

# Cryptographic Commitment and Simultaneous Exchange

Peter Bardsley

Andrew Clausen

Vanessa Teague\*

December 2008

## Abstract

Simultaneous exchange of valuable secrets is problematic without use of a trusted intermediary. Examples include trade between strangers involving the transfer of property rights at a distance by exchange of emails, and simultaneous signing of contracts. There is a danger that the first person to commit to the transaction will lose everything if the other party does not reciprocate. We study information exchange mechanisms that establish credible simultaneous commitment and implement individually rational exchange without using an intermediary.

**Key Words** simultaneous exchange, trade in information, cryptographic cheap talk, incremental commitment

**JEL Classification** C72, D82, D83

Cryptographic techniques are pervasive in the implementation of economic transactions, mostly in forms that are invisible to users<sup>1</sup>. Cryptography has significant implications for how we model and design economic mechanisms. It allows agents to precommit to a message space—the set of messages that they can send; it allows them to make credible assertions about their knowledge without revealing anything about their private information beyond the validity of the assertion; and it allows mechanism designers to control off-equilibrium behaviour by imposing computational costs. Yet relatively little has been written on how this affects the way we model the behaviour of economic agents. The cryptographic literature does not really address incentives or equilibrium behaviour (Dodis & Rabin 2007, Halpern 2008). Cryptographic mechanisms are interesting to economists because they require that agents be computationally bounded, introducing a minimal degree of bounded rationality that fundamentally changes the institutions we can design.

---

\*Peter Bardsley and Vanessa Teague are at the University of Melbourne. Andrew Clausen began this project at the University of Melbourne before moving to the University of Pennsylvania.

<sup>1</sup>For example, all Amazon and eBay transactions use cryptographic communications.

In this paper we consider a specific issue: the implementation of voluntary exchange. Mutually beneficial exchange is a basic building block in the framework through which economists view the world. Yet simultaneous exchange is not entirely unproblematic. Unless the exchange is truly simultaneous, there is a risk that the person who first relinquishes control of an asset is left with nothing if the other person does not reciprocate. Achieving strategic simultaneity is itself problematic when exchange occurs at a distance, and communication is intrinsically asynchronous (for example, through exchange of emails).

Traders might turn to a trusted intermediary, either to carry out the transaction, or to monitor it and punish any infraction. Each party could give their object to the intermediary, safe in the knowledge it would be returned if the exchange fails. But the conditions that would sustain trust in such an intermediary are not straightforward. A useful step in understanding the nature of intermediaries is to understand what can be achieved without them. For this reason we focus on unintermediated trade.

In a message sending environment trade will amount to the exchange of information: secrets which embody ownership of real assets. Examples of such secrets include the number of a Swiss bank account; a digital signature to a legally binding contract; a digital signature to a document that transfers the legal title to an asset to another party; or a digital coin. So the asset trading problem reduces to a secret exchange problem. How general is the secret exchange problem that we address? Our secrets must not only be valuable, but must have an additional property. The value of the secret must be credible to a potential purchaser prior to the trade, before the secret is revealed. We will show how such secrets can arise naturally in a trading environment.

Exchange of secrets has been addressed in the cryptography literature. Blum (1983a) suggested that if secrets could credibly be made divisible, then pieces of the secrets could be exchanged alternately one by one. Divisibility can be achieved by encrypting each secret with a long digital key. Credibility of the pieces can be established through probabilistic proof protocols, an idea that we will describe below. If each key consists of, say, 100 digits, then these digits can be exchanged one by one. Whether this mechanism really resolves the exchange problem is however unclear, since the cryptographic literature on exchange does not address the incentives of agents. Damgård (1995), presenting a more robust protocol to implement Blum's proposal, writes<sup>2</sup> of the exchange problem:

If the two secrets are represented as bit strings of the same length, this can be solved by exchanging the secrets bit by bit; if this is done honestly, no party will be more than one bit ahead of the other; put another way: if at some point  $A$  can compute  $B$ 's secret in time  $T$ , then  $B$  can compute  $A$ 's secret in at most time  $2T$  by guessing the bit he may be missing.

---

<sup>2</sup>The quotation is verbatim, except that  $s_A$  and  $s_B$  were replaced with  $A$ 's secret and  $B$ 's secret, respectively.

This remark provides the starting point for our paper. We inquire whether exchange can be sustained as a subgame perfect equilibrium outcome of a mechanism of the Blum-Damgård type. We find that the Blum-Damgård intuition does not entirely hold up. Despite the divisibility introduced by the bit-by-bit gradual exchange protocol, there remains an intrinsic indivisibility: there is a moment when one of the players becomes irrevocably committed to the exchange, and this is crucial to understanding the equilibrium. In fact, the Blum-Damgård protocol fails for a range of parameters. Despite this negative finding, our overall conclusion is positive. Under weak assumptions we construct mechanisms that support individually rational exchange of secrets as an equilibrium outcome, without the use of any trusted intermediary. We find that if exchange can be supported then it can be implemented by a relatively simple three step protocol.

The structure of the paper is as follows. We discuss related literature in Section 1. In Section 2 we address the question of where do these valuable secrets come from? As mentioned above, we need that the secrets be not only valuable but that the value be credible without revealing the secret. We show that such credible secrets arise very naturally in a simple formalisation of asset exchange. In Section 3 we review the cryptographic primitives that we use throughout the rest of the paper. These primitives relate to the ability to commit to a message space, to establish credibility of certain assertions, and to manipulate computational costs off the equilibrium path. We also discuss our approach to modelling bounded computational ability.

In Section 4 we formalise the class of mechanisms that we study, and embark on our main results. The general approach is that agents exchange partial information by sending credible messages, and in so doing trade reductions in off-equilibrium costs, until full exchange has occurred. We approach this in three steps, weakening the assumptions at each stage. First, in 4.2, we set up an abstract version of the game in computational cost space (the grid game), assuming that agents can trade cost reductions directly. Then, in 4.3, we require that cost reductions must be implemented by sending messages (but at this stage we assume that messages are fully credible). We show how to map strategies, payoffs and equilibria in the grid game to strategies, payoffs and equilibria in the message game. Finally, in 4.4 we weaken the assumption that messages be credible, assuming only that they be  $\varepsilon$ -credible (a concept that we will define below). Then in 4.5 we construct a protocol that delivers the required  $\varepsilon$ -credibility. Conclusions are discussed in Section 5.

## 1 Related Work

In this paper we study a simultaneous commitment issue that lies at the heart of voluntary exchange. Our work is related to several literatures: on mediated and unmediated communication games, on persuasion games, on incremental commitment and incremental investment, and to the computer science literature on exchange protocols.

In the cheap talk literature Forges (1990), BenPorath (2003) and Gerardi (2004) ask when the equilibrium payoffs available in mediated and unmediated communication games coincide, assuming that there are three or more players. Credibility of communication is important but unmodelled in these papers. In contrast to this literature, we introduce a cryptographic communication technology that significantly enlarges what can be achieved through exchange of messages. As in the cheap talk literature, we inquire as to when the function of an intermediary can be replaced by unmediated communication, but we need only two players and we explicitly model the credibility of all messages using cryptographic concepts (Goldreich, Micali & Widgerson 1991). Since we study simple message passing games, in principle this opens the door to the study of credible communication between computationally bounded agents in pure cheap talk games, though we do not pursue this here. Our results also have potential application to the problem of achieving simultaneity of commitment in cheap talk models in general (Aumann & Hart 2003). More recently, Forges & Koessler (2005) study mediated communication with partially verifiable information, but leave the relationship with unmediated communication an open question.

In the literature on persuasion games (see Milgrom (1981), Matthews & Postlewaite (1985), Milgrom & Roberts (1986), Okuno-Fujiwara, Postlewaite & Suzumura (1990), Chen, Kartik & Sobel (2008)) agents play communication games with verifiable or certifiable information, but the nature and origin of these credible messages is unmodelled. We study communication in an environment where the credibility of messages arises endogenously, and where we can precisely characterise these credible messages. Ensuring sufficient credibility of certain messages will be an essential part of our mechanism.

Our work is related to the literature on incremental commitment that derives from Admati & Perry (1991) and Marx & Matthews (2000); see also Gale (2001), Lockwood & Thomas (2002) and Pitchford & Snyder (2004). In this literature incremental commitment emerges through gradual and irreversible investment of some divisible nature. Cost convexity and the existence of positive spillovers are typical in these games. The most closely related paper in this literature is Compte & Jehiel (2004), in which concessions by one party increase the other party's pay-off, and there is an inefficient outside option. They find that in a wide class of situations gradualism is a necessary feature of any non-trivial equilibrium. Our problem differs from those typically considered in this literature because of the fundamental non-convexity of knowledge: I either know something or I don't (on this non-convexity see Stiglitz (2000)). The need to overcome this indivisibility is the chief obstacle to trade, and it imposes constraints on when efficient exchange can be implemented. We find that gradualism per se is not the issue so much as being able to take credible steps of the right size.

The observation that cryptographic communication creates possibilities for unmediated trade in information is due to Blum (1983a), further developed by Damgård (1995) and Jakobsson (1995). The computer science literature on unmediated cryp-

tographic communication is extensively reviewed by Goldreich (2004, Chapter 7). Generally, in this literature agents are not rational in the sense understood by game theorists. They are modelled as either “honest” (meaning that they follow the assigned protocol perfectly) or “malicious” (meaning that they may perform any computationally feasible actions), and the analysis considers whether the malicious agents can cause the honest ones to produce the wrong output or reveal secret information. The relatively small literature on protocols for rational agents, drawing together game theory and computer science is reviewed by Dodis & Rabin (2007); see also Halpern (2008) and Izmalkov, Lepinski & Micali (2007). Within this literature the exchange problem is addressed by Syverson (1998), and further considered by Buttyán & Hubaux (2001) and Buttyán, Hubaux & Capkun (2002), but their analyses have some limitations. Syverson and Buttyán assume that communications can be designed to be fully credible (an assumption that we shall need to consider in detail), Syverson’s protocol relies on an externally imposed punishment for cheating, and none of these authors require the mechanism to be subgame perfect. We find that subgame perfection is essential to credible implementation in our environment. Sandholm’s analysis of unenforced e-commerce transactions (Sandholm (1996, 1997)) does implement a subgame perfect equilibrium. However, he considers only intrinsically divisible goods that can be divided up directly, and the issues of the divisibility of information and the credibility of partial information exchange do not arise.

## 2 The Trading Environment

In this Section we demonstrate that trade in secrets is sufficient to capture a reasonable model of trade in assets. We describe a simple trading environment<sup>3</sup>, in which ownership of an asset is associated with knowledge of a secret, in which the value of knowing a secret is credible even to someone who does not know it, and in which all transactions are simple message passing games. More complex trading environments can readily be simulated in this environment.

To fix ideas, let us first consider for a moment the market for real property (for example houses, cars, shares). An asset can be conceptualised as a bundle of property rights — a set of permissible actions that an agent (the owner of the property rights) can take, including the right to prevent actions by other agents and the right to transfer ownership. These rights are typically embodied in a physical legal document, a title deed, that specifies these rights. If an agent can prove that he or she is the owner of an asset, then they may exercise the rights embodied in the title deed. In the background, there is a legal system (the court) that gives value to assets by enforcing the rights that they embody. The minimal functions required of the legal system are

---

<sup>3</sup>In case the framework that we describe seems artificial, it is worth noting that several jurisdictions are currently working on the implementation of a digital conveyancing framework for land and real property that has most of the features that we set out here in abstract form.

1. To certify the validity and scope of an asset; that is to certify that it embodies a bundle of rights that the court will enforce, and to determine whether an action or set of actions is consistent with these property rights. For example, the court might determine whether a document is a forgery, or a parcel of land falls within the scope of a title deed.
2. To link assets to owners. For example an agent might establish such a link by proving to the court that they can produce a signature (a physical, ink signature on a piece of paper) that links them to the asset. This link might be via a chain of valid transfers of ownership going back to some original well accepted owner, or via a matching signature in a central database (an asset register) of certified owners.
3. To enable transfer of ownership of an asset from one agent to another.

