On Honesty in Sovereign Information Sharing

Rakesh Agrawal[1] Evimaria Terzi[1,2]

[1] IBM Almaden Research Center, San Jose, CA 95120, USA
[2] Department of Computer Science, University of Helsinki, Finland

Abstract. We study the following problem in a sovereign information-sharing setting: How to ensure that the individual participants, driven solely by self-interest, will behave honestly, even though they can benefit from cheating. This benefit comes from learning more than necessary private information of others or from preventing others from learning the necessary information. We take a game-theoretic approach and design a game (strategies and payoffs) that models this kind of interactions. We show that if nobody is punished for cheating, rational participants will not behave honestly. Observing this, our game includes an auditing device that periodically checks the actions of the participants and penalizes inappropriate behavior. In this game we give conditions under which there exists a unique equilibrium (stable rational behavior) in which every participant provides truthful information. The auditing device preserves the privacy of the data of the individual participants. We also quantify the relationship between the frequency of auditing and the amount of punishment in terms of gains and losses from cheating.

1 Introduction

There is an increasing requirement for sharing information across autonomous entities in such a way that only minimal and necessary information is disclosed. This requirement is being driven by several trends, including end-to-end integration of global supply chains, co-existence of competition and co-operation between enterprises, need-to-know sharing between security agencies, and the emergence of privacy guidelines and legislations.

Sovereign information sharing [1, 3] allows autonomous entities to compute queries across their databases such that nothing apart from the result is revealed. For example, suppose the entity R has a set $V_R = \{b, u, v, y\}$ and the entity S has a set $V_S = \{a, u, v, x\}$. As the result of sovereign intersection $V_R \cap V_S$, R and S will get to know the result $\{u, v\}$, but R will not know that S also has $\{a, x\}$, and S will not know that R also has $\{b, y\}$.

Several protocols have been proposed for computing sovereign relational operations, including [1, 3, 6, 8, 16]. In principle, sovereign information sharing can be implemented using protocols for secure function evaluation (SFE) [7].Given two parties with inputs x and y respectively, SFE computes a function $f(x, y)$ such that the parties learn only the result.

The above body of work relies on a crucial assumption, that the participants in the computation are *semi-honest*. This assumption basically says that the participants follow the protocol properly (with the exception that they may keep a record of the intermediate computations and received messages, and analyze the messages). Specifically, it is assumed that the participants will not maliciously alter the input data to gain additional information. This absence of malice assumption is also present in work in which a trusted-third party is employed to compute sovereign operations.

In a real imperfect world, the participants may behave dishonestly particularly when they can benefit from such a behavior. This benefit can come from learning more than necessary private information of others or preventing others from learning the necessary information. In the sovereign intersection example given in the beginning, R may maliciously add x to V_R to learn whether V_S contains x. Similarly, S may exclude v from V_S to prevent R from learning that it has v.

1.1 Problem Addressed

We study the following problem in a sovereign information-sharing setting: *How to ensure that the individual participants, driven solely by self-interest, will behave honestly, even though they can benefit from cheating.*

We take a game-theoretic approach to address the problem. We design a game (i.e. strategies and payoffs) that models interactions in sovereign information sharing. Through this game, we show that if nobody is punished for cheating, it is natural for the rational participants to cheat. We therefore add an auditing device to our game that periodically checks the actions of the participants and penalizes inappropriate behavior. We derive conditions under which a unique equilibrium (stable rational behavior) is obtained for this game such that every participant provides truthful information. We also quantify the relationship between the frequency of auditing and the amount of punishment in terms of gains and losses from cheating.

The auditing device must have the following essential properties: (a) it must not access the private data of the participants and (b) it must be space and time efficient. The auditing device we provide has these properties.

1.2 Related Work

Notions from game theory are used widely in this paper. Game theory was founded by von Neumann and Morgenstern as a general theory of rational behavior. It is a field of study of its own, with extensive literature; see [17] for an excellent introduction.

Games related to our work include the interdependent security(IDS) games [10, 13]. They were defined primarily to model scenarios where a large number of players must make individual investment decisions related to a security - whether physical, financial, medical, or some other type - but in which the ultimate safety of every participant depends on the actions of the entire population. IDS games

are closely related to summarization games [11] in which the players' payoff is a function of their own actions and the value of a global summarization function that is determined by the joint play of the population. Summarization games themselves are extensions of congestion games [15, 19] in which players compete for some central resources and every player's payoff is a decreasing function of the number of players selecting the resources. We have adopted some notions from the IDS games and used them to model information exchange. However, our problem is different from the one presented in [13], while at the same time we are not exploring algorithms for computing the equilibria of the games as in [10].

Inspection games [4, 5, 14, 21] are also related to our work. These are games repeated for a sequence of iterations. There is an inspector responsible for distributing a given number of inspections over an inspection period. Inspections are done so that possible illegal actions of an inspectee can be detected. The inspectee can observe the number of inspections the inspector performs. The question addressed is what are the optimal strategies for the inspector and the inspectee in such a game. The main difference between these games and the game we have designed is that in the inspection games the inspector is a player of the game. This is not true for our game, where the inspector acts as a referee for the players, helping them (via auditing) to achieve honest collaboration.

The modeling of private information exchange using game-theoretic concepts has received some attention recently. In [12], different information-exchange scenarios are considered and the willingness of the participants to share their private information is measured using solution concepts from coalition games. Our study is complementary to this work. We are interested in quantifying when people are willing to participate truthfully in a game, rather than the complementary question of whether they are willing to participate at all.

