

# Games with zero-knowledge signaling

**Abstract.** We observe that in certain two-player repeated games of incomplete information, where information may be incomplete on both sides, it is possible for an informed player to signal his status as an informed player to the other without revealing any information about the choice of chance. The key to obtaining such a class of games is to relax the assumption that the players' moves are observable. We show that in such cases players can achieve a kind of signaling that is “zero-knowledge”, in the sense that the other player becomes convinced that her opponent is informed without ever learning the choice of chance. Moreover, such “zero-knowledge signaling” has all of the statistical properties associated with zero-knowledge proofs in interactive protocols. In particular, under the general assumption that moves are unobservable, such signaling leads to a class of equilibria in repeated games that are *separating* in regard to the status of player 1—informed or uninformed—but only for player 2; any other player in a network, being unable to observe the moves of player 2, remains uncertain as to the status of player 1.

*Keywords:* Repeated games, zero-knowledge, incomplete information, unobservable moves.

## 1. Introduction

We are interested in understanding the conditions under which one player (agent) may convince another that he possesses information without having to reveal what that information is.

The topic arose in the study of cryptographic protocols where it was given the designation “zero-knowledge proof”. The motivation for establishing such a protocol was obvious enough from an operational standpoint: if one agent could convince another that he was in the possession of some sensitive information, he might like to do so and continue making use of the information, while still keeping it secret. In this way, he could benefit both from the information directly and from the leverage that comes from others knowing that he is privy to it. He could, for example, authenticate his identity over a channel (or server, perhaps) without having to reveal that identity.

Interactive proof systems and their algorithms were originally introduced by L. Babai [2] as a way to classify the computational complexity of various problems in group theory. The approach was motivated by the fact that certain well-known complexity classes, like the class **NP** of problems solvable by a non-deterministic Turing machine in polynomial time, could

be regarded as classes of problems decidable by a pair of interacting algorithms possessing different computational resources. For instance, the class **NP** is the class of problems for which purported solution “certificates”, or witnesses, can be checked in polynomial time. Formalizing this a bit, we can say the class is the set of strings  $x$  in a language  $\Sigma^*$  such that there exists a pair of algorithms  $(P, V)$ , where  $P$  has exponential time resources and  $V$  has polynomial time resources, that execute the following protocol: both receive the string  $x$  as input,  $P$  computes a string  $y$  such that  $|y| \leq p(x)$  for some polynomial  $p$ , and sends  $y$  to  $V$ .  $V$  then checks whether  $y = f(x)$  where  $f$  is some computable function. If  $y = f(x)$  then the system as a whole outputs ACCEPT, otherwise it outputs REJECT. Working with this definition, one observes that, as an interactive process, the system is all too special. Certain features need not be so restrictive. For instance, the algorithms could be probabilistic rather than deterministic, they could possess different time resources, and, furthermore, could engage in many rounds of communication rather than simply one. By identifying the model of an interactive system and generalizing in this way, complexity theorists were led to new class hierarchies.

It was soon noticed [3] that these proof systems had the so called zero-knowledge property. The notion arises when one considers that the Prover, given an input string  $x$ , need not produce a new string  $y$  deterministically but may do so in way that gives rise to a probability distribution on the language  $\Sigma^*$  (once  $x$  is fixed, otherwise one thinks of the output string as a random variable). Intuitively, a protocol executed by a pair  $(P, V)$  is zero-knowledge if the Verifier, while ultimately either accepting or rejecting, still learns nothing from the string provided by the Prover. Formally, therefore, we say that a protocol executed by  $(P, V)$  is zero-knowledge if for any  $V'$  there exists a machine  $M$  with polynomial expected running time, such that  $M$  produces the same probability distribution as the one produced by  $P$  in executing the protocol on a system  $(P, V')$ . That is, the Verifier learns nothing by interacting with the Prover that she could not have learned on her own, with her own modest resources.

