Hidden-Role Games: Equilibrium Concepts and Computation

Luca Carminati*†
Politecnico di Milano
luca.carminati@polimi.it

Brian Hu Zhang*
Carnegie Mellon University
bhzhang@cs.cmu.edu

Gabriele Farina
MIT
gfarina@mit.edu

Nicola Gatti
Politecnico di Milano
nicola.gatti@polimi.it

Tuomas Sandholm
Carnegie Mellon University
Strategic Machine, Inc.
Strategy Robot, Inc.
Optimized Markets, Inc.
sandholm@cs.cmu.edu

Abstract

In this paper, we study the class of games known as *hidden-role games* in which players are assigned *privately* to teams and are faced with the challenge of recognizing and cooperating with teammates. This model includes both popular recreational games such as the *Mafia/Werewolf* family and *The Resistance (Avalon)* and many real-world settings, such as distributed systems where nodes need to work together to accomplish a goal in the face of possible corruptions. There has been little to no formal mathematical grounding of such settings in the literature, and it was previously not even clear what the right solution concepts (notions of equilibria) should be. A suitable notion of equilibrium should take into account the communication channels available to the players (*e.g.*, can they communicate? Can they communicate in private?). Defining such suitable notions turns out to be a nontrivial task with several surprising consequences. In this paper, we provide the first rigorous definition of equilibrium for hidden-role games, which overcomes serious limitations of other solution concepts not designed for hidden-role games. We then show that in certain cases, including the above recreational games, optimal equilibria can be computed efficiently. In most other cases, we show that computing an optimal equilibrium is at least NP-hard or coNP-hard. Lastly, we experimentally validate our approach by computing exact equilibria for complete 5- and 6-player *Avalon* instances whose size in terms of number of information sets is larger than 10^{56}.

*Equal contribution; author order randomized
†Work done while at Carnegie Mellon University

Contents

1 Introduction

Consider a multiagent system with communication where the majority of agents share incentives, but there are also hidden defectors who seek to disrupt their progress.

This paper adopts the lens of game theory to characterize and solve a class of games called *hidden-role games*[1]. Hidden-role games model multi-agent systems in which a team of "good" agents work together to achieve some desired goal, but a subset of adversaries hidden among the agents seeks to sabotage the team. Customarily (and in our paper), the "good" agents make up a majority of the players, but they will not know who the adversaries are. On the other hand, the adversaries know each other.

Hidden-role games offer a framework for developing optimal strategies in systems and applications that face deception. They have a strong emphasis on communication: players need to communicate in order to establish trust, coordinate actions, exchange information, and distinguish teammates from adversaries.

Hidden-role games can be used to model a wide range of recreational and real-world applications. Notable recreational examples include the popular tabletop games *Mafia* (also known as *Werewolf*) and *The Resistance*, of which *Avalon* is the best-known variant. As an example, consider the game *Mafia*. The players are split in an uninformed majority called *villagers* and an informed minority called *mafiosi*. The game proceeds in two alternating phases, *night* and *day*. In the night phase, the mafiosi privately communicate and eliminate one of the villagers. In the day phase, players vote to eliminate a suspect through majority voting. The game ends when one of the teams is completely eliminated.

We now provide several non-recreational examples of hidden-role games. In many cybersecurity applications [Garcia Morchon et al., 2007, Garcia-Morchon et al., 2013, Tripathi et al., 2022], an adversary compromises and controls some nodes of a distributed system whose functioning depends on cooperation and information sharing among the nodes. The system does not know which nodes have been compromised, and yet it must operate in the presence of the compromised nodes.

Another instance of problems that can be modeled as hidden-role games arises in *AI alignment*, *i.e.*, the study of techniques to steer AI systems towards humans' intended goals, preferences, or ethical principles [Ziegler et al., 2019, Ji et al., 2023, Hubinger et al., 2024]. In this setting, there is risk that a misaligned AI agent may attempt to deceive a human user into trusting its suggestions [Park et al., 2023, Scheurer et al., 2023]. AI debate [Irving et al., 2018] aims at steering AI agents by using an adversarial training procedure in which a judge has to decide which is the more trustful between two hidden agents, one of which is a deceptor trained to fool the judge. Miller et al. [2021] proposes an experimental setting consisting of a chess game in which one side is controlled by a player and two advisors, which falls directly under our framework. The advisors pick action suggestions for the player to choose from, but one of the two advisors has the objective of making the team lose.

Hidden-role games also include general scenarios where agents receive inputs from other agents which may be compromised. For example, in *federated learning* (a popular category of distributed learning methods), a central server aggregates machine learning models trained by multiple distributed local agents. If some of these agents are compromised, they may send doctored input with the goal of disrupting the training process [Mothukuri et al., 2021].

Our paper aims to characterize optimal behavior in these settings, and analyze its computability.

Related work. To the best of our knowledge, there have been no previous works on general hidden-role games. On the other hand, there has been a limited amount of prior work on solving specific hidden-role games. Braverman et al. [2006] propose an optimal strategy for *Mafia*, and analyze the win probability when varying the number of players with different roles. Similarly, Christiano [2018] proposes a theoretical analysis for *Avalon*, investigating the possibility of *whispering*, *i.e.* any two players being able to communicate without being discovered. Both of those papers describe game-specific strategies that can be adopted by players to guarantee a specific utility to the teams. In contrast, we provide, to our knowledge, the first rigorous

[1]These games are often commonly called *social deduction games*.

definition of a reasonable solution concept for hidden-role games, an algorithm to find such equilibria, and an experimental evaluation with a wide range of parameterized instances.

Deep reinforcement learning techniques have also been applied to various hidden-role games [Aitchison et al., 2021, Kopparapu et al., 2022, Serrino et al., 2019], but with no theoretical guarantees and usually with no communication allowed between players. A more recent stream of works focused on investigating the deceptive capabilities of large language models (LLMs) by having them play a hidden-role game [Xu et al., 2023, O'Gara, 2023]. The agents, being LLM-based, communicate using plain human language. However, as before, these are not grounded in any theoretical framework, and indeed we will illustrate that optimal strategies in hidden-role games are likely to involve communication that does not bear resemblance to natural language, such as the execution of cryptographic protocols.

1.1 Main Modeling Contributions

We first here give an informal, high-level description of our game model. We also introduce our main solution concept of interest, called *hidden-role equilibrium*, and discuss the challenges it addresses. We will define these concepts in more formality beginning in Section 3.

We define a (finite) *hidden-role game* as an n-player finite extensive-form game Γ in which the players are partitioned at the start of the game into two teams. Members of the same team share the same utility function, and the game is zero-sum, *i.e.* any gain for one team means a loss for the other. We thus identify the teams as MAX and MIN, since teams share the same utility function, but have opposite objectives. At the start of the game, players are partitioned at random into two teams. A crucial assumption is that one of the two teams is *informed*, *i.e.* all the members of that particular team know the team assignment of all the players, while this is not true for all players belonging to the other team. Without loss of generality, we use MAX to refer to the uninformed team, and MIN to refer to the informed one.[2]

To allow our model to cover *communication* among players, we formally define the *communication extensions* of a game Γ. The communication extensions are games like Γ except that actions allowing messages to be sent between players are explicitly encoded in the game. In the *public communication extension*, players are able to publicly broadcast messages. In the *private communication extension*, in addition to the public broadcast channel, the players have pairwise private communication channels.[3] In all cases, communication channels are *synchronous* and *authenticated*: messages sent on one timestep are received at the next timestep, and are tagged with their sender. Communication presents the main challenge of hidden role games: MAX-players wish to share information with teammates, but *not* with MIN-players.

In defining communication extensions, we must bound the *length* of the communication, that is, how many rounds of communication occur in between every move of the game, and how many distinct messages can be sent on each round. To do this, we fix a finite message space[4] of size M and length of communication R, and in our definition of equilibrium we will take a supremum over M and R. This will allow us to consider arbitrarily complex message spaces (*i.e.*, M and R arbitrarily large) while still only analyzing finite games: for any *fixed* M and R, the resulting game is a finite hidden-role game. We will show that our positive results (upper bounds) only require $\log M = R = \mathsf{polylog}(|H|, 1/\varepsilon)$, where $|H|$ is the number of nodes (histories) in the game tree and ε is the desired precision of equilibrium.

We characterize optimal behavior in the hidden-role setting by converting hidden-role games into team games in a way that preserves the strategic aspect of hidden-roles. This team game is called *split-personality form* of a given hidden-role game. Given a (possibly communication-extended) hidden-role game Γ, we define and analyze two possible variants:

- the *uncoordinated split-personality form* $\textsc{USplit}(\Gamma)$ is a team games with $2n$ players, derived by splitting each player i in the original game in two distinct players, i^+ and i^-, that pick actions for i in Γ

[2] For example, in *Mafia*, the villagers are MAX while mafiosi are MIN.

[3] If players are assumed to be computationally bounded, pairwise private channels can be created from the public broadcast channels through public-key cryptography. However, throughout this paper, for the sake of conceptual cleanliness, we will not assume that players are computationally bounded, and therefore we will distinguish the public-communication case from the private-communication case.

[4] Note that, if the message space is of size M, a message can be sent in $O(\log M)$ bits.

if the player is assigned to team MAX or MIN respectively.

- the *coordinated split-personality form* CSplit(Γ) is the $(n+1)$-player team game in which the additional player, who we refer to simply as the *adversary* or MIN-*player*, takes control of the actions of all players who have been assigned to the MIN-team. On the contrary, the players from 1 to n control the players as usual only if they belong to the team MAX.

The coordinated split-personality variant encodes an extra assumption on MIN's capabilities, namely, that the MIN-team is controlled by a single player and is therefore perfectly coordinated. Trivially, when only one player is on team MIN, the uncoordinated and coordinated split-personality forms coincide.

In either case, the resulting game is a team game in which each player has a fixed team assignment. We remark that the split-personality form maintains the strategic aspects of hidden roles, since i^+ and i^- share identity when interacting with the environment. For example, players may observe that i has done an action, but do not know if the controller was i^+ or i^-. Similarly, messages sent by i^+ and i^- are signed by player i, since the communication extension is applied on Γ *before* splitting personalities.

Picking which split-personality variant to use is a modeling assumption that depends on the game instance that one wants to address. For example, in many recreational tabletop games, USplit is the more reasonable choice because MIN-players are truly distinct; however, in a network security game where a single adversary controls the corrupted nodes, CSplit is the more reasonable choice. The choice of which variant also affects the complexity of equilibrium computation: as we will detail in later, CSplit yields a more tractable solution concept. In certain special cases, however, CSplit and USplit will coincide. For example, we will later show that this is the case in *Avalon*, which is key to allowing our algorithms to work in that game.

With these pieces in place, we define the *hidden-role equilibria* (HRE) of a hidden-role game Γ as the *team max-min equilibria* (TMEs) of the split-personality form of Γ. That is, the hidden-role equilibria are the optimal joint strategies for team MAX in the split-personality game, where optimality is judged by the expected value against a jointly-best-responding MIN-team. The *value* of a hidden-role game is the expected value for MAX in any hidden-role equilibrium. If communication (private or public) is allowed, we define hidden-role equilibria and values by taking the supremum over M and R of the expected value at the equilibrium, that is, the MAX-team is allowed to set the parameters of the communication.

Our new solution concept encodes, by design, a pessimistic assumption for the MAX-team. MAX picks M, R and its strategy considering a worst-case MIN adversary that knows this strategy and best-responds to it. Throughout our proofs, we will heavily make use of this fact. In particular, we will often consider MIN-players that "pretend to be MAX-players" under certain circumstances, which is only possible if MIN-players know MAX-players' strategies. It is *not* allowed for MAX-players to know MIN-players' strategies in the same fashion. This is in stark contrast to usual zero-sum game analysis, where various versions of the *minimax theorem* promise that the game is unchanged no matter which side commits first to a strategy. Indeed, we discuss in Section 6.2 the fact that, for hidden-role games, the asymmetry is in some sense necessary: a minimax theorem *cannot* hold for nontrivial hidden-role games. We argue, however, that this asymmetry is natural and *inherent* in the the hidden-role setting. If we assumed the contrary and inverted the order of the teams so that MIN commits first to its strategy, MAX could discover the roles immediately by agreeing to message a passphrase unknown to MIN in the first round, thus spoiling the whole purpose of hidden-role games. This argument will be made formal in Section 6.2.

Existing solution concepts failures We defined our equilibrium notion as a team max-min equilibrium (TME) of the split-personality form of a communication-extended hidden role game. Here, we will argue why some other notions would be less reasonable.

- *Nash Equilibrium.* A *Nash equilibrium* [Nash, 1950] is a strategy profile for all players from which no player can improve its own utility by deviating. This notion is unsuitable for our purposes because it fails to capture *team coordination*. In particular, in pure coordination games (in which all players have the same utility function), which are a special case of hidden-role games (with no hidden roles and no adversary team at all), a Nash equilibrium would only be locally optimal in the sense that no player

can improve the team's value alone. In contrast, our notion will lead to the optimal team strategy in such games.

- *Team-correlated equilibrium.* The *team max-min equilibrium with correlation* [von Stengel and Koller, 1997, Celli and Gatti, 2018] (TMECor), is a common solution concept used in team games. It arises from allowing each team the ability to communicate in private (in particular, to generate correlated randomness) *before* the game begins. For team games, TMECor is arguably a more natural notion than TME, as the corresponding optimization problem is a bilinear saddle-point problem, and therefore in particular the minimax theorem applies, avoiding the issue of which team ought to commit first. However, for hidden-role games, TMECor is undesirable, because it does not make sense for a team to be able to correlate with teammates that have not even been assigned yet. The *team max-min equilibrium with communication* (TMECom) [Celli and Gatti, 2018] makes an even stronger assumption about communication among team members, and therefore suffers the same problem.

1.2 Main Computational Contributions

We now introduce computational results, both positive and negative, for computing the hidden-role value and hidden-role equilibria of a given game.

Polynomial-time algorithm Our main positive result is summarized in the following informaion theorem statement.

Theorem 1.1 (Main result, informal; formal result in Theorem 4.3)**.** *If the number of players is constant, private communication is available, the* MIN*-team is a strict minority (*i.e.*, strictly less than half of the players are on the* MIN*-team), and the adversary is coordinated, there is a polynomial-time algorithm for exactly computing the hidden-role value.*

This result should be surprising, for multiple reasons. First, team games are generally hard to solve [von Stengel and Koller, 1997, Zhang et al., 2022], so any positive result for computing equilibria in team games is fairly surprising. Further, it is *a priori* not obvious that the value of any hidden-role game with private communication and coordinated adversary is even a *rational* number[5], much less computable in polynomial time: for example, there exist adversarial team games with no communication whose TME values are irrational [von Stengel and Koller, 1997].

There are two key ingredients to the proof of Theorem 1.1. The first is a special type of game which we call a *mediated* game. In a mediated game, there is a player, the *mediator*, who is always on team MAX. MAX-players can therefore communicate with it and treat it as a trusted party. We show that, when a mediator is present (and all the other assumptions of Theorem 1.1 also hold), the hidden-role value is computable in polynomial time. To do this, we state and prove a form of *revelation principle*. Informally, our revelation principle states that, without loss of generality, it suffices to consider MAX-team strategies in which, at every timestep of the game,

1. all MAX-players send their honest information to the mediator,
2. the mediator sends action recommendations to all players (regardless of their team allegiance; remember that the mediator may not know the team assignment), and
3. all MAX-players play their recommended actions.

MIN-team players are, of course, free to pretend to be MAX-team players and thus send false information to the mediator; the mediator must deal with this possibility. However, MIN-team players cannot just send *any* message; they must send messages that *are consistent with some* MAX*-player*, lest they be immediately revealed as MIN-team. These observations are sufficient to construct a *two-player zero-sum game* Γ_0, where the mediator is the MAX-player and the coordinated adversary is the MIN-player. The value of Γ_0 is the value of the original hidden-role game, and the size of Γ_0 is at most polynomially larger than the size of the original game. Since two-player zero-sum extensive-form games can be solved in polynomial time [Koller et al., 1994, von Stengel, 1996], it follows that mediated hidden-role games can also be solved in polynomial time.