The work presented in [20] models information exchange between a consumer and a web site. Consumers want to interact with web sites, but they also want to keep control of their private information. For the latter, the authors empower the consumers with the ability to test whether a web site meets their privacy requirements. In the proposed games, the web sites signal their privacy policies that the consumers can test at some additional cost. The main conclusion of the study is that such a game leads to cyclic instability. The scenario we are modeling is completely different. Our players are all empowered with the same set of strategies. Our games also admit multiple players.

A recent work [22] addresses the problem of an adversary maliciously changing his input to obtain the private information from another party in a sovereign-intersection computation. They use concepts from non-cooperative games to derive optimal countermeasures for a defendant (and optimal attacking methods for the adversary) that balance the loss of accuracy in the result and the loss of privacy. These countermeasures involve the defendant also changing his input. Our approach is entirely different. We are interested in creating mechanisms so that the participants do not cheat and provide truthful information.

1.3 Road Map

The rest of the paper is structured as follows. In Section 2, we formally define the problem addressed in the paper, and also review the main game-theoretic concepts. In Section 3, we construct our initial game that captures two-party interactions in the absence of auditing and study its equilibria. The auditing device is introduced in Section 4 and its influence on the equilibria of the game is discussed. Section 5 shows how the observations from the two-player game generalize to multiple participants. An implementation of the auditing device is provided in Section 6. We conclude with a summary and directions for future work in Section 7.

2 Definitions

We first formally define the problem the paper addresses. We then review some basic concepts from game theory.

2.1 Problem Statement

First we give the classical sovereign information-sharing problem, which provided the setting for this work. Then we define the honest version of this problem, which is the concern of this paper. Finally, we specify the honest set-intersection problem, which is an important instantiation of the general problem.

Problem 1. [Sovereign information sharing] Let there be n autonomous entities. Each entity i holds a database of tuples D_i. Given a function f defined on D_i's, compute $f(D_1, \ldots, D_n)$ and return it to each entity. The goal is that in the end of the computation each entity knows $f(D_1, \ldots, D_n)$ and no additional information regarding the data of its peers.

The problem we are trying to tackle is more difficult. We want not only to guarantee that each participant in the end knows nothing more than the result, but also that each participant reports his true dataset. More formally:

Problem 2. [Honest sovereign information sharing] Let there be n autonomous entities. Each party i holds a database of tuples D_i. Each entity i reports a dataset $\hat{D}_i$ so that a function $f(\hat{D}_1, \ldots, \hat{D}_n)$ is computed. The goal in the honest information sharing is to find a mechanism that can guarantee that all entities report $\hat{D}_i$ such that $\hat{D}_i = D_i$. As in Problem 1, in the end of the computation each entity knows only $f(\hat{D}_1, \ldots, \hat{D}_n)$ and no additional information regarding the data of its peers.

We use game-theoretic concepts to develop a general framework that can model different information-exchange scenarios and guarantee honest information exchange. For concreteness, we also consider:

Problem 3. [Honest computation of set intersection] Special case of Problem 2 in which $f(\hat{D}_1, \ldots, \hat{D}_n) = \cap_{i=1,\ldots,n} \hat{D}_i$.

The problem of honest computation of other relational operations (e.g. join, set-difference) can be defined analogously; the techniques presented in the paper apply to them as well.

2.2 Games and Equilibria

We mainly focus on *strategic games*. In each game there are n players that can choose among a set of strategies S_i, $i = 1, 2, \ldots, n$. A function u_i is associated with each player i with $u_i : S_1, \ldots, S_n \rightarrow \mathbb{R}$. This is called a *payoff function* since it assigns a payoff to player i, for each combined strategy choices of the n players. The basic question in game theory is what constitutes a rational behavior in such a situation. The most widely-used concept of rationality is the *Nash equilibrium*:

Definition 1 (Nash equilibrium). *A Nash equilibrium (NE) is a combination of strategies: $x_1 \in S_1 \ldots x_n \in S_n$ for which*

$$u_i(x_1, \ldots, x_i, \ldots x_n) \geq u_i(x_1, \ldots, x'_i, \ldots, x_n),$$

for all i and $x'_i \in S_i$.

That is, a Nash equilibrium is a combination of strategies from which no player has the incentive to deviate. A game can have zero, one, or more than one Nash equilibrium and the payoffs of a player can be different in two different equilibria.

Another rationality concept is that of *dominant-strategy equilibrium*:

Definition 2 (Dominant-strategy equilibrium). *A dominant-strategy equilibrium (DSE) is a combination of strategies: $x_1 \in S1, \ldots, x_n \in S_n$ for which*

$$u_i(x'_1, \ldots, x_i, \ldots, x'_n) \geq u_i(x'_1, \ldots, x''_i, \ldots x'_n),$$

for all i and $x''_i \in S_i$ and for all $j \neq i$ and $x'_j \in S_j$.

That is, the strategy of every player in a dominant-strategy equilibrium is the most profitable one (gives the highest payoff to every player) irrespective of what the other players' strategies are. A game need not have a dominant-strategy equilibrium. A dominant-strategy equilibrium is always a Nash equilibrium. The opposite is not true. Nash and dominant-strategy equilibria capture the behavior of selfish players who only care about maximizing their own payoffs without caring about the payoffs of the rest of the players. Nash equilibrium is widely used in many settings. However, there is no consensus on the best concept for rationality.

