Essays in Mechanism Design

Inauguraldissertation
zur Erlangung des akademischen Grades
eines Doktors der Wirtschaftswissenschaften
der Universität Mannheim

vorgelegt von

Alia Gizatulina

Mai 2009

Abteilungssprecher:	Prof. Dr. Enno Mammen
Referent:	Prof. Martin Hellwig, Ph.D.
Korreferent:	Prof. Dr. Ernst-Ludwig von Thadden
Tag der Verteidigung:	13.07.2009

Contents

Acknowledgements

I would like to thank Martin Hellwig for everything I have learnt from him, for his confidence and unmatched patience with my ideas, questions and struggles. The depth and width of his own knowledge in economics and his enthusiasm about a whole variety of issues have been providing the freedom to explore any question and let me never be ennuied by doing economics.

I also would like to thank Ernst-Ludwig von Thadden as the director of the doctoral school for his openness to discuss any possible matter, his solutions to those matters, his support of conference trips and academic visits and for all encouraging feedback on my research.

Of course I would like to thank all my friends who have been with me on this path along the years. All "Les Russes" in Toulouse and especially Elena and Igor who, during my studies there, were transmitting to me tirelessly their own knowledge in math and economics; without it I think I would have never dared to start a PhD. Sava and Xuanlai for their camaraderie in sharing years in the terrific L13 of Mannheim. Nina and Evguenia for our fun during lunches in the Ehrenhof. Hélène and Jan for all their support and a countless number of warm dinners at their home in Heidelberg. Johannes for all care and for being a big source of rational decisions. Giulio who has been reminding me nicely of a bigger picture. And I owe a lot to Aude for indiscribable amounts of all-dimensional help and for having shown me all those unforgettable mountains that made many week-ends so beautiful.

But most of all I would like to thank my mother for her optimism and strength which, despite the distance between us, have supported me during these years like nothing else.

Introduction

The subject of the thesis

The main purpose of my thesis is to study different aspects of design of mechanisms or games that economic agents could play in order to achieve outcome that would improve welfare of everyone.

Because agents' selfish and uncoordinated behaviour may result in socially and individually inefficient outcomes, all agents may agree that institutions or collective mechanisms, curbing individual selfish interests should be introduced in a way that does not damage freedoms of anyone on the other hand. The mechanism design literature working on efficient mechanisms analyzes exactly this problem – which economic institution is the best from the perspective of each individual to achieve a given allocative goal in a given environment. Among examples of such designed institutions are auctions, voting mechanisms or tolls on highways introduced to cover the costs of highway.

As quite often agents payoffs from their behaviour and hence their incentives to behave that or another way are known only by agents themselves the optimal design of a mechanism should account for such unknowns – any allocation that could result from agents behaving within the rules of the designed institution should be incentive compatible, i.e. optimal from agent's individual perspective given his payoff from this allocation, even if this payoff is known only by each agent himself.

Also, as in most of reasonable situations agents have a freedom to opt away from the allocation that they may expect from the mechanism, the anticipated allocation should be individually rational, i.e., not worse than what they could achieve by not participating in the proposed game.

Finally, as often a group of agents may not count to have additional

resources from an external world any outcome resulting from a play within the designed institution should satisfy the budget balance condition.

The three works in the thesis contribute exactly to this branch of the literature – design of socially optimal mechanisms which would satisfy incentive compatibility, individual rationality and budget balance conditions. Though, being unified by this theme, three chapters touch quite different sides of this vast subject.

Chapter I

In the first chapter I study a problem of optimal design of a contribution game which agents should play in order to collect resources to cover up-front costs of a common punishment system that would enforce their trades with each other.

In many economic transactions agents have some scope for behaviour which is individually profitable but damaging for a trading partner. For example, a buyer, once he receives a purchased good could decide to skip paying some part of the price if it should be made over time. The seller, in turn, may renege on promised warranty services once the buyer has paid all what is due. Anticipating such behaviour most of agents are likely not to start trading at all. Nowadays, for most of such transactions, trading agents may conclude a contract specifying obligations of parties to each other and which would be subject to enforcement by the official legal system.

