that extends Γ such that there exists a $(k+t)$-punishment strategy with respect to σ for which there do not exist a natural number c and a strategy σ_{CT} in the cheap talk game extending Γ such that the running time of σ_{CT} on the equilibrium path is at most c and σ_{CT} is a (k,t)-robust implementation of σ.*

Proof. It remains to show that we can use σ' as defined in the main text to get a $(1,0)$-robust implementation in the 3-player mediator game $\Gamma_{3,d}^{n,k+t}$, contradicting the argument above. The idea is straightforward. Player i in the 3-player game simulates the players in B_i in the n-player game, assuming that player $j \in B_i$ has type $(j, f(j))$. (Recall that, in the 3-player game, player i's type is a tuple consisting of $(j, f(j))$ for all $j \in B_i$.) In more detail, consider the strategy σ_i'' where, in each round of the 3-player game, player i sends player j all the messages that a player in B_i sent to a player in B_j in the n-player game (noting what the message is, to whom it was sent, and who sent it). After receiving a round k message, each player i in the 3-player game can simulate what the players in B_i do in round $k+1$ of the n-player game. If a player i does not send player j a message of the right form in the 3-player game, then all the players in B_j are viewed as having sent no message in the simulation. If all players in B_i decide on the same value in the simulation, then player i decides on that value in the 3-player game; otherwise, player i decides 0.

It is easy to see that σ'' implements σ in $\Gamma_3^{n,k+t}$ and there is a bound c such that all executions of σ take at most c rounds, because $\vec{\sigma}'$ implements $\vec{\sigma}^n$ in the n-player game and takes bounded time. It follows from the argument above that $\vec{\sigma}''$ cannot be (1,0)-robust. Thus, some player i must have a profitable deviation. Suppose without loss of generality that it is player 3, and 3's strategy when deviating is τ_3. Note that τ_3 can be viewed as prescribing what messages players in B_3 send to the remaining players in the game $\Gamma_{CT}^{n,k,t}$. (Recall, that if player 3 does not send a message to player j in the 3-player game that can be viewed as part of such as description, and player j is running σ_j'', then player j acts as if all the players in B_3 had sent the players in B_j no message at all. Thus, all messages from player i to player j in the 3-player game can be interpreted as messages from B_i to B_j in the n-player game.) Note that $t < |B_3| \leq k + t$. (We must have $k \geq 1$, since otherwise we cannot have $2k + 3t < n \leq 3k + 3t$, so $n \geq 3 + 3t$.) Choose a subset T of B_3 such that $|T| = t$. Let $\hat{\tau}_{B_3}$ be the strategy in the n-player cheap-talk game whereby the players in B_3 simulate τ_3, the players in $B_3 - T$ make the same decision as player 3 makes using τ_3, and the players in T make the opposite decision. It suffices to show that if all players in B_3 play $\hat{\tau}$, then the players in $B_3 - T$ are better off than they are playing σ'.

Note that every execution r of $(\sigma'_{N-B_3}, \hat{\tau}_{B_3})$ in the cheap-talk extension of $\Gamma^{n,k,t}$ corresponds to a unique execution r' of $(\sigma''_{\{1,2\}}, \tau_3)$ in the cheap-talk extension $\Gamma_3^{n,k,t}$. Thus, it suffices to show that the players in $B_3 - T$ do at least as well in r as player 3 does in r'. Let R_0', R_1', and R_2' be the set of executions of $(\sigma''_{\{1,2\}}, \tau_3)$ where player 3 gets payoff -3, 1, and 2, respectively. Let R_j be the set of executions of $(\sigma'_{N-B_3}, \tau'_{B_3})$ that correspond to an execution of R_j'. If $r' \in R_0'$, then clearly a player in $B_3 - T$ does at least as well in r as player 3 does in r'. If $r' \in R_1'$, then all three players play the secret in r'. Thus, in r, all the players in $B_3 - T$ play the secret, so they all get at least 1. Finally, if $r' \in R_2'$, then in r', player 3 plays the secret, and either player 1 or 2 does not. Hence some player in B_1 or B_2 does not play the secret in the cheap-talk extension of $\Gamma^{n,t,k}$. Moreover, all the players in T do not play the secret. Thus, at least $t+1$ players do not play the secret, so the players in $B_3 - T$ get 2.

This completes the argument.

□

A.4 Proof of Theorem 6

Theorem 6. *If $2k + 2t < n \leq 2k + 3t$ and $t \geq 1$, there exists a game Γ, an $\epsilon > 0$, a strong (k, t)-robust equilibrium σ of a game Γ_d with a mediator d that extends Γ, and a $(k + t)$-punishment strategy with respect to σ, such that there does not exist a strategy σ_{CT} in the CT extension of Γ which is an ϵ–(k, t)-robust implementation of σ.*

The proof uses a reduction to a generalization of the Byzantine agreement problem called the (k, t)-Detect/Agree *problem*, which, as we said, is closely related to the problem of *broadcast with extended consistency* introduced by Fitzi et al. [FHHW03]. We have two parameters, k and t. Each process has some initial value, either 0 or 1. There are at most $k + t$ Byzantine processes. Each non-Byzantine process must decide $0, 1$, or DETECT. An execution is *successful for Detect/Agree* if the following three conditions all hold (the first two of which are slight variants of the corresponding conditions in weak Byzantine agreement):

I. (Agreement:) If there are t or fewer Byzantine processes, then all non-Byzantine processes decide on the same value, and it is a value in $\{0, 1\}$.
II. (Nontriviality:) If all non-Byzantine processes have the same initial value v and no non-Byzantine process decides DETECT, then all the non-Byzantine processes must decide v.
III. (Detection validity:) If there are more than t Byzantine processes, then either all the non-Byzantine processes decide DETECT, or all non-Byzantine processes decide on the same value in $\{0, 1\}$.

Note that if $k = 0$, then clause III is vacuous, so the $(0, t)$-*Detect/Agree* problem is equivalent to Byzantine agreement with t faulty processes [LSP82]. Note that the non-triviality condition for Byzantine agreement requires all processes to decide v if they all had initial value v, even if there are some faulty processes. Thus, it is a more stringent requirement than the weak nontriviality condition of weak Byzantine agreement.

The following argument, from which it follows that there does not exist a protocol for the (k, t)-*Detect/Agree* problem if $n \leq 2k + 3t$, is based on a variant of the argument used for Proposition 1. If $T \subseteq N$, let $\vec{1}_T\vec{0}_{N-T}$ denote the input vector where the players in T get an input of 1 and the players in $N - T$ get an input of 0.

Proposition 4. *If $\max\{2, t\} < n \leq 2k + 3t$ and $t \geq 1$, then for all joint protocols P, there exist six scenarios $S_0, \ldots, S_5$, six protocols P_h^j, for $j = 0, 1$, $h = 0, 1, 2$, and a partition of the players into three nonempty sets B_0, B_1, and B_2 such that $|B_0| \leq t$, $|B_1| \leq k + t$, and $|B_2| \leq k + t$, and*

- *in scenario S_0, the input vector is $\vec{0}$, in S_1, it is $\vec{0}_{B_1 \cup B_2}\vec{1}_{B_0}$; and in S_2, it is $\vec{0}_{B_2}\vec{1}_{B_0 \cup B_1}$; in S_{3+h}, the input vector is the complement of the input vector in S_h, for $h = 0, 1, 2$ (that is, if process i has input ℓ in S_h, it has input $1 - \ell$ in S_{3+h}); for all protocols P^j, for $j = 0, 1$;*
- *in S_{3j+h} the processes in B_h are faulty and use protocol P_h^j, for $j = 0, 1$ and $h = 0, 1, 2$, while the remaining processes are correct and use protocol P;*
- *the processes in $B_{h \oplus_3 2}$ receive exactly the same messages in every round of both both S_{3j+h} and $S_{3j+h \oplus_6 1}$ (where we use $\oplus_\ell$ and $\ominus_\ell$ to denote addition and subtraction mod ℓ).*

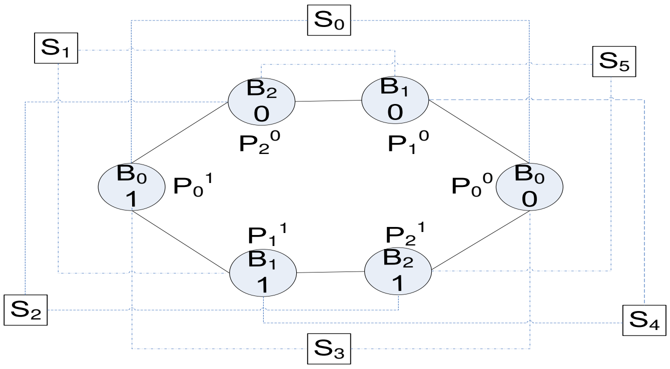

Figure 1: The construction for the proof of Proposition 4.