It is tempting to search for such a phenomenon within the study of game theory, and in particular within the subject of games of incomplete information. As a subject, it has much in common with the study of cryptographic protocols—both are concerned with strategic interaction between agents who, lacking critical information about the other player(s), must act so as to avoid being taken advantage of and to secure their own interests. The main difference between the two subjects is in the modeling resources that each employs. In cryptography the focus is on the resources of compu-

tation and communication available to an agent, whereas in game theory one only makes assumptions about an agent's psychology (that she is rational) and knowledge (which can be limited in many different ways).

Without being able to make explicit assumptions about the availability of communication and computational resources, the barrier to realizing the phenomenon of zero-knowledge interactive proofs within game theory seems considerable. As with all cryptographic results, zero-knowledge interactive proofs seem to rely on one-way operations, operations that are themselves easily carried out but for which the inverse operation is difficult. A mundane example of a one-way operation is dropping a letter into a side-walk mailbox: it is easy to drop the letter in, but difficult to get it out (unless one has the key, of course). The one-way operations used in cryptography are mathematical, and rely on both facts about the nature of mathematical objects and the current state of human knowledge about those objects. For example, the operation of multiplying two large primes is a familiar case of an operation believed to be one-way. Similarly, the problem of determining whether two graphs are isomorphic is not known to be in the class  $\mathbf{P}$ , the class of problems decidable in polynomial time (and the paradigm of tractability), though it is in the class  $\mathbf{NP}$ , and in fact the problem can be decided by an interactive proof system with the zero-knowledge property. The "one-wayness" of an operation is what allows for credible certificates of honesty, as when the Verifier checks  $y = f(x)$ , i.e., whether the certificate  $y$  matches the result of her independent computation  $f(x)$ .

Insofar as it does not offer players the resources to encode and decode mathematical objects as strings and to transmit them to one another, game theory would seem ill-equipped to provide players with one-way operations, and so with the means for establishing credible certificates of the authenticity of information. It would be mistaken, however, to think that the only way for players in a game to engage in a zero-knowledge proof is to assume that the players have explicit material means for communication. Indeed, due to the fact that game theory works directly with the knowledge states (or *information* states) of its players, zero-knowledge proofs can be realized in a much "purer" way, without relying on mathematics or computation, but only on the structure and features of the game itself. Indeed, the phenomenon arises when one considers repeated games of incomplete information, where in addition to one or more players being uninformed about which game is being played, players are unable to observe each other's moves.

## 2. Repeated games of incomplete information

The study of repeated games of incomplete information was initiated in the 1960's by Aumann, Maschler and Stearns [1] under a commission of the U.S. Arms Control and Disarmament Agency. Their project was to undertake a game-theoretic study of *secret information*, i.e. information possessed by an agent that is not even known to be possessed. More specifically, their charge was to study the ways in which such information can be revealed, and the conditions under which it is (or is not) advantageous to do so. The U.S. government was concerned that during arms negotiations with the Soviet Union they might inadvertently reveal information about their weapons stockpile. Alternatively, they wished to know when it would be advantageous to reveal such information. That the authors turned to the study of repeated games of incomplete information was a result of their effort to find a model with features that were closer to the reality of arms negotiations than those found in classical game theory. Games in reality tend to be repeated rather than one-shot, and often players are ignorant about the payoffs associated with a particular course of action. If the U.S. could reach an agreement with Russia that obligated the Russians to dismantle 100 warheads, the associated payoff for the U.S. would depend upon what fraction of Russia's total weapons supply 100 warheads constituted—information that the U.S. was likely to lack.

The authors were primarily interested in two-player games, again due to the cold war context of the study. They maintained the assumption that games were positive sum, however, in order to make use of the classical minimax theorem. Under such a model, the authors obtained the theorem that the question of whether an informed player should reveal or conceal his information exactly depends on whether his payoff function is a convex or, respectively, concave function of the initial probability that the actual state (or his type) is, say, the first of two possibilities\*.

The work presented here is intended to be continuous with the original AMS study and share the same goal, broadly speaking, of understanding how secret information can be managed. AMS discovered how secret information can be indirectly revealed to an uninformed player. Our study extends these results by showing how the secret can be shared without the information being revealed.