3 Dishonest Information Sharing

We now describe a real-world situation, but of course simplified, and use it to motivate the definition of a two-player game that can be used to analyze sovereign information-sharing interactions. Our goal is to formally show that when there is

benefit from cheating that is not accompanied with any bad consequences, there is no guarantee for honesty. In fact, rational players driven solely by self-interest will cheat in such a situation.

Rowi and Colie are successful competitors. Though their products cover all segments of their industry, Rowi has a larger coverage in some while Colie is stronger in others. By finding the intersection of their customer lists, they both can benefit by jointly marketing to their common customers. This benefit accrues from business expansion as well as reduction in marketing costs with respect to these customers. Rowi has estimated that the benefit he will realize is B_1, whereas Colie's estimate is B_2.[3] Clearly, it is in the interest of both Rowi and Colie that they find their common customers without revealing their private customers, and can use sovereign set intersection for this purpose.

In practice, Rowi might be tempted to find more than just common customers. Rowi might try to find private customers of Colie by inserting some additional names in his customer database. By doing so, Rowi estimates that his benefit can increase to F_1. This temptation to cheat and find more holds for Colie too, and Colie's estimate of the increased benefit is F_2. Clearly, it must be that $F_1 > B_1$ and $F_2 > B_2$. We carry this assumption throughout the paper.

However, both Rowi and Colie may also incur some loss due to cheating. For example, from Rowi's perspective, Colie might succeed in stealing some of his private customers. Also, Rowi's customer database has become noisy as it now has some fake names. We use L_{21} (L_{12}) to represent the player's estimate of the loss that Colie (Rowi) causes to Rowi (Colie) due to his cheating.

For now, let us consider the symmetric case: $B_1 = B_2 = B$, $F_1 = F_2 = F$, and $L_{12} = L_{21} = L$, and $F > B$.

We model the above situation as a two-player strategic game with payoffs described in Table 1. Both players have the same set of strategies: "Play Honestly" (**H**) or "Cheat" (**C**). Honest playing corresponds to reporting the true set of tuples, while cheating corresponds to alternating the reported dataset by adding extra tuples or removing real tuples.

Rowi \ Colie	Play Honestly (**H**)	Cheat (**C**)
Play Honestly (**H**)	B / B	F / $B - L$
Cheat (**C**)	$B - L$ / F	$F - L$ / $F - L$

Table 1. Payoff matrix for the two-player game where there is no punishment for cheating. Each entry lists the payoff of Rowi at the left-bottom, and the payoff of Colie at the right-top corner of the cell for the corresponding combination of strategies.

[3] If the benefit is considered to be a function of the number of common customers, the latter can be determined (without revealing who the common customers are) by using the sovereign set intersection size operation.

Observation 1 *For the strategic game described in Table 1 and given that there is extra benefit from cheating ($F > B$), the pair of strategies (***C, C***) is the* only *equilibrium (NE as well as DSE).*

To see that (**C, C**) is a Nash equilibrium, note that for Rowi $u(\mathbf{C}, \mathbf{C}) > u(\mathbf{H}, \mathbf{C})$ and for Colie $u(\mathbf{C}, \mathbf{C}) > u(\mathbf{C}, \mathbf{H})$. On the other hand, (**H, H**) is not a Nash equilibrium since $u(\mathbf{C}, \mathbf{H}) > u(\mathbf{H}, \mathbf{H})$ for Rowi.

Similarly, (**C,C**) is a dominant-strategy equilibrium since for Rowi $u(\mathbf{C}, \mathbf{C}) > u(\mathbf{H}, \mathbf{C})$ and $u(\mathbf{C}, \mathbf{H}) > u(\mathbf{H}, \mathbf{H})$ and for Colie $u(\mathbf{C}, \mathbf{C}) > u(\mathbf{C}, \mathbf{H})$ and $u(\mathbf{H}, \mathbf{C}) > u(\mathbf{H}, \mathbf{H})$. It is easy to see that (**H,H**) is not a dominant-strategy equilibrium.

Note that the above observation holds irrespective of the value of L. In other words, both Rowi and Colie will find it rational to cheat even if the loss from cheating makes $F - L$ less than B for both of them.

4 Enforcing Honesty

We now extend the game described in the previous section with an auditing device that can check whether any player has cheated by altering the input. An implementation of such a device is discussed later in Section 6. Whenever the device finds out that a player has cheated, it penalizes the player. For a fixed penalty amount, we address the question of how often should the auditing be performed. We find a lower bound on the auditing frequency that guarantees honesty. Such a lower bound is important particularly in cases where auditing is expensive. Conversely, for fixed frequency of auditing we calculate the minimum penalty that guarantees honest behavior.

An auditing device can be characterized as follows, depending on the degree of honesty it can guarantee:

1. *Transformative:* It can induce equilibrium states where all players being honest is a dominant-strategy equilibrium (DSE). Recall that every dominant-strategy equilibrium is also a Nash equilibrium (NE), though the opposite is not true.
2. *Highly Effective:* It can induce equilibrium states where all participants being honest is the only Nash equilibrium of the game.
3. *Effective:* It can induce equilibria where all participants being honest is a Nash equilibrium of the game.
4. *Ineffective:* Nothing can be guaranteed about the honest behavior of the players. That is, the auditing device cannot induce equilibrium states where all players are honest.

We first study the symmetric case in which the players have identical payoffs. We then extend the analysis to study asymmetric payoffs.

4.1 The symmetric case

Consider the game with the payoff matrix given in Table 2. The semantics of the parameters B, F and L are the same as in the game described in Section 3.