In a small group of agents, repeating interactions and reputation concerns usually allow agents to sustain mutually beneficial behaviour without any threat of recurrence for arbitration to some third party like a legal system. However, as Dixit (2003b) has argued, when an economy's size grows on the one hand there are likely to be large gains from trade expansion beyond a small group but on the other hand there is a decrease in possibility for repeated interactions. As a result, in order to sustain cooperation and to perceive gains from trade a third party enforcement services should become available.

In order to resolve any single dispute a legal or other contract enforcement system needs to invest a substantial amount of resources into its capacity up-front, which no single individual is likely to be able to cover on his own. For example, if it is the legal system, it needs resources to generate

laws, to establish a court to adjudicate and a police to enforce the court's decisions. As those are done by agents, their incentives to do it properly should be aligned accordingly. Dixit (2003a) demonstrates that curbing opportunistic incentives of adjudicators has non-negligible costs which a single individual may not be able to bear.

That is why it seems to be important to understand how a large group of agents may organize itself in order to cover all necessary costs and to introduce a collective trade enforcement institution. Obviously, each agent's incentives to contribute to a collective system which would enforce his trades depends on how much he gains from trading and what are alternative ways for him to enforce his rights within a contract. Usually agents are heterogeneous in these gains and so they differ in their valuation for a given contract enforcement system. To capture this, in the paper I model explicitly details of two different contract enforcement systems through which agents may enforce their trades.

The first system is called asymmetric. It enforces agreements of agents unequally and depending on their exogenously heterogeneous resource endowments. Namely, in a match of any two agents such that one has more resources than another, an agent who is stronger, when he cheats on his weaker encounter, is able to avoid punishment. Whereas if it is a weaker agent who cheats on the strong one he is punished by the latter who receives from the weak a dedamaging payment. This model arguably captures many inefficient legal systems that exist in the world.

The second system is symmetric, in the sense that it punishes for misbehaviour independently of agents' private resources. Whereas the asymmetric system does not have any fixed costs to function, this system, to be efficient and impartial, requires an up-front investment. Because of its fixed cost and given that it is largely non-rival (as once it is on place no agent has incentives to cheat on his encounter anticipating punishment) it has properties of a public good. And moreover as it is possible to exclude agents from its services, it is an excludable public good.

The analysis of distribution of gains from trade under two systems gives the following. If the asymmetric system is efficient and capable to impose high punishments for misbehaviour these are the strong who prefer that the only contract enforcement system that exists is the asymmetric one. For relatively low levels of imposed punishment these are the weak who benefit

from the asymmetric system. Hence two systems cannot be Pareto ranked, though on aggregate the symmetric system brings a higher social welfare.

Turning to the contribution game, the first observation is that because agents have to collect resources before the very first trade is made, i.e. before a trading partner and his resource strength are known, it is impossible to condition directly each agent's contribution to the system on his valuation for the symmetric system, i.e. on his resources as those are unobservable at the ex ante stage. Hence this is the game of incomplete information and given that I consider a large economy, as it is known from the existing literature (see Hellwig (2007)), the only mechanism which may collect a positive amount of resources for a public good is the mechanism based on exclusion of those who do not pay for the public good from its consumption. In other words, I could directly search for an optimal mechanism within a class of mechanisms called "the admission fee mechanisms", where each agent pays an admission fee for being allowed to use the public good.

The following observation is that because admission to punishing by its very nature is two dimensional, i.e. each agent could be allowed to punish his trading encounter and he could be allowed to be punished himself, exclusion of non-payers takes also a two-dimensional form. Hence, in a trade of anyone who is admitted to the symmetric punishment system and the one who is not, there may be three possible way to exclude non-payer: 1. "Full" – he may be excluded from both ability to punish and ability to be punished; 2. "Partial" – he may be excluded only from ability to punish, but he could be yet punished by anyone who has contributed to the costs; 3. "Quasi-none" – if at least one agent in a match has contributed to the costs of the symmetric system in their trade both of them could punish and being punished by that system.