Proof. We explicitly construct the scenarios and protocols. Given P, we simultaneously describe the protocols P_h^j and scenarios S_{3j+h}, for $j = 0, 1$, $h = 0, 1, 2$, by induction, round by round. Suppose we have defined the behavior of protocol P_h^j and the scenario S_{3j+h} for the first ℓ rounds. As required by the proposition, in scenario S_{3j+h}, the processes in B_h are faulty and use protocol P_h^j, the correct processes use protocol P, and the inputs are as required by the protocol. The hexagon in Figure 1 implicitly defines the scenarios and what happens in round $(\ell+1)$. In round $\ell+1$ of scenario S_{3j+h}, the processes in B_h are faulty; each process $i \in B_h$ sends messages to the processes in $B_{h\oplus_3 1}$ as if i is using P_i and in the first k rounds has received exactly the messages it would have received in scenario $S_{3j+h\ominus_6 1}$, and sends messages to the processes in $B_{3j+h\ominus_3 1}$ as if i is using P_i and has received exactly the same messages it would have received in the first k rounds of $S_{3j+h\oplus_6 1}$. Note that in scenarios $S_{3j+h\oplus_6 1}$ and $S_{3j+h\ominus_6 -1}$, a process $i \in B_h$ uses protocol P_i. Thus, $i \in B_h$ sends the same message to processes in $B_{h\oplus_3 1}$ in both scenarios S_{3j+h} and $S_{3j+h\oplus_6 1}$. Finally, note that there is no need for processes in B_h to send messages to other processes in B_h in scenario S_{3j+h}. The behavior of the processes in B_h does not depend on the messages they actually receive in scenario S_{3j+h}; they are simulating scenarios $S_{3j+h\oplus_6 1}$ and $S_{3j+h\ominus_6 1}$. This behavior characterizes protocol P_j^h.

The six scenarios and protocols are implicitly defined by the hexagon in Figure 1. For example, scenario S_0 is defined by the four nodes in the hexagon starting with the one labeled P_0^0 and going counterclockwise. The inputs for the processes in B_h are defined by the numbers in circle in the first three nodes; in this scenario, all processes have input 0, because these numbers are all 0. The processes in B_0, which are the faulty ones in this scenario, behave to the processes in B_1 as if they have input 0 and to the processes in B_2 as if they have input 1. (This is indicated by the edge joining the node labeled $B_0(0)$ and $B_1(0)$ and the edge joining $B_0(1)$ and $B_2(0)$).

By construction, the processes in $B_{h\oplus_3 2}$ are correct in both S_{3j+h} and $S_{3j+h\oplus_6 1}$; an easy proof by induction on rounds shows that they receive exactly the same messages in every round of both scenarios. $\square$

The following is a generalization of the well-known result that Byzantine agreement (which, as we have observed, is just the $(0, t)$-*Detect/Agree* problem) cannot be solved if $n \leq 3t$.

Proposition 5. *If $n \leq 2k + 3t$ and $t \geq 1$, then there is no protocol that solves the (k, t)-*Detect/Agree *problem. Moreover, if there is a partition of the players into three nonempty sets B_0, B_1, and B_2, as in* Proposition 4*, and the four input vectors $\vec{0}$, $\vec{1}_{B_0}\vec{0}_{B_1 \cup B_2}$, $\vec{1}_{B_0 \cup B_1}\vec{0}_{B_2}$, and $\vec{1}$ each have probability $1/4$ then, for any protocol P, there exists a set T with $|T| \leq k + t$ and protocol P' such that the probability that an execution of (P_{N-T}, P'_T) is unsuccessful for (k, t)-Detect/Agree is at least $1/20$.*

Proof. Suppose that $n \leq 2k + 3t$ and that P solves the (k, t)-*Detect/Agree* problem. Consider the six scenarios from Proposition 4. Since $t \geq 1$, each of B_0, B_1, and B_2 is nonempty. In scenario S_0, all the correct processes must decide on 0, by the nontriviality condition. By Proposition 4, the processes in B_2 cannot distinguish S_0 from S_1 and are correct in both, therefore the correct processes must decide 0 in S_1. Thus, by the agreement property (if $|B_0| \leq t$) or the detection validity property (if $|B_2| > t$), all the processes in B_1 must also decide 0 in S_1. A similar argument shows that all the processes in $B_0 \cup B_1$ must decide 0 in S_2 and all the processes in $B_1 \cup B_2$ must decide 0 in S_3. But in S_3 the nonfaulty processes must decide 1. This proves the first part of the claim.

Now suppose that each of $\vec{0}$, $\vec{1}_{B_0}\vec{0}_{B_1 \cup B_2}$, $\vec{1}_{B_0 \cup B_1}\vec{0}_{B_2}$, and $\vec{1}$ have probability $1/4$. Again, consider the scenarios $S_0, \ldots, S_3$. We claim that, for some $j \in \{0, 1, 2, 3\}$, if the faulty processes use the strategies prescribed for scenario S_j, then conditional on the input vector being that of scenario S_j, the probability that an execution is unsuccessful for (k, t)-Detect/Agree is at least $1/5$. We prove by induction on j that either the claim holds for some scenario S_j with $0 \leq j \leq 3$ or the probability that each correct process in scenario S_j decides 0 is greater than $1 - (j + 1)/5$. For the base step, suppose that in scenario S_0 some correct process decides 1 with probability greater than $1/5$. Then if processes in B_0 are faulty and use protocol P_0^0, no matter what their input, if the input vector is actually $\vec{0}$ the execution will be unsuccessful with probability greater than $1/5$. Assume now that the claim does not hold for processes in S_0. Since processes in B_2 are correct in both S_0 and S_1, and cannot distinguish S_0 from S_1, they must decide 0 with probability at least $4/5$ in S_1. A similar argument now shows that either the claim holds for S_1, or the processes in B_0 must decide 0 with probability at least $3/5$ in S_1. The inductive step is similar, and left to the reader. We complete the proof by observing that in S_3, if the claim does not hold, then all the processes must decide 0 with probability at least $1/5$. But in S_3, all processes are correct and have initial value 1. Thus, again the probability of an unsuccessful execution is at least $1/5$. Since each relevant input vector has probability $1/4$, the probability of an unsuccessful execution is at least $1/20$. □

Proof of Theorem 6. Suppose that $2k+2t < n \leq 2k+3t, t \geq 1$. Consider the same game and strategies as in the proof of Theorem 3 with utility vector u^M, where $M > 20(1 + \epsilon)$. By Lemma 2, σ is a (k, t)-robust strategy in the mediator game Γ_d. Suppose, by way of contradiction, that σ' is an ϵ–(k, t)-robust implementation of σ in the CT extension Γ_{CT} of $\Gamma_d(u^M)$. Let P_i be the protocol for process i that sends the same messages and makes the same decisions as σ'_i. By Proposition 5, there is a protocol P' and a set T with $|T| \leq k + t$ such that the probability that an execution of (P_{N-T}, P'_T) is unsuccessful for (k, t)-Detect/Agree is at least $1/20$. Let σ'' be the strategy where the players choose characteristic B and play according P' in the cheap-talk game, no matter what their actual input. With probability $1/20$, the outcome will be bad (since an unsuccessful outcome with (P_{N-T}, P'_T) corresponds to a bad outcome with $(\sigma'_{N-T}, \sigma''_T)$. Thus, the expected utility for the players in T is greater than $1 + \epsilon$, so σ' is not ϵ–(k, t)-robust. □

	L	R
U	$(3,3)$	$(1,4)$
D	$(4,1)$	$(0,0)$

A simple 2-player game.

