Int. J. Inf. Secur. (2008) 7:85–100
DOI 10.1007/s10207-007-0039-5

SPECIAL ISSUE PAPER

Bayesian rational exchange

Juan M. Estevez-Tapiador · Almudena Alcaide · Julio C. Hernandez-Castro · Arturo Ribagorda

Published online: 10 August 2007

Abstract The notion of rational exchange introduced by Syverson (Proceedings of the 11th IEEE Computer Society Foundations Workshop, pp. 2–13, 1998) is a particularly interesting alternative when an efficient scheme for fair exchange is required but the use of a trusted third party is not allowed. A rational exchange protocol cannot provide fairness, but it ensures that rational (i.e., self-interested) parties would have no reason to deviate from the protocol. Buttyán et al. (J. Comput. Security 12(3/4), 551–588, 2004) have recently pointed out how rationality in exchange protocols can be formalized and studied within the framework provided by Game Theory. In particular, Buttyan's formal model was used to analyze Syverson's rational exchange protocol. In this paper, we identify a series of drawbacks in Buttyan's model which make it somewhat restrictive and unrealistic. We propose an extension to the model which enables us to consider different classes of protocol parties (e.g., honest and dishonest parties), as well as modeling attributes such as reputation or any other participant beliefs that could have an effect on the protocol outcome. The resulting new model enables us to reason rational exchange protocols from the point of view of Bayesian rationality, a notion that may be in some scenarios more appropriate than that defined in terms of Nash equilibrium.

J. M. Estevez-Tapiador (✉) · A. Alcaide ·
J. C. Hernandez-Castro · A. Ribagorda
Computer Science Department, Carlos III University of Madrid,
Avda. Universidad 30, 28911, Leganés, Madrid, Spain
e-mail: jestevez@inf.uc3m.es

A. Alcaide
e-mail: aalcaide@inf.uc3m.es

J. C. Hernandez-Castro
e-mail: jcesar@inf.uc3m.es

A. Ribagorda
e-mail: arturo@inf.uc3m.es

Keywords Rational exchange · Syverson's protocol · Cryptographic protocols · Automated validation · Game theory

1 Introduction

The problem of how to design a general procedure for two parties to exchange items in a *fair* manner has attracted much attention lately. Interest in this class of protocols stems from its importance in many applications that rely on electronic transactions where disputes among parties can take place. Examples of these applications include digital contract signing, certified e-mail, exchange of digital goods and payment, etc. In particular, assurance of fairness is fundamental when the exchanged items include any kind of evidences of non-repudiation, for this constitutes a key service in most of the previously mentioned applications. As a result, fair non-repudiation has experienced an explosion of proposals in recent years (see [13] for a survey on non-repudiation protocols).

Roughly, the property of fairness means that no party should reach the end of the protocol in a disadvantageous position, e.g., having sent her item without having received anything valuable in return [4]. Interested readers can find an introductory treatment of the fair-exchange problem in [16].

Formally, there is no protocol according to which a number of parties can exchange items in a fair manner, exclusively by themselves, and assuming that misbehaving parties can take part in the protocol. To the best of our knowledge, the first result slightly related with the impossibility of fairness was provided by Even and Yacobi in [9]. Pagnia and Gärtner [15] provide a formal treatment of this problem in, with an approach based on establishing a relation between the fair

exchange problem and decision problems. In particular, they first argued that (strong) fair exchange can be viewed as an instance of a decision problem. Subsequently, it is proven that fair exchange is at least as hard to solve as the consensus problem. This is a remarkable result, since Fischer et al. [10] showed in 1985 that distributed consensus with one faulty process is impossible. The underlying idea can be intuitively sketched avoiding the technical details: during the protocol execution, one of the parties has eventually to go first in providing her item to the other party. At that point, she incurs in a unfair condition of which a misbehaving party can take advantage.

The simplest protocol that can provide true fairness relies on the use of a trusted third party (TTP). The role of the TTP varies from one class of fair-exchange protocols to another according to its involvement. In schemes based on an *in-line* TTP (e.g., [5]), this acts as a delivery authority involved in each message between parties. The main drawback of these schemes rely on the high involvement of the TTP, which can become a bottleneck.

To avoid the inefficiency of the previous schemes, some authors proposed the use of an *on-line* TTP (e.g., [1]). Here, the TTP is involved during each protocol execution, but not necessarily in every message exchanged between parties. A third step towards reducing the role of the TTP was the introduction of *off-line* TTPs (see [2,3]). In these protocols, sometimes referred to as *optimistic fair exchange*, parties try to carry out the exchange by themselves, and only appeal to the TTP in case of misbehavior of a dishonest party, or whenever a failure occurs during the protocol execution.

Recent computing paradigms, such as ad hoc and peer-to-peer networks, pose a challenge from the point of view of the security mechanisms that should be applied. In many cases, the operation of these systems rely on a complete lack of fixed infrastructure. Generally, it is not realistic to assume that services such as those provided by a TTP will be available in these environments. In fact, in extreme cases we would probably be obliged to renounce to properties such as strong fairness. It is precisely in this context where notions such as *rationality* becomes particularly interesting. This concept, widely known by game theorists, was first applied to security protocols by Syverson in [17]. Informally, a rational exchange protocol cannot provide fairness, but it ensures that rational (i.e., self-interested) parties would have no reason to deviate from the protocol. In other words, a rational exchange protocol ensures that misbehavior is not worthwhile.

Buttyán et al. [8] introduced in a mathematical model based on Game Theory under which rationality can be formally defined, and properties of rational exchange protocols can be analyzed. As an example, the model is used to analyze Syverson's rational exchange protocol [17]. We have found that some aspects of this model seem too restrictive and unrealistic. Additionally, we have also identified some vulnerabilities in Syverson's protocol which were not detected by Buttyán et al.'s analysis. These facts motivated us to outline a few enhancements to the formal model and also to extend some other features by introducing the notion of Bayesian rational exchange. A Bayesian rational exchange will be defined in terms of the perfect Bayesian equilibrium of a dynamic game of incomplete information. Our model enables us to consider different classes of protocol parties within the protocol game (e.g., honest and dishonest parties) as well as modeling attributes such as reputation or any other belief that could have an effect on the protocol outcome. The resulting new model let us reason about exchange protocols from the point of view of Bayesian rationality, a notion that may be in some scenarios more appropriate than that defined in terms of Nash equilibrium.

The rest of this paper is organized as follows. Section 2 provides a brief overview of how Game Theory can be applied to formalize and analyze exchange protocols. Subsequently, Syverson's rational exchange protocol is introduced and discussed in Sect. 3. In Sect. 4, we identify a number of limitations in Buttyán et al's model which motivate us to propose an enhanced model. In Sect. 5, we present a new formal tool for automatically reasoning about exchange protocols: dynamic games of incomplete information and their associated perfect Bayesian equilibrium. Section 6 serves to illustrate how this framework can be used to analyze Syverson's protocol from an enhanced point of view. Finally, we summarize the most relevant conclusions of this work in Sect. 7.

2 Game theory and protocol games

For readability and completeness, we first provide a brief review of the game-theoretical model of rational exchange introduced by Buttyán et al. in [8]. Since our work is intimately related to this model, where possible we will adopt the same notation.

An extensive-form game is defined by the tuple:

$$\langle P, A, Q, p, (\mathcal{I}_i)_{i\in P}, (\preceq_i)_{i\in P}\rangle \tag{1}$$

where:

- P is a set of players.
- A is a set of actions.
- Q is a set of action sequences, satisfying:
  - $\epsilon \in Q$, where ϵ is the empty sequence.
  - if $(a_k)_{k=1}^{w} \in Q$ and $0 < v < w$, then $(a_k)_{k=1}^{v} \in Q$
  - if $(a_k)_{k=1}^{v} \in Q\ \forall v \geq 1$, then $(a_k)_{k=1}^{\infty} \in Q$

  If q is a sequence of actions and a is an action, then $q.a$ denotes the action composed by q followed by a. A finite sequence of actions $q \in Q$ is said to be *terminal* if there is no a such that $q.a \in Q$. The set of terminal sequences of actions is denoted by Z. Finally,

$A(q) = \{a \in A : q.a \in Q\}$ denotes the set of available actions after $q \in Q \backslash Z$.

- p is the player function. It assigns a player $p(q) \in P$ to every non-terminal sequence $q \in Q \backslash Z$. The interpretation is that player $p(q)$ has the turn after the sequence of actions q.
- $\mathcal{I}_i$ is an information partition for player $i \in P$. It is a partition of the set $\{q \in Q \backslash Z : p(q) = i\}$ preserving the property that if the sequences q and q' are in the same information set $I_i \in \mathcal{I}_i$, then $A(q) = A(q')$.
- $\preceq_i$ is a preference relation of player $i \in P$ on Z.