If the project of AMS was to establish the indirect ways in which informa-

\*See Theorem 3.2 of [1], as well as the generalization of the theorem to the case of unobservable moves in II.5.4.

tion can be revealed, our goal is to establish the ways in which the *possession of information* can be demonstrated, *without* revealing the choice of chance. That is, we show how secret information (information held by one player that is not even known to be held by others) can cease to be secret, and at the same time remain concealed. This becomes possible if we represent stage games in extended form, so that only the terminal reached is reported to the players, preventing the uninformed player from knowing player 1's move in general.

We study a class of repeated games where player 1 is either informed or uninformed about which game is being played (the state variable, the choice of chance in Harsanyi's treatment, etc.), and player 2 believes him to be informed with some initial probability  $0 < q_0 < 1$  (player 1 of course knows whether or not he is informed). Furthermore, we represent stage games in extended form so that moves are not observable in general. Rather, at the end of each stage game players are told which terminal was reached, but not in which game (i.e. players are told the information set to which the terminal node belongs, but sets need not be singletons.) The possibility of the players playing a set of strategies leading to a history that constitutes a zero-knowledge proof of player 1's status immediately presents itself. Indeed, in certain games such strategies are an equilibrium where  $\lim_{n \rightarrow \infty} q_n = 1$ , and yet player 2's conditional probability on the choice of chance remains unaltered.

## 2.1. The Model

We now present the model, which is just AMS' model for repeated games of incomplete information, with incomplete information on one side and with incomplete knowledge of moves.

As in any constant-sum two-person repeated game of incomplete information, we assume that two players, Player 1 and Player 2, are playing a single game (called the 'stage game') over and over again for a potentially infinite number of rounds. Player 1 may or may not know the payoffs that result from the choices of strategies and Player 2 is certainly ignorant of them. Player 2 is then aware only that she is playing one of the stage games  $\{G_1, \dots, G_k\}$ . To simplify our presentation, we will assume that there are but two possibilities  $G_1$  or  $G_2$ , though all results extend in a straightforward way to the more general case. Player 2 attaches a probability distribution  $p^0 = (p_1^0, p_2^0)$  to the two alternatives  $G_1, G_2$ . Each stage is given as an  $m \times n$  matrix whose entries are known to both players. At the end of each round, players receive a payoff based on their actions in the round and also on the

“state variable” or state of nature, i.e. the fact about which of  $G_1, G_2$  is the true stage game. Neither player is informed of his/her payoff, however. We imagine that payoffs are “credited” to a bank account that players are not able to check while playing the stage games.

If Player 1 is an informed player, then, knowing which is the true stage game and also his own move, he will know the move of player 2. Otherwise he will be in the situation of Player 2, who at the end of each round knows her own move and partial information about the move of her oppoent—partial because she is told only which information set the outcome is element of (as is the case for Player 1, though if he knows the true stage game this is often enough to infer the move of Player 2). Depending on the moves available to the players, we have information represented by two  $m \times n$  matrices with entries  $a_{i,j}$  and  $b_{i,j}$  taken from a set  $\{a, b, c, \dots\}$ , such that two cells contain the same letter only if they belong to the same column either in the same or in the two matrices, i.e.

$$\begin{aligned} a_{i_1, j_1} = a_{i_2, j_2} &\Rightarrow j_1 = j_2, \\ b_{i_1, j_1} = b_{i_2, j_2} &\Rightarrow j_1 = j_2, \\ a_{i_1, j_1} = b_{i_2, j_2} &\Rightarrow j_1 = j_2. \end{aligned}$$

We wish to study games in which Player 1 is only potentially informed about which is the true stage game. However, unless otherwise noted, we specify elements of the model as if Player 1 is informed and player 2 is not. Of course, if Player 1 is also uninformed, then his history, strategy, expected distribution, etc., are all exactly as described for Player 2 (with the obvious symmetric changes).

A *history*  $h$  for Player 2 is a sequence  $h = (\alpha_1, \alpha_2, \dots)$ , where  $\alpha_j \in \{a, b, c, \dots\}$ , such that for each  $j \geq 1$ ,  $\alpha_j$  is the information set to which the node reached in stage  $j$  belongs. The *k-stage history*  $h_k = (\alpha_1, \dots, \alpha_k)$  is the truncation of  $h$  to the first  $k$  elements.  $h_k$  can be thought of as the history of signals<sup>†</sup> Player 2 has received about the nodes reached (in the extended form representation of the stage game) through  $k$  rounds of play; the sequence therefore represents her partial knowledge of the moves of Player 1.