[5] assuming all game values and chance probabilities are also rational numbers

The second ingredient is to invoke results from the literature on *secure multi-party computation* to *simulate* a mediator in the case that one is not already present. A well-known result from that literature states that so long as strictly more than half of the players are honest, essentially any interactive protocol—such as the ones used by our mediator to interact with other players—can be simulated efficiently such that the adversary can cause failure of the protocol or leakage of information [Beaver, 1990, Rabin and Ben-Or, 1989].[6] Chaining such a protocol with the argument in the previous paragraph concludes the proof of the main theorem.

Related works on MPC and communication equilibria. The *communication equilibrium* [Forges, 1986, Myerson, 1986] is a notion of equilibrium with a mediator, in which the mediator has two-way communication with all players, and players need to be incentivized to report information honestly and follow recommendations. Communication equilibria include all Nash equilibria, and therefore are unfit for general hidden-role games for the same reason as Nash equilibria, as discussed in the previous subsection.

However, when team MIN has only one player and private communication is allowed, the hidden-role equilibria coincide with the MAX-team-optimal communication equilibria in the original game Γ. Our main result covers this case, but an alternative way of computing a hidden-role equilibrium in this special case is to apply the optimal communication equilibrium algorithms of Zhang and Sandholm [2022] or Zhang et al. [2023]. However, those algorithms either involve solving linear programs, solving many zero-sum games, or solving zero-sum games with large reward ranges, which will be less efficient than directly solving a single zero-sum game Γ_0.

We are not the first to observe that multi-party computation can be used to implement a mediator for use in game theory. In various settings and for various solution concepts, it is known to be possible to implement a mediator using only cheap-talk communication among players [*e.g.*, Urbano and Vila, 2002, Liu, 2017, Abraham et al., 2006, Izmalkov et al., 2005]. For additional reading on the connections between game theory and cryptography, we refer the reader to the survey of Katz [2008], and papers citing and cited by this survey.

Lower bounds. We also show *lower bounds* on the complexity of computing the hidden-role value, even for a constant number of players, when any of the assumptions in Theorem 1.1 is broken.

Theorem 1.2 (Lower bounds, informal; formal statement in Theorems 4.5 and 4.6)**.** *If private communication is disallowed, the hidden-role value problem is* NP*-hard. If the* MIN*-team is uncoordinated, the problem is* coNP*-hard. If both, the problem is* Σ_2^{P}*-hard. All hardness reductions hold even when the* MIN*-team is a minority and the number of players is an absolute constant.*

Price of hidden roles. Finally, we define and compute the *price of hidden roles*. It is defined (analogously to the price of anarchy and price of stability, which are common quantities of study in game theory) as the ratio between the value of a hidden-role game, and the value of the same game with team assignments made public. We show the following:

Theorem 1.3 (Price of hidden roles; formal statement in Theorem 6.3)**.** *Let D be a distribution of team assignments. For the class of games where teams are drawn according to distribution D, the price of hidden roles is equal to $1/p$, where p is the probability of the most-likely team in D.*

Intuitively, in the worst case, the MAX-team can be forced to guess at the beginning of the game all the members of the MAX-team, and win if and only if its guess is correct. In particular, for the class of n-player games with k adversaries, the price of hidden roles is exactly $\binom{n}{k}$.

[6] In this part of the argument, the details about the communication channels become important: in particular, the MPC results that we use for our main theorem statement assume that the network is synchronous (*i.e.*, messages sent in round r arrive in round $r+1$), and that there are pairwise private channels and a public broadcast channel that are all authenticated (*i.e.*, message receivers know who sent the message). This is enough to implement MPC so long as $k < n/2$, where k is the number of adversaries and n is the number of players. Our results, however, do not depend on the specific assumptions about the communication channel, so long as said assumptions enable secure MPC with guaranteed outcome delivery. For a recent survey of MPC, see Lindell [2020]. For example, if $k < n/3$ then MPC does not require a public broadcast channel, so neither do our results. For cleanliness, and to avoid introducing extra formalism, we will stick to one model of communication.

1.3 Experiments: *Avalon*

We ran experiments on the popular tabletop game *The Resistance: Avalon* (or simply *Avalon*, for short). As discussed earlier, despite the adversary team in *Avalon* not being coordinated in the sense used in the rest of the paper, we show that, at least for the 5- and 6-player variants, the adversary would not benefit from being coordinated; therefore, our polynomial-time algorithms can be used to solve the game. This observation ensures that our main result applies. Game-specific simplifications allow us to reduce the game tree from roughly 10^{56} nodes [Serrino et al., 2019] to 10^8 or even fewer, enabling us to compute exact equilibria. Our experimental evaluation demonstrates the practical efficacy of our techniques on real game instances. Our results are discussed in Section 7, and further detail on the game-specific reductions used, as well as a complete hand-analysis of a small *Avalon* variant, can be found in Appendix E.

1.4 Examples

In this section we present three examples that will hopefully help the reader in understanding our notion of equilibrium and justify some choices we have made in our definition.

A hidden-role matching pennies game. Consider a n-player version of matching pennies (with $n > 2$), which we denote as $\mathsf{MP}(n)$. One player is chosen at random to be the adversary (team MIN). All n players then simultaneously choose bits $b_i \in \{0, 1\}$. Team MAX wins (gets utility 1) if and only if all n bits match; else, team MAX loses (gets utility 0).

With no communication, the value of this game is $1/2^{n-1}$: it is an equilibrium for everyone to play uniformly at random. Public communication does not help, because, conditioned on the public transcript, bits chosen by players must be mutually independent. Thus, the adversary can do the following: pretend to be on team MAX, wait for all communication to finish, and then play 0 if the string of all ones is more conditionally likely than the string of all 1s, and vice-versa.

With private communication, however, the value becomes $1/(n + 1)$. Intuitively, the MAX-team should attempt to guess who the MIN-player is, and then privately discuss among the remaining $n - 1$ players what bit to play. We defer formal proofs of the above game values to Section 5, because they rely on results in Section 4.1.

Simultaneous actions. In typical formulations of extensive-form imperfect-information games, it is without loss of generality to assume that games are *turn-based*, *i.e.*, only one player acts at any given time. To simulate simultaneous actions with sequential ones, one can simply forbid players from observing each others' actions. However, when communication is allowed arbitrarily throughout the game, the distinction between simultaneous and sequential actions suddenly becomes relevant, because *players can communicate when one—but not all—the players have decided on an action.*

To illustrate this, consider the game $\mathsf{MP}(n)$ defined in the previous section, with public communication, except that the players act in sequence in order of index $(1, 2, \ldots, n)$. We claim that the value of this game is not $1/2^{n-1}$, but at least $1/2n$. To see this, consider the following strategy for team MAX. The MAX players wait for P1 to (privately) pick an action. Then, P2 publicly declares a bit $b \in \{0, 1\}$, and all remaining players play b if they are on team MAX. If P1 was the MIN player, this strategy wins with probability at least $1/2$, so the expected value is at least $1/2n$. This example illustrates the importance of allowing simultaneous actions in our game formulations.

Correlated randomness matters. We use our third and final example to discuss a nontrivial consequence of the definition of hidden-role equilibrium that may appear strange at first: it is possible for seemingly-useless information to affect strategic decisions and the game value.

To illustrate, consider the following simple game Γ: there are three players, and three role cards. Two of the three cards are marked MAX, and the third is marked MIN. The cards are dealt privately and randomly to the players. Then, after some communication, all three players simultaneously cast votes to elect a winner. If no player gains a majority of votes, MIN wins. Otherwise, the elected winner's team wins. Clearly, MAX

can win no more than 2/3 of the time in this game: MIN can simply pretend to be on team MAX, and in that case MAX cannot gain information, and the best they can do is elect a random winner.

Now consider the following seemingly-meaningless modification to the game. We will modify the two MAX cards so that they are distinguishable. For example, one card has MAX written on it, and the other has MAX′. We argue that this, perhaps surprisingly, affects the value of the game. In fact, the MAX team can now win deterministically, even with only public communication. Indeed, consider following strategy. The two players on MAX publicly declare what is written on their cards (*i.e.*, MAX or MAX′). The player elected now depends on what the third player did. If one player does not declare MAX or MAX′, elect either of the other two players. If two players declared MAX, elect the player who declared MAX′. If two players declare MAX′, elect the player who declared MAX. This strategy guarantees a win: no matter what the MIN-player does, any player who makes a unique declaration is guaranteed to be on the MAX-team.

What happens in the above example is that making the cards distinguishable introduces a piece of *correlated randomness* that MAX can use: the two MAX players receive cards whose labels are (perfectly negatively) correlated with each other. Since our definition otherwise prohibits the use of such correlated randomness (because players cannot communicate only with players on a specific team), introducing some into the game can have unintuitive effects. In Section 6.2, we expand on the effects of allowing correlated randomness: in particular, we argue that allowing correlated randomness essentially ruins the point of hidden-role games by allowing the MAX team to learn the entire team assignment.

2 Preliminaries

Our contributions are based on prior work on extensive-form adversarial team games. In this section, we introduce the main definitions related to this class of games.

Definition 2.1. A (perfect-recall, timeable) extensive-form game[7] consists of the following.

1. A set of n players, identified with the integers $[n] = \{1, \dots, n\}$, and an extra player C denoting the *chance* player modeling stochasticity.
2. A directed tree of *nodes* H (also called *histories*). The root of the tree is indicated with $\varnothing$ while the set of leaves, or *terminal nodes*, in H is denoted Z.
3. At each node $h \in H \setminus Z$, for each player $i \in [n] \cup \mathrm{C}$, an *action set* $A_i(h)$. The edges leaving h are identified with the joint actions $a \in \bigtimes_{i \in [n] \cup \mathrm{C}} A_i(h)$.
4. For each player $i \in [n]$, an *observation function* $o_i : H \setminus Z \to O$, where O is some set of observations. We assume that the observation uniquely determines the set of legal actions, that is, if $o_i(h) = o_i(h')$ then $A(h) = A(h')$.
5. For each player $i \in [n]$, a *utility function* $u_i : Z \to \mathbb{R}$.
6. For each node $h \in H \setminus Z$, a probability distribution $p_h \in \Delta(A_{\mathrm{C}}(h))$ denoting how the chance player picks its action.

The *information state* of player i at node $h \in H$ is the tuple $s_i(h) := (o_i^0 = o_i(\varnothing), a_i^0, o_i^1, a_i^1, \dots, o_i(h))$ of observations made by player i and actions played by player i on the path to node h. We will use Σ_i to denote the set of all information states of player i. A *pure strategy* x_i of player i is a mapping from information states to actions. That is, it is a map $x_i : \Sigma_i \to \cup_{h \in H} A_i(h)$ such that $x_i(s_i(h)) \in A_i(h)$ for all h. A *mixed strategy* is a probability distribution over pure strategies. A tuple of (possibly mixed) strategies $x = (x_1, \dots, x_n)$ is a *profile* or *strategy profile*, and the expected utility of profile x for player i is denoted $u_i(x)$. We will use X_i to denote the set of pure strategies of player i.

[7] Our definition uses slightly different notation than most works on extensive-form games: for example, we use observations in place of information sets, and we allow for simultaneous moves. In our case, both of these choices will be useful later on in the paper, as we will explicitly make reference to observations and simultaneous actions. We discussed in Section 1.4 why it is important in our model to allow simultaneous actions.

An extensive-form game is an *adversarial team game* (ATG) if there is a team assignment $t \in \{\textsf{MAX}, \textsf{MIN}\}^n$ and a team utility function $u : Z \to \mathbb{R}$ such that $u_i(z) = u(z)$ if $t_i = \textsf{MAX}$, and $u_i(z) = -u(z)$ if $t_i = \textsf{MIN}$. That is, each player is assigned to a team, all members of the team get the same utility, and the two teams are playing an adversarial zero-sum game[8]. In this setting, we will write $x_i \in X_i$ and $y_j \in Y_j$ for a generic strategy of a player on team MAX and MIN respectively. ATGs are fairly well studied. In particular, Team Maxmin Equilibria (TMEs) [von Stengel and Koller, 1997, Celli and Gatti, 2018] and their variants are the common notion of equilibrium employed. The *value* of a given strategy profile x for team MAX is the value that x achieves against a best-responding opponent team. The *TME value* is the value of the best strategy profile of team MAX. That is, the TME value is defined as

$$\textsf{TMEVal}(\Gamma) := \max_{x \in \times_i \Delta(X_i)} \min_{y \in \times_j \Delta(Y_j)} u(x, y), \tag{1}$$

and the *TMEs* are the strategy profiles x that achieve the maximum value. Notice that the TME problem is nonconvex, since the objective function u is nonlinear as a function of x and y. As such, the minimax theorem does not apply, and swapping the teams may not preserve the solution. Computing an (approximate) TME is Σ_2^{P}-complete in extensive-form games [Zhang et al., 2022].

3 Equilibrium Concepts for Hidden-Role Games

While the notion of TME is well-suited for ATGs, it is not immediately clear how to generalize it to the setting of hidden-role games. We do so by formally defining the concepts of *hidden-role game*, *communication* and *split-personality form* first introduced in Section 1.1.

Definition 3.1. An extensive-form game is a *zero-sum hidden-role team game*, or *hidden-role game* for short, if it satisfies the following additional properties:

1. At the root node, only chance has a nontrivial action set. Chance chooses a string $t \sim \mathcal{D} \in \Delta(\{\textsf{MAX}, \textsf{MIN}\}^n)$, where t_i denotes the team to which player i has been assigned. Each player i privately observes (at least[9]) their team assignment t_i. In addition, MIN-players privately observe the entire team assignment t.
2. The utility of a player i is defined completely by its team: there is a $u : Z \to \mathbb{R}$ for which $u_i(z) = u(z)$ if player i is on team MAX at node z, and $-u(z)$ otherwise.[10]

In some games, players observe additional information beyond just their team assignments. For example, in *Avalon*, one MAX-player is designated *Merlin*, and Merlin has additional information compared to other MAX-players. In such cases, we will distinguish between the *team assignment* and *role* of a player: the team assignment is just the team that the player is on (MAX or MIN), while the role encodes the extra private information of the player as well, which may affect what actions that player is allowed to legally take. For example, the team assignment of the player with role *Merlin* is MAX. We remark that additional imperfect information of the game may be observed after the root node.[11]

Throughout this paper, we will use k to denote the largest number of players on the MIN-team, that is, $k = \max_{t \in \text{supp}(\mathcal{D})} |\{i : t_i = \textsf{MIN}\}|$.

[8] This is a slight abuse of language: if the MAX and MIN teams have different sizes, the sum of all players' utilities is not zero. However, such a game can be made zero-sum by properly scaling each player's utility. The fact that such a rescaling operation does not affect optimal strategies is a basic result for von Neumann–Morgenstern utilities [Maschler et al., 2020, Chapter 2.4]. We will therefore generally ignore this detail.

[9] It is allowable for MAX-players to also have more observability of the team assignment, *e.g.*, certain MAX-players may know who some MIN-players are.

[10] While at a first look this condition is similar to the one in ATGs, we remark that in this case the number of players in a team depends on the roles assigned at the start. The same considerations as Footnote 8 on the zero-sum rescaling of the utilities hold.

[11] This is an important difference with respect to Bayesian games [Harsanyi, 1967–68], which assume all imperfect information to be the initial *types* of the players. Conversely, we have an imperfect information structure that evolves throughout the game, while only the teams are assigned and observed at the start.

3.1 Models of Communication

The bulk of this paper concerns notions of equilibrium that allow communication between the players. We distinguish in this paper between *public* and *private communication*:

1. *Public communication*: There is an open broadcast channel on which all players can send messages.
2. *Private communication*: In addition to the open broadcast channel, each pair of players has access to a private communication channel. The private communication channel reveals to all players when messages are sent, but only reveals the message contents to the intended recipients.