Two more parameters appear here. The first one, P, represents the penalty that the auditing device imposes on the cheating player once it detects the cheating. Parameter f, with $0 \leq f \leq 1$, corresponds to the *relative frequency* of auditing, and represents how often the device checks truthfulness of the data provided by the players. For brevity, from now on, we will use the term *frequency* to refer to relative frequency.

Colie / Rowi	Play Honestly (**H**)	Cheat(**C**)
Play Honestly (**H**)	B	$(1-f)F - fP$
	B	$B - (1-f)L$
Cheat (**C**)	$B - (1-f)L$	$(1-f)F - fP - (1-f)L$
	$(1-f)F - fP$	$(1-f)F - fP - (1-f)L$

Table 2. Payoff matrix for the symmetric two-player game enhanced with the auditing device.

In Table 2, when both players play honestly they each have benefit B. Since the auditing device checks with frequency f, the expected gain of a player that cheats is $(1-f)F$. That is, a cheating player gains amount F only when he is not caught, which happens with probability $1-f$. A player who cheats and is not caught causes expected loss $(1-f)L$ to the other player. Finally, a cheating player may be caught with probability f and pays penalty P, which gives an expected loss of fP to the cheating player.

When both players are cheating their payoff is the expected cheating benefit $(1-f)F$ minus the expected cost of paying a penalty fP as well as the expected loss caused from other player cheating $(1-f)L$. Note that $(1-f)L$ is the loss of a player due to the cheating behavior of the opponent, multiplied by the probability that the latter is not caught.

We now give some important observations from the analysis (details omitted) of this game. Assume first that all parameters are fixed except for f. In that case the auditing device gets as input the penalty amount P. The goal is to determine the corresponding frequency of auditing that can guarantee honest behavior. The following statement can be made in this case.

Observation 2 *For any fixed penalty amount P, there exists a checking frequency for which the auditing device is both transformative and highly effective. More specifically for fixed P, the equilibria of the game for different values of frequency $f \in [0,1]$ are:*

- *For $0 \leq f < \frac{F-B}{P+F}$, (**C**,**C**) is the only DSE and NE of the game. That is, for those frequencies the auditing device is ineffective.*
- *For $\frac{F-B}{P+F} < f \leq 1$, (**H**,**H**) is the only DSE and NE of the game. That is, for those frequencies the auditing device is transformative and highly effective.*
- *For $f = \frac{F-B}{P+F}$, (**H**,**H**) is among the NE of the game and therefore the auditing device is effective.*

The above observation is rather intuitive. The key quantity is $f = \frac{F-B}{P+F}$ that can be rewritten as $fP = (1-f)F - B$. The left-hand side corresponds to the expected loss due to the penalty imposed by the auditing device. The right-hand side is the net expected gain from cheating. Therefore the first case in observation 2 says that (**C,C**) is DSE and NE only when $fP < (1-f)F-B$; that is when the expected loss from the penalty is less than the expected gain from cheating. In this case, the auditing device does not provide enough deterrence to keep off the players from cheating. However, when the expected loss due to the penalty imposed by the device exceeds the expected gain, the players start behaving honestly.

The landscape of the equilibria for the different values of the checking frequency is shown in Figure 1. Notice that the above game for all the values of $f \neq \frac{F-B}{P+F}$ has only two equilibria in which either both players are honest or both of them are cheating.

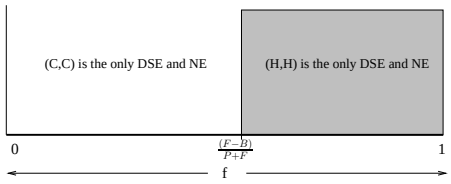

Fig. 1. Equilibria of the two-player symmetric game with auditing device for the different values of checking frequency f and for fixed penalty amount P. Shaded region corresponds to (**H, H**) being both DSE and NE.

Alternatively, we can study the penalty-for-cheating versus frequency-of-checking trade off the other way round. What happens in the case where the auditing device is instructed to check at the specified frequencies? What is the minimum penalty amount it has to impose on cheating players so that honesty is ensured?

Observation 3 *For any fixed frequency $f \in [0,1]$, the auditing device can be transformative and highly effective for wise choices of the penalty amount. Specifically:*

- *For $P > \frac{(1-f)F-B}{f}$, (**H, H**) is the only DSE and NE, and therefore the auditing device is both transformative and highly effective.*
- *For $P < \frac{(1-f)F-B}{f}$, (**C,C**) is the only DSE and NE, and therefore the auditing device is ineffective.*
- *For $P = \frac{(1-f)F-B}{f}$,(**H, H**) is among the NE of the game. That is for this penalty amount the auditing device is effective.*

The above observation is also intuitive as it says that the players will not be deterred by an auditing device that imposes penalties such that the expected

loss due to them is smaller than the expected additional benefit from cheating. This is true no matter how often this device performs its checks.

On the other hand, note the following special case. When $f > \frac{F-B}{F}$, the auditing device does not have to impose any penalty on the cheating participants. The fact that the participants are aware of its existence is daunting by itself. Notice that this happens particularly in high frequencies and for the following reason. Due to high checking frequency, the expected gain from cheating $(1-f)F$ becomes lower than the gain from honest collaboration B. Therefore, the players have incentive to play honestly.

The equilibria of the game as a function of the penalty amount P are given in Figure 2.