A history  $e$  for (an informed) Player 1 is a sequence of pairs  $e = ((\mu_1, \nu_1), (\mu_2, \nu_2), \dots)$ , with  $\mu_i \in \{1, \dots, m\}$  and  $\nu_i \in \{1, \dots, n\}$ , such that for each

<sup>†</sup>We use the term ‘signal’ to mean any of three different things, though which should be clear from context. A signal to a player can be information about the other player’s move, the choice of chance (i.e. the true stage game), or the other player’s status as either informed or uninformed.

$i \geq 1$ ,  $\mu_i$  is the row chosen by Player 1 and  $\nu_i$  is the column chosen by Player 2 in stage  $i$ . A  $k$ -stage history  $e_k = ((\mu_1, \nu_1), \dots, (\mu_k, \nu_k))$  is the truncation of  $e$  to the first  $k$  stages.

A  $k$ -stage history  $h_k$  for Player 2 determines sets of move sequences for Player 1 through  $k$  stages,  $M_1(h_k)$  and  $M_2(h_k)$ , that are consistent with the sequence  $h_k$ . If  $h_k = (\alpha_1, \dots, \alpha_k)$ , then

$$\begin{aligned} M_1(h_k) &= \{(\mu_1, \dots, \mu_k) : a_{\mu_1, \nu_1} = \alpha_1, \dots, a_{\mu_k, \nu_k} = \alpha_k\} \\ M_2(h_k) &= \{(\mu_1, \dots, \mu_k) : b_{\mu_1, \nu_1} = \alpha_1, \dots, b_{\mu_k, \nu_k} = \alpha_k\} \end{aligned}$$

Thus  $M_1(h_k)$ , respectively  $M_2(h_k)$ , is the set of moves that could have been taken by Player 1 in the first  $k$  stages if the true stage game were  $G_1$ , respectively  $G_2$ , given that Player 2 has received the sequence of signals  $h_k$ .

A *strategy* for Player 2 is a sequence  $(\lambda^1(h_0), \lambda^2(h_1), \dots)$ , where  $\lambda^k(h_{k-1})$  is a probability distribution over the columns  $\{1, \dots, n\}$  as a function of the history  $h_{k-1}$  leading up to stage  $k$ . A strategy for Player 1 is a *pair* of sequences  $(\sigma^1(e_0), \sigma^2(e_1), \dots)$ ,  $(\tau^1(e_0), \tau^2(e_1), \dots)$ <sup>‡</sup>, where  $\sigma^k(e_{k-1})$  and  $\tau^k(e_{k-1})$  are probability distributions over  $\{1, \dots, m\}$  as a function of the  $k-1$ -stage history  $e_{k-1}$ , with the interpretation that Player 1 plays  $\sigma^k(e_{k-1})$  ( $\tau^k(e_{k-1})$ ) if the true stage game is  $G_1$  ( $G_2$ ).

Even if Player 2 knows the strategy  $(\sigma, \tau) = ((\sigma^1, \sigma^2, \dots), (\tau^1, \tau^2, \dots))$  of Player 1, as is ordinarily assumed in two-player repeated games of incomplete information, Player 2 still cannot compute a probability distribution over Player 1's moves  $\{1, \dots, m\}$  in stage  $k$ , for any  $k \geq 2$ . This is because, for all  $k \geq 2$ ,  $(\sigma^k(e_{k-1}), \tau^k(e_{k-1}))$  depend on  $e_{k-1}$ , which is unknown to Player 2. Player 1 only uses the probability distribution  $(\sigma^k(e_{k-1}), \tau^k(e_{k-1}))$  over moves  $\{1, \dots, m\}$  if the history of moves up to stage  $k$  is  $e_{k-1}$ . Instead, Player 2 can only compute the *a priori* expected distributions  $s^k, t^k$  given her partial information of the preceding moves  $h_{k-1}$ . These are defined as