	L	R
U	$1/3$	$1/3$
D	$1/3$	0

A correlated equilibrium.

Figure 2: The game used in the proof of Proposition 6

A.5 Proof of Theorem 7

To prove Theorem 7, we start with the case that $n = 2$, $k = 1$, and $t = 0$. The ideas for this proof actually go back to Shamir, Rivest, and Adleman [SRA81]; a similar result is also proved by Heller [Hel05]. However, these earlier proofs assume that in the cheap-talk protocol, the players first exchange messages and then, after the message exchange, make their decision (perhaps using some randomization). That is, they are implicitly assuming that when the cheap-talk phase of the strategy has ended, it is common knowledge that it has ended (although when it ends may depend on some random choices). While this is a reasonable assumption if we have a bounded cheap-talk protocol, our possibility results involve cheap-talk games with no a priori upper bound on running time. We do not want to assume that the players receive a signal of some sort to indicate that the message exchange portion of the cheap-talk has ended. Our lower bound proof does not make this assumption. We can, of course, find a round such b that with high probability. This solves part of the problem. However, the earlier proofs also took advantage of the fact that players decide *simultaneously*, after the cheap-talk phase ends. Since we do not make this assumption, our proof requires somewhat more delicate techniques than the proofs in the earlier papers.

Proposition 6. *If $n = 2$, then there exist a game Γ, $\epsilon > 0$, a mediator game Γ_d extending Γ, a Nash equilibrium σ of Γ_d, and a punishment strategy ρ with respect to σ such that there is no strategy σ' that is an ϵ–$(1,0)$-robust implementation of σ.*

Proof. Let Γ be the game described in the left table of Figure 2, where player 1 is Alice and player 2 is Bob, Alice can choose between actions U and D, and Bob can choose between L and R. The players all have a single type in this game, so we do not describe the types. The boxes in the left table describe the utilities of Alice and Bob for each action profile. The right table describes a correlated equilibrium of this game, giving the probabilities that each action profile is played.

Consider the mediator game Γ_d extending Γ, where the mediator recommends the correlated equilibrium described in the table on the right of Figure 2 (that is, the mediator recommends choosing an action profile with the probability described in the table, and recommends that each player play his/her component of the action profile). Let σ be the strategy profile of following the mediator's recommendation. It is easy to see that σ is a Nash equilibrium of the mediator game; moreover, (D, R) is a punishment strategy with respect to σ. Also, note that the requirement of a broadcast channel trivially holds if $n = 2$.

Suppose, by way of contradiction, that $\sigma' = (\sigma'_1, \sigma'_2)$ is a $(1/10)$-Nash equilibrium that implements σ in a CT extension Γ_{CT} of Γ_d. Since σ' implements σ, it must be the case that, with probability 1, an execution of σ' terminates. Hence, there must be some round b such that, with probability at least $1 - \frac{1}{\beta^2}$, σ' has terminated (with both players choosing an action in the underlying game) by the end of round b. (We determine β shortly.)

The execution of σ' is completely determined by the random choices made by the players. Let $\mathcal{R}_i$ denote the set of possible sequences of random choices by player i. (For example, if player i randomizes by tossing coins, then $r \in \mathcal{R}_i$ can be taken to be a countable sequence of coin tosses.) For ease of exposition, we assume that i makes a random choice at every time step (if the move at some time step in the cheap-talk game is deterministic, then i can ignore the random choice at that time step). If r is a sequence of random choices, we use r^ℓ to denote the subsequence of r consisting of the random choices made in the first ℓ steps. Since σ' implements σ, with probability 1, both players choose an action in the underlying game in an execution of σ'. (That is, while it is possible that there are infinite executions of σ' where some party does not choose an action, they occur with probability 0.) Let $\mathcal{R} = \mathcal{R}_1 \times \mathcal{R}_2$. Note that the probability on the random sequences in $\mathcal{R}$ determines the probability of outcomes according to σ'.

Suppose that r is a finite random sequence of length ℓ for player i. We take $\Pr(r) = \Pr(\{s \in \mathcal{R}_i : s^\ell = r\})$. We similarly define $\Pr(r_1, r_2)$ for a pair (r_1, r_2) of finite sequences of equal length. The random sequences determine the message history and the actions. Given random sequences r_1 and r_2 of length ℓ, let $H(r_1, r_2)$ be the pair of message histories (h_1, h_2) determined by (r_1, r_2), and let $A(r_1, r_2) = (a_1, a_2)$ be the action profile chosen as a result of (r_1, r_2) (where we take a_i to be $\perp$ if player i has not yet taken an action); in this case, we write $A_i(r_1, r_2) = a_i$, for $i = 1, 2$.

A pair of histories (h_1, h_2) of equal length is *deterministic for player i* if it is not the case that both of player i's actions have positive probability, conditional on the message history being (h_1, h_2). We claim that all history pairs that arise with positive probability must be deterministic for some player i. For suppose that a history pair (h_1, h_2) is not deterministic for either player. Let $\mathcal{R}'_1 = \{(r_1, r_2) : A_1(r_1, r_2) = D, H(r_1, r_2) = (h_1, h_2)\}$ and let $\mathcal{R}'_2 = \{(r_1, r_2) : A_2(r_1, r_2) = R, H(r_1, r_2) = (h_1, h_2)\}$. By assumption, $\Pr(\mathcal{R}'_1 \mid (h_1, h_2)) > 0$ and $\Pr(\mathcal{R}'_2 \mid (h_1, h_2)) > 0$. Now let $\mathcal{R}'_3 = \{(r_1, r_2) : \exists r'_1, r'_2((r_1, r'_2) \in \mathcal{R}'_1, (r'_1, r_2) \in \mathcal{R}'_2\}$. It is easy to see that for $(r_1, r_2) \in \mathcal{R}'_3$, we have $H(r_1, r_2) = (h_1, h_2)$ (we can prove by a straightforward induction that $h(r_1^\ell, r_2^\ell) = (h_1^\ell, h_2^\ell)$ for each round ℓ less than or equal to the length of r_1). Hence $A(r_1, r_2) = (D, R)$. Moreover, $\Pr(\mathcal{R}'_3 \mid (h_1, h_2)) \geq \Pr(\mathcal{R}'_1 \mid (h_1, h_2)) \times \Pr(\mathcal{R}'_2 \mid (h_1, h_2)) > 0$. Thus, if (h_1, h_2) has positive probability, then the outcome (D, R) has positive probability, which contradicts the assumption that σ' implements σ.

Consider the following two strategies σ''_1 and σ''_2 for players 1 and 2, respectively. According to σ''_1, player 1 sends exactly the messages that he would have sent according to σ'_1 until round b, but does not take an action until the end of round b. If player 1 observes the histories (h_1, h_2) at the end of b rounds, and if the probability of player 2 playing L conditional on having observed (h_1, h_2) is at least $1 - 1/\beta$, then player 1 decides D, but keeps sending messages according to σ'_1; otherwise, player 1 plays exactly according to σ'_1. (Note that computing whether to play U may be difficult, but we are assuming computationally unbounded players here.) The strategy σ''_2 is similar, except now player 2 will play R if conditional on (h_1, h_2), player 1 plays U with probability at least $1 - 1/\beta$. It is easy to see that if player 1 plays a different action with σ''_1 when observing (h_1, h_2) than with σ'_1, it must be because σ'_1 recommends U and player 1 plays D. Moreover, player 1's expected gain in this case, conditional on observing (h_1, h_2) (given that player 2 plays σ'_2) is at least $1 \cdot (1 - 1/\beta) - 3 \cdot (1/\beta) = 1 - 4/\beta$. Similarly, if player 2 plays a different action when observing (h_1, h_2), then player 2's expected gain conditional on observing (h_1, h_2) and that player 1 plays σ'_1 is at least $1 - 4/\beta$.