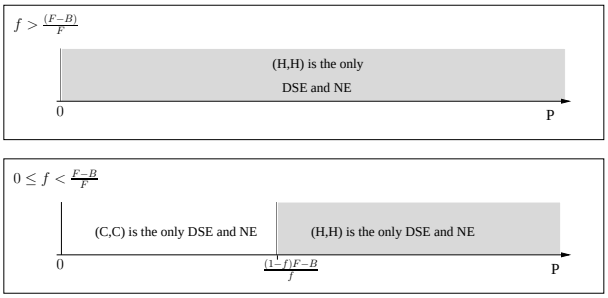

Fig. 2. Equilibria of the two-player symmetric game with auditing device for the different values of penalty regions P for fixed checking frequency f. Shaded region corresponds to (**H, H**) being DSE as well as NE.

The above observations provide the game-designer the chance to decide, based on estimations of the players losses and gains, the minimum checking frequencies or penalty amounts that can guarantee the desired level of honesty in the system.

Colie Rowi	Play Honestly (**H**)	Cheat (**C**)
Play Honestly (**H**)	B_2 B_1	$(1-f_2)F_2 - f_2P_2$ $B_1 - (1-f_2)L_{21}$
Cheat (**C**)	$B_2 - (1-f_1)L_{12}$ $(1-f_1)F_1 - f_1P_1$	$(1-f_2)F_2 - f_2P_2 - (1-f_1)L_{12}$ $(1-f_1)F_1 - f_1P_1 - (1-f_2)L_{21}$

Table 3. Payoff matrix for the asymmetric two-player game enhanced with the auditing device.

4.2 The asymmetric case

We now turn to the study of the asymmetric case where the payoffs of the two players are not necessarily the same. The payoff matrix of the game is given in Table 3. The easiest way to visualize the equilibria of such a game is by fixing the penalty amounts imposed on each player (P_i) and giving to the auditing device the freedom to select the frequency of checking each player (f_i). In this case, we get the landscape of the equilibria shown in Figure 3.

Again the auditing device becomes transformative and highly effective when it checks frequently enough so that the players cannot tolerate the extra losses from being caught cheating. Similar observations can be made by studying the game using the penalty amounts as the free parameters of the auditing device and fixing the checking frequencies.

Note that in contrast to the symmetric case, the current game exhibits equilibria in which the two players do not pick the same strategy. This is the case, for example, when the auditing device checks Colie very frequently and Rowi quite rarely (upper left-hand corner of the figure); the Nash equilibrium has poor Colie playing honestly while Rowi is cheating. This example brings out the need for careful choice of penalties and frequencies; otherwise, the rational players may be forced into unintuitive behaviors.

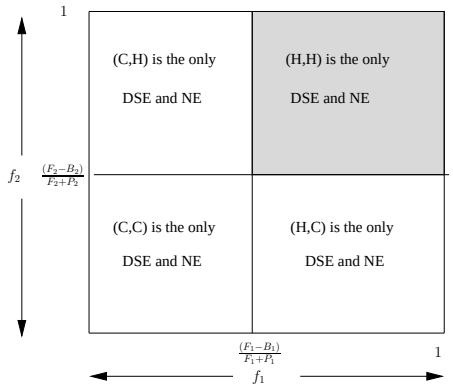

Fig. 3. Equilibria of two-player asymmetric game with auditing device for the different values of penalties (P_1, P_2). Shaded region corresponds to (**H,H**) being both DSE and NE.

5 Generalization to multiple participants

More than two entities are often involved in an information-sharing situation. To model such situations we extend our two-player game to n players.

Each player has again two possible strategies: to play honestly (**H**) or to cheat (**C**). We use indicator variable h_i to denote the strategy of player i:

$$h_i = \begin{cases} 1, \text{ if player i is playing honestly} \\ 0, \text{ otherwise.} \end{cases}$$

We use vector $\boldsymbol{h}$ to represent the strategies of all n players. The vector $\boldsymbol{h}_{-i}$ represents the strategies of all players except for player i. Motivated by [10], we design the n-player game by forming a *payoff function* that adequately describes: (a) the gains/losses a player has due to his own actions, and (b) the gains/losses due to the behavior of others.

The notation is along the same lines as used in the two-player game. We again assume the existence of an auditing device that checks on players with frequency f and imposes penalty P for cheating. We consider the case where the values of f and P are the same for all players. Assume that the benefit from honest collaboration for each player is B. The increased benefit of player i due to his cheating is given by function $\mathcal{F}$, which is assumed to be the same for all players. The specific form of function $\mathcal{F}$ depends on the application domain. However, we do assume that it is monotonically increasing in the number of players that play honestly. That is, the larger the number of honest players in the game, the more the dishonest player gains by exploiting their honesty. Finally, assume that the loss a player i experiences due to the cheating of another player j is given by L_{ji}. The payoff of player i is thus a function $u_i : \{\mathbf{H},\mathbf{C}\}^n \rightarrow \mathbb{R}$, which can be written as:

$$\begin{aligned} u_i(\boldsymbol{h}) = h_i B + (1-h_i)(1-f)\mathcal{F}(\|\boldsymbol{h}_{-i}\|) - (1-h_i)fP \\ - \sum_{j=1, j\neq i}^{n} (1-h_j)(1-f)L_{ji} \end{aligned} \tag{1}$$

The payoff u_i of player i depends on the strategies picked by the participating players and it consists of four terms. The first two terms correspond to the gains of the player and the last two correspond to his losses. The losses are due to either his own choices or the choices of the rest of the participants. More specifically, the first term is the gain player i has in isolation (irrespective of the strategies of the rest $n-1$ players) when he plays honestly. The second term is his gain when he decides to cheat. This gain depends on the strategies of others as well. The third term, $(1-h_i)fP$, corresponds to his loss when he decides to cheat and he is caught. In that case, he experiences an expected loss of fP. The last term represents his loss due to the behavior of the other participants.