In our environment, agents prove that they own an asset by proving knowledge of a hidden secret (a password or, more loosely, a digital signature<sup>4</sup>). Indeed, Goldreich (2004) points out the fundamental role of *proof of knowledge* in almost all authentication systems. This applies even to traditional ink signatures: you know how to produce your written signature, and it is unlikely that any else knows how to do so. More specifically, we assume:

1. There are one or more<sup>5</sup> asset registers. An asset register is a database of title deeds linked to secrets. A title deed is just an (unencrypted) digital document registered in an asset register, specifying a set of property rights. Secrets are published in an encrypted form, so that anyone who knows the secret can easily verify that it corresponds to the published secret, but people who do not know it cannot guess the secret without conducting infeasibly expensive computations. The asset register is public, and its entries can be examined and copied by anybody.
2. An agent owns an asset if they know the secret password associated with it.
3. Each asset register provides a safe and reliable method by which one agent may prove to any other agent (including the law court) that they are the owner of an asset. That is, if and only if they are the owner, they can prove to anybody that they know the secret password associated with the asset without actually revealing the password. Details of how this can be implemented using standard cryptographic techniques are discussed in Section 3.

---

<sup>4</sup>Our digital identities are very stripped back and simplistic. The requirements for real world digital signatures are somewhat more demanding.

<sup>5</sup>We allow the possibility that there may be several asset registers both for realism (assets may be in different jurisdictions) and because we do not want the asset register to act in some way as a trusted intermediary facilitating exchange.

4. Each asset register provides a safe and reliable method by which the owner of an asset can change the secret associated with an asset.
5. To facilitate transfers, an agent who owns an asset can lock the database for a brief fixed time period in favour of a new password known only by another agent. If the database is locked then the only transaction that is permitted is to execute the password change, which transfers the object to the other agent. If the time limit expires without a transfer occurring, the password is automatically updated to a fresh one that was generated beforehand (so the same agent still owns the asset, but has a new secret associated with it).
6. It is common knowledge that passwords generated by an asset register will be independently drawn, random, uniformly distributed integers between 1 and some large fixed integer  $p$ , which is different for each agent, and that a password is known only by the owner of the identity. Details of how this can be implemented are discussed in Section 3; it requires that the asset register has access to a random number generator such as a thermal noise device.

The last assumption is for convenience when we model the exchange game. It ensures that it is common knowledge that the secrets are independent and uniformly random when the exchange game starts.

We note that the role of the legal system is minimal. It plays no part in the exchange and exists in the background only to give value to the assets, and hence to the secrets associated with them. We also note that, unlike what we assume about assets, we do not need the passwords to be authenticated by a trusted third party. Nor do we need passwords to be permanently or transparently associated with agents. Knowing a secret is valuable because this knowledge can be used to prove one's identity to the court, which will then enforce the property rights set out in the title deed (this is the analogue of demonstrating that one can produce a matching pen and ink signature). The value of knowing the secret is credible to others because the property rights are set out in public in the title deed.

Exchange of assets can now be implemented purely by passing messages that exchange secrets. The first agent locks her asset in favour of the second agent (who has generated a new password known only to himself, for the purpose of this transaction). In a similar transaction the second agent locks his asset in favour of the first. They then exchange secrets: their original passwords (not the new passwords that they have just generated). They each now have the authority to unlink the other agent from their newly acquired asset, and the exchange is complete.

Given this framework, we consider protocols or mechanisms whereby agents may exchange messages through a secure but asynchronous communication system. We have in mind a secure email link where delivery of messages within a fixed finite time is reliable, but subject to a random delay with some known upper bound. It is thus difficult to achieve simultaneous action in such a system. We assume that there is no legal framework governing the messages that may be sent.

### 3 Cryptographic Communication

The exchange mechanism that we study in this paper relies on the existence of cryptographic techniques that allow a player to pre-commit to the messages that they will send at a later stage in the game, and to make credible statements about the nature of those messages before they are sent. In this section we explain how this can be done, and define some fundamental concepts that will be used throughout the paper.

#### 3.1 Committing to the message space: one way functions

The basic idea can be explained through an analogy. As is conventional in this literature, we call our protagonists Alice and Bob. Alice commits to a message by writing it on a piece of paper, putting it in a locked box or sealed envelope, and passing it to Bob. The nature of her commitment is that she cannot later change her message, or claim it to be different from what she committed to, since Bob can challenge her to give him the key that opens the box, and he can then read what she wrote. This technique allows the players to pre-commit to the *all* messages that they might choose to send in a communication game. They seal the possible messages in locked boxes, which they then exchange. To send a message, Alice gives Bob the key to one of the boxes. He opens the box and reads the message.

The most straight forward approach to this type of cryptographic commitment involves the use of *one way functions*. A function is one way if it can be computed efficiently but computing an inverse is computationally infeasible (Goldreich 2008). In order to be clear about what we are doing, we describe how Alice can commit to a message (lock it in a box) using little more than simple arithmetic. We will not need to use the precise details of the calculation later, but it will be difficult to understand the nature of our results or the notation that we will use without seeing how this simple example works.

Given a large prime  $p$ , let  $\mathbb{S} = \{1, 2, \dots, p-1\}$ . Then  $\mathbb{S}$  is a cyclic group under multiplication mod  $p$ . This means that there is an element  $g \in \mathbb{S}$ , a generator, such that the powers of  $g$  generate  $\mathbb{S}$ . Every element  $x \in \mathbb{S}$  is of the form  $x = g^n \bmod p$  for some unique  $n \in \mathbb{S}$ . The function  $f : \mathbb{S} \rightarrow \mathbb{S}$  defined by  $f(n) = g^n \bmod p$  (the *discrete exponential* to base  $g$ ) is commonly assumed by cryptographers to be a one way function<sup>6</sup>. The inverse, the *discrete logarithm*, is well defined (it can in principle be calculated by an exhaustive search over all the elements of  $\mathbb{S}$ ) but it is infeasible to compute if  $p$  is large enough. We choose a large fixed prime  $p$  that be held constant throughout this section. If  $p$  is chosen properly, it is easy to find a generator  $g$  (Menezes, van Oorschot & Vanstone 1997). We assume that such a generator has been fixed and is common knowledge.

---

<sup>6</sup>This conjecture, widely accepted in cryptography, is related to but distinct from the  $\mathcal{P} \neq \mathcal{NP}$  conjecture. See Koblitz (1994) for evidence of the computational intractability of the discrete logarithm problem.

We use the notation  $[x] = g^x$  for the application of the discrete exponential to  $x$ . The notation  $[x]$  is chosen to suggest that  $x$  has been placed in a locked box or sealed envelope. For example, if  $p = 11$  and  $g = 7$  then  $[8] = 9$  since  $7^8 \bmod 11 = 5764801 \bmod 11 = 9$ . The discrete exponential  $[x]$  can thus be used as a cryptographic commitment<sup>7</sup>. Here  $x$  serves as its own key. To decommit, or open a commitment  $X$ , Alice sends the value  $x$ ; Bob can then check that this is really what is inside the box by computing  $[x]$ , which is easy to do, and comparing with  $X$ .

The important properties of  $[x]$  are as follows.

- given  $x$ , it is very easy to calculate  $X = [x]$ ; however if  $p$  is large, given  $X$  it is very difficult to find the  $x$  such that  $X = [x]$ . In principle one can search for  $x$  over the whole search space  $\mathbb{S}$  but this is very expensive.
- if  $[x] = [y]$  then  $x = y \bmod p$
- $[x + y] = [x][y]$ ; this is just the law of exponents.

The discrete exponential is a simple version of a Schnorr commitment. It is perfectly binding, but only computationally hiding. This means that Alice cannot cheat when opening  $[x]$  (she cannot find  $y \neq x \bmod p$  such that  $[y] = [x]$ ); however the value  $x$  is hidden from Bob only because it is computationally infeasible for him to find it. It is useful to note that there is a commitment mechanism (the Pedersen commitment (Pedersen 1991)) with converse properties. It is perfectly hiding but only computationally binding. Alice can in principle cheat when opening the commitment (but to do so is computationally infeasible) but Bob cannot find  $x$  even with unbounded computational resources (and even if he knows some extra information about the value of  $x$ ). The Pedersen commitment, which we write  $[x; r]$  depends on an auxiliary parameter  $r$ . We need to use both types of commitment<sup>8</sup>.

### 3.2 Establishing credibility: zero knowledge proofs

The value of committing to a hidden potential message  $s$  is enhanced if Alice can make a credible assertion  $A(s)$  about the nature of the message  $s$  before it is sent. For example, she might assert that the message  $s$  contains information that will halve Bob's cost of searching for some secret that has been hidden. Clearly, one way to convince him of the truth of her assertion is to open the box and reveal  $s$ , but she might not want to do that. Another way is to provide him with a probabilistic proof.

---

<sup>7</sup>Any text string can be encoded as a number, so it is sufficient to be able to encrypt numbers.

<sup>8</sup>Computational difficulty is an asymptotic concept. The computational cost of an intractable problem rises very rapidly with the value of some parameter, known as the security parameter. In the case of the Schnorr commitment described above, the security parameter is  $\log_2 p$ . We will assume that the security parameters of all encryptions that we will use are set so that cracking them is infeasibly hard. The cost of guessing a Schnorr commitment, or of falsely opening a Pedersen commitment, is set higher than any possible payoff in the exchange game.

A *probabilistic proof* is a question-and-answer game, played between Alice and Bob, that Alice can reliably win if and only if her assertion is true. Bob asks Alice a series of tricky questions. If her assertion is indeed true then Alice can always answer his questions satisfactorily, but if not she will eventually be caught out, even though she might guess the right answer some of the time. An important property that we require of this proof game is that no other information, beyond the truth of her assertion, is revealed. Such a proof is called a *zero-knowledge proof* (zero-knowledge meaning zero leakage of knowledge). For precise definitions see (Goldreich 2008). We illustrate the concept of a zero knowledge probabilistic proof by describing how Alice can convince Bob that she knows what is inside a locked box. That is, she knows an  $s$  such that  $[s] = S$ .

Alice and Bob both know  $S$ . Alice claims that she knows  $s$  such that  $[s] = S$ . The zero knowledge proof is an interactive game, played in multiple rounds as follows.

### Protocol 1 (Proof of knowledge of logarithms)

1. Alice chooses a uniformly distributed random number  $x \in \mathbb{S}$  and reveals the commitment  $X = [x]$  to Bob.
2. Bob tosses a coin.
3. If Bob's coin comes up heads, then he challenges Alice to reveal a number  $x$  such that  $[x] = X$ . If she does so then the round ends; if Alice cannot meet the challenge then she must have cheated and Bob wins the game.
4. If Bob's coin comes up tails, he challenges Alice to reveal  $s + x$ ; that is to say, a number  $z$  such that  $[z] = SX = [s + x]$ . If she does so then the round ends; otherwise Bob wins the game.

The first branch of the protocol guarantees that she does not cheat and release a fake commitment  $X \neq [x]$ . The second branch guarantees that she knows  $s$ , since if she has not released a fake commitment  $X$  then she knows the  $x$  such that  $[x] = X$ , and she has just shown that she knows the  $z$  such that  $[z] = SX$ . But  $[z] = SX = [s][x] = [s + x]$  so  $z = s + x$ . Thus she must know  $s$ . If Alice is lying then she may try to guess which challenge Bob will make, and choose to fake  $X$  accordingly, but eventually she will be caught out. If enough rounds are played then the probability that she is just bluffing can be made as low as desired.

At each stage Bob learns only a random number  $x$ , or a random number  $s + x$  (but not both!). So what he sees is just a series of independently drawn random numbers. He can generate such a random series for himself, so he cannot learn anything from the transcript of the game apart from the fact that Alice seems always to be able to win. We note that this proof game has the following form. Alice provides convincing evidence not only that there exists an  $s$  with a particular property, but that she

knows such an  $s$ . An interactive proof of this kind is called a *zero-knowledge proof of knowledge* (see Goldreich (2001) for full details).