The common interpretation of an extensive-form game is the following. The game can be thought of as a tree, where the edges and the vertices are associated to actions and sequences of actions, respectively. The empty sequence ϵ represents the root of the tree. The game begins at ϵ and ends at a terminal node. After any non-terminal sequence of actions $q \in Q \backslash Z$, the player given by $p(q)$ chooses an available action from the set $A(q)$. Next, q is extended with a, and the current history of the game becomes $q.a$. Terminal vertices are those that cannot be followed by any other actions. When a sequence of actions q reaches a terminal vertex, the game ends.

The sequences $q \in Z$ are the possible outcomes of the game. The preference relations $\preceq_i$ establishes which outcomes are preferred by player i. Thus, if $q, q' \in Z$ and $q \preceq_i q'$, then player i prefers q' to q. The most usual form of representing preference relations are payoffs. A vector $y(q) = (y_i(q))_{i \in P}$ of real numbers is assigned to every terminal sequence of actions $q \in Z$, in such a way that $q \preceq_i q' \Leftrightarrow y_i(q) \leq y_i(q')$. The value $y_i(q)$ can be interpreted as a measure of how much player i gains when the game is developed as described by q.

Information sets represent the information available to players at every stage of the game. When an information set covers several nodes, then the player does not know in which part of the tree she is—or, equivalently, she does not know the last action of her rival. Usually, nodes belonging to the same information set are graphically represented by a dashed line that links them together. When an information set is not a singleton (i.e., it has more than one node), it is necessary to specify the player beliefs. Formally, beliefs are represented by a probability distribution over the nodes belonging to the information sets.

If there exists at least one information set $I_i \in \mathcal{I}_i$ such that $|I_i| > 1$, then the game is called a game of imperfect information. On the contrary, if for all players every information set is a singleton, then the game is called a game of perfect information.

2.1 Strategies and Nash equilibrium

A strategy for player i is a function s_i specifying what action i should carry out at each of her information sets. A strategy s_i assigns an action $A(q)$ to each non-terminal sequence of actions q, preserving that if q and q' are in the same information set of player i, then s_i assigns the same action to both sequences. The set of all strategies of player i is denoted by S_i.

A strategy profile is a vector $(s_i)_{i \in P}$ of individual strategies, one for each player. Specifying completely a strategy profile determines univocally the outcome of the game. Sometimes we will denote by $(s_j, (s_i)_{i \in P \backslash \{j\}})$ a strategy profile, in order to emphasize the specification of strategy s_j for player j.

Let $o((s_i)_{i \in P})$ denote the outcome (i.e., the sequence of actions) of a game when players follow the strategies given by the strategy profile $(s_i)_{i \in P}$. A strategy profile $(s_i^*)_{i \in P}$ is a Nash equilibrium iff for every player $j \in P$ and every strategy $s_j \in S_j$, it holds that:

$$o\left(s_j, (s_i^*)_{i \in P \backslash \{j\}}\right) \preceq_j o\left(s_j^*, (s_i^*)_{i \in P \backslash \{j\}}\right) \tag{2}$$

Intuitively, if every player $i \neq j$ follows the strategy s_i^*, then player j should not deviate from s_j^*, for she does not gain anything by doing so. In general, it is possible for a game to have multiple Nash equilibria.

2.2 Protocol games

The protocol game of an exchange protocol is intended to model all possible interactions of the protocol participants, even the potentially misbehaving actions (i.e., those different from the ones prescribed by the protocol). A protocol game is constructed from the protocol description. Each of the parties involved in the protocol becomes a player of the protocol game, including the network. The network is considered to be reliable, which means that it delivers messages to their intended destinations within a constant time interval. Therefore, it has only one fixed strategy, consisting of delivering messages to players. The rest of the participants have the strategies to quit, do nothing, send a message following the protocol steps or send a message deviating from the steps of the original protocol. Each player can send messages which have been defined as *compatible* with the protocol, this is, messages which are within the context of the protocol. The set M_π of compatible messages with a protocol π is formally defined within the model. Although the participants can alter the order in which those messages are sent, the model does not allow the protocol parties to run multiple instances of the protocol in parallel (i.e., they do not consider interleaving attacks), neither to eavesdrop nor modify messages sent between other parties of the protocol. In general, no other limitation is imposed on the participants.

Information sets for players are defined in terms of their local state. Formally, $\Sigma_i(q)$ denotes the information that player i has obtained after the sequence of actions q.

Buttyán et al. [8] formally define both the structure of such sets, as well as the way they are updated according to the actions observed during the game.

Finally, we will describe how the *preference relation* $(\preceq_i)_{i\in P}$ or, equivalently, the payoff function, is established. In the simpler case (a two-party protocol), we consider the two players p_1 and p_2 and the items γ_{p_1} and γ_{p_2} that they want to exchange. The values that γ_{p_1} is worth to p_1 and p_2 are denoted by $u^-_{p_1}$ and $u^+_{p_2}$, respectively. Likewise, the values that γ_{p_2} is worth to p_1 and p_2 are denoted by $u^+_{p_1}$ and $u^-_{p_2}$, respectively. In this way, the values u^+_i and u^-_i can be viewed as the potential gain and loss of player $i \in \{p_1, p_2\}$ in the game.

When the protocol game is over, every participant can assess the profit or the loss they have incurred by using a payoff function. This function takes the local state of every participant at the time the game is over and calculates an outcome value (the highest profit represents the most preferable protocol outcome). Buttyán et al introduce this concept in their model as follows. Given a terminal sequence of actions q, the payoff function for player i is defined as $y_i(q) = y^+_i(q) - y^-_i(q)$, where functions $y^+_i(q)$ and $y^-_i(q)$ represent the gain and the loss player i has incurred, respectively. In general, these functions can be defined as follows:

$$y^{\oplus}_i(q) = \begin{cases} u^{\oplus}_i & \text{if } \phi^{\oplus}_i(q) = \texttt{true} \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where $\oplus \in \{+, -\}$. The purpose of boolean functions $\phi^{\oplus}_i(q)$ is to capture those conditions under which each partner gains/losses control over the items. Thus, $\phi^+_i(q) = \texttt{true} \Leftrightarrow$ player i gains access to γ_j, and $\phi^-_i(q) = \texttt{true} \Leftrightarrow$ player i losses control over γ_i.

2.3 Protocol properties

Informally, a two-party rational exchange protocol is an exchange protocol in which both main parties are motivated to behave correctly and to follow the protocol faithfully. If one of the parties deviates from the protocol, then she may bring the other, correctly behaving party to a disadvantageous situation, but she cannot gain any advantages by her misbehavior. Buttyán et al define the concept of rationality in terms of a Nash equilibrium in the protocol game. It is required that the strategies that correspond to the behavior described by the protocol form a Nash equilibrium in the protocol game and that no other Nash equilibrium is strongly preferable for any of the participants.

Other properties, described below, such as fairness, effectiveness, termination, gains closed, and safe back-out are also formally defined within the model. In particular, the proof of the protocol rationality relies on the fact that it must be closed for gains and it must also satisfy the safe back-out property. In our opinion the model is consistent and correct, even though, as we will see, it is easy to step out of it and break that way the rationality property. It is also possible to break the closed for gains property to attack rationality.

$$\begin{array}{ll} A \rightarrow B: & m_1 = (desc_{item_A}, enc(k, item_A), w(k), \sigma_1) \\ B \rightarrow A: & m_2 = (item_B, m_1, \sigma_2) \\ A \rightarrow B: & m_3 = (k, m_2, \sigma_3) \end{array}$$

where:

$$\begin{aligned} \sigma_1 &= sig(k_A^{-1}, (desc_{item_A}, enc(k, item_A), w(k))) \\ \sigma_2 &= sig(k_B^{-1}, (item_B, m_1)) \\ \sigma_3 &= sig(k_A^{-1}, (k, m_2)) \end{aligned}$$

Fig. 1 Syverson's rational exchange protocol

The closed for gains property is satisfied when for every possible outcome of the game q, it holds that $y^+_A(q) \geq 0 \Rightarrow y^-_B(q) \geq 0$ and $y^+_B(q) \geq 0 \Rightarrow y^-_A(q) \geq 0$. Put simply, this property establishes that if a party A gains access to an item of the other party B ($y^+_A(q) \geq 0$), then B must lose control over the same item ($y^-_B(q) \geq 0$), and vice versa.

The safe back-out property is satisfied when for every possible sequence of actions q, if party A's strategy was to always quit, then A loses nothing by doing so (i.e., $y^-_A(q) = 0$). In the same way, if B's strategy is to always quit, then $y^-_B(q) = 0$.

3 Syverson's rational exchange protocol

The Syverson's protocol was proposed in [17] and analyzed by Buttyán et al. in [8] using their definition of rationality within the game-theoretical model introduced in the same work. Fig. 1 illustrates the protocol.

A and B denote the two protocol parties, with private keys k_A^{-1} and k_B^{-1}, respectively. The protocol assumes that $item_A$ and $item_B$ are the items they would like to exchange, being $desc_{item_A}$ a description of $item_A$. (There is no equivalent description for $item_B$ because the scheme was introduced to serve as a payment protocol, in such a way that $item_B$ has the role of the payment for buying $item_A$). Moreover, $enc(k, m)$ is a symmetric encryption algorithm that encrypts message m with key k. Likewise, $sig(k_i^{-1}, m)$ provides a digital signature on m by using secret key k_i^{-1}. Finally, $w(\cdot)$ is a temporarily secret commitment function (see [17]). For our analysis, it suffices to know that $w(m)$ keeps m secret, but it can be broken in acceptable bounds on time.