Let $\mathcal{R}'(U, L) = \{(r_1, r_2) : A(r_1, r_2) = (U, L),\ r_1,\ r_2 \text{ have length } b\}$. Since σ' implements σ, the probability of the outcome (U, L) with σ' must be $1/3$. Since b was chosen such that the probability of not terminating within b rounds is less than $1/\beta^2$, we must have $\Pr(\mathcal{R}'(U, L)) > 1/3 - 1/\beta^2$.

Let H' consists of all message histories (h_1, h_2) of length b such that, with probability at least $1/\beta$, at least one player does not terminate by the end of round b (where the probability is taken over pairs (r_1, r_2) such that $H(r_1, r_2) = (h_1, h_2)$). The probability of H' must be at most $1/\beta$ (otherwise the probability of not terminating by the end of round b would be greater than $1/\beta^2$). Thus, for any history that is not in H', the probability that both players take an action in σ' is at least $1 - 1/\beta$.

Let $\mathcal{R}''(U, L) = \{(r_1, r_2) \in \mathcal{R}'(U, L) : H(r_1, r_2) \notin H'\}$. The discussion above implies that the inequality $\Pr(\mathcal{R}''(U, L)) > 1/3 - 1/\beta^2 - 1/\beta$ holds. By the arguments above, at least half of the histories in $\mathcal{R}''(U, L)$ are deterministic for one of the players. Without loss of generality, let it be player 1 (Alice). With probability at least $1/2 \cdot (1/3 - 1/\beta^2 - 1/\beta)$, Alice would have made a choice of U by the end of round b, if she had taken an action. With probability $1 - 1/\beta$ she will take an action, and therefore Bob can play σ_2'' and will gain an expected utility $(1 - 4/\beta) \cdot 1/2 \cdot (1/3 - 1/\beta^2 - 1/\beta)$. We can choose β such that the expected gain is at least $1/10$. Thus, σ' is not a $(1/10)$-equilibrium. By multiplying all utilities in Γ by 10ϵ, we get a game Γ^ϵ such that Γ_d^ϵ has a (1,0)-robust equilibrium that has no ϵ-implementation. □

We now prove Theorem 7 by generalizing this result to arbitrary k and t.

Theorem 7. *If $k + 2t < n \leq 2(k + t)$ there exist a game Γ, an $\epsilon > 0$, a mediator game Γ_d extending Γ, a strong (k, t)-robust equilibrium σ of Γ_d, and a $(k + t)$-punishment strategy ρ with respect to σ such that there is no strategy σ_{CT} that is an ϵ–(k, t)-robust implementation of σ in the cheap-talk extension of Γ, even with broadcast channels.*

Proof. Divide the players into three disjoint groups: group A_1 and A_2 have each have $n - (k + t)$ members, and group B has $2k + 2t - n$ members. It is immediate that this can be done, since $2(n - (k - t)) + 2k + 2t - n = n$. Moreover, $|A_1 \cup B| = |A_2 \cup B| = k + t$. Note that neither A_1 nor A_2 is empty; however, B may not have any members.

Players do not get any input (i.e., there is only one type). Intuitively, a player must output a value in field F, with $|F| \geq 6$, signed by a check vector. More precisely, a player $i \in A_1 \cup A_2$ outputs $8(n - 1) + 2$ elements of F, and optionally either U or D if $i \in A_1$ or L or R if $i \in A_2$. The first two elements of F output by i are denoted a_{i1} and a_{i2}. We think of these as i's share of two different secrets. The remaining $8(n - 1)$ elements of F consist of $n - 1$ tuples of 8 elements, denoted $(y_{1ij}, y_{2ij}, b_{1ji}, b_{2ji}, b_{3jij}, c_{1ji}, c_{2ji}, c_{3ji})$. We have one such tuple for each player $j \neq i$. We require that neither b_{1ji} nor b_{2ji} is 0. A player i in group B must output $3 + 9(n - 1)$ numbers, denoted $a_{hi}, y_{hij}, b_{hji}, c_{hji}$, for $h = 1, 2, 3$ and for each player $j \neq i$ (again, we require that $b_{hji} \neq 0$); together with an optional element, which can be any of U, D, L, or R. For each player $j \neq i$, we would like to have

$$a_{hi} + y_{hij} b_{hij} = c_{hij}, \tag{1}$$

for $h = 1, 2$; if $i \in B$, then we also would like to have $a_{3i} + y_{3ij} b_{3ij} = c_{3ij}$. Note that the field elements a_{hi} and y_{hij} are output by player i, while b_{hij} and c_{hij} are output by player j. As we said, we think of the values a_{hi} to be shares of some secret. Thus, we would like there to be a function f_1 which interpolates the values a_{1i}; $f_1(0)$ is intended to encode an action in $\{U, D\}$ for the players in A_1 to play. That is, if $f_1(0) = 0$, then the players in A_1 should play U; if $f_1(0) = 1$, then they should play D.) Similarly we would like there to be a function f_2 that encodes an action in $\{L, R\}$ for the players in A_2 to play. Finally, the values a_{3i} should encode a pair of values in $\{U, D\} \times \{L, R\}$, where (U, L) is encoded by 0, (U, R) by 1, (D, L) by 2, and (D, R) by 3.

The payoffs are determined as follows, where we take a player to be *reliable* if there exists a set N' of $n-1-t$ agents $j \neq i$ such that Eq. (1) holds for $k=1$ and $k=2$, and for $k=3$ if $i \in B$.

- If more than t players are unreliable, then all players get 0.
- If all the players in group $A_i \cup B$ play the same optional value and some are unreliable, then all players get 0.
- If t or fewer players are unreliable, but either (a) there does not exist a unique polynomial f_1 of degree $k+t-1$ that interpolates the values a_{1i} sent by the reliable players in $i \in A_1 \cup B$ such that $f_1(0) \in \{0,1\}$; (b) there does not exist a unique polynomial f_2 of degree $k+t-1$ that interpolates the values a_{1i} sent by reliable players i in group A_2 and the values a_{2i} sent by reliable players in group B such that $f_2(i) \in \{0,1\}$; (c) $(f_1(0), f_2(0))$ encodes (D,R); or (d) there does not exist a unique polynomial f_3 of degree $k+t-1$ that interpolates the values a_{2i} sent by reliable players $i \in A_1 \cup A_2$ and the values a_{3i} sent by reliable players $i \in B$ such that $f_3(0)$ encodes $(f_1(0), f_2(0))$, in the sense described above, then all players get $8/3$;
- if (a), (b), (c), and (d) above all do not hold, then suppose that $g(0)$ encodes (x,y). Let o_1 be x unless everyone in $A_1 \cup B$ plays x' as their optional value, where $x' \in \{U,D\}$, in which case $o_1 = x'$. Similarly, let o_2 be y unless everyone in groups $A_2 \cup B$ plays y' as their optional value, where $y \in \{L,R\}$, in which case $y' = o_2$. Let (p_1,p_2) be the payoff according to (o_1,o_2), as described in Figure 2. Then everyone in group A_i gets p_i, for $i=1,2$. Payoffs for players in B are determined as follows: if everyone in $A_1 \cup B$ played $x' \in \{U,D\}$ as their optional value, then players in B gets p_1; if everyone in $A_2 \cup B$ played $y' \in \{L,R\}$ as their optional value, then everyone in B gets p_2; otherwise, everyone in B gets $8/3$.