The knowledge of logarithms problem is not the only problem that has a zero-knowledge proof. The Bellare Goldreich Theorem (see Goldreich et al. (1991), Bellare & Goldreich (1992) and Goldreich (2001, Theorem 4.7.7)) provides a systematic way to generate zero knowledge proofs of knowledge. In general, Alice and Bob know  $x$ , and Alice wants to convince Bob that she knows a witness, that is a secret  $s$  with the property that  $(x, s) \in R$ , where  $R$  is some relation, without revealing what  $s$  is. In the discrete log example, where she wants to prove that she knows a logarithm of  $x$ , the relation  $R = \{(x, \log_g(x)) : x \in \mathbb{S}\}$  is just the graph of the logarithm function. The Bellare Goldreich Theorem requires that  $R$  be an  $\mathcal{NP}$  relation. This means that if Alice told Bob her secret  $s$ , then Bob could check it in polynomial time<sup>9</sup>. If  $R$  meets this criterion then, under standard intractability assumptions, the theorem claims that one can construct zero-knowledge proofs of knowledge for  $R$  with arbitrarily low bluffing probabilities  $\varepsilon > 0$ .

We use this result as follows. Let  $S = [s]$  be a hidden message to which Alice has committed. She wishes to make a credible assertion that the envelope  $S$  contains an  $s$  with property  $A(s)$ , without revealing  $s$ . By the Bellare Goldreich Theorem one can construct a zero knowledge proof of knowledge for the combined proposition: that  $S = [s]$  and that  $s$  has property  $A(s)$ , and she can use this interactive proof to convince Bob, to any required standard of proof  $\varepsilon$ , that she knows such an  $s$ . If the commitment  $S = [s]$  is computationally binding, then it is computationally infeasible for Alice to find an  $s' \neq s$  such that  $S = [s']$ , so Bob can be confident that if Alice opens the box for him it will contain an  $s$  with the property  $A(s)$ . We will use this and similar techniques repeatedly when Alice wishes to make credible assertions about the potential messages that she can send. The general theorem is constructive but not very practical. The zero knowledge proofs that we need can all be constructed by elaborations of the proof of knowledge of logarithms protocol, but we omit the details<sup>10</sup>.

It is now clear how Alice can manipulate Bob's expected costs. She hides her secret  $s$  in some large search set  $\mathbb{S}$ . She commits to a message  $M = [\mu]$  that contains information about  $s$  such that knowing  $\mu$  would reduce the size of the search space. Using a zero knowledge proof, she provides evidence that  $\mu$  has this property. If Bob receives message  $\mu$  (that is, if Alice opens the box for him) then he believes (or at least he has good evidence) that his search cost has been reduced. By choosing messages appropriately she can exercise very fine control over his expected search cost.

We can now be more specific about the asset database described in Section 2. Entries consist of pairs  $(d, S)$  where  $d$  is a document that will be regarded by the

---

<sup>9</sup>Formally,  $R$  is an  $\mathcal{NP}$  relation if there is a polynomial time Turing machine that determines whether its input  $(x, s)$  lies in  $R$ .

<sup>10</sup>Details are available in an earlier working paper, available at SSRN: <http://ssrn.com/abstract=1153162>.

court as an authentic title deed, and  $S$  is an encrypted secret password (specifically, a discrete exponential  $[s]$ ). Alice can prove that she owns the asset described by  $d$  by proving that she knows  $s$ , using the knowledge of logarithms protocol set out above. Using this protocol, and some help from the asset register<sup>11</sup>, Alice can at any time renew her password (this is called refreshing her identity) and do so in a way that guarantees that her secret is uniformly distributed, and commonly known to be so.

### 3.3 Underlying assumptions

We now discuss the assumptions that are required to support these cryptographic mechanisms. The first is the existence of computationally intractable problems<sup>12</sup>. Encryption relies on the ability to create specific computational tasks that are beyond the capacity of any feasible computing resources.

The second is that agents are computationally bounded, and cannot carry out such tasks. At a fundamental level computation and rational inference are indistinguishable, at least within standard first order logic, so an assumption that agents are computationally bounded is an assumption of bounded rationality. This leads to deep issues about the nature of equilibrium that we prefer to avoid, so we introduce computational intractability into our model through an assumption on the technology rather than as an explicit bounded rationality assumption. That is, we regard encrypting a message as a technological primitive, just as we treat transmission of a message by email. We also treat testing the validity of a proposition by running a proof protocol (an interactive computer program) as a technological primitive.

Let us be clear about what this means. The process of encrypting, decrypting or executing a proof protocol will involve the exchange of various intermediate messages between computers. These encrypted messages are opaque to computationally bounded agents who do not know the keys, but could be informative to hyper-rational, computationally unbounded agents (such a hyper-rational agent can immediately crack any standard encryption and see through any encrypted message). Our assumption is that agents do not have access to the contents of such computationally opaque messages, whether they arise as explicit encryptions or as intermediate messages in proof protocols, and do not treat them as strategically relevant. The strategies and beliefs of agents do not depend on information that is cryptographi-

---

<sup>11</sup>Alice and the authority generate her secret jointly. First, she generates some  $s'$ , independently and uniformly distributed in the required range and sends the authority  $[s']$ . Then the authority generates a “blinding factor”  $b$  (also supposed to be independently and uniformly distributed in the same range) and sends it to Alice. Alice’s secret would then be defined as  $s = s' + b \bmod p - 1$ , which the authority could generate directly from  $[s']$  and  $b$ . The secret is correctly distributed if either Alice or the authority generated their contribution correctly and keep it secret; for it to be common knowledge it is necessary that the authority be trusted to generate the blinding factor correctly. Note that the password is known only to Alice, and is not known by the asset register. See (Goldreich 2004) for more details.

<sup>12</sup>See (Goldreich 2001) for a more precise statement of what must be assumed.

cally hidden from them<sup>13</sup>. We believe that this is a reasonable approach to modelling the decisions available to our agents. We make explicit our assumption that zero knowledge proofs do not release strategically relevant information, apart from the proposition that they assert:

**Assumption 1** *Let  $\mathcal{Z}(R)$  be a zero knowledge proof of knowledge for a relation  $R$ . Then the agents have access to an oracle  $\mathcal{O}(\mathcal{Z})$  that reproduces the outcome of  $\mathcal{Z}(R)$ . Suppose  $x$  is publicly known, and Alice sends the oracle some message  $s$ . If  $(x, s) \in R$  then the oracle publicly announces 1, just as would the zero-knowledge proof  $\mathcal{Z}(R)$ . Otherwise, after receiving message  $s$ , the oracle announces 1 or 0 according to the same probability distribution as would the zero-knowledge proof  $\mathcal{Z}(R)$ . Beyond the announcement of 0 or 1, the oracle releases no other information.*

We call such an oracle a *cryptographic audit technology*<sup>14</sup> for  $R$ .

## 4 The Exchange Game

INSERT FIGURES 1 AND 2 ABOUT HERE

Using this machinery, we now describe the exchange game. Alice and Bob have valuable secrets  $s_A$  and  $s_B$ , which we will also refer to as assets. These secrets have been randomly encoded as numbers  $s_A \in \mathbb{S}$  and  $s_B \in \mathbb{S}$  in some suitable large set  $\mathbb{S}$ , and cryptographic commitments  $S_A = [s_A]$  and  $S_B = [s_B]$  to these secrets are common knowledge, having been published in a public database.

The values of the assets, to Alice and Bob respectively, we write as  $(\alpha_A, \alpha_B)$  and  $(\beta_A, \beta_B)$ . We assume that the trade surpluses  $\beta_A - \alpha_A$  and  $\alpha_B - \beta_B$  are positive, so that they both wish to trade. Throughout the paper, whenever the situation is similar for each agent we will discuss the exchange from Alice's point of view, and in the interest of readability write  $s$  rather than  $s_A$  (and similarly for other variables) when the subscript is clear from the context.

Alice breaks up her secret  $s$  into pieces  $s^1, \dots, s^N$ , using a *decomposition method II*. She writes each piece  $s^i$  of her secret on a slip of paper, puts it in a cryptographically locked envelope  $S^i$ , and passes all the locked envelopes to Bob<sup>15</sup>. Since Bob now has,

---

<sup>13</sup>To implement this encryption technology, agents just use a standard encryption method as described above with the security parameter set sufficiently high that the cost of cracking the encryption exceeds any possible gains from the exchange game.

<sup>14</sup>As outlined above, we imagine agents executing the  $\mathcal{Z}(R)$  protocol but not paying attention to cryptographically opaque intermediate messages. We do not model these intermediate messages as part of the game.

<sup>15</sup>For technical reasons we make slightly different assumptions about the way that Alice encrypts her secret  $s$ , and the way that she encrypts the pieces  $s^i$ . Her secret will be encrypted using a Schnorr encryption (specifically the discrete logarithm) that is perfectly binding but only computationally hiding, as described in Section 3.1. In contrast, the pieces of her secret are encrypted using a Pedersen

in encrypted form, both her secret and the pieces of her secret he can challenge her to provide evidence that she has acted honestly: that she can in fact open the envelopes, and that the contents are in fact a valid  $\Pi$  decomposition of her secret. We model this by giving Bob access to an audit technology, as described in Section 3.2, which gives him a signal as to whether Alice has acted honestly. Bob goes through a similar set up procedure, and they then play a message exchange game.

In the message game they exchange pieces of their secrets by unlocking envelopes. They may halt the exchange of messages, if they so wish, at any stage. If the message exchange has halted, they may, if they wish, try to guess the other player’s secret. Once some envelopes have been opened, Bob will have partial information about Alice’s secret. This information may not be enough to allow him to reconstitute the entire secret  $s$ , but it will reduce his computation cost if he tries to guess it. As information is passed back and forth, the expected guessing costs move along a zigzag exchange path on a grid in computation cost space (see Figure 1). The information that Bob can extract from Alice’s messages, and the path in the cost grid that the exchange moves along, depend on the rules (the decomposition method  $\Pi$ ) and on the credibility of Alice’s assertion that she has followed these rules.

After establishing some notation, we analyse this game in three stages. First, we ignore the message structure and assume that the agents can manipulate each other’s expected costs directly. Then we require that they manipulate expected costs by sending credible messages that change the other agent’s beliefs. Finally we consider the implications of the fact that in reality messages can be made only  $\varepsilon$ -credible, not fully credible.

## 4.1 Decomposing secrets and manipulating computational cost

Alice breaks up her secret  $s$  into a list  $\sigma = (s^1, \dots, s^N) \in \mathbb{S}^N$  of pieces. Let  $\Omega = \mathbb{S} \times \mathbb{S}^N$ .

**Definition 1 (decomposition method)** *A decomposition method is a set  $\Pi \subset \Omega$  of secrets  $s$  and allowable decompositions  $\sigma = (s^1, \dots, s^N)$  of  $s$ . If  $(s, \sigma) \in \Pi$  we say that  $\sigma$  is a  $\Pi$ -decomposition of  $s$ . We write  $\pi : \Pi \rightarrow \mathbb{S} : (s, \sigma) \mapsto s$  for the projection.*

We say that Alice has acted honestly if  $(s, \sigma) \in \Pi$ . That is, she has decomposed her secret as specified by the decomposition method  $\Pi$ .

**Example 1 (Blum Damgård decomposition)** *Blum and Damgård specify a decomposition in which the pieces  $s^i$  are the digits in a binary expansion of  $s$ .*

$$\Pi = \left\{ (s, (z_1, \dots, z_N)) \in \Omega : s = \sum_i^N z_i 2^i, z_i \in \{0, 1\} \right\}.$$

---

encryption that is perfectly hiding but only computationally binding. We set up the commitments in this way in order to restrict the strategy spaces: Bob can only try to guess Alice’s secret  $s$ , not the pieces  $s^i$  into which it has been decomposed.

**Example 2 (block decomposition)** We will use a decomposition<sup>16</sup> based on dividing the search space into  $n$  blocks of length  $\delta$ .

$$\Pi = \left\{ (s, (b, a_0, \dots, a_n)) \in \Omega : s = b + \delta \sum_{i=0}^n a_i, b \in [0, \delta), a_i \in \{0, 1\}, \sum_i a_i = 1 \right\}.$$

**Example 3 (uninformative decomposition)** In the uninformative decomposition  $\Pi = \Omega = \mathbb{S} \times \mathbb{S}^N$  the pieces  $s^i$  are completely unrelated to the secret  $s$ .