Assuming that public-key cryptography is possible (*e.g.*, assuming the discrete logarithm problem is hard) and players are polynomially computationally bounded, public communication and private communication are equivalent, because players can set up pairwise private channels via public-key exchange. However, in this paper, we assume that agents are computationally unbounded and thus treat the public and private communication cases as different. Our motivation for making this distinction is twofold. First, it is conceptually cleaner to explicitly model private communication, because then our equilibrium notion definitions do not need to reference computational complexity. Second, perhaps counterintuitively, equilibria with public communication only may be *more* realistic to execute in practice in human play, precisely *because* public-key cryptography breaks. That is, the computationally unbounded adversary renders more "complex" strategies of the MAX-team (involving key exchanges) useless, thus perhaps resulting in a *simpler* strategy. We emphasize that, in all of our positive results in the paper, the MAX-team's strategy *is* efficiently computable.

To formalize these notions of communication, we now introduce the *communication extension*.

Definition 3.2. The *public* and *private* (M, R)*-communication extensions* corresponding to a hidden-role game Γ are defined as follows. Informally, between every step of the original game Γ, there will be R rounds of communication; in each round, players can send a public broadcast message and private messages to each player. The communication extension starts in state $h = \varnothing \in H_\Gamma$. At each game step of Γ:

1. Each player $i \in [n]$ observes $o_i(h)$.
2. For each of R successive communication rounds:
   (a) Each player i simultaneously chooses a message $m_i \in [M]$ to broadcast publicly.
   (b) If private communication is allowed, each player i also chooses messages $m_{i\to j} \in [M] \cup \{\perp\}$ to send to each player $j \neq i$. $\perp$ denotes that the player does not send a private message at that time.
   (c) Each player j observes the messages $m_{i\to j}$ that were sent to it, as well as all messages m_i that were sent publicly. That is, by notion of communication, the players observe:
      - *Public*: player j observes the ordered tuple $(m_1, \ldots, m_n)$.
      - *Private*: player j also observes the ordered tuple $(m_{1\to j}, \ldots, m_{n\to j})$, and the set $\{(i,k) : m_{i\to k} \neq \perp\}$. That is, players observe messages sent to them, and players see when other players send private messages to each other (but not the contents of those messages)
3. Each player, including chance, simultaneously plays an action $a_i \in A_i(h)$. (Chance plays according to its fixed strategy.) The game state h advances accordingly.

We denote the (M, R)-extensions as $\textsc{Comm}^{M,R}_{\mathsf{priv}}(\Gamma)$, and $\textsc{Comm}^{M,R}_{\mathsf{pub}}(\Gamma)$. To unify notation, we also define $\textsc{Comm}^{M,R}_{\mathsf{none}}(\Gamma) = \Gamma$. When the type of communication allowed and number of rounds are not relevant, we use $\textsc{Comm}(\Gamma)$ to refer to a generic extension.

3.2 Split Personalities

We introduce two different *split-personality* forms $\textsc{USplit}(\Gamma)$ and $\textsc{CSplit}(\Gamma)$ of a hidden-role game Γ, The split-personality forms are adversarial team games which preserve the characteristics of Γ.

Definition 3.3. The *uncoordinated split-personality form*[12] of an n-player hidden-role game Γ is the $2n$-player adversarial team game $\textsc{USplit}(\Gamma)$ in which each player i is split into two players, i^+ and i^-, which control player i's actions when i is on team MAX and team MIN respectively.

Unlike the original hidden-role game Γ, the split-personality game is an adversarial team game without hidden roles: players i^+ are on the MAX team, and i^- are on the MIN-team. Therefore, we are able to apply notions of equilibrium for ATGs to $\textsc{USplit}(\Gamma)$. We also define the *coordinated split-personality form*:

Definition 3.4. The *coordinated split-personality form* of an n-player hidden-role game Γ is the $(n+1)$-player adversarial team game $\textsc{CSplit}(\Gamma)$ formed by starting with $\textsc{USplit}(\Gamma)$ and merging all MIN-players into a single adversary player, who observes all their observations and chooses all their actions.

Assuming MIN to be *coordinated* is a worst-case assumption for team MAX, which however can be justified. In many common hidden-role games, such as the *Mafia* or *Werewolf* family of games and most variants of *Avalon*, such an assumption is not problematic, because the MIN-team has essentially perfect information already. In Appendix E.1, we justify why this assumption is safe also in some more complex *Avalon* instances considered. The coordinated split-personality form will be substantially easier to analyze, and in light of the above equivalence for games like *Avalon*, we believe that it is important to study it.

When team MIN in Γ is already coordinated, that is, if every MIN-team member has the same observation at every timestep, the coordinated and uncoordinated split-personality games will, for all our purposes, coincide: in this case, any strategy of the adversary in $\textsc{CSplit}(\Gamma)$ can be matched by a joint strategy of the MIN-team members in $\textsc{USplit}(\Gamma)$. This is true in particular if there is only one MIN-team member. But, we insert here a warning: even when the base game Γ has a coordinated adversary team, the private communication extension $\textsc{Comm}_{\mathsf{priv}}(\Gamma)$ will not. Thus, with private-communication extensions of Γ, we must distinguish the coordinated and uncoordinated split-personality games even if Γ itself is coordinated.

3.3 Equilibrium Notions

We now define the notions of equilibrium that we will primarily study in this paper.

Definition 3.5. The *uncoordinated value* of a hidden-role game Γ with notion of communication c is defined as

$$\mathsf{UVal}_c(\Gamma) := \sup_{M,R} \mathsf{UVal}_c^{M,R}(\Gamma)$$

where $\mathsf{UVal}_c^{M,R}(\Gamma)$ is the TME value of $\textsc{USplit}(\textsc{Comm}_c^{M,R}(\Gamma))$. The *coordinated value* $\mathsf{CVal}_c(\Gamma)$ is defined analogously by using CSplit.

Definition 3.6. An *ε-uncoordinated hidden-role equilibrium* of Γ with a particular notion of communication $c \in \{\mathsf{none}, \mathsf{pub}, \mathsf{priv}\}$ is a tuple (M, R, x) where x is a MAX-strategy profile in $\textsc{USplit}(\textsc{Comm}_c^{M,R}(\Gamma))$ of value at least $\mathsf{UVal}_c(\Gamma) - \varepsilon$. The *$\varepsilon$-coordinated hidden-role equilibria* is defined analogously, again with CSplit and CVal instead of USplit and UVal.

As discussed in Section 1.1, our notion of equilibrium is inherently asymmetric due to its max-min definition. The MAX-team is the first to commit to a strategy and a communication scheme, and MIN is allowed to know both how much communication will be used (*i.e.*, M and R) as well as MAX's entire strategy x. As mentioned before, this asymmetry is fundamental in our setting, and we will formalize it in Section 6.2.

4 Computing Hidden-Role Equilibria

In this section, we show the main computational results regarding the complexity of computing an hidden-role equilibrium in different settings. We first provide positive results for the private-communication case in Section 4.1 while the negative computational results for the no/public-communications cases are presented in Section 4.2. The resultsfhhsrvgjgvbvgg are summarized in Table 1.

[12] In the language of Bayesian games, the split-personality form would almost correspond to the *agent form*.

4.1 Computing Private-Communication Equilibria

In this section, we show that it is possible under some assumptions to compute equilibria efficiently for hidden-role games. In particular, in this section, we assume that

1. there is private communication,
2. the adversary is coordinated, and
3. the adversary is a minority ($k < n/2$).

Games with a publicly-known MAX-player. First, we consider a special class of hidden-role games which we call *mediated*. In a mediated game, there is a player, who we call the *mediator*, who is always assigned to team MAX. The task of the mediator is to coordinate the actions and information transfer of team MAX. Our main result of this subsection is the following:

Theorem 4.1 (Revelation Principle)**.** *Let Γ^* be a mediated hidden-role game. Then, for $R \geq 2$ and $M \geq |H|$, there exists a coordinated private-communication equilibrium in which the players on MAX have a TME profile in which, at every step, the following events happen in sequence:*

1. *every player on team MAX sends its observation privately to the mediator,*
2. *the mediator sends to every player (MAX and MIN) a recommended action, and*
3. *all players on team MAX play their recommended actions.*

Players on team MIN, of course, can (and will) lie or deviate from recommendations as they wish. The above revelation principle imples the following algorithmic result:

Theorem 4.2. *Let Γ^* be a mediated hidden-role game, $R \geq 2$, and $M \geq |H|$. An (exact) coordinated private-communication hidden-role equilibrium of Γ^* can be computed by solving an extensive-form zero-sum game Γ_0 with at most $|H|^{k+1}$ nodes, where H is the history set of Γ^*.*

Proofs of Theorems 4.1 and 4.2 are deferred to Appendix A.

We give a sketch of how the two-player zero-sum game is structured. Theorem 4.1 allows us to simplify the game by fixing the actions of all players on team MAX, leaving two strategic players, the mediator and the adversary. Any node from the original game is expanded into three levels:

1. the adversary picks messages on behalf of all MIN-players to send to the mediator,
2. the mediator picks recommended actions to send to all players, and
3. the adversary acts on behalf of all MIN-players.

The key to proving Theorem 4.2 is that, in the first step above, the adversary's message space is not too large. Indeed, any message sent by the adversary must be a message that *could have plausibly been sent by a MAX-player*: otherwise the mediator could automatically infer that the sender must be the adversary. It is therefore possible to exclude all other messages from the game since they belong to dominated strategies. Carefully counting the number of such messages would complete the proof.

It is crucial in the above argument that the MIN-team is coordinated; indeed, otherwise, it would not be valid to model the MIN-team as a single adversary in Γ_0. For more elaboration on the case where the MIN-team is not coordinated, we refer the reader to Appendix E.1.

In practice, zero-sum extensive-form games can be solved very efficiently in the tabular setting with linear programming [Koller et al., 1994], or algorithms from the counterfactual regret minimization (CFR) family [Brown and Sandholm, 2019, Farina et al., 2021, Zinkevich et al., 2007]. Thus, Theorem 4.2 gives an efficient algorithm for solving hidden-role games with a mediator.

Simulating mediators with multi-party computation In this section, we show that the previous result essentially generalizes (up to exponentially-small error) to games *without* a mediator, so long as the MIN team is also a minority, that is, $k < n/2$. Informally, the main result of this subsection states that, when private communication is allowed, one can efficiently *simulate* the existence of a mediator using secure multi-party computation (MPC), and therefore team MAX can achieve the same value. The form of secure MPC that we use is *information-theoretically* secure; that is, it is secure even against computationally-unbounded adversaries.

Theorem 4.3 (Main theorem)**.** *Let Γ be a hidden-role game with $k < n/2$. Then $\mathsf{CVal}_{\mathsf{priv}}(\Gamma) = \mathsf{CVal}_{\mathsf{priv}}(\Gamma^*)$, where Γ^* is Γ with a mediator added, and moreover this value can be computed in $|H|^{O(k)}$ time by solving a zero-sum game of that size. Moreover, an ε-hidden-role equilibrium with private communication and $\log M = R = \mathsf{polylog}(|H|, 1/\varepsilon)$ can be computed and executed by the MAX-players in time $\mathsf{poly}(|H|^k, \log(1/\varepsilon))$.*

The proof uses MPC to simulate the mediator and then executes the equilibrium given by Theorem 4.2. The proof of Theorem 4.3, as well as requisite background on multi-party computation, are deferred to Appendix B. We emphasize that Theorems 4.2 and 4.3 are useful not only for algorithmically computing an equilibrium, but also for manual analysis of games: instead of analyzing the infinite design space of possible messaging protocols, it suffices to analyze the finite zero-sum game Γ_0. Our experiments on *Avalon* use both manual analysis and computational equilibrium finding algorithms to solve instances.

Comparison with communication equilibria. As mentioned in Section 1.2, our construction simulating a mediator bears resemblance to the construction used to define *communication equilibria* [Forges, 1986, Myerson, 1986]. At a high level, a communication equilibrium of a game Γ is a Nash equilibrium of Γ augmented with a mediator that is playing according to some fixed strategy μ. Indeed, when team MIN has only one player, it turns out that the two notions coincide:

Theorem 4.4. *Let Γ be a hidden-role game with $k < n/2$. Then $\mathsf{CVal}_{\mathsf{priv}}(\Gamma)$ is exactly the value for MAX of the MAX-optimal communication equilibrium of Γ.*

However, in the more general case where MIN can have more than one player, Theorem 4.4 does not apply: in that case, communication equilibria include all Nash equilibria in particular, and therefore fail to enforce *joint* optimality of the MIN-team, so our concepts and methods are more suitable. The proof is deferred to Appendix C.

4.2 Computing No/Public-Communication Equilibria

In this section, we consider games with no communication or with public-communication and a coordinated minority. Conversely to the private-communication case of Section 4.1, in this case the problem of computing the value of a hidden-role equilibrium is in general NP-hard.

For the remainder of this section, when discussing the problem of "computing the value of a game", we always mean the following promise problem: given a game, a threshold v, and an allowable error $\varepsilon > 0$ (both expressed as rational numbers), decide whether the hidden-role value of Γ is $\geq v$ or $\leq v - \varepsilon$.

Theorem 4.5. *Even in 2-vs-1 games with public roles and $\varepsilon = 1/\mathsf{poly}(|H|)$, computing the TME value (and hence also the hidden-role value, since adversarial team games are a special case of hidden-role games) with public communication is NP-hard.*

Since there is only one MIN-player in the above reduction, the result applies regardless of whether the adversary is coordinated.

Theorem 4.6. *Even with a constant number of players, a minority adversary team, and $\varepsilon = 1/\mathsf{poly}(|H|)$, computing the uncoordinated value of a hidden-role game is coNP-hard with private communication and Σ_2^{P}-hard with public communication or no communication.*

Proofs of results in this section are deferred to Appendix D. Intuitively, the proofs work by constructing gadgets that prohibit any useful communication, thus reducing to the case of no communication.

Adversary Team Assumptions	Communication Type: None	Public	Private
Coordinated, Minority	NP-complete [von Stengel and Koller, 1997]	NP-hard [Thm. 4.5]	P [Thm. 4.3]
Coordinated			*open problem*
Minority	Σ_2^{P}-complete [Thm. 4.6] and [Zhang et al., 2022]	Σ_2^{P}-hard [Thm. 4.6]	coNP-hard [Thm. 4.6]
None			

Table 1: Complexity results for computing hidden-role value with a constant number of players, for various assumptions about the adversary team and notions of communication. The results shaded in green are new to our paper.

5 Worked Example

This section includes a worked example of value computation to illustrate the differences among the notions of equilibrium discussed in the paper and illustrates the utility of having a mediator for private communication. Consider a n-player version of matching pennies $\mathsf{MP}(n)$ as defined in Section 1.4.

Proposition 5.1. *Let $\mathsf{MP}(n)$ be the n-player matching pennies game.*

1. *The TMECor and TMECom values of* PublicTeam($\mathsf{MP}(n)$) *are both* $1/2$.
2. *Without communication or with only public communication, the value of $\mathsf{MP}(n)$ is $1/2^{n-1}$.*
3. *With private communication, the value of $\mathsf{MP}(n)$ is $1/(n+1)$.*

Proof. The first claim, as well as the no-communication value, is known [Basilico et al., 2017].

For the public-communication value, observe that, conditioned on the transcript, the bits chosen by the players must be mutually independent of each other. Thus, the adversary can do the following: pretend to be on team MAX, wait for all communication to finish, and then play 0 if the string of all ones is more conditionally likely than the string of all 1s, and vice-versa[13].

It thus only remains to prove the third claim.