For building some intuition, consider the following special cases. When all players except player i cheat, then the payoff of player i would be:

$$u_i(\boldsymbol{h}_{-i} = \mathbf{0}, h_i = 1) = B - \sum_{j=1, j\neq i}^{n} (1-f)L_{ji}.$$

If player i decides to cheat as well, his gain is:

$$u_i(\boldsymbol{h} = \mathbf{0}) = \mathcal{F}(0) - fP - \sum_{j=1, j\neq i}^{n} (1-f)L_{ji}.$$

Although it seems that the analysis of the auditing device in the presence of n players could be more demanding, it turns out that some intuition and the results from the two-player game carry over.

Assume we fix the checking frequency f with which the auditing device checks the participating players.

Proposition 1. *For the n-player game where the payoff of each player i is given by u_i as defined in equation 1, the following is true: For fixed frequencies $f \in [0, 1]$ an auditing device that imposes penalty $P > \frac{(1-f)\mathcal{F}(n-1)-B}{f}$ is transformative and highly effective. That is, for those values of f and P, (***H, H, ..., H***) is the only combination of strategies that is DSE and NE.*

Proof. (Sketch) First we show that the auditing device is transformative. For this, we have to show that when $P > \frac{(1-f)\mathcal{F}(n-1)-B}{f}$ each player i prefers $h_i = 1$ irrespective of the strategies of the other $n-1$ players. This comes down to proving that the inequality:

$$u_i(\boldsymbol{h}_{-i} = \mathbf{1}, h_i = 1) > u_i(\boldsymbol{h}_{-i} = \mathbf{1}, h_i = 0) \tag{2}$$

is true for player i (and thus for every player). If inequality 2 holds for $\boldsymbol{h}_{-i} = \mathbf{1}$, then it would also hold for any other $\boldsymbol{h}_{-i} \neq \mathbf{1}$. This means that even in the worst-case, where all $n-1$ other players are playing honestly (this is the case where player i has the highest benefit from cheating), player i still has more benefit from being honest than from cheating. This makes $h_i = 1$ dominant strategy. Indeed by solving inequality 2, we end up with a true statement.

Then we have to show that the auditing device is also highly effective. For this we need to show that when $P > \frac{(1-f)\mathcal{F}(n-1)-B}{f}$ there does not exist an equilibrium other than (**H,H,...,H**).

The proof is by contradiction. Assume there exists another equilibrium where x players are playing honestly and $n-x$ players are cheating, with $x \neq n$. Now consider a player i with $h_i = 1$. Since we have assumed an equilibrium state, the following should be true:

$$\begin{aligned} &u_i(h_1 = 1, \ldots, h_i = 1, \ldots, h_x = 1, h_{x+1} = 0, \ldots, h_n = 0) > \\ &u_i(h_1 = 1, \ldots, h_i = 0, \ldots, h_x = 1, h_{x+1} = 0, \ldots, h_n = 0). \end{aligned}$$

This would mean that

$$B - \sum_{j=1, j\neq i}^{n} (1-h_j)(1-f)L_{ji} > (1-f)\mathcal{F}(x-1) - fP - \sum_{j=1, j\neq i}^{n} (1-h_j)(1-f)L_{ji},$$

and thus

$$P > \frac{(1-f)\mathcal{F}(x-1) - B}{f}. \tag{3}$$

Now consider a player j from the set of $n-x$ cheating players. Due to the equilibrium assumption, the following should also hold:

$$u_j(h_1 = 1, \ldots, h_x = 1, h_{x+1} = 0, \ldots, x_j = 0, \ldots, h_n = 0) > \\ u_j(h_1 = 1, \ldots, h_x = 1, h_{x+1} = 0, \ldots, x_j = 1, \ldots, h_n = 0).$$

This would mean that

$$(1-f)\mathcal{F}(x) - fP - \sum_{i=1, l \neq j}^{n} (1-h_i)(1-f)L_{ij} > B - (1-f) \sum_{i=1, l \neq j}^{n} (1-h_i)(1-f)L_{ij}$$

and thus

$$P < \frac{(1-f)\mathcal{F}(x) - B}{f}. \tag{4}$$

However, inequalities 3, 4 and the constraint $P > \frac{(1-f)\mathcal{F}(n-1)-B}{f}$ cannot be satisfied simultaneously, due to the monotonicity property of $\mathcal{F}$. Therefore the auditing device is also highly effective. □

In a similar manner we can show the following proposition:

Proposition 2. *For the n-player game where the payoff of each player i is given by u_i as defined in equation 1, the following is true: For fixed frequencies $f \in [0,1]$ an auditing device that imposes penalty $P < \frac{(1-f)\mathcal{F}(0)-B}{f}$ is ineffective. That is, for those values of f and P, (***C, C, ..., C***) is the only combination of strategies that is NE and DSE.*

Finally we can generalize the above propositions in the following theorem:

Theorem 1. *For the n-player game where the payoff of each player i is given by u_i, as defined in equation 1, the following is true: For $x \in 1, \ldots, n-1$ and for any $f \in [0,1]$, when the auditing device imposes penalty $\frac{(1-f)\mathcal{F}(x-1)-B}{f} < P < \frac{(1-f)\mathcal{F}(x)-B}{f}$, then the n-player game is in an equilibrium state where x players are honest and $n-x$ players are cheating.*

Consequently, the equilibria landscape looks as in Figure 4.