We now discuss the messages that may be sent. We continue to think of the commitment  $S^i$  as an envelope labelled with the number  $i$  and containing  $s^i$ , and write  $\Sigma = (S^1, \dots, S^N)$  for the list of these commitments. A message  $\tau$  is a set of envelopes that have been opened, together with their contents<sup>17</sup>. Clearly messages are ordered by inclusion: if  $\tau \subset \tau'$  then we say that  $\tau'$  extends  $\tau$ . A conversation  $h$  is an increasing sequence of messages  $(\tau_A^0 \subset \tau_A^1 \subset \tau_A^2 \subset \dots \subset \tau_A^k)$  from Alice and a similar sequence  $(\tau_B^0 \subset \tau_B^1 \subset \tau_B^2 \subset \dots \subset \tau_B^l)$  from Bob; they speak alternately so either  $k = l$  or  $k = l \pm 1$ . It is useful to establish some notation:

1.  $\Pi_\tau = \{(s, \sigma) \in \Pi : \tau \subset \sigma\}$  is the set of all secrets  $s$  and decompositions  $\sigma$  that are consistent with the message  $\tau$ .
2.  $\mathbb{S}_\tau = \{s \in \mathbb{S} : \exists \sigma : (s, \sigma) \in \Pi_\tau\}$  is the set of all secrets  $s$  that are consistent with the message  $\tau$ . It is the image of  $\Pi_\tau$  under the projection  $\pi$ .
3.  $c(\tau) = |\mathbb{S}_\tau| - 1$ . This is the number of candidates that might need to be inspected if one were conducting an exhaustive search for  $s$ , knowing that  $s \in \mathbb{S}_\tau$ .

The cost of guessing is the expected search cost: this is proportional to  $c(\tau)$  if the secret is uniformly distributed in  $\mathbb{S}_\tau$ , or if Bob uses a random search strategy. We discuss the precise connection between  $c(\tau)$  and expected search costs more carefully below, when we make discuss the exchange game in Section 4.3. Abusing notation slightly, we will in the mean time call  $c(\tau)$  the guessing cost. If the decomposition  $\Pi$  is not clear from the context, we write  $\mathbb{S}_{\Pi, \tau}$  and  $c_\Pi(\tau)$  respectively. In principle, the set  $G_B = \{c(\tau) : \tau \subset \sigma\}$  of guessing costs (Bob's cost grid) that can occur as a result of Alice's messages may depend on  $s$  and  $\sigma$ , but we make the following assumptions (easily checked in the examples above).

**Assumption 2** The number of ways to decompose a secret  $s$  does not depend on the value of  $s$  (in the first two examples above this number is 1; in the last it is  $|\mathbb{S}|^N$ ).

<sup>16</sup>For a related construction see Schoenmakers (2005).

<sup>17</sup>A message is thus a partial function  $\tau : \{1, \dots, N\} \rightarrow \mathbb{S}$ . That is, it is a set of ordered pairs  $(i, s^i) \in \{1, \dots, N\} \times \mathbb{S}$  such that  $(i, s^i) \neq (j, s^j) \Rightarrow i \neq j$ . The domain of the partial function is just the set of envelopes that have been opened.

**Assumption 3**  $G_B$  does not depend on the secret  $s$  or its encoding  $\sigma$ .

The first assumption is useful to ensure that, for any  $A \subset S$ , if  $(s, \sigma)$  is uniformly distributed in  $\pi^{-1}(A)$  then  $s$  is uniformly distributed in  $A$ . The second says that the cost path can be chosen independently of what the secret is. This is important, because otherwise the cost path chosen could leak information about the contents of the secret.

As Alice and Bob send messages and exchange pieces of their secrets, their computation costs move between points on the feasible cost grid<sup>18</sup>  $G = G_A \times G_B$ . The Blum Damgård decomposition provides an exponentially spaced grid, since under their decomposition revealing any piece of the secret halves the search space. The block decomposition contains a linear grid that can be made as fine as desired. The uninformative decomposition contains a single (uninformative) point.

Our objectives are, first, to characterise decomposition methods  $\Pi$ , and the cost grids  $G$  that they induce, that will support exchange as an equilibrium outcome of the message game and, second, to show that exchange can be supported by constructing such a  $\Pi$ . We approach this task in three stages. In Section 4.2 we study an abstract version of the game, which we call the grid game, played directly on the cost grid  $G$ . In this game we assume that agents can manipulate each other’s off-equilibrium expected costs directly, rather than by sending messages. We characterise grids  $G$  that support exchange as a subgame perfect equilibrium. In Section 4.3 we study the message game, but under the assumption that all messages are fully credible. We impose structural assumptions on  $\Pi$  that allow us to map beliefs, strategies and payoffs in the message game to strategies and payoffs in the grid game. This allows us to characterise a class of decomposition methods  $\Pi$  that support exchange as a perfect Bayes equilibrium (under the assumption that all messages are fully credible). In Section 4.4 we relax the full credibility assumption, assuming only that messages are  $\varepsilon$ -credible. We show that an exchange equilibrium is supported under essentially the same conditions provided that  $\varepsilon$  is small enough. Finally in Section 4.5 we display a decomposition  $\Pi$  that delivers the  $\varepsilon$ -credible fine control of costs that we need in order to support exchange.

## 4.2 The Grid Game

The grid game is a cut down version of the full communication game that exposes the main features of the equilibrium. We return later to the full communication game, and discuss how the grid game may be implemented there by sending messages.

A cost path is a decreasing sequence of costs  $c_B^0 > c_B^1 > \dots > c_B^k$  chosen by Alice and a similar sequence  $c_A^0 > c_A^1 > \dots > c_A^l$  chosen by Bob, defining a zig-zag path in  $G = G_A \times G_B$  (see Figure 1). We say that the cost path is *complete* if  $c_B^k = c_B^l = 0$ . If

---

<sup>18</sup>For simplicity, we will assume that the values  $\alpha_A, \alpha_B, \beta_A, \beta_B$  do not lie on the grid  $G$ . This means we can ignore cumbersome knife-edge cases.

a complete cost path is successfully traversed then exchange has been implemented, since the agents then know each other's secrets (the search space has been reduced to a singleton). We note that Alice chooses Bob's costs, and in exchange Bob chooses Alice's. These off-equilibrium costs are incurred only if the players deviate from the exchange path.

The grid game is played as follows. At each stage where it is her turn, Alice may choose either to extend the cost path (in the full message game she will do this by sending Bob a message that reduces his guessing cost), or not to extend it, which terminates the exchange. We call a path that has been terminated a terminal path. Given a path that has been terminated (by either herself or by Bob) she can either do nothing, or she can choose to guess Bob's secret. The cost of guessing is  $c_A^l$ , the last cost chosen by Bob. Her strategy  $(m, g)$  is thus in two parts:

1. Given a cost path, her *path strategy*  $m$  determines whether and how she will extend the path (by choosing a suitable point in the cost grid that extends the path)
2. Her *guessing strategy*  $g$  determines, at any terminal path, whether she will search.

Payoffs are determined by the guessing costs, and by the values of the objects. We now study which complete cost paths can be supported in a subgame perfect equilibrium in the grid game. Since we consider subgame perfect equilibria, the continuation game depends only on the grid point resulting from the last message that has been sent. We call this grid point  $(c_A, c_B)$  the state of the game.

**Lemma 1** *In the guessing stage Alice will attempt to guess Bob's secret if and only<sup>19</sup> if  $c_A < \beta_A$  and Bob will attempt to guess Alice's secret if and only if  $c_B < \alpha_B$ .*

Since the only incentive the players have to give each other information is the promise of more information in the future, no trade is an equilibrium.

**Lemma 2** *Silence is an equilibrium. That is, the strategies of always terminating the exchange and guessing as specified in Lemma 1 form a subgame perfect equilibrium.*

**Proof.** By the one-deviation principle, we only need to consider single deviations from silence. Assume that Alice is the active player. Since Bob never says anything, there is no change to her information and her guessing costs remain the same. By monotonicity, Bob's guessing cost can only decrease if she deviates. If it decreases enough, then under the equilibrium, he will choose to search and acquire Alice's object. Otherwise, payoffs are the same, so deviating can only reduce her payoff. ■

It is not possible for a player to get ripped off in equilibrium, because they can opt-out.

---

<sup>19</sup>By the assumption that none of the valuations lie on a grid point, we do not need to consider cases of equality.

**Lemma 3** *In no equilibrium does only one player learn their counterpart's secret.*

**Proof.** If Alice surrenders her secret without learning Bob's, her payoff is  $-\alpha_A$  which is worse than her abort payoff of 0. So Alice would deviate from any such equilibrium by aborting at the start. Similarly, Bob would abort at his first chance. ■

Since Alice knows anything that Bob tries to compute, computation is wasteful. Instead, an efficient equilibrium requires that Alice and Bob tell each other their entire secrets. This raises the question: when would Alice prefer to abort the conversation rather than complete it? If Bob's computational cost is sufficiently low that he would guess Alice's secret anyway, no matter what she does, then she cannot prevent Bob from learning her secret and she has nothing to gain by aborting the conversation. In this situation we say that she is committed to the transaction in the sense that she has lost control of her own object and it is (weakly) dominant for her to continue the conversation. We define commitment regions for Alice and Bob to be the sets  $X_A = [0, \infty) \times [0, \alpha_B]$  and  $X_B = [0, \beta_A] \times [0, \infty)$ .

Alice's position is at risk if she is committed to the transaction but Bob is not. In Alice's unilateral commitment region  $X_A - X_B$ , Bob can walk away with both objects, leaving her with nothing. Whether he would do so depends upon his guessing cost. We define  $Y_A = [\beta_A, \infty) \times [0, \beta_B]$  and  $Y_B = [0, \alpha_A] \times [\alpha_B, \infty)$  to be their danger regions. At a point in in Alice's danger region  $Y_A$ , Alice is committed to the transaction but Bob is not; furthermore, it is clearly in his interest to abort immediately and seize her object since the benefit of getting Alice's object and not losing his (even after paying the computation costs that he will incur) exceed the trade surplus (see Figure 2).

**Lemma 4** *If the conversation enters Alice's danger region  $Y_A$  then she loses her object but Bob does not lose his. Hence, every equilibrium cost path avoids the danger regions  $Y_A$  and  $Y_B$ .*

**Proof.** We note first that if a conversation enters Alice's danger region then it will never leave it. Consider a conversation  $h$  that terminates in  $Y_A$  and a conversation  $h'$  that extends  $h$  and terminates outside  $Y_A$ . We consider Bob's guessing behaviour at  $h$  and  $h'$ . At  $h$  Alice is committed to the transaction but Bob is not, so his payoff would be  $\alpha_B - c_B$ . The conversation can only move out of  $Y_A$  by moving into Bob's commitment region, so at  $h'$  both agents are committed to the transaction and Bob's payoff is  $\alpha_B - \beta_B - c'_B$ , which is strictly less than  $\alpha_B - c_B$  since  $c_B < \beta_B$ . Thus Bob's payoff is strictly lower if the conversation were to leave  $Y_A$ . This could only happen by his sending a message that committed him to the exchange, which he would never do. Since the conversation must eventually terminate (there is only a finite number of messages to be sent), it must terminate in  $Y_A$ . The result then follows. ■

Let  $Z_A = X_A - X_B - Y_A$  and  $Z_B = X_B - X_A - Y_B$  be the agents' safe unilateral commitment regions (see Figure 2).

**Lemma 5** *Every complete equilibrium cost path must enter a safe unilateral commitment region, either  $Z_A$  or  $Z_B$ .*

**Proof.** Consider a complete equilibrium cost path such that at the beginning, when costs are  $(c_A^0, c_B^0)$ , neither agent is committed to the exchange. At the end, when costs are  $(0, 0)$ , both are committed. Since the agents move alternately, at some point one agent must be committed while the other remains uncommitted. Moreover, this point lies outside the danger regions by Lemma 4, and therefore lies inside a safe unilateral commitment region. ■

**Theorem 1** *A complete cost path is supported by a subgame perfect equilibrium in the grid game if and only if it avoids the danger regions.*

**Proof.** The necessity of avoiding danger regions is clear from Lemma 4. It is also a sufficient condition. The guessing strategy is completely determined by Lemma 1. On the equilibrium path the message strategy is determined by the cost path. Off the equilibrium path, we fix the message rules to silence. In these subgames, silence is an equilibrium by Lemma 2.