$$\begin{aligned} s^k &= \frac{\sum_{\mu^{k-1} \in M_1(h_{k-1})} Pr(\mu^{k-1}, \nu^{k-1}) \sigma^{k-1}(\mu^{k-1}, \nu^{k-1})}{\sum_{\mu^{k-1} \in M_1(h_{k-1})} Pr(\mu^{k-1}, \nu^{k-1})} \\ t^k &= \frac{\sum_{\mu^{k-1} \in M_2(h_{k-1})} Pr(\mu^{k-1}, \nu^{k-1}) \sigma^{k-1}(\mu^{k-1}, \nu^{k-1})}{\sum_{\mu^{k-1} \in M_2(h_{k-1})} Pr(\mu^{k-1}, \nu^{k-1})}, \end{aligned}$$

where  $\mu^{k-1} = (\mu^1, \dots, \mu^{k-1})$  and  $\nu^{k-1} = (\nu^1, \dots, \nu^{k-1})$ . Notice that since  $h_{k-1}$  completely determines  $\nu_{k-1}$ ,  $(s^k, t^k)$  is a function of  $h_{k-1}$  only. For <sup>‡</sup> $h_0$  and  $e_0$  are the histories at the start of the first stage of play, and so are empty sequences.

Player 2, knowing  $(s, t) = ((s^1(h_0), \dots), (t^1(h_0), \dots))$  is as good as knowing  $(\sigma, \tau)$ . Both lead to the same probability distributions on possible sequences  $h^k = (\alpha_1, \dots, \alpha_k)$ , for all  $k$ , and thus both give rise to the same expected payoffs for Player 2, for some given strategy  $\lambda$ .

For strategies  $(s, t)$  and  $\lambda$ , the conditional probability of a history  $h_k = (\alpha_1, \dots, \alpha_k)$ , given that the true stage game is  $G_1$  ( $G_2$ ), is  $w(h_k|G_1)$  ( $w(h_k|G_2)$ ), which is just the product of the individual probabilities of each member of the sequence  $(\alpha_1, \dots, \alpha_k)$  arising, i.e.,

$$w(h_k|G_1) = \left[ \sum_{\mu_1: a_{\mu_1, \nu_1} = \alpha_1} s_{\mu_1}^1(h_0) [\lambda_{\nu_1}^1(h_0)] \cdots \left[ \sum_{\mu_k: a_{\mu_k, \nu_k} = \alpha_k} s_{\mu_k}^k(h_{k-1}) [\lambda_{\nu_k}^k(h_{k-1})] \right] \right] \quad (1)$$

$$w(h_k|G_2) = \left[ \sum_{\mu_1: a_{\mu_1, \nu_1} = \alpha_1} t_{\mu_1}^1(h_0) [\lambda_{\nu_1}^1(h_0)] \cdots \left[ \sum_{\mu_k: a_{\mu_k, \nu_k} = \alpha_k} t_{\mu_k}^k(h_{k-1}) [\lambda_{\nu_k}^k(h_{k-1})] \right] \right]. \quad (2)$$

Recall that  $p^0 = (p_1^0, p_2^0)$  is Player 2's prior probability on the choice of chance over  $\{G_1, G_2\}$ . It follows that her prior probability that a history  $h_k$  will take place is

$$w(h_k) = p_1^0 w(h_k|G_1) + p_2^0 w(h_k|G_2).$$

If she knows  $(s, t)$  and  $\lambda$ , and history  $h_k$  results through  $k$  stages of play, Player 2 can then compute the conditional probability distribution on the choice of chance  $(p_1^0(h_k), p_2^0(h_k))$ , where

$$p_1^0(h_k) = \frac{p_1^0(h_k) w(h_k|G_1)}{w(h_k)} \quad (3)$$

$$p_2^0(h_k) = \frac{p_2^0(h_k) w(h_k|G_2)}{w(h_k)} \quad (4)$$

**DEFINITION 2.1.** We say that a strategy  $(s, t)$  for Player 1 is *non-revealing* if for all  $h_k$  and all strategies  $\lambda$ ,  $p_1^0(h_k) = p_1^0$ .