The context in which the protocol is executed should be carefully checked. Syverson identifies scenarios where such a protocol could be used for: (1) If the vendor A is selling relatively low value items, so it is not worth for the customer (in terms of computational cost or the inconvenience of delay) to break the encryption to recover the item; (2) the vendor A might be selling some item that might be of timely and diminishing value, such as short term investment advice or regularly changing lists of bargain items for sale; (3) the

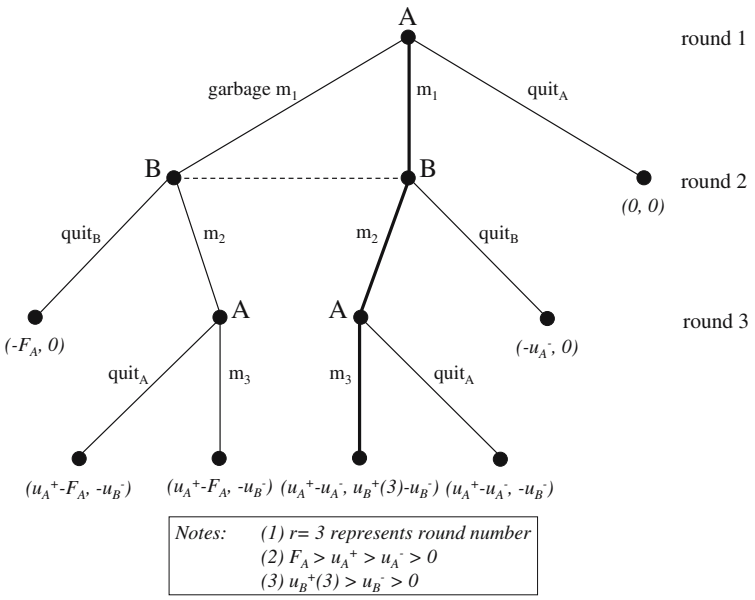

Fig. 2 Partial representation of Syverson's protocol game in extensive form

protocol might begin one step earlier with a signed customer request for $item_A$. The vendor A can then assume the risk of trading with unknown customers and refuse to service customers who repeatedly fail to pay.

3.1 Overview of the protocol analysis

We will only describe those aspects of Buttyán's protocol analysis indispensable to motivate the discussion below. Please refer to [8] for a detailed description of the whole analysis. We have chosen to represent the Syverson's protocol game in an extensive form (see Fig. 2). The tree represents the different moves each participant can make and the possible outcomes. The vectors assigned to each terminal node represent the outcome for A (first value) and B (second value) when A and B follow the path of actions to terminate the protocol at that final node.

The values u_A^+ and u_B^+ denote the values that $item_B$ and $item_A$ are worth to parties A and B respectively. In a similar way, the values u_A^- and u_B^- denote the values that $item_A$ and $item_B$ are worth to A and B, respectively. The diminishing value of $item_A$ for B is modeled as a function $u_B^+(r)$, which decreases as the round number r increases. Note that A loses control over $item_A$ and $u_A^- > 0$ when A sends m_1. If this wasn't the case, Syverson's protocol would not be the most appropriate protocol to use in the exchange. Finally, punishment for misbehavior is denoted by F_A.

We have highlighted in the tree the strategy profile for A and B which would result in a rational exchange of items $item_A$ and $item_B$. This profile constitutes a Nash equilibrium so, by definition, neither of the players would want to deviate from such strategies. In this regard, Buttyán et al's model serves to formally prove that Syverson's protocol is rational and that two rational—selfinterested—protocol participants would terminate the protocol having exchanged their items.

4 Weaknesses of Buttyán et al.'s model

In our opinion, Buttyán et al.'s model has as major contribution the identification of Game Theory as an appropriate framework in which rational exchange protocols can be formally analyzed. However, we have identified a few drawbacks in its definition which lead to a very restricted model. The first two observations are related to the way participants actions and behavior are modeled, i.e., the computational model according to which messages are analyzed. Furthermore, we find basic Game Theory as a too narrow formalism for real-world protocols. In particular, we strongly believe that uncertainty plays a major role in rational exchange and, therefore, it should be somehow incorporated into the reasoning model. Despite the main contribution of this paper is related to this last issue, it is first necessary to amend the flaws encountered in the original model before introducing any extension.

4.1 Flaws in local computations and history records

In Appendix 7, we describe a number of vulnerabilities in Syverson's protocol that allow to mount several attacks against the scheme. The reason why these flaws were not detected by Buttyán et al.'s analysis are not related to the reasoning model itself (i.e., Game Theory), but to a inherent hypothesis: the model assumes that messages are well constructed from a security point of view. In other words, in the process of formalizing the participants local actions,

the logic applied is limited to message compatibility and not message content. We further elaborate on this topic in what follows.

In Buttyán et al.'s model, an exchange protocol is considered to describe a set of local computations Π_j, one for each participant j of the protocol. Typically, each program Π_j contains instructions to wait for messages that satisfy certain conditions or to generate events such as to send a message m to a given participant p. When the program Π_B is described for entity B, no test is defined to determine whether m_1 is an old message being reused. In this regard, the model assumes that evidences are fresh and well constructed, thus failing to reflect the actual content of message m_1 as it is described in the protocol. This erroneous assumption is subsequently reflected in the protocol game. As we can see in Fig. 2, the value of F_A represents the penalty A has to pay when proved to be the *author* and *sender* of a forged message m_1. However, there is nothing in m_1 ensuring that A is *both* its author and its sender. The result is that the formal notation and structures used do not allow entities to verify essential properties such as freshness of a message, originator, sender or intended receiver.

Furthermore, it is specified that each player creates a history record of all the events that were generated by her and the round number of their generation. Possible entries in the history record file of protocol participant A would be $\mathsf{send}(m_1, B)$ or $\mathsf{rcv}(m_2)$, in round r. Based on the entries stored in this record, each player is or not allowed to send a particular compatible message. For instance, a valid digital signature $sig(k_A^{-1}, m)$ can only be generated by A. Therefore, B can send a message containing $sig(k_A^{-1}, m)$ iff B received a message containing $sig(k_A^{-1}, m)$ *earlier* in the current protocol execution.

As the model was defined, this history record is newly created for each protocol run, so information received in previous runs is discarded at the end of each execution. However, any (malicious) participant of the protocol could have compatible messages from previous runs and will be able to use them, as they are perfectly compatible with the protocol. In particular, attacks 1 and 2 described in Appendix 7 are based in the use of evidences obtained in different protocol runs.

Summarizing, there are two aspects of the model that must be improved. First, the construction of the protocol game (denoted G_π by Buttyán et al.) must take into account the fact that messages could not be well constructed. For this, a number of additional tests on messages have to be carried out, apart from those aimed at ensuring compatibility with the current protocol round. Second, these security checks must embrace messages received in previous runs, and not only those corresponding to the current protocol execution.

Improving the model with respect to the above aspects is out of the scope of this paper. For our purposes, we will use the corrected version of Syverson's protocol described at the end of Appendix 7.

4.2 Limitations in expressiveness

In the previous section, we have pointed out some weaknesses of Buttyán et al's model and suggested how it can be corrected to overcome them. However, there is a fundamental aspect that remains unaddressed by this formalism: the lack of support for dealing with uncertainty, especially to model contextual information. Due to their very nature, uncertainty plays a major role in rational protocols. Not in vain, it has been stated in several occasions that the *context* in which the protocol is to be executed should be checked carefully. Roughly, this means that some environmental factors, such as how much we trust other(s) participant(s) or the degree of reliability of the network, should be in some way incorporated into the security analysis.

4.2.1 Example 1

To illustrate this idea, consider the following example. As we can see in Fig. 2, A is not motivated to be fair to B in the last round of the protocol. Therefore, A could threaten B to execute $quit_A$ or to delay sending m_3 to B. B would then be safer quitting the protocol before round 2 and aborting the exchange. The best response that A can give to B's quit strategy is to quit as well. Therefore $s^* = (quit_A, quit_B)$ is also a Nash equilibrium in the protocol game of the model, which could be the most preferable protocol outcome under threaten situations. In order to resolve this issue, A should have an incentive to be fair to B in the last round of the protocol, so the exchange takes place when the set of strategies executed constitute a perfect Nash equilibrium (Nash equilibrium in each of the subgames). This incentive may be a kind of "reputation factor", securely managed by external parties, publicly known and which would have an effect on entity A's payoff function.