We now check for profitable deviations on the equilibrium path. Since silence is played after any deviation from the cost path, all deviations terminate the exchange of messages. By Lemma 5, the cost path goes through three stages when neither player is committed, one player is committed, and both are committed. If Alice terminates the exchange before either player is committed, then she does not learn Bob's secret, and her deviation payoff is  $0 < \beta_A - \alpha_A$ . If she terminates the exchange when only she is committed, then she loses her object without gaining Bob's, so her payoff is  $-\alpha_A$ . If she terminates when only Bob is committed, then she guesses Bob's secret without revealing her own for a payoff  $\beta_A - c_A < \beta_A - \alpha_A$ . Finally, if she terminates after both players are committed, then both players learn each others secrets after some computation, so Alice's payoff is  $\beta_A - \alpha_A - c_A$ . ■

If there are any grid points available in a safe unilateral commitment region, then trade can be implemented.

**Theorem 2** *There is a trade equilibrium if and only if the players' safe unilateral commitment region  $Z_A \cup Z_B$  overlaps with the grid  $G$  of feasible computation costs.*

The decomposition that we develop below will provide a linear grid that can be made as fine as desired.

**Corollary 1** *If trade can be implemented, then it can be implemented in three steps<sup>20</sup>. In particular, if  $(\tilde{c}_A, \tilde{c}_B)$  lies in Alice's commitment region, then the 3-turn cost path*

$$\begin{aligned} c_A &= (c_A^0, 0) \\ c_B &= (c_B^0, \tilde{c}_B, 0) \end{aligned}$$

*is supported in equilibrium. In this equilibrium Alice speaks twice, Bob only once.*

---

<sup>20</sup>This three-turn secret exchange protocol is reminiscent of the Jakobsson (1995) note-ripping exchange, and also of a blackmail mechanism described by Anton & Yao (1994).

**Corollary 2** *The Blum Damgård protocol may fail to implement trade as a subgame perfect equilibrium of the grid game.*

The Blum Damgård protocol, which exchanges one bit of a binary expansion at each point, halves the search space at each step. The exchange conversation thus moves along a fixed path in a rather coarse, exponentially spaced grid. Whether this supports exchange depends on whether it provides enough points to allow one player to commit safely. Given the fixed Blum Damgård path there is freedom to choose the asset value parameters and move the points  $(\alpha_A, \alpha_B)$  and  $(\beta_A, \beta_B)$  around as one wishes, and for many reasonable parameter values the exchange path will intersect the danger region.

**Remark 1** *The result is robust to a degree of uncertainty about the values  $\alpha_A, \alpha_B, \beta_A, \beta_B$ . It is sufficient that they lie in commonly known intervals such that the gains from trade are bounded away from zero.*

#### 4.2.1 Discussion

There are thus three distinct phases in the exchange. Initially neither player is committed, with the initial stages of the game occurring outside  $X_A \cup X_B$ . The game could be aborted at this stage with no loss to either player. Eventually one player makes a safe unilateral commitment and play enters  $Z_A \cup Z_B$ . It then progresses to a stage of mutual commitment  $X_A \cap X_B$ . There is an intrinsic indivisibility here, into no matter how many steps the transaction is divided. The crux is the moment when one player makes a unilateral commitment, which they will do only if they can do so safely. In fact we notice that if exchange can be implemented, then it can be done very simply in a three step exchange. This emphasises the fact that the apparent divisibility introduced by proceeding in steps does not buy us as much as may appear at first glance.

### 4.3 Full Credibility Exchange

We return to the general message passing game. In this section we study exchange under the assumption that each player acts honestly, and this honesty is fully credible. That is, lies are detected with complete certainty (we relax this assumption in Section 4.4). We implement this assumption simply by assuming that is impossible to lie. We will impose structural assumptions on  $G$  and  $\Pi$  that allow us to map beliefs, strategies and payoffs in the message game to strategies and payoffs in the grid game. Under these assumptions we have a close relationship between equilibria in the message exchange game and the grid game.

The exchange game proceeds in two stages: a setup stage and an exchange stage. First, Alice and Bob announce a cost path: feasible cost sequences  $(c_B^0 > c_B^1 > \dots > c_B^k)$  and  $(c_A^0 > c_A^1 > \dots > c_A^l)$  that define a zig-zag path in the cost grid  $G$  (see Figure 1).

Then they go back to the asset registrar and refresh their secrets (see Section 3.2). This ensures that it is common knowledge that their secrets are hidden uniformly, and that the announced cost path (which was chosen before the secrets were refreshed) does not contain any information about the secrets.

The game then proceeds as follows. Recall that a message is a set of open envelopes and a conversation  $h$  is a strictly increasing sequence of messages  $(\tau_A^0 \subset \tau_A^1 \subset \tau_A^2 \subset \dots \subset \tau_A^k)$  from Alice, and a similar sequence  $(\tau_B^0 \subset \tau_B^1 \subset \tau_B^2 \subset \dots \subset \tau_B^l)$  from Bob, where they speak alternately. Alice chooses a strategy  $(\sigma, \mu, \gamma, \phi)$  as follows.

1. Knowing her secret  $s$  she chooses a decomposition  $\sigma = \sigma(s) = (s^1, \dots, s^N)$  such that  $(s, \sigma) \in \Pi$  (this is where we impose the full credibility assumption) and commits to the decomposition by sending Bob  $\Sigma = (S^1, \dots, S^N)$ , a list of the encrypted versions of the pieces  $s^i$ . They then begin the conversation phase.
2. Conditional on  $s$  and  $\sigma$  she chooses a conversation strategy  $\mu$ . This is a strategy for continuing conversations. At a conversation  $h$  where it is her turn to speak Alice either extends the conversation by sending a (possibly random) message<sup>21</sup>  $\mu(h)$  to Bob (which she does by opening a non-empty set of envelopes that have not already been opened and letting him read the contents), or she terminates the exchange (by sending an empty message).
3. If the conversation has been terminated (by either herself or by Bob) then, conditional on  $s$ ,  $\sigma$ , and the conversation  $h$ , she uses a guessing strategy  $\gamma$  which determines whether or not she attempts to guess Bob's secret.
4. If she chooses to guess Bob's secret then, conditional on the entire history, she chooses a search strategy  $\phi$ : an order in which to search for  $s$ .

Payoffs depend upon the values of the objects, and on the computation costs encountered; we assume that these computation costs are negligible except in the search phase. Alice's decisions will be influenced by her beliefs about  $(s_B, \sigma_B)$ , and Bobs by his beliefs about  $(s_A, \sigma_A)$ . We look for perfect Bayes-Nash equilibria.

### 4.3.1 Assumptions on $G$ and $\Pi$

Before discussing the equilibrium, it will be useful to make some definitions and assumptions that clarify the connection between strategies in the message game and in the grid game. These assumptions are readily checked for the Blum Damgård and uninformative decompositions. They will be confirmed in Section 4.5 for the block decomposition.

Clearly every conversation  $h = ((\tau_A^0 \subset \tau_A^1 \subset \dots), (\tau_B^0 \subset \tau_B^1 \subset \dots))$  in the message game induces a monotone cost path  $p = ((c_A^0 \geq c_A^1 \dots), (c_B^0 \geq c_B^1 \dots))$  in the grid  $G$ ,

---

<sup>21</sup>Since this choice is conditional on  $s$  and  $\sigma$ , we should really write  $\mu_{s,\sigma}(h)$  or  $\mu(h; s, \sigma)$ . We omit these conditioning arguments from our notation in the interest of simplicity.

with  $c_B^i = c(\tau_A^i)$  and  $c_A^i = c(\tau_B^i)$  — note the reversal of subscripts: Bob’s cost depends on Alice’s message. However the converse need not be true: it is not necessarily the case that every cost path can be lifted to a conversation and implemented by sending messages. All of the paths that we need will have this property, and it is useful to make a definition that captures this.

**Definition 2 (permissible paths)** *A path  $p$  in  $G$  is permissible if it is induced by a conversation history  $h$ . A grid path strategy is permissible if it always extends a permissible path permissibly or terminates it, and terminates any impermissible path. A subgrid  $G_0 \subset G$  is permissible if every monotone decreasing path in  $G_0$  is permissible.*

We note that a permissible grid path strategy is determined entirely by its values on permissible paths. It is easy to check that the uninformative decomposition and the Blum-Damgård decomposition have the path lifting property, so for these the full cost grid  $G_0 = G$  is permissible. In Section 4.5 we characterise permissible paths and identify a large permissible subgrid  $G_0 \subset G$  of the block decomposition grid that is sufficient for our needs. From here on we assume that we have identified a fixed non-empty permissible subgrid  $G_0$ .

**Remark 2** *Nothing in Section 4.2 changes if we restrict ourselves to permissible paths and strategies, provided that the initial cost path is permissible. In particular, every complete permissible cost path that avoids the danger regions is supported by a permissible subgame perfect equilibrium (Theorem 1), since the equilibrium that we construct there is permissible, and there is a permissible trade equilibrium if the safe unilateral commitment region  $Z_A \cup Z_B$  overlaps the permissible subgrid  $G_0$ .*

We now consider the connection between (mixed) cost path strategies in the grid game and conversation strategies in the message game. Recall that two messages  $\tau$  and  $\tau'$  are cost equivalent if  $c(\tau) = c(\tau')$ . Conversations  $h$  and  $h'$  are cost equivalent if the corresponding messages are cost equivalent. First we get rid of messages that contain redundant, cost-irrelevant information.

**Definition 3 (minimal extension)** *A message  $\tau' \supset \tau$  is a minimal extension of  $\tau$  if there is no smaller cost equivalent extension  $\tau'' \supset \tau$  (that is, such that  $\tau \subset \tau'' \subsetneq \tau'$  and  $c(\tau') = c(\tau'')$ ). A non-minimal extension contains cost irrelevant information that could be omitted.  $\tau'$  is a minimal extension of a conversation  $h$  if it is a minimal extension of the appropriate preceding message.*

Next we impose a path independence assumption that says, in essence, that the cost relevant message strategies available in any cost state do not depend on the prior messages.

**Assumption 4 (symmetry)** For any message  $\tau$  write  $T(\tau) = \{\tau' : \tau' \text{ is a minimal extension of } \tau\}$ . The decomposition  $\Pi$  is symmetric if, given any cost equivalent messages  $\tau_0$  and  $\tau_1$ , there is a cost preserving bijection between the sets  $T(\tau_0)$  and  $T(\tau_1)$ .

We now identify a natural class of message strategies that induce path strategies in the grid game.

**Definition 4 (symmetric conversation strategies)** A conversation strategy  $\mu$  is symmetric if it has the following properties.

1. It always selects messages that are minimal conversation extensions
2. Let  $h, h'$  be cost equivalent conversations, and let  $\tau, \tau'$  be cost equivalent messages extending  $h$  and  $h'$  respectively. Then the probability that  $\tau$  is selected as the continuation of  $h$  is the same as the probability that  $\tau'$  is selected as the continuation of  $h'$ .

**Lemma 6** Under the symmetry assumption, there is a bijective correspondence between symmetric conversation strategies  $\mu$  and permissible mixed cost path strategies  $m$ .

**Proof.** If  $\mu$  is a symmetric conversation strategy, then it induces a permissible path strategy in a natural way. We need only consider how to extend permissible paths. Given such a path  $p$ , let  $h$  be a conversation inducing  $p$ . Let  $\tau$  be the (random) continuation of  $h$  chosen by  $\mu$ . Then the probability that the continuation of the path is  $m(p) = c_0$  is the probability that  $c(\tau) = c_0$ . This construction is independent of the choice of  $h$  and  $\tau$  by the symmetry definition. Conversely, given a permissible path strategy  $m$  we construct a symmetric message strategy  $\mu$  as follows. Given a conversation  $h$  with induced cost path  $p$ ,  $\mu$  randomly chooses a minimal extension that has cost  $m(p)$ . ■

We now turn to the inferences that may be made after receiving messages.