(*Lower bound*) The players simulate a mediator using multi-party computation (see Theorems 4.2 and 4.3). Consider the following strategy for the mediator. Sample a string $b \in \{0,1\}^n$ uniformly at random from the set of $2n+2$ strings that has at most one mismatched bit. Recommend to each player i that they play b_i.

Consider the perspective of the adversary. The adversary sees only a recommended bit b_i. Assume WLOG that $b_i = 0$. Then there are $n+1$ possibilities:

1. b is all zeros (1 way)
2. All other bits of b are 1 (1 way)
3. Exactly one other bit of b is 1 ($n-1$ ways).

The adversary wins in the third case automatically (since the team has failed to coordinate), and, regardless of what the adversary does, it can win only one of the first two cases. Thus the adversary can win at most $n/(n+1)$ of the time, that is, this strategy achieves value $1/(n+1)$.

(*Upper bound*) Consider the following adversary strategy. The adversary communicates as it would do if it were on team MAX. Let b_i be the bit that the adversary would play if it were on team MAX. The adversary plays b_i with probability $1/(n+1)$ and $1-b_i$ otherwise. We need only show that no pure strategy of the medaitor achieves value better than $1/(n+1)$ against this adversary. A strategy of the mediator is identified by a bitstring b. If b is all zeros or all ones, the team wins if and only if the adversary plays b_i (probability

[13] In general, computing the conditional probabilities could take exponential time, but when defining the notion of value here, we are assuming that players have unbounded computational resources. This argument not work for computationally-bounded adversaries. Indeed, if the adversary were computationally bounded, MAX would be able to use cryptography to build private communication channels and thus implement a mediator, allowing our main positive result Theorem 4.3 to apply.

$1/(n+1)$). If b has a single mismatched bit, the team wins if and only if the mismatched bit is the adversary (probability $1/n$) and the adversary flips b_i (probability $n/(n+1)$). □

6 Properties of Hidden-role Equilibria

In the following, we discuss interesting properties of hidden-role equilibria given the definition we provided in Section 1.1, and that make them fairly unique relative to other notions of equilibrium in team games.

6.1 The Price of Hidden Roles

One interesting question arising from hidden-role games is the *price* of having them. That is, how much value does MAX lose because roles are hidden? In this section, we define this quantity and derive reasonably tight bounds on it.

Definition 6.1. The *public-team refinement* of an n-player hidden-role game Γ is the adversarial team game PUBLICTEAM(Γ) defined by starting with the (uncoordinated) split-personality game, and adding the condition that all team assignments t_i are publicly observed by all players.

Definition 6.2. For a given hidden-role game Γ in which MAX is guaranteed a nonnegative value (*i.e.*, $u_i(z) \geq 0$ whenever i is on team MAX), the *price of hidden roles* $\mathsf{PoHR}(\Gamma)$ is the ratio between the TME value of PUBLICTEAM(Γ) and the hidden-role value of USPLIT(Γ).

For a given class of hidden-role games $\mathcal{G}$, the price of hidden roles $\mathsf{PoHR}(\mathcal{G})$ is the supremum of the price of hidden roles across all games $\Gamma \in \mathcal{G}$.

Theorem 6.3. *Let $D \in \Delta(\{\text{MAX}, \text{MIN}\}^n)$ be any distribution of teams assignments. Let $\mathcal{G}_{n,D}$ be the class of all hidden-role games with n players and team assignment distribution D. Then the price of hidden roles of $\mathcal{G}_{n,D}$ is exactly the largest probability assigned to any team by D, that is,*

$$\mathsf{PoHR}(\mathcal{G}_{n,D}) = \max_{t \in \{\text{MAX},\text{MIN}\}^n} \Pr_{t' \sim D}[t' = t].$$

The lower bound is achieved even in the presence of private communication.

Proof. Let t^* be the team to which D assigns the highest probability, and let p^* be that probability. Our goal is to show that the price of hidden roles is $1/p^*$.

(*Upper bound*) Team MAX assumes that the true MAX-team is exactly the team t^*. Then MAX gets utility at most a factor of $1/p^*$ worse than the TME value of PUBLICTEAM(Γ): if the assumption is correct, then MAX gets the TME value; if the assumption is incorrect, MAX gets value at least 0 thanks to the condition on MAX's utilities in Definition 6.2.

(*Lower bound*) Consider the following game Γ. Nature first selects a team assignment $t \sim D$ and each player privately observes its team assignment. Then, all players are simultaneously asked to announce what they believe the true team assignment is. The MAX-team wins if every MAX-player announces the true team assignment. If MAX wins, MAX gets utility 1; otherwise MAX gets utility 0.

Clearly, if teams are made public, MAX wins easily. With teams not public, suppose that we add a mediator to the game so that Theorem 4.1 applies. This cannot decrease MAX's value. The mediator's strategy amounts to selecting what team each player should announce. Mediator strategies in which different players announce different teams are dominated. The mediator strategy in which the mediator tells every player to announce team t wins if and only if t is the true team, which happens with probability at most p^* (if $t = t^*$). Thus, even the game with a mediator added has value at most p^*, completing the proof. □

This implies immediately:

Corollary 6.4. *Let $\mathcal{G}_{n,k}$ be the class of all hidden-role games where the number of players and adversaries are always exactly n and k respectively. The price of hidden roles in $\mathcal{G}_{n,k}$ is exactly $\binom{n}{k}$.*

Variant	5 Players	6 Players
No special roles (*Resistance*)	$3 / 10 = 0.3000^*$	$1 / 3 \approx 0.3333^*$
Merlin	$2 / 3 \approx 0.6667^*$	$3 / 4 = 0.7500^*$
Merlin + Mordred	$371 / 1782 \approx 0.4102$	$6543 / 12464 \approx 0.5250$
Merlin + 2 Mordreds	$5 / 18 \approx 0.2778$	$99 / 340 \approx 0.2912$
Merlin + Mordred + Percival + Morgana	$67 / 120 \approx 0.5583$	—

Table 2: Exact equilibrium values for 5- and 6-player *Avalon*. The values marked * were also manually derived by Christiano [2018]; we match their results. '—': too large to solve.

In particular, when $k = 1$, the price of hidden roles is at worst n. This is in sharp contrast to the *price of communication* and *price of correlation* in ATGs, both of which can be arbitrarily large even when $n = 3$ and $k = 1$ [Basilico et al., 2017, Celli and Gatti, 2018].

6.2 Order of Commitment and Duality Gap

In Definition 3.5, when choosing the TME as our solution concept and defining the split-personality game, we explicitly choose that MAX should pick its strategy before MIN—that is, the team committing to a strategy is the same one that has incomplete information about the roles. One may ask whether this choice is necessary or relevant: for example, what happens when the TME problem (1) satisfies the minimax theorem? Perhaps surprisingly, the answer to this question is that, at least with private communication, *the minimax theorem in hidden-role games only holds in "trivial" cases*, in particular, when the hidden-role game is equivalent to its public-role refinement (Definition 6.1).

Proposition 6.5. *Let Γ be any hidden-role game. Define $\mathsf{UVal}'_{\mathsf{priv}}(\Gamma)$ identically to $\mathsf{UVal}_{\mathsf{priv}}(\Gamma)$, except that* MIN *commits before* MAX*—that is, in (1), the maximization and minimization are flipped. Then $\mathsf{UVal}'_{\mathsf{priv}}(\Gamma)$ is equal to the TME value of* PublicTeam(Γ) *with communication—that is, the equilibrium value of the zero-sum game in which teams are public and intra-team communication is private and unlimited.*

Proof. It suffices to show that team MAX can always cause the teams to be revealed publicly if MIN commits first. Let s be a long random string. All members of team MAX broadcast s publicly at the start of the game. Since MIN commits first, MIN cannot know or guess s if it is sufficiently long; thus, with exponentially-good probability, this completely reveals the teams publicly. Then, using the private communication channels, team MAX can play a TMECom of PublicTeam(Γ). □

Therefore, the choice of having MAX commit to a strategy before MIN is forced upon us: flipping the order of commitment would ruin the point of hidden-role games.

7 Experimental Evaluation: *Avalon*

In this section, we apply Theorem 4.2 to instances of the popular hidden-role game *The Resistance: Avalon* (hereafter simply *Avalon*). We solve various versions of the game with up to six players.

A game of *Avalon* proceeds, generically speaking, as follows. There are n players, $\lceil n/3 \rceil$ of which are randomly assigned to team MIN and the rest to team MAX. Team MIN is informed. Some special rules allow players observe further information; for example, *Merlin* is a MAX-player who observes the identity of the players on team MIN, except the MIN-player *Mordred*, and the MAX-player *Percival* knows *Merlin* and *Morgana* (who is on team MIN), but does not know which is which. The game proceeds in five rounds. In each round, a *leader* publicly selects a certain number of people (defined as a function of the number of players and current round number) to go on a *mission*. Players then publicly vote on whether to accept the leader's choice. If a strict majority vote to accept, the mission begins; otherwise, leadership goes to the player to the left. If four votes pass with no mission selected, there is no vote on the fifth mission (it automatically gets accepted). If a MIN-player is sent on a mission, they have the chance to *fail* the mission. The goal of MAX is to have three

missions pass. If *Merlin* is present, MIN also wins by correctly guessing the identity of Merlin at the end of the game. *Avalon* is therefore parameterized by the number of players and the presence of the extra roles *Merlin, Mordred, Percival*, and *Morgana*.

Avalon is far too large to be written in memory: Serrino et al. [2019] calculates that 5-player *Avalon* has at least 10^{56} information sets. However, in *Avalon* with ≤ 6 players, many simplifications can be made to the zero-sum game given by Theorem 4.2 without changing the equilibrium. These are detailed in Appendix E, but here we sketch one of them which has theoretical implications. Without loss of generality, in the zero-sum game in Theorem 4.2, the mediator completely dictates the choice of missions by telling everyone to propose the same mission and vote to accept missions, and MIN can do nothing to stop this. Therefore, team MIN always has symmetric information in the game: they know each others' roles (at least when $n \leq 6$), and the mediator's recommendations to the players may as well be public. Therefore, *Avalon* is already natively without loss of generality a game with a coordinated adversary in the sense of Section 3.2, so the seemingly strong assumptions used in Definition 3.3 are in fact appropriate in *Avalon*. Even after our simplifications, the games are fairly large, *e.g.*, the largest instance we solve has 2.2 million infosets and 26 million terminal nodes.

Our results are summarized in Table 2. Games were solved using a CPU compute cluster machine with 64 CPUs and 480 GB RAM, using two algorithms:

1. A parallelized version of the PCFR+ algorithm [Farina et al., 2021], a scalable no-regret learning algorithm. PCFR+ was able to find an approximate equilibrium with exploitability $< 10^{-3}$ in less than 10 minutes in the largest game instance, and was able to complete 10,000 iterations in under two hours for each game.
2. An implementation of the simplex algorithm with exact (rational) precision, which was warmstarted using incrementally higher-precision solutions obtained from configurable finite-precision floating-point arithmetic implementation of the simplex algorithm, using an algorithm similar to that of Farina et al. [2018]. This method incurred significantly higher runtimes (in the order of hours to tens of hours), but had the advantage of computing *exact* game values at equilibrium.

Table 2 shows exact game values for the instances we solved.

Findings We solve *Avalon* exactly in several instances with up to six players. In the simplest instances (*Resistance* or only Merlin), Christiano [2018] previously computed equilibrium values by hand. The fact that we match those results is positive evidence of the soundness of both our equilibrium concepts and our algorithms.

Curiously, as seen in Table 2, the game values are not "nice" fractions: this suggests to us that most of the equilibrium strategies will likely be inscrutable to humans. The simplest equilibrium not previously noted by Christiano, namely Merlin + 2 Mordreds with 5 players, is scrutable, and is analyzed in detail in Appendix E.4.

Also curiously, having Merlin and two Mordreds (*i.e.*, having a Merlin that does not actually know anything) is not the same as having no Merlin. If it were, we would expect the value of Merlin and two Mordreds to be $0.3 \times 2/3 = 0.2$ (due to the 1/3 probability of MIN randomly guessing Merlin). But, the value is actually closer to 0.28. The discrepancy is due to the "special player" implicit correlation discussed in Section 1.4.

8 Conclusions and Future Research

In this paper, we have initiated the formal study of hidden-role games from a game-theoretic perspective. We build on the growing literature on ATGs to define a notion of equilibrium, and give both positive and negative results surrounding the efficient computation of these equilibria. In experiments, we completely solve real-world instances of *Avalon*. As this paper introduces a new and interesting class of games, we hope that it will be the basis of many future papers as well. We leave many interesting questions open.

1. From our results, it is not even clear that hidden-role equilibria and values can be computed in *finite* time except as given by Theorem 4.3. Is this possible? For example, is there a revelation-principle-like characterization for *public* communication that would allow us to fix the structure of the communication? We believe this question is particularly important, as humans playing hidden-role games are often restricted to communicating in public and cannot reasonably run the cryptographic protocols necessary to build private communication channels or perform secure MPC.
2. Changing the way in which communication works can have a ripple effect on the whole paper. One particular interesting change that we do not investigate is *anonymous* messaging, in which players can, publicly or privately, send messages that do not leak their own identity. How does the possibility of anonymous messaging affect the central results of this paper?
3. In this paper, we do not investigate or define hidden-role games where *both* teams have imperfect information about the team assignment. What difference would that make? In particular, is there a way to define an equilibrium concept in that setting that is "symmetric" in the sense that it does not require a seemingly-arbitrary choice of which team ought to commit first to its strategy?

Acknowledgements

The work of Prof. Sandholm's group is supported by the National Science Foundation under grants RI-2312342 and RI-1901403, and by the ARO under award W911NF2210266. Nicola Gatti is supported by FAIR (Future Artificial Intelligence Research) project, funded by the NextGenerationEU program within the PNRR-PE-AI scheme (M4C2, Investment 1.3, Line on Artificial Intelligence). We thank Justin Raizes for useful discussions on multi-party computation.

References

Ittai Abraham, Danny Dolev, Rica Gonen, and Joe Halpern. Distributed computing meets game theory: robust mechanisms for rational secret sharing and multiparty computation. In *ACM Symposium on Principles of Distributed Computing*, pages 53–62, 2006.

Matthew Aitchison, Lyndon Benke, and Penny Sweetser. Learning to deceive in multi-agent hidden role games. In Stefan Sarkadi, Benjamin Wright, Peta Masters, and Peter McBurney, editors, *Deceptive AI*, pages 55–75. Springer International Publishing, 2021. ISBN 978-3-030-91779-1.

Nicola Basilico, Andrea Celli, Giuseppe De Nittis, and Nicola Gatti. Team-maxmin equilibrium: efficiency bounds and algorithms. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

Donald Beaver. Multiparty protocols tolerating half faulty processors. In *Advances in Cryptology—CRYPTO'89 Proceedings 9*, pages 560–572. Springer, 1990.

Mark Braverman, Omid Etesami, and Elchanan Mossel. Mafia: A theoretical study of players and coalitions in a partial information environment. *Annals of Applied Probability*, 18:825–846, 2006.

Noam Brown and Tuomas Sandholm. Solving imperfect-information games via discounted regret minimization. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

Andrea Celli and Nicola Gatti. Computational results for extensive-form adversarial team games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Paul F. Christiano. Solving avalon with whispering. 2018. Blog post. Available at https://sideways-view.com/2018/08/25/solving-avalon-with-whispering/.

Gabriele Farina, Nicola Gatti, and Tuomas Sandholm. Practical exact algorithm for trembling-hand equilibrium refinements in games. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Faster game solving via predictive Blackwell approachability: Connecting regret matching and mirror descent. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

Francoise Forges. An approach to communication equilibria. *Econometrica: Journal of the Econometric Society*, pages 1375–1385, 1986.

Oscar Garcia Morchon, Heribert Baldus, Tobias Heer, and Klaus Wehrle. Cooperative security in distributed sensor networks. In *2007 International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom 2007)*, pages 96–105, Nov 2007. doi: 10.1109/COLCOM.2007.4553817.