Suppose now that, in the past, A has been honest (i.e., she has sent m_3 at step 3) in the 75% of the exchanges performed with her. How can this information be taken into account to decide whether to initiate a new protocol run with A or not? Surely, it will depend not only on this experience, but also on the values that the items are worth to both parties. For instance, if the item is very important to us, we could run the risk of exchanging it with an unpleasant party). In any case, in Buttyán et al.'s model, the protocol participants could not bring their past experience into the current protocol instance. We found this inappropriate and unrealistic.

4.2.2 Example 2

In Buttyán et al.'s model—as in many others reasoning models, the network is considered an additional participant of the protocol. However, its behavior is limited to always deliver messages. Even though a reliable network can be assumed in many circumstances, it would be also interesting to count on a formalism in which more complex behaviors can be analyzed.

This feature is particularly relevant in the case of rational exchange protocols, for their most probable execution environments could be hostile to, at least, one of the participants. For instance, in a mobile ad hoc network, two devices that are not in each other's range must rely on intermediary nodes to carry out a multi-hop communication. In this case, assuming that each node that compose the network will behave well is a hypothesis that simplifies the analysis, but it is unrealistic.

Apart from modeling the actions that the network can perform as a participant (e.g., deliver or not deliver messages), it would be also informative to take into account *beliefs* about its behavior. This would allow us to distinguish between a network which is highly reliable (e.g., 99% of the messages are properly delivered) and a network which is highly non-reliable.

4.2.3 Overview of our approach

Fortunately, Game Theory provides us with a solid body of knowledge capable of modeling features such as those discussed above. In particular, the so-called *games of incomplete information*, or *Bayesian games*, are those in which some player does not know some parameter of the game they are playing. In this type of games, player's beliefs over other participants's real nature, past experiences, reputation factors, etc. can be taken into account when making the optimal decision at any given point during the protocol execution.

Several Game Theory results allow us to predict the outcome of such a game, and therefore the outcome of the protocol execution. Despite the analysis becomes now more complex than by using basic Game Theory, we find it more realistic and more informative. In the rest of the paper, we introduce this proposal.

5 A model based on dynamic games of incomplete information

In the following, we briefly introduce the two most relevant concepts in Bayesian games: player's type and player's beliefs. Subsequently, we discuss the notion of perfect Bayesian equilibrium—a generalization of Nash equilibrium for this kind of games. Throughout this section, we will discuss how this extension of basic Game Theory can be incorporated into the formalism to analyze rational-exchange protocols in a most powerful manner.

5.1 Player's type

Following John C. Harsanyi's framework, a way of modeling uncertainty in a game is by introducing the notion of a player's type [12]. The type of a player determines univocally that player's payoff function, being perfectly possible that different types will be associated with different payoff functions.

The following definitions formalize this concept:

Definition 1 (*player's type and type space*) We will assume that each player $i \in P$ has a type $T_i \in \mathcal{T}_i$, where $\mathcal{T}_i$ is the type space for player i.

Definition 2 (*type profile*) A type profile is a tuple of types $T = (T_1, T_2, \ldots, T_n)$, one for each player, which univocally determines the type of every player involved in a specific game. We denote by $\mathcal{T} = \mathcal{T}_1 \times \cdots \times \mathcal{T}_n$ the type-profile space.

The main advantage of this formalism is that it allows us to introduce diverse forms of uncertainty in the analysis. For instance, in scenarios where a protocol cannot guarantee true fairness, trust among parties plays a significant role and it should be taken into account within the reasoning. In the simplest case, we can conceive that players can be either *honest* or *dishonest*, thus having a type space with just two elements. Of course, executing a protocol against a dishonest party will surely be different that dealing with an honest one. This fact is modeled by means of different payoff functions for each type, in such a way that the strategies to follow during the protocol run will be different in each case.

Note that games of incomplete information are not limited to a discrete type-space profile. In fact, we can conceive continuous type spaces of the general form $\mathcal{T}_i \subseteq \mathbb{R}$. This can be useful to generalize the notion of honesty discussed before: We can assign a type $T_i \in [0, 1]$ to each player i, which can be viewed as her *reputation*.

Types can be also useful for considering features regarding network behavior. In Buttyán et al's model, the network is considered reliable, so it has a fixed strategy: to deliver messages. However, it is not difficult to conceive a more realistic scenario in which messages are not always properly delivered. In this case, the effects of an unreliable network can be easily incorporated into the analysis.

5.2 Player's belief system

In a Bayesian game, the incompleteness of information means that each player does not know the type of the rest of the players with complete certainty. As a result, players have

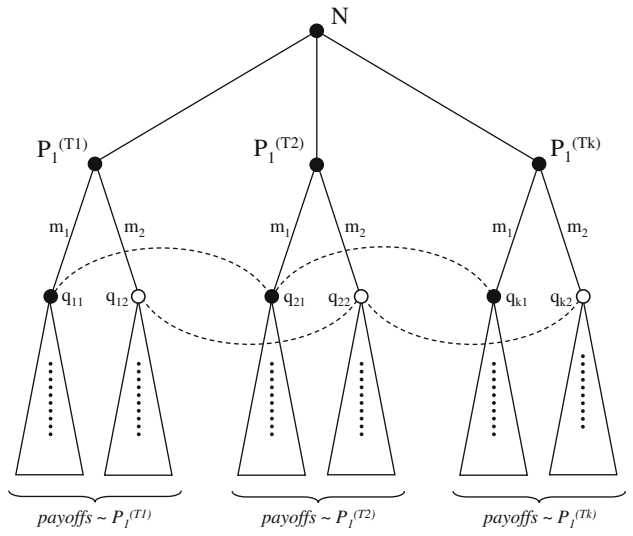

Fig. 3 Illustration of a game of incomplete information

initial beliefs about the type of each player, and can update them according to Bayes' Rule as the game advances and more information is available (e.g., beliefs about a player can change on the basis of the actions she have played).

Next definition formalizes this notion. ($\Delta(X)$ denotes the space of probability distributions over the set X).

Definition 3 (*belief system*) The belief that a player has about the type of player $j \in P$ is represented by a probability distribution over j's type-space $\Delta(\mathcal{T}_j)$. In general, we will denote each belief by a Greek letter $\alpha(\cdot), \beta(\cdot), \ldots$ The set of all beliefs will be termed ρ.

Since each type is associated to a (possibly different) payoff function, uncertainty about the opponent's type has severe implications in the decision-making process, particularly in computing the best strategy during the protocol execution.

In practice, assignment of types to players is carried out by introducing a fictitious player: Nature. In the course of the game, nature randomly chooses a type for each player according to a probability distribution over each player's type space. A typical scenario in a dynamic game of incomplete information is graphically illustrated by Fig. 3, in which $P_i^{T_j}$ means that player P_i has type T_j. Initially, nature (N) "moves", thus selecting a type for player P_1, who moves either m_1 or m_2. Next, it is the turn for player P_2. Despite P_2 has observed the move performed by P_1, she is not sure about P_1's type. Formally, $I_1 = \{q_{11}, q_{21}, \ldots, q_{k1}\}$ and $I_2 = \{q_{12}, q_{22}, \ldots, q_{k2}\}$ form disjoint information sets for P_2, since she knows whether she is in I_1 or I_2, but she does not know the specific node.

From this point on, P_2 will have to include into her analysis a belief system, i.e., the probability of P_1 being of type T_1, T_2, etc. Therefore, strategies must take into account not only the several ways in which other players can play, but also the probability for each player to be of a specific type.

5.3 Perfect Bayesian equilibrium

In 1991, Fudenberg and Tirole [11] formally defined the perfect Bayesian equilibrium (PBE) for extensive Bayesian games. PBE adds to Nash equilibrium the requirement that players choose optimally given their beliefs about the rest of the game. In extensive Bayesian games of incomplete information, each player is not only aware of the informational uncertainties over the other participants, but also analyzes their implications. Thus, each player looks for the best response, anticipating other party's reaction.

We will formally define candidates to be a PBE in an extensive-form game as a pairing strategy-believe profile $(S; \rho)$ such that:

- The profile $(S; \rho)$ does not only represent a Bayesian equilibrium in the whole game, but also in each of the continuation subgames. This is, from each information set, the moving player's strategy maximizes its expected payoff for the remainder of the game, considering its beliefs and all player's strategies.
- On the equilibrium path, Bayes' rule and equilibrium strategies determine believes. An information set is on-the-equilibrium path if, it will be reached with positive probability iff the game is played according to the equilibrium strategies.
- Off-the-equilibrium path, Bayes' rule and equilibrium strategies determine beliefs where possible. A defection from the equilibrium path does not increase the chance that others will play irrationally.

The stated $(S; \rho)$ profile would describe a set of strategies such that for every player $i \in P$ and every information set $I_i \in \mathcal{I}_i$, player's i strategy is her best response, given her believes at set $\mathcal{I}_i$.

Before formally defining the concept of Bayesian perfect equilibrium, a series of requirements are necessary [12].

Definition 4 (*Bayes requirement 1*) Given a strategy profile S, it is required that, for each player $i \in P$, and at each of her information sets $I_i \in \mathcal{I}_i$, player i has beliefs $\rho(I_i) \in \Delta(I_i)$ about the node at which she is located, conditional upon being informed that play has reached the information set I_i.