**Lemma 7 (monotonicity)** Assume that  $\Pi$  satisfies Assumptions 2 to 4 and that a cost path  $p$  has been announced chosen prior to refreshing the secrets. Suppose that Alice plays a symmetric conversation strategy consistent with  $p$  (this means that if the cost state lies on the announced cost path  $p$ , and it is Alice's turn, then she will send a message that implements the pre-announced cost reduction, and that she will terminate the conversation at cost states that are not on  $p$ ), and that Bob has a uniform beliefs over  $\mathbb{S}_\tau$  after every message  $\tau$ . Then:

1. Bob's beliefs are Bayesian
2. After a message  $\tau$  from Alice, Bob's guessing cost is  $c(\tau)$ , exactly as in the grid game

3. *Bob's guessing cost declines monotonically as he receives more messages from Alice.*

**Proof.** When Alice sends a message  $\tau$ , Bob knows that  $s \in \mathbb{S}_\tau$ . In addition, he could get information from her choice of message that will cause him to update his belief that  $s$  is uniformly distributed. Let us consider what he learns from  $\tau$ . He learns the cost state  $c(\tau)$ , and he learns which message  $\tau$ , within the cost equivalence class, Alice has chosen to use. Since the choice of the cost state  $c(\tau)$  is predetermined by the cost path  $p$ , which was announced before Alice's secret was refreshed, he learns nothing from  $c(\tau)$  that would cause him to update his belief that the secret is uniformly distributed. If Alice uses a symmetric strategy then she randomises uniformly between cost equivalent messages, so the choice of the message  $\tau$  within its cost equivalence class is also uninformative. Thus he learns nothing beyond the fact that  $s \in \mathbb{S}_\tau$ , and his posterior is uniform. Given this belief, the best that Bob can do is to search in any random order, and the expected cost of doing so is proportional to the size of the search set; by an appropriate choice of cost units we can arrange that the expected cost is equal to  $c(\tau)$ , the size of the set that must be searched. Monotonicity follows because of the monotonicity of Alice's messages  $\tau$ , and hence of the search sets  $\mathbb{S}_\tau$ . ■

This Lemma justifies our referring to  $c(\tau)$  as the cost function.

### 4.3.2 Equilibrium

We extend the term symmetric, and say that a message game strategy  $(\sigma, \mu, \gamma, \phi)$  is symmetric if: the decomposition strategy  $\sigma$  is uniform random; the search strategy  $\phi$  is uniform random; and the conversation strategy  $\mu$  is symmetric in the sense of Definition 4.

**Theorem 3** *Under Assumptions 2 to 4 and full credibility assumption there is a bijective correspondence between:*

1. *permissible subgame perfect equilibria in the grid game; and*
2. *symmetric uniform-belief perfect Bayes-Nash equilibria in the message game.*

**Proof.** Given a permissible strategy  $(m, g)$  in the grid game, we need to define a symmetric message game strategy  $(\sigma, \mu, \gamma, \phi)$ . There is no choice about  $\sigma$  and  $\phi$ : by the definition of a symmetric strategy they must be uniform random. We have already constructed in Lemma 6 a correspondence between permissible path strategies  $m$  and symmetric message strategies  $\mu$ . The search strategy is clear: search at the terminal conversation history  $h$  if  $g$  recommends searching at the induced cost path  $p$ . By Lemma 7, uniform beliefs are Bayesian and expected costs in the message game are exactly as in the grid game. So the profitability of deviating from the conversation or guessing strategies is the same in either game. Given uniform beliefs, any search order

is optimal so no deviation from the random search strategy is profitable. Similarly, no deviation from the random decomposition  $\sigma$  can be profitable since payoffs are unaffected. ■

As a consequence, under these assumptions the main results from the grid game translate to corresponding results in the full credibility message game.

**Corollary 3** *Assume full credibility. Let  $p$  be a permissible complete cost path in the cost grid  $G$ , chosen independently of the secrets  $s_A$  and  $s_B$  and their encoding. There exists a symmetric uniform-belief perfect Bayes-Nash equilibrium inducing the cost path  $p$  if and only if  $p$  avoids the danger regions.*

**Corollary 4** *Assume full credibility. If there is a permissible subgrid  $G_0 \subset G$  that overlaps with the safe unilateral commitment region  $Z_A \cup Z_B$  then there is a perfect Bayes-Nash equilibrium that supports trade.*

## 4.4 $\varepsilon$ -Credible Exchange

We now relax the unrealistic perfect credibility assumption, allowing the possibility that Alice may be dishonest when she decomposes her secret. She may choose a  $\sigma_A$  such that  $(s_A, \sigma_A) \notin \Pi$ . This is very dangerous for Bob. He may go to a lot of trouble to learn the components of  $\sigma_A$ , but then discover that these have no relation to  $s_A$ .

The most we can hope for in our environment is  $\varepsilon$ -credibility. Our objective in this section is to show that this is sufficient to sustain exchange. While Bob does not know  $(s_A, \sigma_A)$ , he does know the commitments  $(S_A, \Sigma_A)$ . We allow him to use this information to test her honesty, using the cryptographic audit technology discussed in Section 3.3.

**Assumption 5** *The relation  $\Pi$  is in  $\mathcal{P}$  (that is, the assertion that  $(s, \sigma) \in \Pi$  can be verified in polynomial time).*

**Lemma 8** *Let  $(S, \Sigma) \in \Omega$ , and let  $\varepsilon > 0$ . Then there is an audit technology generating a signal  $T_A = T_A(\Pi, S, \Sigma) \in \{0, 1\}$  such that if Alice has been honest (that is, she knows  $s$ ,  $\sigma = (s^1, \dots, s^N)$  and  $r = (r^1, \dots, r^N)$  such that  $(s, \sigma) \in \Pi$ ,  $S = [s]$ , and  $\Sigma^i = [s^i; r^i]$ ) then she has a strategy that guarantees that  $T_A = 1$  and reveals no other information apart from her honesty, while if she has been dishonest  $T_A = 0$  with probability at least  $1 - \varepsilon$ .*

**Proof.** The relationship between the encrypted pieces  $(S, \Sigma)$  and their corresponding secrets  $(s, \sigma)$  can be written formally as  $R = \{(S, \Sigma, s, \sigma, r) : (s, \sigma) \in \Pi, S = [s], \Sigma^i = [\sigma^i; r^i], \}$ , where  $\Sigma^i = [\sigma^i; r^i]$  is a Pedersen commitment to  $\sigma^i$ . Since  $\Pi$  is in  $\mathcal{P}$ , and the Schnorr and Pedersen commitments can be verified in polynomial time after they have need opened, the relation  $R$  is in  $\mathcal{NP}$ . The Bellare-Goldreich Theorem asserts that there exists a zero-knowledge proof of knowledge for  $R$ , provided that a one-way function

exists. By Assumption 1 there is an audit technology that implements this proof while hiding all cryptographically opaque intermediate messages. ■

Given an audit technology for the decomposition method  $\Pi$ , the specification of the  $\varepsilon$ -game is now exactly as in 4.3 except for the first step, where we now allow agents to lie, but the audit reveals a signal of whether they have done so. The first stage of the game becomes

1. Knowing  $s_A$  Alice chooses a decomposition  $\sigma_A$  such that  $(s_A, \sigma_A) \in \Omega$ , and commits to the decomposition by sending Bob  $(S_A, \Sigma_A)$ . Bob challenges her with the audit technology and she responds, generating a signal  $T_A \in \{0, 1\}$ . All strategies from this point forward are conditional on both audit reports  $T = (T_A, T_B)$ .

The main complication that  $\varepsilon$ -credibility introduces is that expected costs (which now of course depend in a non-trivial way upon beliefs) may no longer be monotonic. In the full credibility game, any message  $\tau$  from Alice can only decrease the size of the set  $\mathbb{S}_\tau$  where Bob believes her secret to be hidden. Once we admit doubt about Alice's initial assertions as to how she has encoded her secret it is possible that Bob may at some stage judge her message  $\tau$  to be implausible, in the sense that he no longer believes any of her previous assertions. In this case the set of possible hiding places  $\mathbb{S}_\tau$  may expand rather than contract. It is not clear that Alice would not want to send such a message, even if she could avoid it. If Bob were at a history where he may be expected to search for Alice's secret, then Alice might wish deter him from searching by sending a message that he found implausible, causing him to doubt the information on which he had been relying.

Unlike what occurs in the full credibility game, Bob's beliefs may change radically according to what he sees in the course of the game, and we need to keep track of the state of his knowledge. As the game progresses, Bob will know, or progressively learn his secret  $s$ , its decomposition  $\sigma$ , the audit signals  $T = (T_A, T_B)$ , and the conversation history  $h$ . The audit result  $T$  and the conversation history  $h$  are public information, while  $s$  and  $\sigma$  are private. We call the combination  $(T, h)$  the (public) game history, and the sequence  $k_B = (s_B, \sigma_B, T, h)$ , or some initial segment of such a sequence, Bob's knowledge state<sup>22</sup>. Conditional on the state  $k_B$  of his knowledge Bob will have beliefs over  $(s_A, \sigma_A) \in \mathbb{S} \times \mathbb{S}^N$ .

We will say that a message  $\tau$  is *implausible* if  $\mathbb{S}_\tau = \emptyset$ . That is, there exists no secret  $s' \in \mathbb{S}$  compatible with the message  $\tau$ . Alice cannot send such a message if she has acted honestly. Note that plausibility is a property of the message  $\tau$ , not of Alice's secret  $s_A$  (though her ability to send such a message will depend on her secret and its decomposition). A conversation  $h$  is implausible if it contains an implausible message. If  $h$  is an implausible conversation then any extension of  $h$  remains implausible. A state  $k_B = (s_B, \sigma_B, T, h)$  is implausible if it contains an implausible conversation, or

---

<sup>22</sup>By allowing  $h$ ,  $Z$  or  $\sigma$  to be empty sets, we can write any state  $k$  as a sequence of this form.

if the audit transcript  $T$  indicates that Alice lied. Such an implausible state cannot arise if Alice has been honest. Notice that the plausibility of a state depends only the public history  $(T, h)$ .

As in the persuasion games of Matthews & Postlewaite (1985) and Milgrom & Roberts (1986), we specify that agents have sceptical beliefs off the equilibrium path as soon as they see evidence that is incompatible with honest behaviour. In making the following definition recall that  $\Omega = \mathbb{S} \times \mathbb{S}^N$  is the uninformative decomposition. For any message  $\tau$  we have  $\mathbb{S}_{\Omega, \tau} = \mathbb{S}$ . In particular, for  $\tau = \emptyset$  the empty message,  $\mathbb{S}_{\Omega, \emptyset} = \mathbb{S}_{\Pi, \emptyset} = \mathbb{S}$ .

**Definition 5 (sceptical beliefs)** *We say that Bob has sceptical beliefs if*

1. *In any plausible state  $k_B$  he believes that  $(s_A, \sigma_A)$  is distributed uniformly in  $\Pi_{\tau_A}$ , and hence that  $s_A$  is distributed uniformly in  $\mathbb{S}_{\tau_A}$ ; here  $\tau_A$  is the most recent message sent by Alice*
2. *In any implausible state  $k_B$  he believes that  $(s_A, \sigma_A)$  is distributed uniformly in  $\Omega_{\tau_A}$ , and hence that  $s$  is distributed uniformly in  $\mathbb{S} = \mathbb{S}_{\Omega, \tau_A}$ .*

Thus initially he believes Alice's assertion that  $(s_A, \sigma_A) \in \Pi$ , updating his belief using Bayes rule after plausible messages. As soon as she fails to act plausibly he believes nothing that she has said that he cannot check herself and treats her messages as uninformative. He revises his initial belief to  $(s_A, \sigma_A) \in \Omega$ , updating it only with what he can himself verify in Alice's messages.

In the Lemma below, we require that Alice's conversation strategy be symmetric so that it induces a well defined strategy in the grid game, but we need to be a little careful as there are now two decompositions under consideration,  $\Pi$  and the uninformative decomposition  $\Omega$ , both of which could be used to define this concept. By symmetry we will of course mean  $\Pi$ -symmetry, which is consistent with our definition in Section 4.3.