Obviously, any strategy for Player 1 is non-revealing if  $p_1^0 = 0$  or 1. In all other cases we have the following characterization.

PROPOSITION 2.2. *If  $0 < p_1^0 < 1$ , then the following are equivalent:*

- (i) *A strategy  $(s, t)$  is non-revealing.*
- (ii)  *$w(h_k|G_1) = w(h_k|G_2)$ .*
- (iii)  *$\sum_{\mu: a_{\mu\nu}=\alpha} s_\mu^k(h_{k-1}) = \sum_{\mu: a_{\mu\nu}=\alpha} t_\mu^k(h_{k-1})$*

PROOF. (i)  $\Leftrightarrow$  (ii) follows immediately from (3). (ii)  $\Leftrightarrow$  (iii) follows immediately from (1) and (2).  $\blacksquare$

### 3. Zero-knowledge signaling

We now present the main results. We are interested in two-person repeated games of incomplete information with nonobservable moves as described in the previous section, with the provision that Player 1 may be informed or uninformed of the choice of chance.

We suppose that at the start of the first stage game, Player 2 is told that Player 1 is using one strategy  $(s, t)$  (as defined in Section 2) if he is an informed player, and another strategy  $\bar{\sigma}$  if he is uninformed. Here  $\bar{\sigma} = (\bar{\sigma}^1(g_0), \bar{\sigma}^2(g_1), \dots)$  is a sequence of strategies for each stage game, where for all  $k \geq 1$ ,  $\bar{\sigma}^k(g_{k-1}) = (\bar{\sigma}_1(g_{k-1}), \dots, \bar{\sigma}_m(g_{k-1}))$  is a probability distribution over the rows  $\{1, \dots, m\}$ , based on upon a history  $g_{k-1} = (\alpha'_1, \dots, \alpha'_{k-1})$ , where  $\alpha'_j \in \{a, b, c, \dots\}$  for all  $j \geq 1$ . Thus  $\bar{\sigma}$  is just a strategy for each stage game such as might be chosen by any uninformed player who receives only partial information about his opponent's moves. A strategy  $\bar{\sigma}$  leads in turn to a set

$$\bar{M}(h_k) = \{(\mu_1, \dots, \mu_k) : a_{\mu_1, \nu_1}, b_{\mu_1, \nu_1} = \alpha_1, \dots, a_{\mu_k, \nu_k}, b_{\mu_k, \nu_k} = \alpha_k\}$$

of possible moves taken by Player 1, given that Player 2 has received history  $h_k = (\alpha_1, \dots, \alpha_k)$ . We also have

$$\bar{N}(g_k) = \{(\nu_1, \dots, \nu_k) : a_{\mu_1, \nu_1}, b_{\mu_1, \nu_1} = \alpha'_1, \dots, a_{\mu_k, \nu_k}, b_{\mu_k, \nu_k} = \alpha'_k\},$$

the set of moves that could have been taken by Player 2, given that Player 1 has received the history  $g_k = (\alpha'_1, \dots, \alpha'_k)$ . This allows us to define a new expected distribution for Player 2 facing an uninformed Player 1.

For a strategy  $\bar{\sigma}$ , we then have a new expected distribution  $\bar{s}^k$ , given by

$$\bar{s}^k(h_k) = \frac{\sum_{\mu^{k-1} \in \bar{M}(h_{k-1}), \nu^{k-1} \in \bar{N}(g_{k-1})} Pr(\mu^{k-1}, \nu^{k-1}) \bar{\sigma}^k(g_{k-1})}{\sum_{\mu^{k-1} \in \bar{M}(h_{k-1}), \nu^{k-1} \in \bar{N}(g_{k-1})} Pr(\mu^{k-1}, \nu^{k-1})}$$

Finally, if (an uninformed) Player 1 plays a strategy  $\bar{\sigma}$ , giving rise to expected distribution  $\bar{s}$ , and Player 2 plays a strategy  $\lambda = (\lambda^1(h_0), \lambda^2(h_1), \dots)$ ,

we define the probability that history  $h_k$  results to be

$$\bar{w}(h_k) = \left[ \sum_{\mu_1: a_{\mu_1}, \nu_1 = \alpha_1} \bar{s}_{\mu_1}^1(h_0) \right] [\lambda_{\nu_1}^1(h_0)] \cdots \left[ \sum_{\mu_k: a_{\mu_k}, \nu_k = \alpha_k} \bar{s}_{\mu_k}^k(h_{k-1}) \right] [\lambda_{\nu_k}^k(h_{k-1})]$$

Using these definitions and notation, we can describe Player 2's conditional probability distribution on Player 1's status.