This requirement establishes that at every node of any information set, a player should have some beliefs about the node at which she is located, given that she has reached that information set. In fact, the beliefs $\rho(I_i) \in \Delta(I_i)$ are no more than a probability distribution over the nodes in I_i.

Definition 5 (*Bayes requirement 2*) Let us suppose the continuation game defined by the information set $I_i \in \mathcal{I}_i$ of some

player i, and the conditional beliefs $\rho_i(I_i)$. The restriction of the strategy-belief profile $(S; \rho)$ to this game must be a Nash equilibrium of the continuation game starting at information set I_i.

The concept of perfect equilibrium in the continuation game adds to the Nash concept the requirement that players choose optimally in continuation games. More generally, Bayes requirement 2 rejects all strategy profiles which specify at any information set an action which is dominated at that information set.

Definition 6 (*Bayes requirement 3*) Beliefs at any on the equilibrium path information sets must be determined from the strategy profile according to Bayes' rule. This is, if $I_i \in \mathcal{I}_i$ is an information set of player i reached with positive probability when players follow strategy profile S, then $\rho(I_i) \in \Delta(I_i)$ must be computed from S according to Bayes' rule.

Definition 7 (*Bayes requirement 4*) The beliefs at any off the equilibrium path information set must be determined from the strategy profile according to Bayesć6 rule whenever possible.

This requirement establishes that a defection from the equilibrium path does not increase the chance that others will play irrationally. In a PBE, players cannot threaten to play strategies that are strictly dominated beginning at any information set, off the equilibrium path.

Definition 8 (*perfect Bayesian equilibrium*) Given a strategy profile S and a set of beliefs ρ, then the strategy-belief profile $(S; \rho)$ forms a perfect Bayesian equilibrium iff it satisfies Bayes requirements 1 to 4.

6 A Bayesian analysis of Syverson's protocol

To illustrate the application of the proposed Bayesian model, we will analyze Syverson's protocol, though assuming a richer set of environmental hypotheses. We note that in such a protocol entity B always plays a more risky strategy than player A. Entity B is the first in sending her item, $item_B$, hoping that A would not misbehave and that the exchange would take place successfully. Moreover, it is very difficult for B to be able to misbehave, as m_2 is always a *fresh* message (contains m_1) and it is also a token for non-repudiation of origin, so B will always be proven responsible for any malicious action.

By contrast, A could not be held responsible for any malicious action until the end of protocol and, in most cases, provided A issues the proof of such misbehavior. Furthermore, A gains access to $item_B$ before B gains access to $item_A$. It is B then, who has to carefully analyze A's reputation, credibility, surroundings, and her current and real intentions. Therefore, we will only consider the simplest scenario: that in which participant A can be either an honest or a dishonest protocol party, while B is always honest.

6.1 Player and types

As in a regular game, we consider $P = \{A, B, Net\}$ as the set of players. However, since the network always plays following a fixed strategy (the network is considered to be reliable), we only develop specifications for players A and B.

Let $\mathcal{T} = \mathcal{T}_A \times \mathcal{T}_B$ be the type-profile space, where $\mathcal{T}_A = \{A_h, A_d\}$ and $\mathcal{T}_B = \{B_h\}$ are the type spaces for players A and B, respectively. By convention, subscript h denotes an honest participant, while d represents a dishonest one.

We consider the following probability distribution θ over $\mathcal{T}_A$:

$$\begin{aligned} &\theta_h = prob(A_h|B) \\ &\theta_d = prob(A_d|B) \\ &\text{s.t. } \theta_h + \theta_d = 1 \end{aligned} \tag{4}$$

Note that $prob(B|A_h) = prob(B|A_d) = 1$, since $\mathcal{T}_B$ has only one element.

6.2 Strategies and beliefs

The set of actions available to players is $A = A_A \cup A_B$, where $A_B = \{m_2, quit_B\}$ and $A_A = \{m_1, gm_1, quit_A, m3\}$ are the sets of actions for players B and A, respectively. (Note that gm_1 stands for a garbage message compatible with m_1).

Player B has two possible pure strategies. A pure strategy for player B is $s_B \in S_B = \{quit_B, m_2\}$. Alternatively, a pure strategy for player A is a tuple:

$$s_A \in S_A = \{(s_1, s_3)_d, (s_1, s_3)_h\}$$

where $s_1 \in \{m_1, gm_1, quit_A\}$ defines an action at round 1 of the protocol, and $s_3 \in \{quit_A, m_3\}$ defines an action at round 3. The first component represents a strategy for type A honest and the second one for A dishonest.

A strategy profile in the new Bayesian game is a vector $s = (s_A, s_B)$ of individual strategies, one for each player. Specifying a strategy profile determines univocally the outcome of the game.

The following probability distributions represent the set of beliefs each entity holds over the opponent's set of actions, at each particular stage of the protocol.

At stage 2 of the protocol, B's beliefs are represented by the following probability distribution functions over A's set of actions:

$$\alpha_h, \alpha_d : \mathcal{T}_A \rightarrow \Delta(A_A)$$

satisfying:

$$\alpha_h(gm_1) + \alpha_h(m_1) = 1$$
$$\alpha_d(gm_1) + \alpha_d(m_1) = 1$$

and,

$$\beta_h, \beta_d : \mathcal{T}_A \rightarrow \Delta(A_A)$$

satisfying:

$$\beta_h(quit_A|m_1) + \beta_h(m_3|m_1) = 1$$
$$\beta_d(quit_A|m_1) + \beta_d(m_3|m_1) = 1$$

Note that B also holds the following beliefs representing the fact that, when A has cheated, A will never sign the token of non-repudiation of origin to claim responsibility for such misbehavior; instead, A will always quit the protocol. Therefore:

$$Prob_B[quit_A|gm_1 \wedge A_h] = 1$$
$$Prob_B[m_3|gm_1 \wedge A_h] = 0$$
$$Prob_B[quit_A|gm_1 \wedge A_d] = 1$$
$$Prob_B[m_3|gm_1 \wedge A_d] = 0$$

By contrast, A's attempt to anticipate B's behavior in the game is represented by the probability distribution

$$\sigma_B : \mathcal{T}_B \rightarrow \Delta(A_B)$$

satisfying:

$$\sigma_B(m_2) + \sigma_B(quit_B) = 1$$

6.3 Payoff functions

As stated before, one of the key points of Bayesian games is the fact that each type of player is associated with a possibly different payoff function. We define the following payoff functions:

$$U_A, U_B : \mathcal{T}_A \times A_A \times A_A \times A_B \rightarrow \mathbb{R}$$

so for each branch in the tree, a payoff value is defined representing the total outcome players A and B obtain, when taking such a path (see Fig. 4). For instance:

$$U_A(A_h, m_1, m_3, m_2) = u^+_{A_h} - u^-_{A_h}$$
$$U_B(A_h, m_1, m_3, m_2) = u^+_B(3) - u^-_B$$

6.4 Expected payoffs

We denote by $EP(i, s_i)$ the expected payoff for player i when following strategy s_i. We first consider the expected payoffs when players follow pure strategies. For every strategy profile $s_A = ((s_1, s_3)_d, (s_1, s_3)_h)$ for player A, the expected payoff value is:

$$\begin{aligned} EP(A_h, s_A) &= \sigma_B(m_2) * U_A(A_h, (s_1, s_3)_h, m_2) \\ &\quad + \Big(1 - \sigma_B(m_2)\Big) * U_A(A_h, (s_1, s_3)_h, quit_B) \\ EP(A_d, s_A) &= \sigma_B(m_2) * U_A(A_d, (s_1, s_3)_d, m_2) \\ &\quad + \Big(1 - \sigma_B(m_2)\Big) * U_A(A_d, (s_1, s_3)_d, quit_B) \end{aligned}$$

In the case of player B, we have:

$$\begin{aligned} EP(B, quit_B) &= 0 \\ EP(B, m_2) &= \theta_h * \big[\alpha_h(gm_1) * (-u^-_b) + \alpha_h(m_1) \\ &\quad * \big(\beta_h(quit_A|m1) * (-u^-_B) \\ &\quad + (1 - \beta_h(quit_A|m_1)) * (u^+_B(3) - u^-_B)\big)\big] \\ &\quad + \theta_d * \big[\alpha_d(gm_1) * (-u^-_b) + \alpha_d(m_1) \\ &\quad * \big(\beta_d(quit_A|m1) * (-u^-_B) \\ &\quad + (1 - \beta_d(quit_A|m_1)) * (u^+_B(3) - u^-_B)\big)\big] \\ &= -u^-_B + u^+_B(3) * (\theta_h * \alpha_h(m_1) \\ &\quad * \beta_h(m_3|m_1) + \theta_d * \alpha_d(m_1) * \beta_d(m_3|m_1)) \end{aligned}$$

Now, we will denote:

$$\begin{aligned} L_B &= \theta_h * \alpha_h(m_1) * \beta_h(m_3|m_1) + \theta_d * \alpha_d(m_1) \\ &\quad * \beta_d(m_3|m_1) \end{aligned} \tag{5}$$

Note that we can conclude that $EP(B, m_2) \geq EP(B, quit_B)$ iff:

$$-u^-_B + u^+_B(3) * L_B \geq 0 \tag{6}$$

or, equivalently, iff:

$$L_B \geq u^-_B / u^+_B(3) \tag{7}$$

Therefore, B would play action `send(`m_2, A`)` at round 2 of the protocol, instead of action $quit_B$, iff a linear combination, denoted as L_B, of her set of beliefs verifies the relation given by expression (7).