**Lemma 9 (monotonicity II)** *Assume that  $\Pi$  satisfies Assumptions 1 to 4 and that a cost path  $p$  has been chosen prior to refreshing the secrets. Suppose that Alice decomposes her secret honestly, and that Bob has a sceptical beliefs. Then:*

1. *Bob's beliefs are Bayesian;*
2. *at plausible histories his guessing costs are exactly as in the grid game;*
3. *at plausible histories his guessing cost declines monotonically as he receives more messages from Alice.*

**Proof.** Bob’s updating strategy is explicitly Bayesian, except when he receives the first implausible message (either a failed audit or an implausible message  $\tau$  in the conversation stage). But receiving such a message is a zero probability event. Since Alice’s decomposition strategy is honest, the audit never fails and all her messages are plausible. So receiving a first implausible message is a zero probability event, leaving the posterior that follows unconstrained by Bayes’ rule. The rest of the Lemma is then as in Lemma 7. ■

If the agents have sceptical beliefs, then the evolution of their expected search cost is straight forward. While ever the game history remains plausible, costs evolve exactly as in the full credibility game. If Bob sees an implausible event (a failed decomposition audit, or an implausible message) his cost jumps immediately, and non-monotonically, to  $c(\emptyset) = c_{\Pi}(\emptyset) = c_{\Omega}(\emptyset) = |\mathbb{S}| - 1$  where it remains thereafter. Alice’s costs evolve in a similar fashion. Since plausibility of a state depends only on the public history, if agents have sceptical beliefs then they adjust those beliefs only in response to public events. The cost path thus depends only on the public history  $(T, h)$ . As in the full credibility game, a cost path is complete if it declines monotonically to  $(0, 0)$ . Note that any history that induces a complete cost path must be plausible.

**Theorem 4** *Assume that Assumptions 2 to 5 hold, and that the honesty of players is  $\varepsilon$ -credible, with  $0 < \varepsilon_A < 1 - \frac{\alpha_A}{\beta_A}$  and  $0 < \varepsilon_B < 1 - \frac{\beta_B}{\alpha_B}$ . Let  $p$  be a permissible complete cost path in  $G$ , chosen independently of the secrets  $s_A$  and  $s_B$ . We consider perfect Bayes Nash equilibria in the exchange game in which players have sceptical beliefs and play symmetric conversation strategies. There exists such an equilibrium inducing the cost path  $p$  if and only if  $p$  avoids the danger regions.*

**Proof.** The necessity of the condition is immediate. By Lemma 9 any such equilibrium induces an equilibrium in the grid game supporting  $p$ , so  $p$  must avoid the danger regions.

To show the converse we must construct an equilibrium. By theorem 1 there is a permissible subgame perfect equilibrium in the grid game that supports  $p$ . We describe the exchange equilibrium from Alice’s point of view.

1. Her beliefs are sceptical
2. She sets up honestly, choosing  $\sigma_A$  such that  $(s_A, \sigma_A) \in \Pi$ ; if she has a choice in this she chooses uniformly; she responds to Bob’s audit challenge in a way that confirms her honesty, if she is able to do so.
3. If Bob has acted implausibly then she believes that she can never acquire his object. The best that she can do is to discourage him from searching for hers (even if she has set up dishonestly, Bob might stumble upon her object if he does any searching). So she will reveal her dishonesty by sending an implausible message if she can; otherwise she ends the conversation and does not guess.

4. We now consider game histories  $(T, h)$  where Bob is plausible. If Alice is implausible then she has already discouraged Bob from guessing and she can send no messages that he will believe, so she aborts any conversation and plays the guessing strategy recommended in the grid game.
5. If Alice is honest and plausible, she uses the unique symmetric conversation strategy that is consistent with the equilibrium in the grid game.
6. If Alice is dishonest but remains plausible, then she mimics honest play as long as she can. That is, there is a plausible message of maximal length; she sends submessages of the appropriate length until she must send an implausible message. Note that she will not, at this stage, be using a symmetric conversation strategy.
7. If the conversation terminates and Bob is implausible, then Alice does not guess. If Bob is plausible, then she chooses to guess or not as recommended in the grid equilibrium.
8. If she searches, she does so randomly.

We note, by Lemma 9, that beliefs are Bayesian. To show that strategies are rational it suffices to show that there is no strictly profitable one step deviation. First we consider the decision of whether or not to set up honestly (the remaining possible deviations are routine to check). If Alice is honest, then the exchange will be implemented costlessly, giving her an outcome worth  $\beta_A - \alpha_A$ . If she is dishonest, then the best she could do would be to acquire both objects at zero cost, giving her an outcome  $\beta_A$ . But she can achieve this only if she successfully bluffs her way through Bob's audit, which occurs with probability  $\varepsilon_A$ ; otherwise the exchange will fail, leaving her with outcome 0. So her expected gain from deviating from truthfulness is at most  $\varepsilon_A \beta_A + (1 - \varepsilon_A) 0$  which is strictly less than  $\beta_A - \alpha_A$  under our assumption on  $\varepsilon_A$ .

Second, Alice's symmetric conversation strategy is optimal: because each permissible message reveals the same amount of information, Alice is indifferent between them.

Finally, the random search policy is optimal, as each candidate secret is equally likely. ■

**Theorem 5** *Assume that the gains from trade are strictly positive for both agents. Assume that the decomposition protocol  $\Pi$  satisfies Assumptions 2 to 5. Then trade can be supported as a perfect Bayes-Nash equilibrium of the exchange game defined by  $\Pi$ .*

**Proof.** Since the gains from trade are positive, we can find credibility parameters  $\varepsilon_A$  and  $\varepsilon_B$  satisfying the conditions of Theorem 4. Thus there exists a trade equilibrium

provided there is a permissible grid  $G_0 \subset G$ , and a cost path in  $G_0$  that avoids the danger regions. But this is guaranteed by Assumptions 2 to 4. ■

It remains only to show that we can construct a decomposition method  $\Pi$  that satisfies Assumptions 2 to 5. This we do in the next section.

## 4.5 The Block Decomposition

We confirm that the block decomposition  $\Pi$  has all the properties required for Theorem 5. Recall that the block decomposition segments the secret space  $\mathbb{S}$  into  $n + 1$  blocks of size  $\delta$ . We assume that  $n + 1 < \delta$ . In this section  $\Pi$  will refer exclusively to the block decomposition, which is defined as

$$\Pi = \left\{ (s, (b, a_0, \dots, a_n)) \in \Omega : s = b + \delta \sum_{i=0}^n a_i, b \in [0, \delta), a_i \in \{0, 1\}, \sum_i a_i = 1 \right\}.$$

**Lemma 10** *The block decomposition  $\Pi$  satisfies Assumptions 2 to 4.*

**Proof.** Since the block decomposition is unique, Assumption 2 holds.

In this decomposition the secret space  $\mathbb{S}$  is split up into  $n + 1$  disjoint blocks of size  $\delta$ , labeled by the integers 0 to  $n$ . By relabelling the blocks we may assume without loss of generality that the secret is hidden in the first block (that is, that  $a_0 = 1$ ). A message  $\tau$  will open some subset of the commitments  $\{[a_0], [a_1], \dots, [b]\}$ , and we can classify messages according to the pattern of what is opened. Let us write  $z_0 = 1$  if the commitment  $[a_0]$  has been opened, otherwise setting  $z_0 = 0$ . Let us write  $z_b = 1$  if the commitment  $[b]$  has been opened, otherwise setting  $z_b = 0$ . Let  $k$  be the number of commitments  $[a_i]$ , with  $i \neq 0$ , that have been opened.

It is straightforward to check that the computation cost  $c(\tau) = |\mathbb{S}_\tau| - 1$  only depends on  $z_0$ ,  $z_b$ , and  $k$ . In fact,  $c(\tau) + 1 = (z_0 + (1 - z_0)(n + 1 - k))(z_b + (1 - z_b)\delta)$ . This implies Assumption 3, that the grid  $G$  of feasible computation costs does not depend on the secret  $s$  or its decomposition  $\sigma$ .

To verify Assumption 4 we must check that the size of the set  $T(\tau)$  depends only on  $c(\tau)$ . Note that  $c(\tau) + 1$  must lie in one of two disjoint sets  $\{1, \dots, n + 1\}$  and  $\{\delta, 2\delta, \dots, (n + 1)\delta\}$ . If  $c(\tau) + 1 \in \{2, \dots, n + 1\}$  then  $z_0 = 0$  and  $z_b = 1$ , while if  $c(\tau) + 1 \in \{2\delta, \dots, (n + 1)\delta\}$  then  $z_0 = 0$  and  $z_b = 0$ . In these cases we can recover  $z_0$ ,  $z_b$ , and  $k$ , and hence the whole structure of the message, from the cost  $c(\tau)$ . Any two such messages with the same structure are isomorphic under relabelling the blocks 1 to  $n$ , so in particular the sets  $T(\tau)$  are isomorphic.

If  $c(\tau) + 1 = \delta$  then  $[b]$  cannot have been opened. Either  $[a_0]$  has been opened, or  $k = n$  and all the commitments  $[a_i]$ ,  $i \neq 0$ , have been opened. In either case there are exactly two minimal messages extending  $\tau$ ; one is the empty message, the other opens  $[d]$ , since all other envelopes are uninformative. The case where  $c(\tau) + 1 = 1$  is similar; the only minimal extension is the empty message. ■

It follows from the above construction that Alice's cost grid can be written

$$G_A = \{c_{zk} = (n + 1 - k)(z + (1 - z)\delta) - 1 : 1 \leq k \leq n, z \in \{0, 1\}\},$$

and it contains the subgrid  $G_A^0 = \{c_{0k} : 1 \leq k \leq n\} = \{\delta, 2\delta, \dots, (n + 1)\delta\}$ .

**Lemma 11** *A path in  $G_A$  is permissible if and only if it is monotonic in both  $z$  and  $k$ . The subgrid  $G_0 = G_A^0 \times G_B^0 \subset G$  is a permissible subgrid.*

In the exposition above we have assumed that the parameters  $n$  and  $\delta$  are the same for both Alice and Bob. There is however no need that they be the same, and we now allow them to differ.

**Lemma 12** *Assume that trade is individually rational and that there are strictly positive gains for trade for at least one player. Then, for suitable  $n$  and  $\delta$ , the permissible subgrid  $G_0$  meets the safe unilateral commitment region.*

**Proof.** Assume that the gains from trade are strictly positive for Alice. To ensure that there is a permissible gridpoint in Bob's safe unilateral commitment region  $Z_B$  it is sufficient that there be an integral multiple of  $\delta_A$  between  $\beta_A$  and  $\alpha_A$ , and that there be an integral multiple of  $\delta_B$  greater than  $\alpha_B$  (see Lemma 5 and Figure 2). We can ensure this by choosing  $\delta_A < \beta_A - \alpha_A$  and  $n_B \delta_B > \alpha_B$ . ■

Finally, we address Assumption 5, guaranteeing the existence of a zero knowledge  $\varepsilon$ -audit technology  $T$  for  $\Pi$ . we need only to confirm that, given  $s, \sigma$  and the parameters defining  $\Pi$ , the assertion " $(s, \sigma) \in \Pi$ " can be checked in polynomial time. But this is clear, as it involves only a few simple calculations.

## 5 Conclusion

In this paper we study a simultaneous commitment issue that lies at the heart of voluntary exchange.

We study trade in an environment where it can be reduced to the asynchronous exchange of information. Following protocols suggested in the computer science literature, we note that information can be made divisible by encrypting it with a long divisible key, which can be exchanged piece by piece. More generally, information can be exchanged by hiding it and exchanging hints on where to find it. Clearly the viability of this approach depends on the credibility of these hints. Computer science again shows the way, providing general techniques to establish  $\varepsilon$ -credibility of a well defined class of propositions.

Once it is recognised that information can be made divisible, trade becomes an incremental commitment problem analogous to those studied by Admati & Perry (1991) and subsequent literature. We study our problem first in computational cost space, where the exchange of information leads to a zig-zag exchange path. We characterise

the exchange paths that can be implemented as subgame perfect equilibria of a game in computational cost space, assuming that costs can be controlled as finely as may be required, and that all communications are fully credible. We find that trade can always be implemented if it is individually rational; that is, there are positive gains from trade for both agents. The bit-by-bit exchange protocol that has been discussed in the computer science literature fails however for a range of parameters. We observe that the analogy with the incremental investment literature is imperfect. Despite the apparent divisibility introduced by allowing arbitrarily fine control of computational cost, there remains an intrinsic indivisibility: there is a moment when one of the players becomes irrevocably committed to the exchange, and this is crucial to understanding the equilibrium. Gradualism per se is not the issue so much as being able to take credible steps of the right size.