The mediator chooses an output (o_1,o_2) according to the distribution described in Figure 2. The mediator then encodes o_i as the secret of a degree $k+t-1$ polynomial f_i; that is, $o_i = f_i(0)$ and encodes (o_1,o_2) as the secret of a degree $k+t$ polynomial g. Suppose players $1,\dots n-(k+t)$ are in group A_1; players $n-(k+t)+1,\dots,2(n-(k+t))$ are in group A_2; and players $2(n-(k+t))+1,\dots,n$ are in group B. The mediator sends each player j in group A_i $(f_i(j), g(j))$, for $i=1,2$, and sends each player j in group B $(f_1(j), f_2(j), g(j))$. In addition, the mediator sends all players consistent check vectors such that $a_1(i) + y_{1ij}b_{1ij} = c_{1ij}$, $a_2(i) + y_{2ij}b_{2ij} = c_{2ij}$ and $a_3(i) + y_{3ij}b_{3ij} = c_{3ij}$ for all $i,j \in N$. If the players play the message sent by the mediator (and do not play the optional value), then they get expected payoff $8/3$.

We now show that this strategy is (k,t) robust. For t-immunity, note that the t players cannot take over all of $A_i \cup B$ for $i=1$ or $i=2$ (since both of these sets have cardinality $k+t$), so they cannot take advantage of sending the optional element of $\{U,D,L,R\}$. If any of the t players are shown to be unreliable, it is easy to see that this cannot hurt the other players, since there will not be more than t unreliable players. If the reliable players do not send values that pass checks (a), (b), (c), and (d) above, then each player gets $8/3$, which is the expected payoff of playing the recommended strategy. Finally, if a large enough subset of the t players manage to guess the check vectors and send values that satisfy (a), (b), (c), and (d) above, because they do not know any of the secrets, they are effectively making a random change to the output, so the expected payoff is still $8/3$.

For (k,t) robustness, note that a set of $k+t$ that consist of all of $A_i \cup B$ for $i=1$ or $i=2$ can change the output by all playing the same optional value, but they cannot improve their payoff this way, since we have a correlated equilibrium in the 2-player game. If all $k+t$ players are in group $A_1 \cup B$, and their value is U, they can try to change the outcome to (D, L) by all playing D and guessing shares and check values in the hope of changing them to the (D, L) outcome. But if they are caught, they will all get 0. Since $|F| \geq 6$, the probability of getting caught is greater than $1/4$, so they do not gain by deviating in this way.

Note that if $n > k + 2t$, then if $n - (k+t)$ players deliberately do not play the values sent them by the mediator, then this is a $(k+t)$-punishment strategy with respect to σ, since $n - (k+t) > t$.

Finally, the same argument as in the 2-player game shows that, by taking over either $A_1 \cup B$ or $A_2 \cup B$, a set of size $k+t$ can improve their outcome by deviating. □

A.6 Proof of Theorem 8

Theorem 8. *If $\max(2, k+t) < n \leq k+3t$, then there is a game Γ, a strong (k,t)-robust equilibrium σ of a game Γ_d with a mediator d that extends Γ for which there does not exist a strategy σ_{CT} in the CT game that extends Γ such that σ_{CT} is an ϵ–(k,t)-robust implementation of σ even if players are computationally bounded and we assume cryptography.*

Proof. We consider a relaxation of Byzantine agreement that we call the *(k,t)-partial broadcast problem*. There are n processes and process 1 is designated as leader. The leader has an initial value 0 or 1. Each process must decide on a value in $\{0, 1, \text{PASS}\}$. An execution of a protocol P is *successful for (k,t)-partial broadcast* if the following two conditions hold:

I. (Agreement): If there are t or fewer Byzantine processes and the leader is non-Byzantine then all non-Byzantine processes decide on the leader's value.
II. (No disagreement): If there are $k+t$ or fewer Byzantine processes, then there do not exist two non-Byzantine processes such that one decides 0 and the other decides 1.

Note that if the leader is faulty, it is acceptable that some non-Byzantine processes decide on a common value $v \in \{0,1\}$ and all other non-Byzantine processes decide PASS. Observe that the $(0,t)$-partial broadcast problem is a relaxation of the well-known *Byzantine generals* problem [LSP82]. We provide probabilistic lower bounds for this problem, which also imply known probabilistic lower bounds for the Byzantine generals problem.

Proposition 7. *If $\max(2, k+t) < n \leq k+3t$ and each input for the leader is equally likely. Then there exists a function Ψ that maps protocols to protocols such that for all joint protocols P, there exists a set T of processes such that either*

(a) $|T| \leq t$, $1 \notin T$ and the execution $(P_{N-T}, \Psi(P)_T)$ fails the agreement property with probability at least 1/6; or

(b) $|T| \leq k+t$, $1 \in T$ and the execution $(P_{N-T}, \Psi(P)_T)$ fails the no-disagreement property with probability at least 1/6.

The running time of $\Psi(P)$ is polynomial in the number of players and the running time of P.

Proof. Partition the n players into 3 nonempty sets B_0, B_1, and B_2 such that $|B_0| \leq t$, $|B_1| \leq k+t$, and $|B_2| \leq t$. Assume that B_1 contains process 1 (the leader). Consider the scenario consisting of $2n$ processes partitioned into six sets $A_0, A_1, \ldots, A_5$ such that the processes of A_i have the indexes of the processes of $B_{i \pmod 3}$; messages sent by processes in A_i according to P reach the appropriate recipients in $A_{i-1 \pmod 6}, A_i, A_{i+1 \pmod 6}$. For example, if a processes $\ell \in A_i$ has index $j \in N$ and is supposed to send a messages to a processes with index $j' \in N$ such that $j' \in B_{i' \pmod 3}$ and $i' \in \{i-1, i, i+1\}$ then this message will reach the process $\ell' \in A_{i'}$ whose index is j'; the process with index 1 (the leader) starts with initial value 1 in A_1 with with initial value 0 in A_4.

Any two consecutive sets $A_i, A_{i+1 \pmod 6}$ define a possible scenario denoted S_i, where the non-faulty processes, $A_i \cup A_{i+1 \pmod 6}$, execute P, and the faulty processes simulate the execution of all the 4 remaining sets of processes. For $i \in \{0,1,3,4\}$, let e_i denote the probability that protocol P fails the agreement condition in scenario S_i; for $i \in \{2,5\}$ let e_i denote the probability that protocol P fails the no-disagreement condition in scenario S_i.

We claim that $e_2 \geq 1-(e_1+e_3)$. Indeed with probability $1-e_1$ processes in A_2 succeed in S_1 which implies that processes in A_2 must decide 1 in this case. Similarly, with probability $1-e_3$ processes in A_3 must decide 0 due to S_3, hence in S_2, the non-faulty processes must reach disagreement with probability at least $1-(e_1+e_3)$. A symmetric argument gives $e_5 \geq 1-(e_4+e_0)$.

Therefore it cannot be the case that for all $i \in \{0,1,2\}$, $e_i + e_{i+3} < 2/3$. Let i be an index such that $e_i + e_{i+3} \geq 2/3$ and consider the set $B_{i-1 \pmod 3}$ of processes that are Byzantine and simulate four sets of processors according to scenario S_i or $S_{i+3 \pmod 6}$ with uniform probability.

Given that $e_i + e_{i+3 \pmod 6} \geq 2/3$ then the expected probability of failure is $1/3$ if the faulty players know the initial value. Since they can guess the initial value, the faulty players in $B_{i-1 \pmod 3}$ will cause P to fail with probability at least $1/6$. If $i=1$ then $|B_i| \leq t+k$, $1 \in B_i$ and the no-disagreement condition fails, and if $i \in \{0,2\}$ then $|B_i| \leq t$, $1 \notin B_i$ and the agreement condition fails. □

We now construct a game that captures the (k,t)-partial broadcast problem. Given k and t, consider the following game Γ with n players. Player 1 is the *broadcaster* and has two possible types, 0 or 1, both equally likely. Each player must choose a characteristic G or B and output a value in $\{0, 1, \text{PASS}\}$. Let $M > 6 + 6\epsilon$. We define the utility function u as follows.