6.5 PBE candidates

We introduced the notion of strategy-belief profile in Sect. 5.3 as the tuple $(S; \rho)$. Candidates to be PBE (perfect Bayesian equilibrium) in the Syverson's extensive-form game will be of the form strategy-belief profile $(S; \rho)$ where:

$$S = (s_A, s_B) \quad \text{where } s_A \in S_A, s_B \in S_B.$$

and

$$\rho = (\alpha_h, \alpha_d, \beta_h, \beta_d, \sigma_B, \theta)$$

is a tuple containing the probability distribution functions representing the set of beliefs described above.

A given profile $(S^*; \rho^*)$ represents a Perfect Bayesian equilibrium if it defines a set of strategies such that, for every

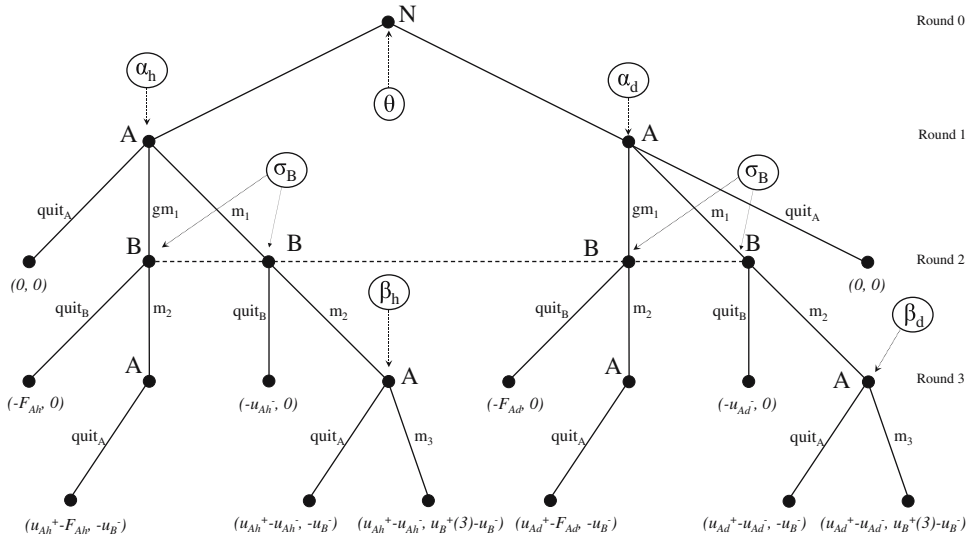

Fig. 4 Syverson's protocol Bayesian game in extensive form

player i and every information set I_i, player's i strategy is her best response to the opponent's strategy, given her believes at set $\mathcal{I}_i$. In other words, to be a PBE candidate, the strategies and beliefs defined by $(S^*; \rho^*) = (s_A^*, s_B^*; \alpha_h^*, \alpha_d^*, \beta_h^*, \beta_d^*, \sigma_B^*, \theta^*)$ must satisfy the following formulae for every α_d, β_d, α_h, β_h and σ_B probability distributions:

– For every strategy s_A of A_d:

$$EP(A_d, s_A^*) * \alpha_d^*(s_1^*) * \beta_d^*(s_3^*) \geq EP(A_d, s_A) * \alpha_d(s_1) * \beta_d(s_3)$$

– For every strategy s_A of A_h:

$$EP(A_h, s_A^*) * \alpha_h^*(s_1^*) * \beta_h^*(s_3^*) \geq EP(A_h, s_A) * \alpha_h(s_1) * \beta_h(s_3)$$

– And finally:

$$\begin{aligned} &EP(B, quit_B) * \sigma_B^*(quit_B) \\ &\quad + EP(B, m_2) * \sigma_B^*(m_2) \\ &\geq EP(B, quit_B) * \sigma_B(quit_B) \\ &\quad + EP(B, m_2) * \sigma_B(m_2) \end{aligned} \tag{8}$$

Since $EP(B, quit_B) = 0$, expression (8) can be reduced to:

$$EP(B, m_2) * \sigma_B^*(m_2) \geq EP(B, m_2) * \sigma_B(m_2) \tag{9}$$

for every probability distribution σ_B.

In fact, when $EP(B, m_2) > 0$, expression (9) is satisfied iff $\sigma_B^*(m_2) = 1$. Likewise, when $EP(B, m_2) < 0$, expression (9) is satisfied iff $\sigma_B^*(m_2) = 0$. In previous calculations, see expressions (5) and (7), we have computed a value L_B which represents a threshold to determine a value for $EP(B, m_2)$. B's best response to all possible A's strategies will be determined by the value of the linear combination L_B. Note that this is a stronger result than Eq. (6) as we are now considering all possible strategies for A and B. Therefore, we will present the following strategy-belief profile $(S^*; \rho^*)$ as our first candidate to PBE for the game in Fig. 4.

$$(S^*; \rho^*) = \big((m_1, m_3)_h, (m_1, m_3)_d, m_2; \alpha_h^*, \alpha_d^*, \beta_h^*, \beta_d^*, \sigma_B^*, \theta^*\big)$$

with

$$\sigma_B^*(m_2) = 1 \quad \text{and} \quad L_B^* \geq u_B^-/u_B^+(3)$$

where L_B^* will be the linear combination defined as in Eq. (5). Note that the presented candidate expresses the participant B's intention to succeed in the exchange of $item_A$ and $item_B$.

The next PBE candidate to be considered represents the set of strategies for A and B when A believes that B is likely to leave the protocol at round 2. Then, A's best response is to quit too:

$$(S^0; \rho^0) = \big((quit_A, quit_A)_h, (quit_A, quit_A)_d, quit_B; \alpha_h^0, \alpha_d^0, \beta_h^0, \beta_d^0, \sigma_B^0, \theta^0\big)$$

with

$$\sigma_B^0(m_2) = 0 \quad \text{and} \quad L_B^0 < u_B^-/u_B^+(3)$$

where L_B^0 will be the linear combination defined as in Eq. (5).

In order to prove that each one of the previous profiles conforms a PBE, they must satisfy Bayesian requirements 1 to 4. We will commence with the case of $(S^*; \rho^*)$, as $(S^0; \rho^0)$ can be trivially derived from the following steps.

Lemma 1 *The strategy-believe profile* $(S^*; \rho^*)$ *in Syverson's Bayesian game satisfies Bayes requirement* 1.

Proof It is required that each player A and B assigns a probability distribution over the nodes in each information set

$I_i \in \mathcal{I}_i$. The set $I_B \in \mathcal{I}_B$ identified at round 2 of the protocol is the only information set with more than one element. Indeed, the requirement 1 is satisfied as, considering the belief profile ρ^*, B defines the distributions α_d^* and α_h^* to satisfy $\alpha_h^*(gm_1)+\alpha_h^*(m_1)=1$ and $\alpha_d^*(gm_1)+\alpha_d^*(m_1)=1$. Since $\theta_h^*+\theta_d^*=1$ (see Eq. (4)), then

$$\theta_h^* * \alpha_h^*(gm_1) + \theta_h^* * \alpha_h^*(m_1) + \theta_d^* * \alpha_d^*(gm_1) + \theta_d^* * \alpha_d^*(m_1) = 1$$

□

Lemma 2 *The strategy-belief profile* $(S^*; \rho^*)$ *in Syverson's Bayesian game satisfies Bayes requirement* 2.

Proof Requirement 2 forces B to be rational and to behave accordingly to its beliefs once B has reached the information set I_B and for the rest of the game. Let CG be the continuation game starting at the information set $I_B \in \mathcal{I}_B$, and let $\rho^*(I_B)$ be the conditional beliefs at I_B. Then, the restriction of the strategy-belief profile $(S^*; \rho^*(I_B))$, must be a Nash equilibrium of the continuation game CG. Evaluating the payoff vectors we obtain:

- $EP(B, quit_B, CG) = 0$
- $EP(B, m_2, CG) \geq 0$ iff

$$\theta_h * \alpha_h^*(m_1) * \beta_h^*(m_3|m_1) + \theta_d * \alpha_d^*(m_1) * \beta_d^*(m_3|m_1) \geq u_B^-/u_B^+(3)$$

which it is, in fact, the value of L_B described in the candidate's definition.

Therefore, the profile strategy given by $(S^*; \rho^*(I_B))$ constitutes a Nash equilibrium in the continuation game CG. □

Lemma 3 *The strategy-believe profile* $(S^*; \rho^*)$ *in Syverson's Bayesian game satisfies Bayes requirement* 3.