We then consider implementation of trade through message exchange, augmenting a cheap talk environment with two cryptographic primitives: the ability to lock a secret in a box, and the ability to probabilistically audit certain statements about what is in the box without opening it. Agents break up their secrets into tradeable pieces using a decomposition method  $\Pi$ , and trade occurs through the simple exchange of messages. The set of possible messages delineates a grid  $G$  in computational cost space. Assuming that messages are perfectly credible, we identify properties of  $\Pi$  that allow us to interpret the message exchange game as a game in computational cost space. We find that trade can be implemented if and only if the feasible cost grid  $G$  (which depends on  $\Pi$ ) intersects the *safe unilateral commitment region* (which depends only on the values of the secrets being traded).

We then consider what happens if statements are only  $\varepsilon$ -credible, which is the most we can achieve through probabilistic audit protocols. We find that there is an  $\varepsilon_0 > 0$ , which depends only on the values of the secrets being traded, such that if  $\varepsilon \leq \varepsilon_0$  then  $\varepsilon$ -credibility delivers exactly the same as perfect credibility. Standard theorems ensure the existence of such  $\varepsilon$ -audit protocols for any  $\varepsilon > 0$ .

Finally we construct a decomposition method  $\Pi$  that does the job. We display a decomposition, with all the necessary properties to allow the message exchange game to be interpreted as a game in computational cost space, that delivers a feasible grid  $G$  that is as fine as desired. We thus find that trade can be implemented purely through the exchange of messages, and without the intervention of a trusted intermediary, provided that there are strictly positive gains from trade for both agents.

We have ignored the fixed cost of running this mechanism. This cost will depend on  $\varepsilon$ , but not on the value of the assets traded or the gains from trade except in so far as they impinge on the value of  $\varepsilon$ . When these fixed costs are taken into account then trade is individually rational provided that the gains from trade are sufficient to pay this fixed cost. Since we have relied on a general (though constructive) existence theorem for our audit technology, this does not immediately provide us with a good bound on this cost. In an earlier working version of this paper we construct an explicit audit technology which can provide such a bound. Since trade can be implemented

in three steps, the overhead is not very high and the fixed cost not large<sup>23</sup>.

## References

- Admati, A. R. & Perry, M. (1991), ‘Joint projects without commitment’, *Review of Economic Studies* **58**(2), 259–356.
- Anton, J. J. & Yao, D. A. (1994), ‘Expropriation and inventions: Appropriable rents in the absence of property rights’, *American Economic Review* **84**(1), 190–209.
- Aumann, R. J. & Hart, S. (2003), ‘Long cheap talk’, *Econometrica* **71**(6), 1619–1660.
- Bellare, M. & Goldreich, O. (1992), ‘On defining proofs of knowledge’, *Springer Verlag Lecture Notes in Computer Science* **740**, 390–420.
- BenPorath, E. (2003), ‘Cheap talk in games with incomplete information’, *Journal of Economic Theory* **108**(1), 45–71.
- Blum, M. (1983a), ‘How to exchange (secret) keys’, *ACM Transactions on Computer Systems* **1**(2), 175–193.
- Buttyán, L. & Hubaux, J. P. (2001), Rational exchange: A formal model based on game theory, in ‘Proc. 2nd International Workshop on Electronic Commerce; Lecture Notes in Computer Science Vol. 2232’, Springer-Verlag, pp. 114–126.
- Buttyán, L., Hubaux, J. P. & Capkun, S. (2002), A formal analysis of syverson’s rational exchange protocol, in ‘Proc. 15th IEEE Computer Security Foundations Workshop’, IEEE Computer Society Press, pp. 181–193.
- Chen, Y., Kartik, N. & Sobel, J. (2008), ‘Selecting cheap-talk equilibria’, *Econometrica* **76**(1), 117–136.
- Compte, O. & Jehiel, P. (2004), ‘Gradualism in bargaining and contribution games’, *Review of Economic Studies* **71**(4), 975–1000.
- Damgård, I. B. (1995), ‘Practical and provably secure release of a secret and exchange of signatures’, *Journal of Cryptology* **8**(4), 201–222.
- Dodis, Y. & Rabin, T. (2007), Cryptography and game theory, in N. Nisan, T. Roughgarden, E. Tardos & V. Vazirani, eds, ‘Algorithmic Game Theory’, Cambridge University Press, New York, pp. 181–205.

---

<sup>23</sup>Under reasonable assumptions on the security parameters, if each player’s surplus is at least 0.1% of their secret’s value then in total the transaction would require transmission of around 20 megabytes of data and around 5 seconds of waiting for messages to arrive, assuming that the transaction is between Sydney and New York. These costs are virtually all incurred in the set up phase, before the agents begin exchange of secrets.

- Forges, F. (1990), ‘Universal mechanisms’, *Econometrica* **58**(6), 1341–1364.
- Forges, F. & Koessler, F. (2005), ‘Communication equilibria with partially verifiable types’, *Journal of Mathematical Economics* **41**(7), 793–811.
- Gale, D. (2001), ‘Monotone games with positive spillovers’, *Games and Economic Behavior* **37**(2), 295–320.
- Gerardi, D. (2004), ‘Unmediated communication in games with complete and incomplete information’, *Journal of Economic Theory* **114**(1), 104–131.
- Goldreich, O. (2001), *Foundations of Cryptography*, Vol. 1, Cambridge University Press.
- Goldreich, O. (2004), *Foundations of Cryptography*, Vol. 2, Cambridge University Press.
- Goldreich, O. (2008), *Computational Complexity A Conceptual Perspective*, Cambridge University Press.
- Goldreich, O., Micali, S. & Wigderson, A. (1991), ‘Proofs that yield nothing but their validity or all languages in np have zero-knowledge proofs’, *Journal of the ACM* **38**(1), 691–729.
- Halpern, J. (2008), Computer science and game theory: A brief survey, in S. N. Durlauf & L. E. Blume, eds, ‘Palgrave Dictionary of Economics’, Palgrave MacMillan.
- Izmalkov, S., Lepinski, M. & Micali, S. (2007), ‘Perfect implementation of normal form mechanisms’, *CSAIL Technical Report* **40**.
- Jakobsson, M. (1995), Ripping coins for a fair exchange, in L. C. Guillou & J. J. Quisquater, eds, ‘Proc. Eurocrypt 1995; Lecture Notes in Computer Science Volume 921’, Springer-Verlag, pp. 220–230.
- Koblitz, N. I. (1994), *A Course in Number Theory and Cryptography*, Springer-Verlag.
- Lockwood, B. & Thomas, J. P. (2002), ‘Gradualism and irreversibility’, *The Review of Economic Studies* **69**(2), 339–356.
- Marx, L. M. & Matthews, S. A. (2000), ‘Dynamic voluntary contribution to a public project’, *The Review of Economic Studies* **67**(2), 327–358.
- Matthews, S. & Postlewaite, A. (1985), ‘Quality testing and disclosure’, *RAND Journal of Economics* **16**(3), 328–340.
- Menezes, A., van Oorschot, P. & Vanstone, S. (1997), *Handbook of Applied Cryptography*, CRC Press.

- Milgrom, P. R. (1981), ‘Good news and bad news: Representation theorems and applications’, *Bell Journal of Economics* **12**(2), 380–391.
- Milgrom, P. & Roberts, J. (1986), ‘Relying on the information of interested parties’, *RAND Journal of Economics* **17**(1), 18–32.
- Okuno-Fujiwara, M., Postlewaite, A. & Suzumura, K. (1990), ‘Strategic information revelation’, *Review of Economic Studies* **57**(1), 25–47.
- Pedersen, T. (1991), Non-interactive and information-theoretic secure verifiable secret sharing, in ‘Proc. of Crypto 1991; Lecture Notes in Computer Science Vol. 576’, Springer-Verlag, pp. 129–140.
- Pitchford, R. & Snyder, C. M. (2004), ‘A solution to the hold-up problem involving gradual investment’, *Journal of Economic Theory* **114**(1), 88–103.
- Sandholm, T. (1996), Negotiation among self-interested computationally limited agents, PhD thesis, University of Massachusetts Amherst.
- Sandholm, T. (1997), ‘Unenforced e-commerce transactions’, *IEEE Internet Computing* **1**(6), 47–54.
- Schoenmakers, B. (2005), Interval proofs revisited, in ‘International Workshop on Frontiers in Electronic Elections’, Milan.
- Stiglitz, J. E. (2000), ‘The contributions of the economics of information to twentieth century economics’, *The Quarterly Journal of Economics* **115**(4), 1441–1478.
- Syverson, P. (1998), Weakly secret bit commitment: Applications to lotteries and fair exchange, in ‘Proc. 11th Computer Security Foundations Workshop’, IEEE Computer Society Press, pp. 2–13.

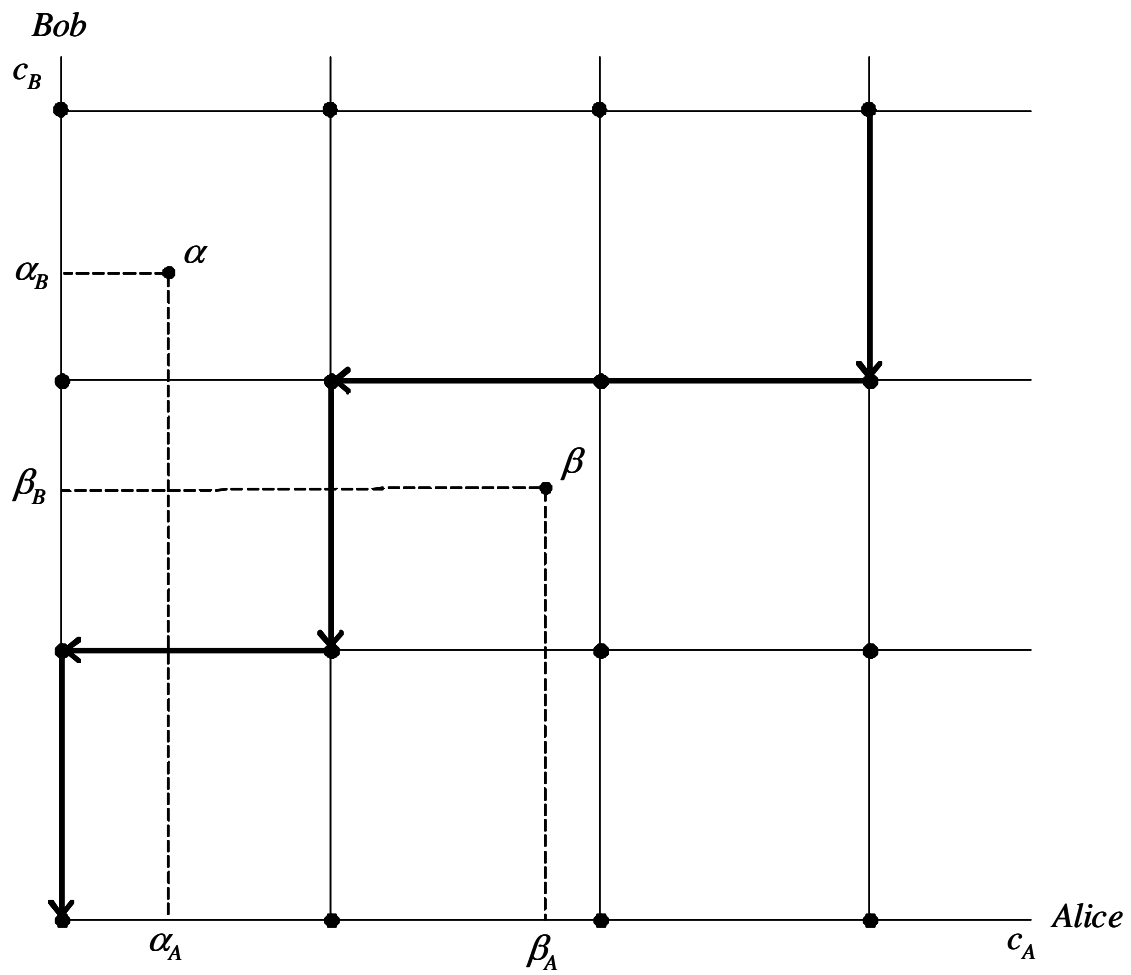


Figure 1: The exchange path must lie in the grid  $G$ .