Oscar Garcia-Morchon, Dmitriy Kuptsov, Andrei Gurtov, and Klaus Wehrle. Cooperative security in distributed networks. *Computer Communications*, 36(12):1284–1297, 2013. ISSN 0140-3664. doi: https://doi.org/10.1016/j.comcom.2013.04.007. URL https://www.sciencedirect.com/science/article/pii/S0140366413001072.

John Harsanyi. Game with incomplete information played by Bayesian players. *Management Science*, 14: 159–182; 320–334; 486–502, 1967–68.

Johan Håstad. Some optimal inapproximability results. *Journal of the ACM (JACM)*, 48(4):798–859, 2001.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermyn, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank Sharma, Nova DasSarma, Roger Grosse, Shauna Kravec, Yuntao Bai, Zachary Witten, Marina Favaro, Jan Brauner, Holden Karnofsky, Paul Christiano, Samuel R. Bowman, Logan Graham, Jared Kaplan, Sören Mindermann, Ryan Greenblatt, Buck Shlegeris, Nicholas Schiefer, and Ethan Perez. Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv: 2401.05566*, 2024.

Geoffrey Irving, Paul F. Christiano, and Dario Amodei. AI safety via debate. *CoRR*, abs/1805.00899, 2018. URL http://arxiv.org/abs/1805.00899.

Sergei Izmalkov, Silvio Micali, and Matt Lepinski. Rational secure computation and ideal mechanism design. In *Symposium on Foundations of Computer Science (FOCS)*, pages 585–595, Washington, DC, 2005.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Kwan Yee Ng, Juntao Dai, Xuehai Pan, Aidan O'Gara, Yingshan Lei, Hua Xu, Brian Tse, Jie Fu, Stephen McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, and Wen Gao. AI alignment: A comprehensive survey. *CoRR*, abs/2310.19852, 2023. doi: 10.48550/ARXIV.2310.19852. URL https://doi.org/10.48550/arXiv.2310.19852.

Jonathan Katz. Bridging game theory and cryptography: Recent results and future directions. In *Theory of Cryptography: Fifth Theory of Cryptography Conference, TCC 2008, New York, USA, March 19-21, 2008. Proceedings 5*, pages 251–272. Springer, 2008.

Daphne Koller and Nimrod Megiddo. The complexity of two-person zero-sum games in extensive form. *Games and Economic Behavior*, 4(4):528–552, October 1992.

Daphne Koller, Nimrod Megiddo, and Bernhard von Stengel. Fast algorithms for finding randomized strategies in game trees. In *ACM Symposium on Theory of Computing (STOC)*, 1994.

Kavya Kopparapu, Edgar A. Duéñez-Guzmán, Jayd Matyas, Alexander Sasha Vezhnevets, John P. Agapiou, Kevin R. McKee, Richard Everett, Janusz Marecki, Joel Z. Leibo, and Thore Graepel. Hidden agenda: a social deduction game with diverse learned equilibria. *CoRR*, abs/2201.01816, 2022. URL https://arxiv.org/abs/2201.01816.

Yehuda Lindell. Secure multiparty computation (mpc). *Cryptology ePrint Archive*, 2020.

Heng Liu. Correlation and unmediated cheap talk in repeated games with imperfect monitoring. *International Journal of Game Theory*, 46:1037–1069, 2017.

Michael Maschler, Shmuel Zamir, and Eilon Solan. *Game Theory*. Cambridge University Press, 2020.

James D. Miller, Roman Yampolskiy, Olle Häggström, and Stuart Armstrong. Chess as a testing grounds for the oracle approach to AI safety. In Huáscar Espinoza, John A. McDermid, Xiaowei Huang, Mauricio Castillo-Effen, Xin Cynthia Chen, José Hernández-Orallo, Seán Ó hÉigeartaigh, Richard Mallah, and Gabriel Pedroza, editors, *Proceedings of the Workshop on Artificial Intelligence Safety 2021 co-located with the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI 2021), Virtual, August, 2021*, volume 2916 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021. URL https://ceur-ws.org/Vol-2916/paper_13.pdf.

Viraaji Mothukuri, Reza M Parizi, Seyedamin Pouriyeh, Yan Huang, Ali Dehghantanha, and Gautam Srivastava. A survey on security and privacy of federated learning. *Future Generation Computer Systems*, 115:619–640, 2021.

Roger B Myerson. Multistage games with communication. *Econometrica: Journal of the Econometric Society*, pages 323–358, 1986.

John Nash. Equilibrium points in n-person games. *National Academy of Sciences*, 36:48–49, 1950.

Aidan O'Gara. Hoodwinked: Deception and cooperation in a text-based game for language models. *CoRR*, abs/2308.01404, 2023. doi: 10.48550/ARXIV.2308.01404. URL https://doi.org/10.48550/arXiv.2308.01404.

Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. AI deception: A survey of examples, risks, and potential solutions. *CoRR*, abs/2308.14752, 2023. doi: 10.48550/ARXIV.2308.14752. URL https://doi.org/10.48550/arXiv.2308.14752.

Tal Rabin. Robust sharing of secrets when the dealer is honest or cheating. *Journal of the ACM (JACM)*, 41(6):1089–1109, 1994.

Tal Rabin and Michael Ben-Or. Verifiable secret sharing and multiparty protocols with honest majority. In *Proceedings of the twenty-first annual ACM symposium on Theory of computing*, pages 73–85, 1989.

Jérémy Scheurer, Mikita Balesni, and Marius Hobbhahn. Technical report: Large language models can strategically deceive their users when put under pressure. *CoRR*, abs/2311.07590, 2023. doi: 10.48550/ARXIV.2311.07590. URL https://doi.org/10.48550/arXiv.2311.07590.

Jack Serrino, Max Kleiman-Weiner, David C Parkes, and Josh Tenenbaum. Finding friend and foe in multi-agent games. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

Dipty Tripathi, Amit Biswas, Anil Kumar Tripathi, Lalit Kumar Singh, and Amrita Chaturvedi. An integrated approach of designing functionality with security for distributed cyber-physical systems. *J. Supercomput.*, 78(13):14813–14845, sep 2022. ISSN 0920-8542. doi: 10.1007/s11227-022-04481-9. URL https://doi.org/10.1007/s11227-022-04481-9.

Amparo Urbano and Jose E Vila. Computational complexity and communication: Coordination in two–player games. *Econometrica*, 70(5):1893–1927, 2002.

Bernhard von Stengel. Efficient computation of behavior strategies. *Games and Economic Behavior*, 14(2): 220–246, 1996.

Bernhard von Stengel and Daphne Koller. Team-maxmin equilibria. *Games and Economic Behavior*, 21(1): 309–321, 1997. ISSN 0899-8256. doi: https://doi.org/10.1006/game.1997.0527.

Zelai Xu, Chao Yu, Fei Fang, Yu Wang, and Yi Wu. Language agents with reinforcement learning for strategic play in the werewolf game. *CoRR*, abs/2310.18940, 2023. doi: 10.48550/ARXIV.2310.18940. URL https://doi.org/10.48550/arXiv.2310.18940.

B. Zhang, Gabriele Farina, and Tuomas Sandholm. Team belief dag: Generalizing the sequence form to team games for fast computation of correlated team max-min equilibria via regret minimization. In *International Conference on Machine Learning*, 2022.

Brian Hu Zhang and Tuomas Sandholm. Polynomial-time optimal equilibria with a mediator in extensive-form games. In *NeurIPS*, 2022.

Brian Hu Zhang, Gabriele Farina, Ioannis Anagnostides, Federico Cacciamani, Stephen Marcus McAleer, Andreas Alexander Haupt, Andrea Celli, Nicola Gatti, Vincent Conitzer, and Tuomas Sandholm. Computing optimal equilibria and mechanisms via learning in zero-sum extensive-form games. *arXiv preprint arXiv:2306.05216*, 2023.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *CoRR*, abs/1909.08593, 2019. URL http://arxiv.org/abs/1909.08593.

Martin Zinkevich, Michael Bowling, Michael Johanson, and Carmelo Piccione. Regret minimization in games with incomplete information. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2007.

A Mediated Games and the Revelation Principle

In this section, we prove Theorem 4.1 and Theorem 4.2.

A.1 Theorem 4.1

Theorem 4.1 (Revelation Principle)**.** *Let Γ^* be a mediated hidden-role game. Then, for $R \geq 2$ and $M \geq |H|$, there exists a coordinated private-communication equilibrium in which the players on* MAX *have a TME profile in which, at every step, the following events happen in sequence:*

1. *every player on team* MAX *sends its observation privately to the mediator,*
2. *the mediator sends to every player (*MAX *and* MIN*) a recommended action, and*
3. *all players on team* MAX *play their recommended actions.*

Proof. We follow the usual proof structure of revelation principle proofs. Let $x = (x_1, \ldots, x_n)$ be any strategy profile for team MAX in $\textsc{CSplit}(\textsc{Comm}_{\mathsf{priv}}(\Gamma^*))$. Consider the strategy profile $x' = (x'_1, \ldots, x'_n)$ that operates as follows. For each player i, the mediator instantiates a simulated version of each player i playing according to strategy x_i. These simulated players are entirely "within the imagination of the mediator".

1. When a (real) player i sends an observation $o_i(h)$ to the mediator, the mediator forwards observation $o_i(h)$ to the simulated player i.
2. When a simulated player i wants to send a message to another player j, the mediator forwards the message to the *simulated* player j.
3. When a simulated player i plays an action a_i, the mediator forwards the action as a message to the real player i.

Since the strategy x_i is only well-defined on sequences that can actually arise in $\textsc{CSplit}(\textsc{Comm}_{\mathsf{priv}}(\Gamma^*))$, the simulated player i may crash if it receives an impossible sequence of observations. If player i's simulator has crashed, it will no longer send simulated messages, and the mediator will no longer send messages to player i.

It suffices to show that team MIN cannot exploit x' more than x. Let y' be any best-response strategy profile for MIN against x'.

We will show that there exists a strategy y such that (x', y') is equivalent to (x, y). Consider the strategy y for team MIN in which each player i maintains simulators of both x_i and y'_i, and acts as follows.

1. Upon receiving an observation or message, forward it to y'_i
2. If y'_i wants to send an observation to the mediator, forward that observation to x_i.
3. If x_i sends a message, send that message.
4. If x_i plays an action a_i, forward that action to y'_i as a message from the mediator. If x_i crashes, send empty messages to y'_i from the mediator. In either case, when y'_i outputs an action, play that action.

By definition, the profiles (x, y) and (x', y') have the same expected utility (in fact, they are equivalent, in the sense that they induce the same outcome distribution over the terminal nodes of Γ), so we are done. □

A.2 Theorem 4.2

Theorem 4.2. *Let Γ^* be a mediated hidden-role game, $R \geq 2$, and $M \geq |H|$. An (exact) coordinated private-communication hidden-role equilibrium of Γ^* can be computed by solving an extensive-form zero-sum game Γ_0 with at most $|H|^{k+1}$ nodes, where H is the history set of Γ^*.*

Proof. Consider the zero-sum game Γ_0 that works as follows. There are two players: the *mediator*, representing team MAX, and a single *adversary*, representing team MIN. The game state of Γ_0 consists of a history h in Γ (initially the root), and n *sequences* s_i. Define a sequence s_i to be *consistent* if it is a prefix of a terminal sequence $s_i(z)$ for some terminal node z of Γ. For each sequence s_i which ends with an action, let $O(s_i)$ be the set of observations that could be the next observations of player i in Γ.

1. For each player i on team MAX, let $\tilde{o}_i = o_i(h)$ be the true observation of player i. The adversary observes all recommendations $o_i(h)$ for players i on its team. Then, for each such player, the adversary picks an observation $\tilde{o}_i \in O(s_i) \cup \{\perp\}$. Each $\tilde{o}_i$ is appended to the corresponding s_i.

2. The mediator observes $(\tilde{o}_1, \ldots, \tilde{o}_n)$, and picks action recommendations $a_i \in A_i(\tilde{o}_i)$ to recommend to each player i, and appends each a_i to the corresponding s_i.

3. Players on team MAX automatically play their recommended actions. The adversary observes all actions $(a_i : t_i = \text{MIN})$ recommended to players on team MIN, and selects the action played by each member of team MIN.

The size of this game is given by the number of tuples of the form $(h, s_1, \ldots, s_k)$ where s_i is the sequence of adversary i and h is a node of the original game Γ. There are at most $|H|^{k+1}$ of these, so we are done. □

B Multi-Party Computation and Proof of Theorem 4.3

We first formalize the usual setting of multi-party computation (MPC). Let X be the set of binary strings of length ℓ, and let λ be a security parameter. Rabin and Ben-Or [1989] claims in their Theorem 4 that essentially any protocol involving a mediator can be efficiently simulated without a mediator so long as more than half the players follow the protocol and we allow some exponentially small error. However, they do not include a proof of this result. In the interest of completeness, we prove the version of their result that is needed for our setting, based only on the primitives of *secure multi-party computation* and *verifiable secret sharing*.

B.1 Secure MPC

In *secure MPC*, there is a (possibly randomized) function $f : X^n \to X^n$ defined by a circuit with N nodes. A subset $K \subset [n]$ of size $< n/2$ has been *corrupted*. Each *honest* player $i \in [n] \setminus K$ holds an input $x_i \in X$. The goal is to design a randomized messaging protocol, with *private* communication, such that, regardless of what the corrupted players do, there exist inputs $\{x_j : j \in K\}$ such that:

1. (Output delivery) At the end of the protocol, each player i learns its own output, $y_i := f(x_1, \ldots, x_n)_i$, with probability $1 - 2^{-\lambda}$.

2. (Privacy) No subset of $< n/2$ players can learn any information except their own output y_i. That is, the players cannot infer any extra information from analyzing the transcripts of the message protocol than what they already know.

   This can be intuitively modeled as follows. If any minority of colluded players were to analyze the empirical distributions of transcripts of the protocol for any input-output tuple, then such distribution would be fully explained in terms of their input-outputs, leaving no conditional dependence on other players' input-outputs. Formally, for any such subset $K \subseteq [n]$ of size $< n/2$, there exists a randomized

algorithm Sim_K that takes the inputs and outputs of the players in set K, and reconstructs transcripts T, such that for all $x \in X^n$, we have

$$\sum_T |\Pr[\mathsf{Sim}_K(x_K, y_K) = T] - \Pr[\mathsf{View}_K(x) = T]| \leq 2^{-\lambda}$$

where $\mathsf{View}_K(x)$ is the distribution of transcripts observed by players in set K when running the protocol with input x.

In other words, the players in set K cannot do anything except pass to f inputs of their choice. For our specific application, this implies that introducing an MPC protocol to simulate the mediator is equivalent to having a mediator, because no extra information aside from the intended function will be leaked to the players.

Theorem B.1 ([Beaver, 1990, Rabin and Ben-Or, 1989])**.** *Secure MPC is possible, with polytime (in ℓ, λ, n, N) algorithms that take at most polynomially many rounds and send at most polynomially many bits in each round.*

B.2 Verifiable Secret Sharing

In the *verifiable secret sharing* (VSS) problem, the goal is to design a function $\mathsf{Share} : X \to \bar{X}^n$, where $\bar{X} = \{0,1\}^{\mathsf{poly}(\ell,\lambda,n)}$, that *shares* a secret $x \in X$ by privately informing each player i of its piece $\mathsf{Share}(x)_i$, such that:

1. (Reconstructibility) Any subset of $> n/2$ players can recover the secret fully, even if the remaining players are adversarial. That is, there exists a function $\mathsf{Reconstruct} : X^n \to X$ such that, where $(x_1, \ldots, x_n) \leftarrow \mathsf{Share}(x)$, we have $\mathsf{Reconstruct}(x'_1, \ldots, x'_n) = x$ so long as $x'_i = x_i$ for $> n/2$ players i.