Proof Requirement 3 forces B to establish *sensible* beliefs at the on-equilibrium-path information set I_B. These set of believes must be determined from the strategy profile according to Bayes' rule. Therefore, B has to establish probability distributions α_d^* and α_h^* in terms of the different actions that A can take at round 1 of the protocol. This is, if B believes that A_d would choose the action of sending m_1 with probability q_{d1}, the action of sending gm_1 with probability q_{d2} and action $quit_A$ with probability $1-q_{d1}-q_{d2}$, then $\alpha_d^*(m1)$ and $\alpha_d^*(gm_1)$ must take the following values:

$$\alpha_d^*(m_1) = \frac{q_{d1}}{(q_{d1}+q_{d2})} \quad \alpha_d^*(gm_1) = \frac{q_{d2}}{(q_{d1}+q_{d2})}$$

Likewise, B is forced to define:

$$\alpha_h^*(m_1) = \frac{q_{h1}}{(q_{h1}+q_{h2})} \quad \alpha_h^*(gm_1) = \frac{q_{h2}}{(q_{h1}+q_{h2})}$$

□

Lemma 4 *The strategy-believe profile* $(S^*; \rho^*)$ *in Syverson's Bayesian game satisfies Bayes requirement* 4.

Proof Requirement 4 forces B to establish *sensible* beliefs at any off-equilibrium-path information set. There are no information sets off the equilibrium path, so requirement 4 is trivially satisfied. □

Theorem 1 *The strategy-believe profile* $(S^*; \rho_e^*)$ *is a perfect Bayesian equilibrium.*

Proof Immediate by Definition 8 and Lemmas 1 to 4. □

A similar rationale can be applied to prove the strategy-belief profile $(S_0; \rho_0)$ does also constitute a PBE in the protocol game. In this case, however, the gains obtained are less than that obtained following $(S_e; \rho_e)$. As a result, the strategy to be followed will depend on the specific values of the items and the beliefs that player B holds about the behavior of the other party.

6.6 Discussion

What Theorem 1 tell us is that the strategy S^* is an equilibrium (i.e., will constitute a successful exchange for both parties), but it depends on the specific values taken by the beliefs of each participant, ρ_e^*. To be precise, the series of $B's$ beliefs would have to form a linear combination which would determine the best response participant B can give to a player A. At the same time, A would create its own set of beliefs to confront participant B. There could be more than one equilibrium, as many as linear combinations L_B there are, such that $L_B \geq u_B^-/u_B^+(3)$.

Next we provide an example with a set of values for which an equilibrium is reached when both entities, behaving rationally, carry out a successful exchange (see Fig. 5). We also present a scenario where B does not see justification in exchanging items with A, as the price to pay for $item_A$ is too high to justify the risk.

Let us suppose that entity B has reasons (past experience, reputation factor, etc.) to believe that entity A is not always honest. B estimates A to be honest with probability $\theta_h = 0.85$. Let us suppose that entity B does also hold enough evidence to estimate that, when A is honest, the probability of A misbehaving at step one of the protocol is very low, i.e., the probability of sending the correct message m_1 at step one is very high, $\alpha_h(m_1) = 0.9$. Let B suppose that, by contrast, when A is dishonest, the probability of sending the correct message m_1 is also high, $\alpha_d(m_1) = 0.7$. See Fig. 5. Given the previous set of values, B computes L_B and establishes a decision-making criteria. For instance, when the value $u_B^- = 3.5$, L_B is not big enough to encourage B to follow an exchanging strategy; instead, B would be better off quitting the protocol without sending payment. A, aware of B calculations, would

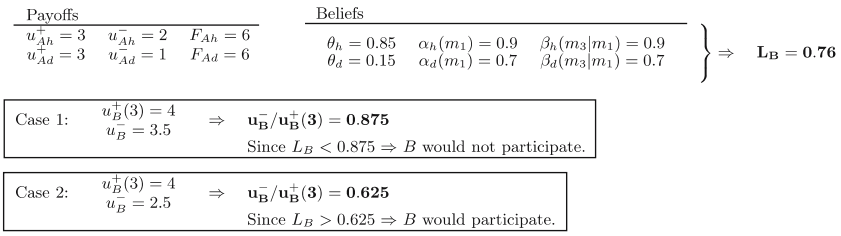

Fig. 5 A numerical example

execute the strategy which best responds to B, this is, to also quit (see the analysis for cases 1 and 2 in Fig. 5). By contrast, when $u^-_B = 2.5$, the payment B is asked to pay for $item_A$ is lower and satisfies the required criteria given by expression (7). Entity B would then participate in the protocol following an strategy to complete the exchange successfully.

7 Conclusions and future work

Several reasons lead to believe that some protocols will be executed within environments where uncertainty plays a significant role. This is the case, for example, with infrastructure-less networks, on which, in general, one cannot assume the existence of elements such as a TTPs. The protocols designed to operate under these circumstances should be analyzed using tools that can support the modeling of some forms of uncertainty in misbehavior. In the case of (rational) exchange protocols, Game Theory seems a good candidate due to the very purpose for which these protocols are designed and the similarities found between a protocol run and a formal game.

Nevertheless, the formal framework provided by basic (i.e., extensive-form) games and related concepts (Nash equilibrium) is somewhat narrow to capture some of the complexities that could be relevant for an in-depth analysis. Even though in our proposal the underlying model remains essentially unaltered with respect to that provided by Buttyán et al., we provide the framework with a theoretical tool –the notion of dynamical games of incomplete information and the associated PBE—that can be easily adapted to a variety of more realistic and complex scenarios.

Our model is not exempted from inconveniences. In practice, dealing with incomplete information involves making hypotheses about facts such as the probability of the other party being dishonest, etc. In this regard, the protocol participants have to take an active role when applying and implemented the model. Each participant has to individually decide whether there exits a secure context in which the rational exchange could be successfully completed.

Moreover, our model offers flexibility in other aspects. It can easily be extended to multiparty protocols. A set of beliefs set upon each party will determine the best possible individual response in multiparty scenarios. In such a type of protocols, some of the participants might follow fixed strategies (for instance, a TTP or the communication network itself), while other participants' payoff functions could differ depending on the state of those entities when the protocol ends (if the TTP has been or not involved in the protocol might have an impact on the participant's payoff functions).

In future works, we will tackle two main aspects of our proposal. First, we will further elaborate on the extensions briefly outlined in Sect. 4.1. A major goal for us is to develop a software prototype to automatically perform the analysis. At this respect, we are currently modifying the set of tools provided by Gambit [14] (specifically the GCL language).

Acknowledgments The authors are grateful to the anonymous reviewers for their insights and fruitful comments during the review process, which have greatly contributed to improve the quality of the original manuscript.

Appendix A: Breaking and repairing Syverson's protocol

The following observations will help in understanding the forthcoming analysis. Some of them are also mentioned by the protocol author, while others are of interest to our study and therefore newly formulated.

1. Message m_1 could be used as a proof of A's misbehavior. Indeed, due to the nature of $w(k)$, if A randomly generates a ciphertext ϵ to include in m_1, A can be penalized whenever the commitment $w(k)$ is broken and k disclosed. Therefore, this kind of misbehavior can always be proven to a judge. However, note that m_1 ensures B that A is the *author* of such a message, but it does not guarantee that A is also the *sender* of such a message. Message m_1 could be used to prove that A once generated a forged message, but m_1 cannot be used to prove that A was the sender of such a message in the *current* instance of the protocol. In this regard, it is not until step 3 of the protocol that B holds a valid NRO (non-repudiation of origin) token for $item_A$. Message m_3 serves as a NRO evidence for $item_A$, ensuring B that A has been a participant in the current protocol execution.
2. Message m_2 could serve as a NRR (non-repudiation of receipt) token for message m_1 as well as a NRO (non-repudiation of origin) of $item_B$. B's signature on message

m_2 ensures A that B received $item_A$ (weakly encrypted) and that B has proceeded sending $item_B$. Message m_2 could always be used as a proof of B's misbehavior in the protocol.

3. Message m_3 could serve as a NRR token for message m_2, as well as a NRO of $item_A$. A's signature on message m_3 ensures B that A received $item_B$ and that A has generated and sent m_3 with the corresponding key. Message m_3 could always be used as a proof of A's misbehavior in the protocol. A might not send the third message, or not do it for a long time, but A gains nothing by doing that apart from a poor reputation that could damage her business. The context in which to execute this protocol should then be a regularly repeated scenario.

The protocol, therefore, when rationally executed could provide with rational exchange of non repudiation evidences which are needed to force the participant's *rational—self-interest*—behavior. However, the non-repudiation evidences would have to be linked to each particular protocol run to serve the purposes of repudiation in future disputes. Since A is asked to generate a fresh key k for each run of the protocol, k could be the unique label to reference each different protocol run and the corresponding evidences.

Nevertheless, the protocol, as it is defined and described by its author, presents several vulnerabilities and some attacks can successfully be carried out.