- If player 1 (the broadcaster) has type v, then
  - if there exists a set S of at least $n-t$ players such that all players in S choose G and output v, and the broadcaster chooses G, then the broadcaster gets 1;
  - if there does not exist a set R of at least $n-t$ players and a value $v' \in \{0,1\}$ such that all players in R choose characteristic G and output a value in $\{v', \text{PASS}\}$, and the broadcaster chooses characteristic B, then the broadcaster gets M;
  - in all other cases, the broadcaster gets 0.
- Utility for player $i \neq 1$:
  - if there exists a set R of at least $n-(k+t)$ players and a value $v' \in \{0,1\}$ such that all players in R choose G and all players in R output a value in $\{v', \text{PASS}\}$, then player i gets 1 if he chooses G and outputs a value in $\{v', \text{PASS}\}$ and gets 0 otherwise;
  - in all other cases, if i chooses B he gets M and if he chooses G he gets 0.

Consider a mediator that receives a value from the broadcaster and sends this value to all players. It is easy to see that the strategy σ where the broadcaster truthfully tells the mediator his type and chooses characteristic G, and all other players choose characteristic G and output the value sent by the mediator, is a (k,t)-robust equilibrium whose payoff is 1 for all players.

We claim that we cannot ϵ-implement this mediator using cheap talk if $n \leq k+3t$. Suppose, by way of contradiction, that there exists a strategy σ' in Γ_{CT} that ϵ-implements σ. Much as in the proof of Theorem 3, we can transform σ' into a protocol P for the (k,t)-partial broadcast problem: take P_i to be the strategy where process i sends messages according to σ'_i, taking its initial value to be its type if $i=1$, and decides on the value output by σ'_i.

Viewing P as a protocol for the (k,t)-partial broadcast problem, let T be the set of processes guaranteed to exist by Proposition 7. If $|T| \leq t$, $1 \notin T$ and the execution $(P_{N-T}, \Psi(P)_T)$ fails the agreement condition with probability at least 1/3, then it is easy to see that σ' is not ϵ–t-immune. Otherwise, if $|T| \leq k+t$, $1 \in T$ and the execution $(P_{N-T}, \Psi(P)_T)$ fails the no-disagreement condition with probability at least 1/6, then it is easy to see that in Γ, deviating to $\Psi(P)$ and choosing B gives the members of T an expected utility greater than $M/6 = 1+\epsilon$, contradicting the assumption the σ' is $\epsilon-(k,t)$-robust. □

A.7 Proof of Theorem 9

Theorem 9. *If $k+t < n \leq 2(k+t)$ and $k \geq 1$, then there exists a game Γ, a mediator game Γ_d that extends Γ, a strategy σ in Γ_d, and a strategy ρ in Γ such that*

(a) for all ϵ and b, there exists a utility function $u^{b,\epsilon}$ such that σ is a (k,t) robust equilibrium in $\Gamma_d(u^{b,\epsilon})$ for all b and ϵ, ρ is a (k,t)-punishment strategy with respect to σ in $\Gamma(u^{b,\epsilon})$ if $n > k+2t$, and there does not exist an ϵ–(k,t)-robust implementation of σ that runs in expected time b in the cheap-talk extension $\Gamma_{\mathrm{CT}}(u^{b,\epsilon})$ of $\Gamma(u^{b,\epsilon})$,

(b) there exists a utility function u such that σ is a (k,t) robust equilibrium in $\Gamma_d(u)$ and, for all b, there exists ϵ such that there does not exist an ϵ–(k,t)-robust implementation of σ^i that runs in expected time b in the cheap-talk extension $\Gamma_{\mathrm{CT}}(u)$ of $\Gamma(u)$.

This is true even if players are computationally bounded, we assume cryptography and there are broadcast channels.

Proof. First assume that $k=1$, $t=0$, and $n=2$. Consider a 2-person secret-sharing game Γ with the secret taken from the field $F=\{0,\ldots,6\}$, and the shares are signed using check vectors. Specifically, nature uniformly chooses a secret $s \in F$ and a degree 2 polynomial f over F such that $f(0)=s$ and all remaining coefficients are uniformly random. Nature also chooses for $i \in \{1,2\}$, check vectors: y_i uniformly in F, b_i uniformly in $F \setminus \{0\}$ and c_i such that $f(i)+y_ib_i = c_i$. Let $\bar{i} = 3-i$; thus, $\bar{i}$ is the player other than i.

Player $i \in \{1,2\}$ gets as input $(f(i), y_i, b_{\bar{i}}, c_{\bar{i}})$ and must guess the secret. Let $u_i^{M,\delta}$ be the following utility function for player $i \in \{1,2\}$: If player i gets the right answer (i.e., guesses the secret) and $\bar{i}$ does not, then i gets M; if both get the right answer, then i gets 1; if $\bar{i}$ gets the right answer and i does not, then i gets $-M+2-2\delta$; finally, if neither get the right answer, then i gets $1-\delta$.

Consider the following mediator. It expects to receive from each player i 4 field values: a share a_i, a signature y_i, and two verification values $b_{\bar{i}}, c_{\bar{i}}$. If any player does not send 4 values then the mediator chooses a value in F at random, and sends that to both players. Otherwise the mediator interpolates the degree 1 polynomial f from the shares a_1 and a_2. Then it checks that $a_i + y_i b_i = c_i$ for $i \in \{1, 2\}$. If both checks are successful, he sends $f(0)$ to both players, otherwise he sends both a value in $F \setminus \{f(0)\}$ chosen uniformly at random.

Consider the truthful strategy σ in the mediator game: players tell the mediator the truth, and the mediator reports the secret. The strategy profile σ gives both players utility 1 for all utility function $u^{M,\delta}$. Truthfulness is easily seen to be a 1-robust strategy in the mediator game (i.e., a Nash equilibrium). If a player i lies and $\bar{i}$ tells the truth, then with probability $6/7$, i will be caught. In this case, i will definitely play the wrong value. (Note that i will play the right value if $f(0)$ is sent, since i will be able to calculate what the true secret should have been, given his lie; his calculation will be incorrect if a value other than $f(0)$ is sent, which is what happens if i's lie is detected.) On the other hand, $\bar{i}$ will play the right value with probability $1/6$. Thus, i's expected utility if he is caught is $5(1-\delta)/6 + (-M + 2 - 2\delta)/6 = (7 - 7\delta - M)/6$. If player i is not caught, then his utility is M. Thus, cheating has expected utility $1 - \delta$, so i does not gain by lying as long as $\delta \geq 0$.

Moreover, it is easy to see that, if $\delta > 0$ and i chooses a value at random, then if player $\bar{i}$ chooses the same value, his expected utility is $1/7 + 6/7(1-\delta) = 1 - 6\delta/7$; if player $\bar{i}$ chooses a different value, then his expected utility is $M/7 + (-M + 2 - 2\delta)/7 + (1-\delta)(5/7) = 1 - \delta$; it follows that player $\bar{i}$'s expected utility is at most $1 - 6\delta/7$. Thus, if $\delta > 0$, then choosing a value at random is a 1-punishment strategy with respect to σ in $\Gamma(u_{M,\delta})$ (even if the other player knows what value is chosen).

For part (a), fix $\epsilon > 0$ and b. We show that there is no cheap-talk strategy σ_{CT} that ϵ-implements σ and has expected running time b in $\Gamma_{\mathrm{CT}}(u^{M,\epsilon})$ if $M > (1-\epsilon) + 28b\epsilon/3$. Suppose, by way of contradiction, that there exists such a cheap-talk strategy σ_{CT}. The key idea is to consider the expected probability that a player will be able to guess the correct answer at any round, assuming that both players use σ_{CT}. With no information, the probability that i guesses the right answer is $1/7$ and all values are equally likely. When the strategy terminates, the probability must be 1 (because we assume that both players will know the right answer at the end if they follow the recommended strategy). In general, at round j, player i has acquired some information I^j. (What I^j is may depend on the outcome of coin tosses, of course.) There is a well-defined probability of guessing the right answer given I^j. Thus, the expected probability of player i guessing the right answer after round j is the sum, taken over all the possible pieces of information I^j that i could have at the end of round j, of the probability of getting information I^j times the probability of guessing the right answer given I^j. By Markov's inequality, both players terminate by round $2b$ with probability at least $1/2$, the expected probability of guessing the right answer by round $2b$ must be at least $4/7$ for both players. (Since if an execution terminates, he can guess the right answer with probability 1; otherwise, he can guess it with probability at least $1/7$.) Thus, for each player i, there must be a round $b' < b$ such that the expected probability of player i getting the right answer increases by at least $3/7b$ between round b' and $b' + 1$. It follows that there must be some round $b' < b$ such that either the expected probability of player 1 guessing the answer after round $b' + 1$ is at least $3/14b$ more than that of player 2 guessing the answer at round after round b', or the expected probability player 2 guessing the answer after round $b' + 1$ is at least $3/32b$ more than that of player 1 guess the answer after round b'. (Proof: Consider a round b' such that player 1's expected probability of guessing the right answer increases by at least $3/7b$. If player 2's probability of guessing the right answer is at least $3/14b$ more than that of player 1 at round b', then clearly player 2's probability of guessing the right answer at round b' is at least $3/14b$ more than player 1's probability of guessing the right answer at $b' - 1$; otherwise, player 1's probability of guessing the right answer at round $b' + 1$ is at least $3/14b$ more than player 2's probability of guessing it at round b'.)