Let  $q_I^0$  be Player 2's initial probability that Player 1 is informed, and  $q_U^0$  be the initial probability that Player 1 is uninformed (Player 1 of course knows whether or not he is informed). Then  $q_I^0(h_k)$ ,  $q_U^0(h_k)$  are Player 2's conditional probabilities on Player 1's status, given she has received the signal history  $h_k$ . They are defined to be

$$q_I^0(h_k) = \frac{q_I^0 w(h_k)}{q_I^0 w(h_k) + q_U^0 \bar{w}(h_k)}$$

$$q_U^0(h_k) = \frac{q_U^0 \bar{w}(h_k)}{q_I^0 w(h_k) + q_U^0 \bar{w}(h_k)}$$

DEFINITION 3.1. A game  $\Gamma_\infty(p^0)$  with equilibrium strategies  $(s, t)$  and  $\lambda$  constitutes a *proof* (that Player 1 is informed) if  $\lim_{k \rightarrow \infty} q_I^0(h_k) = 1$ .

DEFINITION 3.2. A game  $\Gamma_\infty(p^0)$  with strategies  $(s, t)$  and  $\lambda$  constitutes a *zero-knowledge proof* if it is a proof and  $(s, t)$  is a non-revealing strategy.

We can now state and prove the following theorem, which partly draws upon the original analysis<sup>§</sup> of AMS of a certain kind of game that can arise where players are unable to observe moves.

THEOREM 3.3. A game  $\Gamma_\infty(p^0)$  with stage games  $G_1$  and  $G_2$  of the form

$x$	$y$	$z$	$x$
$x$	$z$	$y$	$x$
$G_1$		$G_2$	

with  $y > z > x$ , has equilibrium strategies  $(\sigma, \tau)$  and  $\lambda$  that constitute a zero-knowledge proof.

<sup>§</sup>See II.5 of [1]. The authors take interest in a certain game as an example where their results are invalid. In analyzing the game, the authors state "In order to conceal information, Player 1 must make full use of the information". This is the essence of the idea behind a zero-knowledge proof: one demonstrates one's knowledge of something not by revealing it, but by *doing* something that only someone who had such knowledge could do (with high probability).

PROOF. Let  $\Gamma_\infty(p^0)$  be as above. Recall that Player 2 is told that Player 1 is playing a certain strategy  $(\sigma, \tau)$  if he is informed and another strategy  $\bar{\sigma}$  if he is uninformed. Suppose Player 1 is informed, and consider the strategy  $(\sigma, \tau) = (((1, 0), (1, 0), \dots), ((0, 1), (0, 1), \dots))$ , i.e., “always play row 1 if  $G_1$  is the true stage game, and always play row 2 if  $G_2$  is the true stage game”. Now consider the resulting conditional probability distribution for Player 2 on Player 1’s status.

Without any special information about Player 1’s status, at the beginning of the first stage game the initial probability  $q_I^0 = 1/2$ . The game  $\Gamma_\infty(p^0)$  with stage games  $G_1, G_2$  of the form stated, has outcomes that are partitioned into the following information sets

$a$	$b$	$a$	$c$
$a$	$c$	$b$	$c$
$I_1$		$I_2$	

where  $I_j$  represents the sets that partition the outcomes in the stage game  $G_j$ ,  $j = 1, 2$ . Suppose that for  $k \geq 1$ ,  $h_k = (\alpha_1, \dots, \alpha_k)$ . If  $\alpha_j = b$  for any  $j \leq k$  then  $p_1^0(h_k) = 0$  or 1. That is, if node  $b$  is ever reached then Player 1 will immediately know the choice of chance:  $G_1$  if Player 2 chose column 2,  $G_2$  if she chose column 1. Thus suppose for  $h_k = (\alpha_1, \dots, \alpha_k)$ ,  $\alpha_j \neq b$  for all  $j \leq k$ . For such an  $h_k$  and strategy  $(\sigma, \tau)$  as given above,  $w(h_k|G_1) = w(h_k|G_2) = w(h_k) = 1$ . So  $(\sigma, \tau)$  is the unique non-revealing strategy.