A.1 Attack 1

Consider the following scenario:

$$A \rightarrow B : m_1 = (desc_{item_A}, enc(k, item_A), w(k), \sigma_1)$$
$$B(A) \rightarrow C : m_1 = (desc_{item_A}, enc(k, item_A), w(k), \sigma_1)$$
$$C \rightarrow B(A) : m_2 = (item_C, m_1, \sigma_2)$$

This attack is based on B impersonating A, sending the same message m_1 to C and receiving $item_C$ in return. B would have to quit the protocol after receiving the payment as she has no key to send to C. Although C has paid a full price for $item_A$, by the time that k is disclosed to C, $item_A$ would be of very little value to C. The customer C could only present message m_1 to prove A misbehaved. However, A will claim that m_1 was never intended for C and that she was not part of such a communication. Indeed, there is nothing in m_1 linking A and C as participants on the same protocol run. To overcome this attack, new restrictions would have to be placed over the communicating network or amendments should be made to the structure of m_1.

A.2 Attack 2

Let us suppose the following simplistic scenario: A is selling an access code to enable the viewing of a football match on a private television network. Let us suppose that A and B carried out a successful Syverson's protocol execution and that they properly exchanged the encrypted access code, $item_B$ and the corresponding key k in messages m_{11}, m_{12}, and m_{13}, respectively. The access code that B has bought from A is obviously of timely diminishing value, but B could still have time to impersonate A and sell the access code to other customers, receiving payment in return (note that $m_{21} = m_{11}$ and $m_{23} = m_{13}$):

$$B(A) \rightarrow C : \ m_{21} = (desc_{item_A}, enc(k, item_A), w(k), \sigma_A)$$
$$C \rightarrow B(A) : \ m_{22} = (payment_C, m_{21}, \sigma_C)$$
$$B(A) \rightarrow C : \ m_{23} = (k, m_{12}, \sigma_A)$$

In this scenario, by the time C receives message three and realizes that there is a fraud going on, C has no evidence of such a fraud to present in front of a judge, and has got the key k to decrypt the football match access code and watch the match. However, A could claim that C is watching a program without a license and take action against her. If the number of reselling codes is large, the scale of the fraud would make it impractical to pursue each one of the individuals watching the match with no license. Furthermore, trying to trace back the origin of such messages would be practically impossible. Again, to address this problem the content of message one should be amended.

A.3 Attack 3

If a vendor sends the customer a message m_1 containing garbage (i.e., a ciphertext which does not correspond with the actual $item_A$), the vendor is indeed providing the customer with evidence of such a form of cheating. Message m_1 could be presented to a judge and the vendor would be charged with the appropriate penalty. Such a penalty could greatly exceed the value of the goods, so the vendor is completely discouraged from performing such a scheme. However, the vendor could not be sued and penalized twice for the same offence and, on these terms, a vendor A could carry on sending the forged message m_1 to many others customers, receiving payments in return. These new angry customers would only have message m_1 to inculpate vendor A. Vendor A would claim that she never sent m_1 to them and that they must have got it from the first resentful customer. As a matter of fact, there will be nothing in m_1 to prove that A is using the same forged message all over again.

A.4 Fixing the protocol

Even though the attacks described above correspond to simple deviations from the protocol description, they represent real threats to parties using the scheme to exchange their items. In e-commerce transactions, neither vendor A nor customer B would want to take the risk of being cheated on. However, the previous attacks can be avoided if a better cryptographic evidence is constructed. This can be done in many ways. Probably the easiest one is just by including the identity of B in m_1, thus linking the message with its intended receiver:[1]

$$A \rightarrow B : m_1 = (\mathbf{B}, desc_{item_A}, enc(k, item_A), w(k), \sigma_1)$$

where:

$$\sigma_1 = sig(k_A^{-1}, (\mathbf{B}, desc_{item_A}, enc(k, item_A), w(k)))$$

This modification suffices to prevent attacks one to three.

References

1. Abadi, M., Glew, N., Horne, B., Pinkas, B.: Certified email with a light on-line trusted third party: design and implementation. In: Proceedings of 2002 International World Wide Web Conference, pp. 387–395. ACM Press (2002)
2. Asokan, N., Schunter, M., Waidner, M.: Optimistic protocols for fair exchange. In: Matsumoto, T. (ed.), 4th ACM Conference on Computer and Communications Security, vol. 6, pp. 8–17. ACM Press (1997)
3. Asokan, N., Shoup, V., Waidner, M.: Asynchronous protocols for optimistic fair exchange. In: Proceedings of the IEEE Symposium on Research in Security and Privacy, pp. 86–99. IEEE Press (1998)
4. Asokan, N.: Fairness in electronic commerce. Ph.D. Thesis. University of Waterloo (1998)
5. Bahreman, A., Tygar, J.D.: Certified electronic mail. In: Proceedings of 1994 Symposium on Network and Distributed System Security, pp. 3–19. Internet Society (1994)
6. Buttyán, L., Hubaux, J.P.: Rational exchange—a formal model based on game theory. In: Proceedings of the 2nd International Workshop on Electronic Commerce (2001)
7. Buttyán, L., Hubaux, J.P.: A formal analysis of syverson's rational exchange protocol. In: Proceedings of the 15th IEEE Computer Security Foundations Workshop, pp. 193–205. IEEE Press (2002)
8. Buttyán, L., Hubaux, J.P., Čapkun, S.: A formal model of rational exchange and its application to the analysis of syverson's protocol. J. Comput. Security **12**(3/4), 551–588 (2004)
9. Even, S., Yacobi, Y.: Relations among public key signature system. Technical Report 175 (1980). Computer Science Department, Technion, Haifa, Israel
10. Fischer, M.J., Lynch, N.A., Paterson, M.S.: Impossibility of distributed consensus with one faulty process. J. ACM **32**(2), 374–382 (1985)
11. Fudenberg, D., Tirole, J.: Perfect Bayesian equilibrium and sequential equilibrium. J. Econ. Theory **53**(2), 236–260 (1991)
12. Gibbons, R.: Game Theory for Applied Economists. Princeton University Press, Princeton (1992)
13. Kremer, S., Markowitch, O., Zhou, J.: An intensive survey of fair non-repudiation protocols. Comput. Commun. **25**(17), 1606–1621 (2002)
14. McKelvey, R.D., McLennan, A.M., Turocy, T.L.: Gambit: software tools for game theory, Version 0.2006.01.20 (2006). Available online at: http://econweb.tamu.edu/gambit
15. Pagnia, H., Gärtner, F.C.: On the impossibility of fair exchange without a trusted third party. Darmstadt University of Technology, Department of Computer Science. Technical Report TUD-BS-1999-02 (1999)
16. Pagnia, H., Vogt, H., Gärtner, F.C.: Fair exchange. Comput. J. **46**(1), 55–75 (2003)
17. Syverson, P.: Weakly secret bit commitment: applications to lotteries and fair exchange. In: Proceedings of the 11th IEEE Computer Security Foundations Workshop, pp. 2–13. IEEE Press (1998)

[1] As usual, we assume that A's identity is implicit in m_1, since the message contains A's signature. If this was not the case, then we must include it explicitly to avoid a different class of attacks.

Author Biography

Juan M. Estevez-Tapiador is Associate Professor at the Computer Science Department of Carlos III University of Madrid. He holds a M.Sc. in Computer Science from the University of Granada (2000), where he obtained the Best Student Academic Award, and a Ph.D. in Computer Science (2004) from the same university. His research is focused on cryptography and information security, especially in formal methods applied to computer security, design and analysis of cryptographic protocols, steganography, and some theoretical aspects of network security. In these fields, he has published around 40 papers in specialized journals and conference proceedings. He is member of the program committee of several conferences related to information security and serves as regular referee for various journals.

Almudena Alcaide Raya is Assistant Professor at the Cryptography and Information Security Group of the Computer Science Department of Carlos III University of Madrid. She has a B.Sc. in Mathematics by Complutense University of Madrid and a M.Sc. in Advanced Computing by King's College of London. Currently, she is a Ph.D. student at the Computer Science Department of Carlos III University of Madrid. Her work is focused on formal methods applied to the design and analysis of cryptographic protocols. In particular, her most recent research activity is related to applying Game Theory results to security protocols, and how these methods can assist in developing automated protocol verification tools and techniques.

Julio C. Hernandez-Castro is Associate Professor at the Computer Science Department of Carlos III University of Madrid. He has a B.Sc. in Mathematics, a M.Sc. in Coding Theory and Network Security, and a Ph.D. in Computer Science. His interests are mainly focused in cryptology, network security, steganography and evolutionary computation.

Arturo Ribagorda is Full Professor at Carlos III University of Madrid, where he is also the Head of the Cryptography and Information Security Group and currently acts as the Director of the Computer Science Department. He has a M.Sc. in Telecommunications Engineering and a Ph.D. in Computer Science. He is one of the pioneers of computer security in Spain, having more than 25 years of research and development experience in this field. He has authored four books and more than 100 articles in several areas of information security. Additionally, he is member of the program committee of several conferences related to cryptography and information security.