2. (Privacy) No subset of $< n/2$ players can learn any information about the secret x. That is, for any such subset K and any secret x, there exists a distribution $\mathsf{Sim}_K \in \Delta(X^n)$ such that

$$\|\mathsf{Share}(x)_K - \mathsf{Sim}_K\|_1 \leq 2^{-\lambda},$$

where $\|\cdot\|_1$ denotes the ℓ_1-norm on probability distributions.

Theorem B.2 ([Rabin, 1994, Rabin and Ben-Or, 1989])**.** *There exist algorithms* Share *and* $\mathsf{Reconstruct}$*, with runtime* $\mathsf{poly}(\ell, \lambda, n)$*, that implement robust secret sharing.*

VSS is a primitive used in a fundamental way to build secure MPC protocols; to be formally precise in this paper, we will require VSS as a separate primitive as well to maintain the state of the mediator throughout the game.

B.3 Simulating a Mediator

We will assume that the game Γ_0 in Theorem 4.2 has been solved, and that its solution is given by a (possibly randomized) function

$$f : \Sigma \times O^n \to \Sigma \times A^n \tag{2}$$

where Σ is the set of information states of the mediator in Γ_0. At each step, the mediator takes n observations $o_1, \ldots, o_n$ and its current infostate s as input, and updates its infostate and outputs action recommendations according to the function f.

Consider the function $\hat{f} : (\bar{X} \times O)^n \to (\bar{X} \times A)^n$ defined by

$$\hat{f}((x_1, o_1), \ldots, (x_n, o_n)) = ((x'_1, a_1), \ldots, (x'_n, a_n))$$

where

$$(s', a_1, \ldots, a_n) = f(\mathsf{Reconstruct}(x_1, \ldots, x_n), o_1, \ldots, o_n)$$

$$(x'_1, \ldots, x'_n) = \mathsf{Share}(s').$$

That is, $\hat{f}$ operates the mediator with its state secret-shared across the various players. The players will run secure MPC on $\hat{f}$ at every timestep. By the properties of MPC and secret sharing, this securely implements the mediator in such a way that the players on team MIN can neither break privacy nor cause the protocol to fail, with probability better than $O(|H| \cdot 2^{-\lambda})$, where H is the set of histories of the game.

We have thus shown our main theorem, which is more formally stated as follows:

Theorem B.3 (Formal version of Theorem 4.3)**.** *Let:*

- Γ *be a hidden-role game with a* MIN*-team of size* $k < n/2$*,*
- Γ^* *be identical to* Γ *except that there is an additional player who takes no nontrivial actions but is always on team* MAX*;*
- Γ_0 *be the zero-sum game defined by Theorem 4.2 based on* Γ^**;*
- x *be any strategy of the* MAX *player (mediator) in* Γ_0*, represented by an arithmetic circuit* f *as in (2) with* $N = \mathsf{poly}(|H|^k)$ *gates; and*
- λ *be a security parameter.*

Then there exists a strategy profile x' *of the* MAX *players in* $\textsc{CSplit}(\textsc{Comm}^{M,R}_{\mathsf{priv}}(\Gamma))$*, where* $\log M = R = \mathsf{poly}(\lambda, N, \log |H|)$ *such that:*

1. (Equivalence of value) *the value of* x' *is within* $2^{-\lambda}$ *of the value of* x *in* Γ_0*, and*
2. (Efficient execution) *there is a* $\mathsf{poly}(r)$*-time randomized algorithm* $\mathcal{A}_\Gamma$ *that takes as input an infostate* s_i *of* $\textsc{CSplit}(\textsc{Comm}^{M,R}_{\mathsf{priv}}(\Gamma))$ *that ends with an observation, and returns the (possibly random) action that player* i *should play at* s_i*.*

C Connection to Communication Equilibria

In this section, we prove Theorem 4.4, recalled below.

Theorem 4.4. *Let* Γ *be a hidden-role game with* $k < n/2$*. Then* $\mathsf{CVal}_{\mathsf{priv}}(\Gamma)$ *is exactly the value for* MAX *of the* MAX*-optimal communication equilibrium of* Γ*.*

Before proving this result, we must first formally define a communication equilibrium. Given an arbitrary game Γ with n players, consider the $(n+1)$-player game in which the extra player is the mediator. Consider a private-communication extension $\tilde{\Gamma}$ of that game, with $R = 2$ rounds and $M = |H|$, in which only communication with the mediator is allowed. In Γ^*, each player has a *direct strategy* x^*_i in which, at every timestep, the player sends its honest information to the mediator, interprets the mediator's message in reply as an action recommendation, and plays that action recommendation.

Definition C.1. A *communication equilibrium* of Γ is a strategy profile $\mu = (x_1, \ldots, x_n)$ for the mediator of $\tilde{\Gamma}$ such that, with μ held fixed, the profile $(x_1, \ldots, x_n)$ is a Nash equilibrium of the resulting n-player game.

The *revelation principle for communication equilibria* [Forges, 1986, Myerson, 1986] states that, without loss of generality in the above definition, it can be assumed that $x = x^*$, *i.e.*, players are direct in equilibrium.

We now prove the theorem. Let Γ^* be Γ with a mediator who is always on team MAX, and let Γ_0 be the zero-sum game constructed by Theorem 4.2. Because $\mathsf{CVal}_{\mathsf{priv}}(\Gamma)$ is the zero-sum value of Γ_0 by Theorem 4.3, it is enough to prove the inequality chain

$$\mathsf{CVal}_{\mathsf{priv}}(\Gamma) \le \mathsf{CommVal}(\Gamma) \le \mathsf{Val}(\Gamma_0),$$

where CommVal is the value of the MAX-optimal communication equilibrium and Val is the zero-sum game value.

By Theorem 4.3, the hidden-role equilibria of Γ are (up to an arbitrarily small error $\varepsilon > 0$) TMEs of CSplit(Γ^*). By von Stengel and Koller [1997], since MIN has only one player, the TMEs of CSplit(Γ^*) are precisely the MAX-team-optimal Nash equilibria of CSplit(Γ^*). Thus, for a TME $(\mu, x_1, \dots, x_n)$ of that game (where μ is a strategy of the mediator), there exists an adversary strategy y such that $(\mu, x_1, \dots, x_n, y)$ is a Nash equilibrium. But then $(\mu, x_1, \dots, x_n, y)$ is also a communication equilibrium. Thus TMEVal(CSplit(Γ^*)) $\leq$ CommVal(Γ).

We now show the second inequality. If Γ is an adversarial hidden-role game, each player i's strategy can be expressed as a tuple (x_i, y_i) where x_i is player i's strategy in the subtree where i is on team MAX, and y_i is the same on team MIN. In that case, the problem of finding a MAX-optimal communication equilibrium can be expressed as:

$$\begin{aligned} \max_{\mu} \quad & u(\mu, x^*, y^*) \\ \text{s.t.} \quad \forall i \quad & \max_{x_i} u(\mu, x_i, x^*_{-i}, y^*) \leq u(\mu, x^*, y^*) \\ \forall i \quad & \min_{y_i} u(\mu, x^*, y_i, y^*_{-i}) \geq u(\mu, x^*, y^*) \end{aligned}$$

where u is the MAX-team utility function. That is, no player can increase their team's utility by deviating from the direct profile (x^*, y^*), regardless of which team they are assigned to. Now, using the fact that MIN has only one player, we can write MAX's utility function u as a sum $u = \sum_i u_i$ where u_i is the utility of MAX when the MIN-player is player i (weighted by the probability of that happening). Each term u_i depends only on x and the y_is. Thus, the above program can be rewritten as

$$\begin{aligned} \max_{\mu} \quad & \sum_i u_i(\mu, x^*, y^*_i) \\ \text{s.t.} \quad \forall i \quad & \max_{x_i} u(\mu, x_i, x^*_{-i}, y^*) \leq u(\mu, x^*, y^*) \\ \forall i \quad & \min_{y_i} u_i(\mu, x^*, y_i) \geq u_i(\mu, x^*, y^*_i) \end{aligned}$$

or, equivalently,

$$\begin{aligned} \max_{\mu} \quad & \sum_i \min_{y_i} u_i(\mu, x^*, y_i) \\ \text{s.t.} \quad \forall i \quad & \max_{x_i} u(\mu, x_i, x^*_{-i}, y^*) \leq u(\mu, x^*, y^*). \end{aligned}$$

or, equivalently,

$$\begin{aligned} \max_{\mu} \min_{y} \quad & u(\mu, x^*, y) \\ \text{s.t.} \quad \forall i \quad & \max_{x_i} u(\mu, x_i, x^*_{-i}, y^*) \leq u(\mu, x^*, y^*). \end{aligned}$$

This is precisely the problem of computing a zero-sum equilibrium in the game Γ_0, except with an extra constraint, so the inequality follows. □

D Complexity Bounds and Proofs

Here we state and prove the various lower bounds in Table 1. Before proceeding, we make several remarks about the conventions used in this section.

- Utilities will be given *unnormalized* by chance probability. That is, if we say that a player gets utility 1, what we really mean is that the player gets utility $1/p$, where p is the probability that chance sampled all actions on the path to z. Thus the contribution to the expected value from this terminal node will be 1. This makes calculations easier.

- The utility range of the games used in the reductions will usually be of the form $[-M, M]$ where M is large but polynomial in the size of the game. Our definition of an extensive-form game allows only games with reward range $[-1, 1]$. This discrepancy is easily remedied by dividing all utility values in the proofs by M.

- $\Theta(\cdot)$ hides only an absolute constant.

To recap, we show hardness of *approximating* the value of the game to a desired precision $\varepsilon > 0$. Our hardness results will hold even when $\varepsilon = 1/\mathsf{poly}(|H|)$.

Theorem 4.5. *Even in* 2*-vs-*1 *games with public roles and* $\varepsilon = 1/\mathsf{poly}(|H|)$*, computing the TME value (and hence also the hidden-role value, since adversarial team games are a special case of hidden-role games) with public communication is* NP*-hard.*

Proof. We show that given any graph G, it is possible to construct a hidden-role game based on G whose value correspond to the size the graph's max-cut. This reduces MAX-CUT to the TME value problem.

Let G be an arbitrary graph with n nodes and m edges, and consider the following team game (no hidden roles). There are 3 players, 2 of whom are on team MAX. The game progresses as follows.

1. Chance chooses two vertices v_1, v_2 in G, independently and uniformly at random. The two players on team MAX observe v_1 and v_2 respectively.

2. The two players on team MAX select bits b_1, b_2, and the MIN player selects a pair (v_1', v_2'). There are now several things that can happen: (L is a large number to be picked later)

   (a) (*Agreement of players*) If $v_1 = v_2$ and $b_1 \neq b_2$, team MAX gets utility $-L$.

   (b) (*Objective*) If $v_1 \neq v_2$, $b_1 \neq b_2$, and (v_1, v_2) is an edge in G, then team MAX gets utility 1.

   (c) (*Non-leakage*) If $(v_1', v_2') = (v_1, v_2)$ then team MAX gets utility $-(n^2 - 1)L$. Otherwise, team MAX gets utility L.[14]

Consider a sufficiently large $L = \mathsf{poly}(m)$. The game is designed in such a way that MIN's objective is to guess the vertices v_1, v_2 sampled by chance, but she has no information apart from the transcript of communication to guess it. Therefore, MIN's optimal strategy is to punish any communication attempt between the players and play the most likely pair of vertices v_1, v_2. If no communication happens, her best strategy is to play a random pair of vertices. On the other hand, MAX optimal strategy must ensure that under no circumstance players play the same bit when assigned to the same vertex (lest they incur the large penalty L). Therefore, the strategy of both MAX players is to play a fixed bit in each vertex, and the optimal strategy is the one that assigns a bit to the vertices in such a way that the number of edges connecting vertices with different bits is maximized. This corresponds to finding a max cut and therefore the value of the game is (essentially) c^* where c^* is the true size of the maximum cut. Moreover, any communication attempt would be immediately shut down by MIN strategy since any leak of the observation received on the public channel implies to receive a fraction of the large penalty L.

We now formalize this intuition.

First, note that MAX can achieve utility exactly c^* by playing according to a maximum cut. To see that MAX cannot do significantly better, consider the following strategy for team MIN. Observe the entire transcript τ of messages shared between the two MAX players. Pick (v_1, v_2) maximizing the probability $p(\tau | v_1, v_2)$ that the players would have produced τ.[15]

First, suppose that MAX does not communicate. Then MIN's choice is independent of MAX's, so MAX can WLOG play a pure strategy. If $b_1 \neq b_2$ for any pair (v_1, v_2), then MAX loses utility L in expectation. For

[14] Note that MIN playing uniformly at random means that the expected utility of this term is 0.

[15] Note again, as in Section 5, that this computation may take time exponential in r and the size of G, but we allow the players to perform unbounded computations.

$L > n^2$, this makes any such strategy certainly inferior to playing the maximum cut. Therefore, MAX should play the maximum cut, achieving value c^*.

Now suppose that MAX uses communication. Fix a transcript τ, and let

$$\delta := \max_{v_1, v_2} p(v_1, v_2|\tau) - \frac{1}{n^2}.$$

Now note that we can write $p(\cdot|\tau) = (1-\alpha)q_0 + \alpha q_1$, where $q_0, q_1 \in \Delta(V^2)$, q_0 is uniform, and $\alpha \leq \Theta(n^4\delta)$. Now consider any strategy that MAX could play, given transcript τ. Such a strategy has the form $x_1(b_1|v_1)$ and $x_2(b_2|v_2)$. Since q is α-close to uniform, the utility of MAX under this strategy conditioned on τ must be bounded above by

$$(1-\alpha)u_0 + \alpha \leq (1-\alpha)c^* + \alpha \leq c^* + \alpha \leq c^* + \Theta(n^4\delta)$$

where u_0 is the expected value of profile (x_1, x_2) given τ if $v_1, v_2|\tau$ were truly uniform. But now, in expectation over τ, team MIN can gain utility $Ln^2\delta$ by playing $\operatorname{argmax}_{v_1,v_2} p(v_1, v_2|\tau)$. So, MAX's utility is bounded above by

$$c^* + \Theta(n^4\delta) - Ln^2\delta \leq c^*$$

by taking L sufficiently large. □

The next result illustrates the difference between the uncoordinated hidden-role value and the coordinated hidden-role value which is the focus of the positive results in our paper. Whereas the coordinated hidden-role value with private communication can be computed in polynomial time when k is constant (Theorem 4.3), the uncoordinated hidden-role value cannot, even when $k = 2$:

Theorem D.1. *Even in 3-vs-2 hidden-role games, the uncoordinated hidden-role value problem with private communication is* coNP*-hard.*

Proof. We reduce from UNSAT. Let ϕ be any 3-CNF-SAT formula, and consider the following 5-player hidden-role game. Two players are chosen uniformly at random to be on team MIN; the rest are on team MAX. The players on team MIN know each other. The players on team MIN play the SAT gadget game described by Koller and Megiddo [1992]. Namely:

1. The players on team MIN are numbered P1 and P2, at random, by chance.
2. Chance selects a clause C in ϕ and tells P1.
3. P1 selects a variable x_i in C, and that variable (but not its sign in C, nor the clause C itself) is revealed to P2.
4. P2 selects an assignment $b_i \in \{0, 1\}$ to x_i. MIN wins the gadget game if the assignment b_i matches the sign of x_i in C.

The value of this game is decreasing with M and R since MAX does nothing, so it is in the best interest of MAX to select $R = 0$ (*i.e.*, allow no communication). In that case, the best probability with which MIN can win the game is exactly the maximum fraction of clauses satisfied by any assignment, which completes the proof. □

The above result is fairly straightforward: it is known that optimizing the joint strategy of a team with asymmetric information[16] is hard [Koller and Megiddo, 1992], and private communication does not help if MAX does not allow its use. However, next, we will show that the result even continues to apply when MIN has *symmetric* information, that is, when the original game Γ is coordinated. This may seem mysterious at first, but the intuition is the following. Just because Γ has symmetric information for the MIN-team, does not mean $\textsc{USplit}(\textsc{Comm}_{\mathsf{priv}}(\Gamma))$ does. Indeed, MAX-players can send different private messages to different

[16] For our purposes, we will say that MIN has *symmetric information* if all players MIN have the same observation at every timestep. This implies that they can be merged into a single player without loss of generality.