Additionally, for  $h_k = (\alpha_1, \dots, \alpha_k)$  such that  $\alpha_j \neq b$  for all  $j \leq k$ , such as obtains for the strategy  $(\sigma, \tau)$  mentioned above,  $q_U^0(h_k) = \frac{(\frac{2}{3})^k}{1+(\frac{2}{3})^k}$ , and hence  $\lim_{k \rightarrow \infty} q_I = \lim_{k \rightarrow \infty} 1 - \frac{(\frac{2}{3})^k}{1+(\frac{2}{3})^k} = 1$ . Thus  $(\sigma, \tau)$  is a zero-knowledge proof.

Finally, notice that since  $y > z > x$ , and since  $(\sigma, \tau)$  is the only non-revealing strategy, it is also an optimal strategy for Player 1, because if Player 2 knew the choice of chance she could guarantee herself a mere loss of  $x$ , and so Player 1 would only get a gain of  $x$  in each stage game.  $\square$   $\blacksquare$

#### 4. Conclusion

The phenomenon of zero-knowledge signaling depends first upon the fact that games with unobservable moves allow for the possibility of a unique non-revealing strategy on the part of an informed player 1. The class of games with payoffs structured in the way specified by the theorem is precisely

this class. When the unique non-revealing strategy for an informed player 1 coincides with the player's optimal strategy, the successful execution of an optimal strategy effectively separates player 1 as an informed player. This is because the only way to guarantee that player 2 will not learn the true stage game, and hence guarantee the optimal strategy's expected payoff, is by *using* the information about which is the true stage game. The probability that an uninformed player could persist in the execution of a non-revealing strategy becomes vanishingly small as the number of rounds increases.

With straightforward modifications, the result holds good for games of indeterminate but finite length. It is interesting to consider the implications of these ideas in the setting of social networks. Consider a population of players such that at the beginning of play the status of any player is unknown to all the other players. Suppose that at each increment of time these players pair up and engage in two-person repeated games of indefinite finite length. Then one player can only learn that the status of another player is 'informed' by directly playing against him. That is, there is no way for a third player to learn that the status of another player is informed if the other player is playing against another opponent. This is because we assume that moves are unobservable and so even if a third player observes the partial information afforded to the uninformed player, this information cannot act as a proof to her, since, unlike the players who are directly involved, she does not know the moves of *either* player. Each of the players directly involved, though potentially ignorant of the opponent's moves, at least knows her own moves and so can compute expected distributions on the choice of chance and on the status of her opponent. But a third player cannot determine the move sequence for either player based on a history of signals, and so can make none of the necessary computations. Furthermore, there is no way for one player to "report" the status of an opponent she has faced to some third party, at least not credibly. Without having "seen" the proof, there is no reason for one player to believe the testimony of a second player about the informed status of a third player no matter how convinced the second player is, because the second player herself will not necessarily know the choice of chance!

## References

- [1] AUMANN, R., M. MASCHLER, and R. STEARNS, *Repeated games of incomplete information*, MIT Press, 1995.
- [2] BABAI, L., 'Trading group theory for randomness' *17th ACM Symp. Theory of Computing* 421–429, 1985.

- [3] GOLDWASSER, S., S. MICALI, and C. RACKOFF, ‘The knowledge-complexity of interactive proof systems’, *SIAM Journal on Computing* 18:186–208, 1989.

EDWARD EPSEN  
Department of Philosophy  
University of Pennsylvania  
433 Logan Hall  
Philadelphia, PA, USA  
epsen@sas.upenn.edu