6 Auditing Device

We turn now to a discussion of the feasibility of realizing the auditing device. The auditing service must be space as well as time efficient. It must also not see any private data of any of the participants.

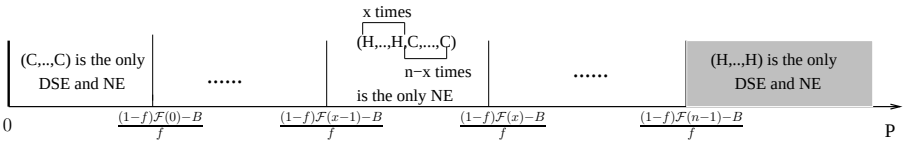

Fig. 4. Equilibria of the n-player symmetric game with auditing device for different values of penalty P. Shaded region corresponds to (**H,H**) being both DSE and NE.

6.1 Incremental multiset hash functions

Our proposed auditing device makes use of incremental multiset hash functions [2], which are hash functions that map multisets of arbitrary finite size to hashes of fixed length. They are incremental in that when new members are added to the multiset, the hash can be quickly updated.

Definition 3 (Multiset hash function [2]). *Let $(\mathcal{H}, +_{\mathcal{H}}, \equiv_{\mathcal{H}})$ be a triple of probabilistic polynomial time algorithms. This triple is an* incremental multiset hash function *if it satisfies:*

- **Compression:** *$\mathcal{H}$ maps multisets of a domain $\mathcal{D}$ into elements of a set with cardinality $\approx 2^m$, where m is some integer. Compression guarantees that hashes can be stored in a small bounded amount of memory.*
- **Comparability:** *Since $\mathcal{H}$ can be a probabilistic algorithm, a multiset need not always hash to the same value. Therefore a means of comparison ($\equiv_{\mathcal{H}}$) is needed to compare hashes. For this it should hold that $\mathcal{H}(M) \equiv_{\mathcal{H}} \mathcal{H}(M)$, for all multisets of M of $\mathcal{D}$.*
- **Incrementality:** *Finally, $\mathcal{H}(M \cup M')$ is computed efficiently using $\mathcal{H}(M)$ and $\mathcal{H}(M')$. The $+_{\mathcal{H}}$ operator makes this possible:*

$$\mathcal{H}(M \cup M') \equiv_{\mathcal{H}} \mathcal{H}(M) +_{\mathcal{H}} \mathcal{H}(M'),$$

  for all multisets M and M' of $\mathcal{D}$. In particular, knowing $\mathcal{H}(M)$ and an element $t \in \mathcal{D}$, one can easily compute $\mathcal{H}(M \cup \{t\}) = \mathcal{H}(M) +_{\mathcal{H}} \mathcal{H}(\{t\})$.

Multiset hash functions are collision resistant in that it is computationally infeasible to find a multiset M of $\mathcal{D}$ and a multiset M' of $\mathcal{D}$ such that $M \neq M'$ and $\mathcal{H}(M) \equiv_{\mathcal{H}} \mathcal{H}(M')$.

6.2 Auditing

Auditing is provided by a secure network service, built using a secure coprocessor [9]. For the purposes of this paper, it is sufficient to observe that a certified application code can be securely installed into a secure coprocessor and, once installed, the application can execute untampered. The remote attestation mechanism provided by the secure coprocessor can be used to prove that it is indeed executing a known, trusted version of the application code, running under a

known, trusted version of the OS, and loaded by a known, trusted version of the bootstrap code. Communication between the auditing device and the participants in the sovereign computation makes use of authenticated encryption that provides both message privacy and message authenticity [18].

The auditing device (AD) periodically checks the integrity of the data reported by the players, and hands over penalties if needed. As we shall see, AD accomplishes this check without accessing the private data of the players.

There is a tuple generator TG_i, associated with each player i. In the scenario given in Section 3, TG_i may correspond to the customer registration process. TG_i provides legal tuples to the player i that should participate in sovereign computations. The player i cannot influence TG_i into generating illegal tuples[4] but can himself fabricate them. Each TG_i operates as follows:

1. TG_i picks $\mathcal{H}_i$ and announces it publicly.
2. For each new tuple t entering the system and to be provided to player i:
   (a) TG_i computes $\mathcal{H}_i(t)$.
   (b) TG_i sends message $(\mathcal{H}_i(t), i)$ to AD.
   (c) TG_i sends t to player i.

AD maintains for each player i a hash value HV_i. This is the hash value of all the tuples that player i has received from TG_i. Upon receiving $(\mathcal{H}_i(t), i)$, AD updates the hash value so that $\mathrm{HV}_i = \mathrm{HV}_i +_{\mathcal{H}_i} \mathcal{H}_i(t)$. Note that the auditing device does not know the actual tuples that each player i has received. It only knows the hash value of this multiset of tuples, which it incrementally updates.

Finally, each player i also maintains locally the hashed value of the set of tuples it has received, $\mathcal{H}(D_i)$. Therefore, upon receiving tuple t from TG_i, the player i updates the hash value so that $\mathcal{H}_i(D_i) = \mathcal{H}_i(D_i) +_{\mathcal{H}_i} \mathcal{H}_i(t)$.

For sovereign information-sharing computation, the players follow one of the standard protocols that guarantee correct and private computation of the result. These protocols require that each player i reports D_i (usually encrypted) to the other players or to a trusted third party. Here, we additionally require that along with the encrypted version of D_i, each player i reports $\mathcal{H}_i(D_i)$.