MIN-players, resulting in asymmetric information among MIN-players. This result illustrates precisely the reason that we define two different split-personality games, rather than simply dealing with the special case where the original game Γ has symmetric information for the MIN-team.

Theorem D.2. *Even in* 3*-vs-*2 *hidden-role games with a mediator in which no* MIN*-player has any information beyond the team assignment, the uncoordinated hidden-role value problem with private communication is* coNP*-hard.*

Proof. We will reduce from (the negation of) MAX-CUT. Let G be an arbitrary graph with n nodes and m edges, and consider the following 5-player hidden-role game with 2 players on the MIN-team and 3 players on the MAX-team. Player 5 is always on team MAX and is the mediator. The other four players are randomly assigned teams so that two are on team MAX and two are on team MIN. The game proceeds as follows.

1. Chance chooses vertices $v_1, \dots, v_4$ uniformly at random from G. The mediator privately observes the whole tuple $(v_1, \dots, v_4)$.

2. For notational purposes, call the players on team MAX 3 and 4, and MIN 1 and 2. (The mediator does not know these numbers.) After some communication, the following actions happen simultaneously:

   (a) P1 and P2 select bits $b_1, b_2 \in \{0, 1\}$;

   (b) P3 and P4 select vertices v_3', v_4' of G and players $i_3, i_4 \in \{1, 2, 3, 4\}$.

3. The following utilities are given: (L is a large number to be picked later, and each item in the list represents an additive term in the utility function)

   (a) (*Correct vertex identification*) For each player $i \in \{3, 4\}$, if $v_i \neq v_i'$ then MAX gets utility $-L^4$

   (b) (*Agreement of* MIN*-players*) If $v_1 = v_2$ and $b_1 \neq b_2$ then team MAX gets utility L.

   (c) (*Objective*) If $v_1 \neq v_2$, $b_1 \neq b_2$, and (v_1, v_2) is an edge in G, then team MAX gets utility -1.

   (d) (*Privacy*) For each $j \in \{3, 4\}$, if i_j^* is on team MIN, then MAX gets utility L. Otherwise, MAX gets utility $-L/2$.

We claim that this game has value (essentially) $-c^*$, where c^* is the actual size of the maximum cut of G. To see this, observe first that the mediator *must* tell all players their true vertices v_i, lest it risk incurring the large negative utility $-L^2$. Further, any player except the mediator who sends a message must be on team MIN. This prevents team MIN from communicating. Thus, the mediator's messages force MIN to play an asymmetric-information identical-interest game, which is hard.

We now formalize this intuition. First, consider the following strategy for team MAX: The mediator sends all players their true types, and MAX-players play their types. If any MAX-player sees a message sent from anyone except the mediator, the MAX-player guesses that that player is on team MIN.

Now consider any (pure) strategy profile of team MIN. First, MIN achieves utility $-c^*$ by observing the mediator's message and playing bits according to a maximum cut. We now show that this is the best that MIN can do. Sending messages is, as before, a bad idea. Thus, a pure strategy profile of MIN is given by four $f_1, f_2, f_3, f_4 : V \to \{0, 1\}$ denoting how player i should pick its bits. But then $f_1 = f_2 = f_3 = f_4$; otherwise, the agreement of MIN-players would guarantee that MIN is not playing optimally for large enough L.

Now, for any (possibly mixed) strategy profile of team MAX, consider the following strategy profile for each MIN-player. Let $f : V \to \{0, 1\}$ be a maximum cut. Pretend to be a MAX-player, and let v_i' be the vertex that would be played by that MAX-player. Play $f(v_i')$.

First, consider any MAX-player strategy profile for which, for some player i and some v_i, the probability that $v_i' \neq v_i$ exceeds $1/L^2$. Then MAX gets a penalty of roughly L^2 in expectation, but now setting L large enough

would forces MAX to have utility worse than -1, so that MAX would rather simply play v_i with probability 1.

Now, condition on the event that $v_i = v_i'$ for all i (probability at least $1 - \Theta(1/L^2)$). In that case, the utility of MAX is exactly $-c^*$, because MIN is playing according to the maximum cut. Thus, the utility of MAX is bounded by $-(1 - \Theta(1/L^2))c^* + \Theta(1/L^2 \cdot L) \leq -c^*/n^2 + \Theta(n/L) < c^* + 1/2$ for sufficiently large L. Thus, solving the hidden-role game to sufficient precision and rounding the result would give the maximum cut, completing the proof. □

Theorem D.3. *Even in 5-vs-4 hidden-role games, the uncoordinated hidden-role value problem with public communication is Σ_2^{P}-hard.*

Proof. We reduce from $\exists\forall$3-DNF-SAT, which is Σ_2^{P}-complete [Håstad, 2001]. The $\exists\forall$3-DNF-SAT problem is the following. Given a 3-DNF formula $\phi(x, y)$ with k clauses, where $x \in \{0,1\}^m$ and $y \in \{0,1\}^n$, decide whether $\exists x \forall y\ \phi(x, y)$. Consider the following game. There are 9 players, 5 on team MAX and 4 on team MIN. One designated player, who we will call P0, is MAX and has no role in the game. (The sole purpose of this player is so that MAX is a majority.) The other players are randomly assigned teams. These other players are randomly assigned teams. For the sake of analysis, we number the remaining players P1 through P8 such that P1, P3, P4, P5 are on team MAX and P2, P6, P7, P8 are on team MIN. MIN knows the entire team assignment, whereas MAX dos not. We will call P3–P8 "regular players", and P1–P2 "guessers". The game proceeds as follows.

1. For each regular MAX-player, chance selects a literal (either x_j or $\neg x_j$), uniformly at random. For each regular MIN-player (P6–8), chance selects a literal (either y_j or $\neg y_j$), also uniformly at random. Each player privately observes the *variable* (index j), but not the sign of that variable.

2. After some communication, the following actions happen simultaneously.
   (a) P3–P8 select assignments (0 or 1) to their assigned variables.
   (b) P1 (who observes nothing) guesses a player (among the six players P3–P8) that P1 believes is on team MIN.
   (c) P2 (who observes nothing) guesses one literal for each MAX-player (there are $K := (2m)^3$ such possible guesses.)

3. The following utilities are assigned. (L is a large number to be picked later, and each item in the list represents an additive term in the utility function)
   (a) (*Satisfiability*) Chance selects three regular players at random. If the three literals given to those players form a clause in ϕ, and that clause is satisfied, MAX gets utility 1.
   (b) (*Consistency*) If chance selected the same variable three times, if the three players did not give the same assignment, then the team (MAX if the variable was an x_i and MIN if the variable was a y_j) gets utility $-L^2$.
   (c) (*Privacy for* MAX) If P2 guesses the three literals correctly, MIN gets utility L^3K; otherwise, MIN gets utility $-L^3$.[17]
   (d) (*Privacy for* MIN) If P1 guesses a MIN-player, MAX gets utility L; otherwise, MAX gets utility $-L$.

We claim that the value of this game with public communication is at least 1 if and only if ϕ is $\exists\forall$-satisfiable. Intuitively, the rest of the proof goes as follows:

[17] These utilities are once again selected so that a uniformly random guess gets utility 0.

1. MAX will not use the public communication channels if ϕ is $\exists\forall$-satisfiable. By the *Privacy for* MIN term, MIN-players therefore cannot do so either without revealing themselves immediately. Thus, MAX will get utility at least 1 if ϕ is $\exists\forall$-satisfiable.

2. If ϕ is not $\exists\forall$-satisfiable, then consider any MAX-team strategy profile. By the *Privacy for* MAX term, team MAX cannot make nontrivial use of the public communication channel without leaking information to P9. By the *Consistency* term, P1 through P3 must play the same assignment, or else incur a large penalty. So, MAX must play essentially an assignment to the variables in x, but such an assignment cannot achieve positive utility because there will exist a y that makes ϕ unsatisfied.

We now formalize this intuition. Suppose first that ϕ is $\exists\forall$-satisfiable, and let x be the satisfying assignment. Suppose that MAX-players never communicate and assign according to x, and P4 guesses any player that sends a message. Then any MIN-strategy that sends a message is bad for sufficiently large L because it guarantees a correct guess from MAX; any MIN-strategy that is inconsistent is bad because it will lose utility at least L; and any MIN-strategy that is consistent will cause MAX to satisfy at least one clause. Thus MAX guarantees utility at least 1.

Now suppose that ϕ is not $\exists\forall$-satisfiable. Consider any strategy profile for MAX. Suppose that the MIN-players play as follows. During the public communication phase, each MIN-player samples a literal from the set $\{x_1, \neg x_1, \ldots, x_m, \neg x_m\}$ uniformly at random and pretends to be a MAX-player given that literal. P8 observes the public transcript, and selects the triplet of literals that is conditionally most likely given the transcript. By an identical argument to that used in Theorem 4.5, MAX then cannot profit from using the communication channel for sufficiently large L. Therefore we can assume that MAX does not use the communication channels, and therefore by the argument in the previous paragraph, neither does MIN.

Now, the strategy of each MAX-player i can be described by a vector $s_i \in [0,1]^m$, where s_{ij} is the probability that player i assigns 1 to variable x_t. For sufficiently large L, we have $s_i \in [0, \varepsilon] \cup [1-\varepsilon, 1]$ where $\varepsilon = 1/L$, because otherwise MAX would incur a penalty proportional to $L^2\varepsilon > k$ and would rather just play (for example) the all-zeros profile, which guarantees value 0. Condition on the event that every player at every variable chooses to play the most-likely assignment according to the s_is. This happens with probability at least $1 - \Theta(m\varepsilon)$. These assignments must be consistent (*i.e.*, every player must have the same most-likely assignment), or else the players would once again incur a large penalty proportional to L^2. Call that assignment x, and let y be such that $\phi(x, y)$ is unsatisfied. Suppose MIN plays according to y. Then MAX's expected utility is bounded above by $\Theta(m\varepsilon k)$: with probability $1-\Theta(m\varepsilon)$ it is bounded above by 0; otherwise it is bounded above by k. For $\varepsilon < \Theta(1/mk)$ this completes the proof. □

E The Game *Avalon*

E.1 Equivalence of Split-Personality Games

In this section, we show that the choice of whether to use USplit or CSplit—that is, whether MIN is coordinated—is irrelevant for the instantiations of *Avalon* that we investigate in this paper. We refer to Section 7 for a complete description of the *Avalon* game.

For this section, we assume that all MIN players know the roles of all other MIN players. This is always true in the instances considered in the experiments. In particular, we considered instances with six or fewer players and no *Oberon* role, which is a MIN member which is not revealed as such to both *Merlin* and the other MIN members. This guarantees that both the MIN-players can deduce their respective roles. Conversely this is not guaranteed, as for example with 7 players (and hence $k = 3$) and Mordred, the two non-Mordred MIN-players do not know the identity of Mordred.

The main observation we make in this subsection states that the choice of whether to use USplit or CSplit does not matter.

Theorem E.1. *In* Avalon *with private communication, assuming that all* MIN *players know each others' roles, every r-round uncoordinated hidden-role equilibrium is also an r-round coordinated hidden-role equilibrium.*

Before proving the result, we first observe that we cannot *a priori* assume the use the results in Section 4.1 for this proof, because they do not apply to the symmetric hidden-role equilibrium.

Fortunately, a variant of the revelation principle for mediated games (Theorem 4.1) and the resulting zero-sum game theorem (Theorem 4.2) still holds *almost* verbatim: the lone change is that the zero-sum game formed from symmetric splitting, which we will denote $\tilde{\Gamma}_0$, is a zero-sum *team* game, where the MIN players cannot perfectly coordinate with each other, and the equality of value becomes an inequality:

Proposition E.2. *Let Γ^* be a mediated hidden-role game, and let Γ_0 be the zero-sum version posed by Theorem 4.2. Let $\tilde{\Gamma}_0$ be the game that is identical to Γ_0, except that the MIN-players are not coordinated and thus cannot communicate except as permitted in Γ. That is, $\tilde{\Gamma}_0$ is a* team game *with the mediator as the MAX-player and the adversaries as the MIN-players. Then $\mathsf{UVal}_{\mathsf{priv}}(\Gamma^*) \leq \mathsf{TMEVal}(\tilde{\Gamma}_0)$, where TMEVal is the TME value function (with MAX committing first).*[18]

Proof. The revelation principle (Theorem 4.1) applies verbatim in this setting, so we may assume that MAX in Γ^* uses the mediator as specified in the revelation principle. With this assumption, the only difference that remains between Γ^* and $\tilde{\Gamma}_0$ is that, in the latter, MIN-players cannot coordinate *at all*, whereas in Γ^*, MIN-players are able to somewhat coordinate by sending private messages (albeit at the cost of revealing themselves as adversaries). But this is a strict disadvantage for MIN in $\tilde{\Gamma}_0$. □

We therefore have the inequality chain

$$\mathsf{Val}(\Gamma_0) = \mathsf{CVal}_{\mathsf{priv}}(\Gamma) \leq \mathsf{UVal}_{\mathsf{priv}}(\Gamma) \leq \mathsf{UVal}_{\mathsf{priv}}(\Gamma^*) \leq \mathsf{TMEVal}(\tilde{\Gamma}_0)$$

where Γ_0 is the zero-sum game that appears in Theorem 4.2, with asymmetric splitting, Val is the zero-sum game value, and Val is the zero-sum value.

So far, this inequality chain would hold for *any* hidden-role game Γ. We will now show that, specifically for *Avalon*, we have $\mathsf{TMEVal}(\tilde{\Gamma}_0) = \mathsf{Val}(\Gamma_0)$, which would complete the proof of Theorem E.1. This part of the proof depends on the special structure of *Avalon*.

The only difference between Γ_0 and $\tilde{\Gamma}_0$ is that, in the latter, MIN-players cannot communicate among each other because Theorem 4.2 only consider communication with the mediator. But, what would MIN-players even need to communicate? Their information is entirely symmetric, except that they do not know the recommendations that were given to their teammates in the mission proposal and voting phases. We will show the following result.

Lemma E.3. *For both $\tilde{\Gamma}_0$ and Γ_0, the game value does not change if we make the assumption that the mediator unilaterally dictates missions (circumventing the mission proposal and voting process).*

Proof Sketch. Make the following restrictions to strategy spaces.

- The mediator picks a single mission proposal in each round, and recommends that proposal and that everyone vote in favor of it, until it is approved, and ignores any information gained during this phase.
- *Spies* follow the above recommendations, and ignore all information except the eventually-approved mission.

Consider any TMECor of the restricted game. We claim that it also must be a TMECor of the full game:

- The mediator cannot do better, because *Spies* are following all recommendations and ignoring any information gained except the eventual actual mission.
- *Spies* cannot do better, because deviations are ignored by the mediator and will ultimately not affect the approved mission. (because the number of mission proposals and the number of *Resistance* players both exceed the number of *Spies*).

[18] Since MIN is the nontrivial team and MAX commits first, the TME and TMECor values of $\tilde{\Gamma}_0$ coincide here.

This completes the proof. □

But, having restricted the strategy spaces in the above manner, $\tilde{\Gamma}_0$ and Γ_0 become identical, because there is no other source of asymmetric information. Therefore, $\mathsf{TMEVal}(\tilde{\Gamma}_0) = \mathsf{Val}(\Gamma_0)$, and we are done.

E.2 Abstractions

In this section we describe the abstractions employed on the original game in order to transforming into a smaller version, which can then be solved tabularly. The abstractions employed are lossless, in the sense that they do not change the value of the game, and that any Nash equilibrium in the abstracted game corresponds to a Nash equilibrium in the original one.