Suppose, without loss of generality, that b' is such that player 1's probability of guessing the right answer at round $b' + 1$ is at least $3/14b$ more than player 2's probability of guessing the right answer at round b'. Then player 1 deviates from σ_{CT} by not sending any messages to player 2 at round b', and then making a decision based on his information at round $b' + 1$, using his best guess based on his information. (Note that player 2 will still send player 1 a message at round b' according to σ_{CT}.) The best player 2 can do is to use his round b' information. If α_1, α_2, α_3, and α_4 are the probabilities of player 1 getting the right answer and player 2 not, both getting the right answer, 2 getting the right answer and 1 not, and neither getting the right answer, we must have $\alpha_1 - \alpha_3 \geq 3/14b$. Moreover, since σ_{CT} is an ϵ-implementation of σ, by assumption, the expected utility of player 1 if he deviates is at least $(1 - \epsilon)(1 - (\alpha_1 - \alpha_3)) + M(\alpha_1 - \alpha_3)$. It easily follows that, since $M > (1 - \epsilon) + 28b\epsilon/3$, then 1's expected utility by deviating at round b' is greater than $1 + \epsilon$. Hence, σ is not an ϵ-equilibrium in $\Gamma_{\mathrm{CT}}(u^{M,\epsilon})$.

For part (b), consider the utility function $u^{2,0}$. Note that there is no 1-punishment strategy $\Gamma_{\mathrm{CT}}(u^{2,0})$ with respect to σ. We show that, for all b, there is no cheap-talk strategy σ_{CT} that ϵ-implements σ and has expected running time b in $\Gamma_{\mathrm{CT}}(u^{2,0})$ if $\epsilon < 3/14b$. Suppose that σ_{CT} is a cheap-talk that ϵ-implements σ and has expected running time b. The argument above shows that there must be a round b' where one of player 1 or player 2 can deviate and have expected utility at least $(1 - (\alpha_1 - \alpha_3)) + 2(\alpha_1 - \alpha_3) = 1 + (\alpha_1 - \alpha_3) > 1 + 3/14b$. The result immediately follows.

For the general argument, we do the proof of part (a) here; the modifications needed to deal with part (b) are straightforward. Consider a $k + t + 1$ out of n secret sharing game, where the initial shares are "signed" using check vectors. Specifically, for each share $f(i)$ and for each player $j \in N \setminus i$, player i is given a uniformly random value y_{ij} in F and player j is given a uniformly random value b_{ij} in $F \setminus \{0\}$ and a value c_{ij} such that $f(i) + y_{ij}b_{ij} = c_{ij}$. In the underlying game, players can either choose a value in the field (intuitively, their best guess as to the secret) or play DETECT. The utility functions are defined as follows:

- if at least $n - t$ players play DETECT, then all players playing DETECT get 1, and all others get 0;
- if fewer than $n - t$ players play DETECT and at least $n - (k + t)$ but fewer than $n - t$ players play the secret, then the players playing the secret get M and the other players get $-M + 2 - 2\delta$
- if fewer than $n - t$ players play DETECT and either $n - t$ or more players or fewer than $n - (k + t)$ players play the secret, then the players playing the secret get 1, and the remaining players get $1 - \delta$.

In the mediator game, each player is supposed to send the mediator his type (share, signatures, and verifications). The mediator checks that all the shares sent pass the checks. Note that each share can be subjected to n checks, one for each player. A share is *reliable* if it passes at least $n - t$ checks. If there are not at least $n - t$ reliable shares, but a unique polynomial f can be interpolated through the shares that are sent, then the mediator sends a random value that is not $f(0)$ to all the players; otherwise, the mediator chooses a value in F at random and sends it to all the players. If there are at least $n - t$ reliable shares, the mediator checks if a unique polynomial of degree $k + t$ can be interpolated through the shares. If so, the mediator sends the secret to all the players; if not, the mediator sends DETECT to all the players. Let σ be the strategy for the players where they truthfully tell the mediator their type, and play what the mediator sends.

We claim that playing 1 is a $(k+t)$-punishment strategy with respect to σ if $\delta > 0$ and $n > k + 2t$, and that σ is a (k,t)-robust equilibrium. The argument that playing 1 is a $(k+t)$-punishment strategy is essentially identical to the argument for the 2-player game. However, note that if $n \leq k + 2t$, then $t \geq n - (k+t)$. If at most t players play 1, this is not a punishment strategy since the remaining $n - t$ players can play DETECT and guarantee themselves a payoff of 1.

To show that σ is a (k,t)-robust equilibrium, we first show it is t-immune. Suppose that a subset T of at most t players attempt to fool the mediator by guessing shares and appropriate check values, and the remaining players play σ. Then at least $n - t$ shares will be reliable. Note that $n - t \geq k + t + 1$. Either the mediator can interpolate a unique polynomial f of degree $k + t$ through the reliable shares or not. In the former case, the good players will learn the secret; in the latter case, the good players will play DETECT. In either case, their payoff is 1. Thus, σ is t-immune.

For robustness, suppose that a set T up to $k + t$ players deviate from the recommended strategy. If the true polynomial is f, the players in T can convince the mediator that some polynomial $f' \neq f$ is the true polynomial, and the players in T know $(k+t)$ of the points on f', then, once the players in T learn $f'(0)$, they will know $k + t + 1$ points on f', and hence will be able to compute f'. They can then compute the shares of all the other players, and thus compute $f(0)$. This can happen only if $|T| = k+t$, $2(k+t) - n$ of the players in T send their true shares (and the correct check vectors), and the remaining players in T send incorrect shares. So we assume that $|T| = k + t$, and $n - (k+t)$ players in T send incorrect values. If the mediator cannot interpolate a unique polynomial through the values sent, then the mediator chooses a value in F at random and sends it to all the players. Even if the players in T know that this is what happened, the players not in T are playing a punishment strategy, so a player in T cannot get expected utility higher than $1 - 6\delta/7$, even if they know that the mediator is sending a random value. If the mediator can interpolate a unique polynomial f through the shares sent, then the mediator will send $f(0)$ if all of the $n - (k+t)$ shares received are reliable, and a value different from $f(0)$ otherwise. In the former case, which occurs with probability $1/7^{n-(k+t)}$, the players in T can compute the true secret, and will get a payoff of M. In the latter case, they compute the wrong value. With probability $(1/6)(1 - 1/7^{n-(k+t)})$, the other players get the right value and the players in T get a payoff of $-M + 2 - 2\delta$; otherwise, they get a payoff of 1. It is easy to see that the expected utility of the players in T is at most $1 - 2\delta/7$, so the rational players will not deviate.