Note that reporting $\mathcal{H}_i(D_i)$, along with the encrypted D_i, does not reveal anything about the actual D_i. This is due to the assumption that for a given multiset hash function $\mathcal{H}_i$, it is computationally infeasible to construct multisets M and M' such that $\mathcal{H}_i(M) \equiv_{\mathcal{H}_i} \mathcal{H}_i(M')$. Secondly, player i will be reluctant to report D_i along with $\mathcal{H}_i(D_i')$ such that $D_i \neq D_i'$ because that will be a violation of the protocol and if the entity that received the encrypted D_i along with $\mathcal{H}_i(D_i')$ takes i to court, the judge will be able to decide in polynomial time whether the hash value $\mathcal{H}_i(D_i') \equiv_{\mathcal{H}_i} \mathcal{H}_i(D_i)$.

Given this communication model, the job of the auditing device is straightforward. If AD decides to audit player i, it requests the hash value that i reported during the set-intersection computation. Let this hash value be $\mathcal{H}_i(D_i)$. Then AD can decide whether i is cheating by checking whether $\mathrm{HV}_i \equiv_{\mathcal{H}_i} \mathcal{H}_i(D_i)$.

[4] If player i can corrupt TG_i into generating illegal tuples on his behalf, it can be shown that no automated checking device can detect this fraudulent behavior.

7 Summary and Future Directions

A key inhibitor in the practical deployment of sovereign information sharing has been the inability of the technology to handle the altering of input by the participants. We applied game-theoretic concepts to the problem and defined a multi-party game to model the situation. The analysis of the game formally confirmed the intuition that as long as the participants have some benefit from cheating, honest behavior cannot be an equilibrium of the game. However, when the game is enhanced with an auditing device that checks at an appropriate frequency the integrity of the data submitted by the participants and penalizes by an appropriate amount the cheating behaviors, honesty can be induced not only as a Nash equilibrium but also as a dominant-strategy equilibrium. We addressed practical issues such as what should be the frequency of checking and the penalty amount and how the auditing device can be implemented as a secure network device that achieves the desired outcome without accessing private data of the participants.

In the future, we would like to study if appropriately designed incentives (rather than penalties) can also lead to honesty. We would also like to explore the application of game theory to other privacy-preservation situations.

Acknowledgment We thank Alexandre Evfimievski for helpful discussions.

References

1. R. Agrawal, A. V. Evfimievski, and R. Srikant. Information sharing across private databases. In *SIGMOD*, 2003.
2. D. Clarke, S. Devadas, M. van Dijk, B. Gassend, and G. E. Suh. Incremental multiset hash functions and their applications to memory integrity checking. In *Asiacrypt*, 2003.
3. C. Clifton, M. Kantarcioglu, X. Lin, J. Vaidya, and M. Zhu. Tools for privacy preserving distributed data mining. *SIGKDD Explorations*, 4(2):28–34, Jan. 2003.
4. T. Ferguson and C. Melolidakis. On the inspection game. *Naval Research Logistics*, 45, 1998.
5. T. Ferguson and C. Melolidakis. Games with finite resources. *International Journal on Game Theory*, 29, 2000.
6. M. J. Freedman, K. Nissim, and B. Pinkas. Efficient private matching and set intersection. In *EUROCRYPT*, Interlaken, Switzerland, May 2004.
7. O. Goldreich. *Foundations of Cryptography*, volume 2: Basic Applications. Cambridge University Press, May 2004.
8. B. A. Huberman, M. Franklin, and T. Hogg. Enhancing privacy and trust in electronic communities. In *Proc. of the 1st ACM Conference on Electronic Commerce*, Denver, Colorado, November 1999.
9. IBM Corporation. IBM 4758 Models 2 and 23 PCI cryptographic coprocessor, 2004.
10. M. Kearns and L. E. Ortiz. Algorithms for interdependent security games. In *NIPS*, 2004.
11. M. J. Kearns and Y. Mansour. Efficient Nash computation in large population games with bounded influence. In *UAI*, 2002.

12. J. Kleinberg, C. Papadimitriou, and P. Raghavan. On the value of private information. In *8th Conference on Theoretical Aspects of Rationality and Knowledge*, 2001.
13. H. Kunreuther and G. Heal. Interdependent security. *Journal of Risk and Uncertainty*, 2002.
14. M. Maschler. A price leadership method for solving the inspector's non-constant sum game. *Princeton econometric research program*, 1963.
15. D. Monderer and L. S. Shapley. Potential games. *Games and Economic Behavior*, 14, 1996.
16. M. Naor and B. Pinkas. Efficient oblivious transfer protocols. In *Proc. of the 12th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 448–457, Washington DC, USA, January 2001.
17. M. Osborne and A. Rubinstein. *A course in game theory*. MIT Press, 1994.
18. P. Rogaway, M. Bellare, and J. Black. OCB: A block-cipher mode of operation for efficient authenticated encryption. *ACM Transactions on Information and System Security*, 6(3):365–403, August 2003.
19. R. W. Rosenthal. A class of games possessing pure-strategy Nash equilibria. *International Journal of Game Theory*, 1973.
20. T.Vila, R. Greenstadt, and D. Molnar. Why we can't be bothered to read privacy policies: Models of privacy economics as a lemons market. In *Second International Workshop on Economics and Information Security*, 2003.
21. B. von Stengel. Recursive inspection games. *Technical Report S-9106, University of the Federal Armed Forces, Munich.*, 1991.
22. N. Zhang and W. Zhao. Distributed privacy preserving information sharing. In *VLDB*, 2005.