In particular, our main interest in the experiments is to compute one strategy belonging to a Nash equilibria for each player. This allows us to iteratively remove many actions without loss of generality.

Note that throughout this section, we interchangeably use the standard names *Resistance* and *Spies* to refer to the teams instead of **MAX** and **MIN**.

We apply Theorems 4.1 to 4.3 to an Avalon game instance in order to compute a hidden-communication equilibria. In the first communication round, anybody can claim to the mediator to be a specific role and to know specific knowledge about other player's roles. (For example, a **MIN**-player, that is, a *Spy*, could claim to be Merlin and lie about who is on team **MAX**.) Thanks to the revelation principle (Theorem 4.1), we can assume that each player in the *Resistance* team truthfully communicates their information, while each *Spies* player may decide correlate with the others to emit a specific fake claim. During the game, the mediator recommends an action to each player anytime it is their move. In this case, the revelation principle allows to assume WLOG that *Resistance* players always obey the recommendations received from the mediator, while *Spies* can decide to deviate. Overall we obtain a two-player zero-sum game in which the *Resistance* team is represented by a mediator and the *Spies* team is a single player.

In the following, we list the abstractions we applied to our Avalon instances, and for each we sketch a proof of correctness. Each proof will have the same structure: we will impose a restriction on the strategy spaces of both teams, resulting in a smaller zero-sum game; we will then show that any equilibrium of this smaller zero-sum game cannot have a profitable deviation in the full game.

1. *The recommendations emitted by the mediator to the player in charge of the mission proposal should never be rejected. That is, the mediator has unilateral power to dictate who goes on missions.*

   This is Lemma E.3.

2. *Call the following information* common knowledge*:*

   - *The sets of good players claimed by players claiming to be Merlin, but* not *the identities of the claimants themselves (because* **MIN** *does not know the true Merlin)*
   - *Mission proposals and results*
   - *Any claimant whose claim is disproven by common-knowledge information is a* Spy

   Call a subset $S \subseteq [n]$ plausible *if, based on common-knowledge information, it is possible that S contains only good players. Every mission proposal should be plausible.*

   Proof sketch. Make the following restrictions to strategy spaces.

   - Every mission proposal by the mediator is plausible.
   - If the mediator proposes an implausible set, the *Spies* pick a plausible set at random, announce it publicly, and act as if the mediator proposed that set instead.

   Consider any Nash equilibrium of the restricted game. We claim that it also must be a Nash equilibrium of the full game.

- The mediator cannot improve, because *Spies* will pretend that the mediator played a plausible set anyway, and the mediator can simulate how the spies will do this.

- *Spies* cannot improve, because against a mediator who always proposes a plausible set, the strategy space of the spies is not limited.

3. *If any player is contained in* every *plausible set S, then that player should always be included on missions if possible. We call such players* safe.

   Proof sketch. Make the following restrictions to strategy spaces.

   - The mediator always sends all *safe* players on missions, if possible.

   - *Spies* pretend that all *safe* players are always sent on missions. More precisely, if i is safe, not sent on a mission, and at least one unsafe player j is sent on that mission, *Spies* pick such a pair (i, j) at random, announce both i and j publicly, and pretend that i was sent in place of j. (If there are multiple such replacements that can be done, this process is simply repeated.)

   Consider any Nash equilibrium of the restricted game. We claim that it also must be a Nash equilibrium of the full game.

   - The mediator cannot improve, because *spies* will pretend that all safe players are always sent regardless of how the mediator actually proposes missions, and the mediator can simulate how *spies* will perform the replacement.

   - *Spies* cannot improve, because against a mediator who always sends all safe players on missions, the strategy space of the spies is not limited.

4. *If a mission of size s passes, all missions will pass until the next mission whose size is $> s$.* (In particular, any maximum-size mission passing implies that all remaining missions pass.)

   Proof sketch. Suppose that a mission M_0 of size s_0 has passed, and suppose the next t missions have sizes $s_1, \ldots, s_t \leq s_0$. Make the following restrictions to strategy spaces.

   - The mediator randomly[19] selects subsets $M'_i \subseteq M_0$ to send on missions s_i for $i = 1, \ldots, t$. If any of them (say, M_i) fails, the mediator pretends that M_0 failed instead. Then, the mediator generates and publicly announces the missions $M'_1, \ldots, M'_i$ that it *would have* proposed had M_0 failed, assuming that each M'_j passes. The mediator then pretends that that is what happened. The mediator does not use the M_is to inform future decisions.

   - For each $i = 1, \ldots, t$, *Spies* automatically pass M_i and ignore what missions are proposed by the mediator.

   Consider any Nash equilibrium of the restricted game. We claim that it also must be a Nash equilibrium of the full game.

   - The mediator has no incentive to propose a mission different from a random $M_i \subseteq M_0$ in the unrestricted game, because against spies who ignore and always pass missions $M_1, \ldots, t$, any mediator mission would lead to the same outcome of having a mission passing, no change in behavior from *Spies*, and no information gained.

   - The *Spies* have no incentive to fail mission M_i for $i > 0$ if they pass mission M_0. This is true because the mediator strategy is such that *Spies* passing missions $M_0, \ldots, M_{i-1}$ and failing mission M_i would be equivalent to *Spies* failing M_0 and passing missions $M'_1, \ldots M'_i$.

[19] The fact that this is *random* instead of dictated by the mediator allows us to avoid issues of imperfect recall.

Christiano [2018] used a very strong reduction that allowed the assumption that the smallest two missions are guaranteed to pass. The reduction is sound in the base game *Resistance*, as well as when the mediator's strategy is manually constrained to *never* reveal information about Merlin, as is the case in Christiano's analysis. It turns out, however, that this reduction is *not* sound in general, because its proof assumes that all mission proposals—even those that occur after *Resistance* has already reached the required number of missions—are public information. However, in our setting, it is possible that each mission proposal leaks more and more information about Merlin, so there is incentive to reveal as few missions as possible. Indeed, in the 6-player variants with Mordred, we observe that removing the two smallest missions causes a small but nonzero change in the game value, whether or not future mission proposals are published. However, a more refined analysis can be used to partially recover a reduction in the number of missions.

With five players, the refined analysis allows the same result Christiano [2018]: the first two missions should automatically pass.

In 5-player *Avalon*, the mission sizes are 2, 2, 3, 3, 3 in that order.

Proposition E.4. *In 5-player Avalon with private communication, the hidden-role value is the same if the two missions of size 2 are automatically passed (hence skipped).*

Proof Sketch. Make the following restrictions to strategy spaces.

- The mediator privately picks three missions M_1, M_2, M_3 of size 3. Then the mediator proposes M_1 (or a subset thereof) until it fails, then M_2 or a subset thereof until that fails, and finally M_3 until that fails.

- *Spies* ignore and automatically pass the first two mission proposals, and fail the last three missions if possible.

Consider any Nash equilibrium of the restricted game. We claim that it also must be a Nash equilibrium of the full game.

- The mediator has no incentive to behave differently, because *Spies* are ignoring the first two missions regardless.

- *Spies* win if either three missions are failed, or they guess Merlin. But if at least one mediator guess is the true set of good players, only two missions can fail, and so *Spies* cannot increase their probability of failing three missions. As for guessing Merlin, *Spies* cannot increase the amount of information gained. Let i be the index such that M_i is the true good set, and $i = 4$ if all mediator guesses were wrong. Spies are currently gaining the information $M_{\leq i}$, and Spies cannot gain more information from deviating because the mediator always plays the missions in order.

This completes the proof. □

Our analysis is refined compared to Christiano [2018] in that we must consider the possibility that *Spies* can gain more information about Merlin by deviating from the restricted strategy space. With five players, this turns out not to have an effect, but with six players, it will. With six players, the refined analysis only allows the first mission to be removed.

In six-player Avalon, the mission sizes are 2, 3, 4, 3, 4. The problem in adapting the analysis of the previous result is that the final size-3 mission comes *between* the size-4 missions, preventing its removal.

Proposition E.5. *In 6-player Avalon with private communication, the hidden-role value is the same if the mission of size 2 is automatically passed (hence skipped).*

Proof Sketch. Consider any strategy x for the mediator in the restricted game in which the first mission is ignored. We will extend x to a full-game strategy x'. Strategy x' works as follows. The mediator maintains a simulator of strategy x.

1. The mediator queries the x-simulator for its second mission M_2 (of size 3), and plays a random size-2 subset of M_2. If this mission passes, the mediator plays M_2 itself on the second mission, then plays the rest of the game according to x.

2. If the first mission fails, the mediator tells the x-simulator that M_2 failed. The x-simulator will then output its third mission M_3 (of size 3). The mediator picks a random size-3 subset of M_3 to send on the second mission. If this mission passes, the mediator plays M_3 again on the third mission, and once again plays the rest of the game according to x.

3. If the second mission fails, the mediator tells the x-simulator that M_3 failed. The x-simulator will then output its fourth mission M_4 (of size 3). The mediator immediately tells the x-simulator that M_4 passed. The x-simulator will then output its fifth mission M_5, of size 4. The mediator plays random subsets of M_5 until the game ends.

We claim that, against this mediator, the adversary cannot do better by failing the first mission than by passing it.

- If the adversary fails the first mission and passes the second, then the adversary will have, in the first three missions, observed a random subset of M_2 and the whole set M_3. But the adversary could have accomplished more by passing the first mission and failing M_2—the behavior of the mediator would be the same in both cases (by construction of the mediator), and the adversary would observe the whole set M_2 instead of merely a random subset.

- If the adversary fails the first two missions, the adversary (regardless of what else happens in the game) will have observed a random subset of M_2, a random subset of M_3, and M_5 by the end of the game. Further, three missions will pass if and only if M_5 is the true good set. The adversary, however, could have accomplished more by passing the first mission, failing the second and third (M_2 and M_3), and passing the fourth—the adversary induces the same outcome in the game, but instead observes all four missions M_2, M_3, M_4, M_5.

Thus, the adversary is always better off failing the first mission, and therefore x and x' have the same value. □

E.3 A Description of *Avalon* in Reduced Representation

The optimized *Avalon* game instances can be succinctly represented as follows:

1. At the start of the game, Chance player deals the roles.

2. *[if Merlin is present in the game]* Merlin reports who the *Spies* are (except *Mordred*). Spies may also (falsely) do the same.

3. *[if Percival is present in the game]* Percival reports who Merlin is. *Spies* may also falsely do the same.

4. Until the number of mission passed is less than 3, the mediator sends a mission, and the *Spies* are offered the option to fail it if any of them is on it. The space of possible missions to be proposed is constrained according to the previous optimizations

5. If the number of missions failed is 3, the game ends with a win for the *Spies*.

6. If the number of mission won is 3 and Merlin is present, the *Spies* have to guess the player with the Merlin role. If they guess correctly, they win, otherwise they lose. If Merlin is absent, the game ends with a win for the *Resistance*.

Variant	5 Players			6 Players		
	$\lvert Z\rvert$	RW	MG	$\lvert Z\rvert$	RW	MG
No special roles (*Resistance*)	5.9×10^3	0.3000	n/a	1.4×10^6	0.3333	n/a
Merlin only	1.5×10^4	1.0000	0.3333	6.0×10^5	1.0000	0.2500
Merlin + Mordred	4.9×10^5	0.6691	0.3869	1.8×10^8	0.7529	0.3046
Merlin + 2 Mordreds	9.0×10^4	0.5000	0.4444	2.8×10^7	0.4088	0.2886
Merlin+Mordred+Percival+Morgana	2.4×10^6	0.9046	0.3829	unknown	unknown	unknown

Table 3: Breakdown of equilibrium outcome probabilities for each variant of *Avalon*. $|Z|$: number of terminal nodes in the reduced game tree. 'RW': probability of Resistance passing three missions. 'MG': probability of Spies guessing Merlin correctly, conditioned on Resistance passing three missions. (The game value, which appears in Table 2, is RW $\cdot$ (1 − MG)).

E.4 Example of Optimal Play in *Avalon*

In this section, we fully describe one of the equilibria claimed in Table 2: the case of 5 players with Merlin and both MIN players "Mordred" (hidden from Merlin). We choose this version because it is by far the easiest to describe the equilibrium: the other new values listed in Table 2 are very highly mixed and difficult to explain.

Similarly to what happens in the example provided in Section 1.4, this case is different from the case with no Merlin (even if Merlin does not have any useful knowledge), due to the effect of added correlation. In particular, if Merlin were not known to the mediator, then the equilibrium value would be $3/10 \cdot 2/3 = 2/10$: $3/10$ from the value of pure *Resistance* (no Merlin), and the factor of $2/3$ from a blind Merlin guess.

By the discussion above, it suffices to analyze the reduced game with three missions, each of size 3, where MAX need only win one mission to win the game.

The following strategy pair constitutes an equilibrium for this *Avalon* variant:

Strategy for MAX The mediator's strategy is different depending on how many players claim to it to be Merlin.

One claim. Randomly label the players A through E such that A is Merlin. Send missions ABC, ABD, and ABE.

There are six possible correct sets (ABC, ABD, ABE, ACD, ACE, ADE), and MAX has three guesses, so MAX wins the regular game with probability $1/2$. If the first mission succeeds, MIN learns nothing about Merlin, so Merlin is guessed correctly with probability $1/3$. Otherwise, MIN knows that Merlin is either A or B, so Merlin is guessed correctly with probability $1/2$. Thus, this strategy attains value $(1/6)(2/3) + (1/3)(1/2) = 5/18$.

Two claims. The non-claimants are guaranteed to be good. Pick a non-claimant at random, pretend that they are Merlin, and execute the strategy from the one-claim case. This does strictly better than the one-claim case, because MIN learns only information about who is *not* Merlin instead of who *is* Merlin. (This strategy may not be subgame-perfect, but for the sake of this analysis it is enough for this strategy to achieve value *at least* $5/18$).

Three claims. Randomly label the players A through E such that A and B are the Merlin claimers. Send missions ACD, BCE, and ADE.

The analysis of the one-claim case applies nearly verbatim: there are six possible correct sets, and the second mission narrows down MIN's list of possible Merlins from 3 to 2 (namely, A or D), so again the value for MIN is $5/18$.

Strategy for MIN Emulate the behavior of a non-Merlin MAX-player (*i.e.*, do not claim to be Merlin).

1. If the first guess by MAX is correct, guess Merlin at random from the good players.

2. If the second guess by **MAX** is correct, guess Merlin at random from the good players included on both guesses.

3. If the third guess by **MAX** is correct, guess Merlin at random among all good players sent on at least two missions, weighting any player sent on all three missions double. (for example, if there was one player sent on all three missions and one player sent on two, guess the former with probability 2/3 and the latter with probability 1/3).

Against this strategy, the mediator's only decision is what three distinct missions to send.

1. If Merlin is sent on *one* mission, the probability of winning is at most 1/6 because only that mission has a chance of being the correct one.

2. If Merlin is sent on *two* missions, then the probability of one mission passing is at most 1/3. If the first mission passes, Merlin is guessed correctly with probability 1/3. If the second mission passes, Merlin may not be guessed correctly at all (it is possible for Merlin to only have been sent on the second mission, but not the first). If the third mission passes, Merlin is guessed correctly with probability at least 1/5. Thus, the probability of winning for good is at most $(1/3)(1-(1/3+0+1/5)/3) = 37/135 < 5/18$.

3. If Merlin is sent on *all three* missions, the probability of a mission passing is at most 1/2. If the first mission passes, Merlin is guessed correctly with probability 1/3. Otherwise, Merlin is guessed correctly with probability at least 1/2, since the only way to decrease this probability under 1/2 would be for the last mission to pass *and* for another player to also have been sent on all three missions, but in that case it is impossible for anyone to have been sent on two missions. Therefore, the probability of winning for good is at most $(1/2)(1-(1/3+1/2+1/2)/3) = 5/18$.