Suppose that this mediator strategy can be implemented using cheap talk. We claim that, as in the proof of Theorem 7, we can use the implementation to give a cheap-talk implementation in the 2-player game. We simply divide the players into three groups: groups A and B both have $n - (k+t)$ members; group K has the remaining $2(k+t) - n$ players (group K may be empty). Notice that $|A \cup B| = |A \cup K| = k + t$. Just as in the 2-player case, we can show that there must be a round b' such that if the players in group A pool their information together, the probability of them guessing the right answer at round $b' + 1$ is at least $3/14b$ more than the probability of the players in group B guessing the answer after round b' even if the players in group B pool their knowledge together, or the same situation holds with the roles of A and B reversed. Assume that A is the group that has the higher probability of guessing the right answer at time b'. Then we assume that the players in $A \cup K$ deviate by not sending a round b' message to the players in B. (Here we use the fact that $|A \cup B| \leq k + t$.) Now the argument continues as in the 2-player case. $\square$

References

[ADGH06] I. Abraham, D. Dolev, R. Gonen, and J. Y. Halpern. Distributed computing meets game theory: Robust mechanisms for rational secret sharing and multiparty computation. In *Proc. 25th ACM Symp. Principles of Distributed Computing*, pages 53–62, 2006.

[ADGH07] I. Abraham, D. Dolev, R. Gonen, and J. Y. Halpern. Distributed computing meets game theory: Robust mechanisms for rational secret sharing and multiparty computation. unpublished manuscript, 2007.

[ADH] I. Abraham, D. Dolev, and J.Y. Halpern. On implementing mediators with asynchronous cheap talk. Unpublished manuscript.

[AH03] R. J. Aumann and S. Hart. Long cheap talk. *Econometrica*, 71(6):1619–1660, 2003.

[Aum59] R. J. Aumann. Acceptable points in general cooperative n-person games. *Contributions to the Theory of Games, Annals of Mathematical Studies*, IV:287–324, 1959.

[Aum87] R. J. Aumann. Correlated equilibrium as an expression of Bayesian rationality. *Econometrica*, 55:1–18, 1987.

[Bar92] I. Barany. Fair distribution protocols or how the players replace fortune. *Mathematics of Operations Research*, 17:327–340, 1992.

[Ben03] E. Ben-Porath. Cheap talk in games with incomplete information. *J. Economic Theory*, 108(1):45–71, 2003.

[BGW88] M. Ben-Or, S. Goldwasser, and A. Wigderson. Completeness theorems for non-cryptographic fault-tolerant distributed computation. In *Proc. 20th ACM Symp. Theory of Computing*, pages 1–10, 1988.

[BN00] Dan Boneh and Moni Naor. Timed commitments. In *CRYPTO '00: Proceedings of the 20th Annual International Cryptology Conference on Advances in Cryptology*, pages 236–254, London, UK, 2000. Springer-Verlag.

[BPW89] B. D. Bernheim, B. Peleg, and M. Whinston. Coalition proof Nash equilibrium: Concepts. *J. Economic Theory*, 42(1):1–12, 1989.

[CCD88] D. Chaum, Claude Crépeau, and I. Damgard. Multiparty unconditionally secure protocols. In *Proc. 20th ACM Symp. Theory of Computing*, pages 11–19, 1988.

[CS82] V. P. Crawford and J. Sobel. Strategic information transmission. *Econometrica*, 50(6):1431–51, 1982.

[DHR00] Y. Dodis, S. Halevi, and T. Rabin. A cryptographic solution to a game theoretic problem. In *CRYPTO 2000: 20th International Cryptology Conference*, pages 112–130. Springer-Verlag, 2000.

[EGL85] S. Even, O. Goldreich, and A. Lempel. A randomized protocol for signing contracts. *Commun. ACM*, 28(6):637–647, 1985.

[Eli02] K. Eliaz. Fault-tolerant implementation. *Review of Economic Studies*, 69(3):589–610, 2002.

[FHHW03] M. Fitzi, M. Hirt, T. Holenstein, and J. Wullschleger. Two-threshold broadcast and detectable multi-party computation. In *Advances in Cryptology — EUROCRYPT '03*, volume 2656 of *Lecture Notes in Computer Science*, pages 51–67. Springer-Verlag, 2003.

[FLM86] M. J. Fischer, N. A. Lynch, and M. Merritt. Easy impossibility proofs for distributed consensus problems. *Distributed Computing*, 1(1):26–39, 1986.

[FLP85] M. J. Fischer, N. A. Lynch, and M. S. Paterson. Impossibility of distributed consensus with one faulty processor. *Journal of the ACM*, 32(2):374–382, 1985.

[For90] Francoise Forges. Universal mechanisms. *Econometrica*, 58(6):1341–64, 1990.

[GK06] D. Gordon and J. Katz. Rational secret sharing, revisited. In *SCN (Security in Communication Networks) 2006*, pages 229–241, 2006.

[GMW87] O. Goldreich, S. Micali, and A. Wigderson. How to play any mental game. In *Proc. 19th ACM Symp. Theory of Computing*, pages 218–229, 1987.

[Gol04] O. Goldreich. *Foundations of Cryptography, Vol. 2*. Cambridge University Press, 2004.

[Hel05] Y. Heller. A minority-proof cheap-talk protocol. Unpublished manuscript, 2005.

[HT04] J. Y. Halpern and V. Teague. Rational secret sharing and multiparty computation: extended abstract. In *Proc. 36th ACM Symp. Theory of Computing*, pages 623–632, 2004.

[IML05] S. Izmalkov, S. Micali, and M. Lepinski. Rational secure computation and ideal mechanism design. In *Proc. 46th IEEE Symp. Foundations of Computer Science*, pages 585–595, 2005.

[KW82] D. M. Kreps and R. B. Wilson. Sequential equilibria. *Econometrica*, 50:863–894, 1982.

[KY84] A. Karlin and A. C. Yao. Probabilistic lower bounds for byzantine agreement. Unpublished manuscript, 1984.

[Lam83] L. Lamport. The weak byzantine generals problem. *J. ACM*, 30(3):668–676, 1983.

[LMPS04] M. Lepinski, S. Micali, C. Peikert, and A. Shelat. Completely fair SFE and coalition-safe cheap talk. In *Proc. 23rd ACM Symp. Principles of Distributed Computing*, pages 1–10, 2004.

[LMS05] M. Lepinksi, S. Micali, and A. Shelat. Collusion-free protocols. In *Proc. 37th ACM Symp. Theory of Computing*, pages 543–552, 2005.

[LSP82] L. Lamport, R. Shostak, and M. Pease. The Byzantine Generals problem. *ACM Trans. on Programming Languages and Systems*, 4(3):382–401, 1982.

[LT06] A. Lysyanskaya and N. Triandopoulos. Rationality and adveresarial behavior in multi-party comptuation. In *CRYPTO 2006*, pages 180–197, 2006.

[MW96] D. Moreno and J. Wooders. Coalition-proof equilibrium. *Games and Economic Behavior*, 17(1):80–112, 1996.

[Mye97] Roger B. Myerson. *Game Theory: Analysis of Conflict*. Harvard University Press, September 1997.

[PW96] B. Pfitzmann and M. Waidner. Information-theoretic pseudosignatures and byzantine agreement for $t >= n/3$. Technical Report RZ 2882 (#90830), IBM Zurich Research Laboratory, 1996.

[Rab] M. Rabin. How to exchange secrets with oblivious transfer. 1981. http://eprint.iacr.org/2005/187.

[RB89] T. Rabin and M. Ben-Or. Verifiable secret sharing and multiparty protocols with honest majority. In *Proc. 21st ACM Symp. Theory of Computing*, pages 73–85, 1989.

[Sel75] R. Selten. Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory*, 4:25–55, 1975.

[SRA81] A. Shamir, R. L. Rivest, and L. Adelman. Mental poker. In D. A. Klarner, editor, *The Mathematical Gardner*, pages 37–43. Prindle, Weber, and Schmidt, Boston, Mass., 1981.

[UV02] A. Urbano and J. E. Vila. Computational complexity and communication: Coordination in two-player games. *Econometrica*, 70(5):1893–1927, 2002.

[UV04] A. Urbano and J. E. Vila. Computationally restricted unmediated talk under incomplete information. *Economic Theory*, 23(2):283–320, 2004.

[Yao82] A. Yao. Protocols for secure computation (extended abstract). In *Proc. 23rd IEEE Symp. Foundations of Computer Science*, pages 160–164, 1982.