The result for the contribution game are the following. If the costs of the symmetric punishment system are very high, it is optimal to employ the admission mechanism based on the full exclusion rule. For an intermediary level of costs, the partial exclusion rule would suffice. And finally for low costs it is the quasi-none exclusion rule that is optimal. The optimal choice among these exclusion rules happens not to be fully free and dependent only on the costs of the efficient contract enforcement system. Each of the exclusion rules allows for a multiplicity of equilibria and different equilibria result in different levels of the social welfare. For a given exclusion rule, the

worst equilibrium brings lower or higher welfare than the worst equilibrium under another exclusion rule depending on the parameters of the trade (i.e. gains from cheating on the trading partner) and the level of punishment imposed by the systems.

To summarize, in general, for each level of the costs of the symmetric system which it is efficient to cover from a collective perspective, there is a an exclusion rule which induces agents to pay their share to the cost. But this exclusion may generate coordination failures and demand for services of the asymmetric contract enforcement institutions may persist even if each of agents who remains under the asymmetric system would gain if he starts to govern his trades via the impartial institution.

Chapters II and III

A Unifying Theme

The second and the third chapters of my thesis are the joint project of Martin Hellwig and me. A main idea of the project is to study robustness of mechanism design results obtained in environments where agents types are modelled by subset of types from the payoff-based universal type space.

Harsanyi (1967/68) has suggested and Mertens and Zamir (1985) have formalized that all strategically relevant information of an agent, i.e. his payoff characteristic and his beliefs about others' payoffs and others' beliefs could be compactly represented by his "type". Thus, for any one who wishes to analyze a given problem in a given environment with incomplete information it suffices to specify a profile of "types" and two mappings, one defining how each agent's types map into his payoffs and another defining how each agent's types map into his beliefs about others types, and so, recursively, into his beliefs about others' payoffs and others' beliefs about everyone else's payoff and so on. A collection of all possible payoffs and belief mappings together with corresponding abstract types constitutes what is called the payoff-based universal type space.

Neeman (2004) and Heifetz and Neeman (2006) have demonstrated that the result of Crémer and McLean (1988) holds true only for a very special and "small" set of types from the entire universal type space. Specifically, it is valid only in type spaces where payoff and belief mappings are such

that with each possible belief type there associated a unique payoff type, as Neeman (2004) and Heifetz and Neeman (2006) name them – the beliefs-determine-preferences (BDP) environments. Heifetz and Neeman (2006) parametrize all type spaces by priors that generate those types and show in geometric and measure theoretic senses that priors generating BDP types in the universal type space are non-generic.

In the papers we argue that the results, including the ones of Heifetz and Neeman (2006) and Neeman (2004) obtained under the payoff-based universal type space assumption, may lead to misleading conclusions. The main reason for this is that types from the payoff-based universal type spaces do not account explicitly for correlation in agents' information. In the paper constituting the Chapter II we show that the BDP property is generic if agents' heterogeneity in belief hierarchies is due to heterogeneity in information, i.e. if beliefs mappings are endogenous – they are derived from a common prior conditionally on agents' information. In the paper of Chapter III we show that for a given subset of types from the payoff based universal type space there may exist a variety of models accounting explicitly for agents payoff characteristic and informative signals and where agents' hierarchies of beliefs about payoffs would be exactly the same. However the implementation possibility results may vary depending on the details of the model.

Chapter II

We prove that in most of quasi-linear environments where agents share a common prior about payoff uncertainty, any objective which is efficient from a social point of view could be achieved via a lottery mechanism where each agent is volunteer to participate. We allow agents' private information to be multidimensional – each agent possesses private information about his payoff characteristic (which is one-dimensional) and he also receives a finite number of signals giving him information about others' payoffs and others' signals.

The key feature why the mechanism designer is able to achieve any socially efficient outcome is due to the fact that under most of common priors each agent's beliefs about others payoffs and others' signals conditional on his own information are fully informative about that information, i.e. about his payoff characteristic and his signals. Thus, as it is known, in quasi-linear

environments there exists a mechanism, developed by Crémer and McLean (1988), where by offering a menu of lotteries whose payoffs are conditional on the reports of all other agents, the mechanism designer could extract perfectly agents' beliefs, by observing individual choices among these lotteries. Consequently, if agents beliefs are unique to their private information, the mechanism designer could learn it automatically as well from beliefs and achieve allocative goals as if he had full information from the beginning.

Hence the main purpose of the paper is two show that for "most" of common priors agents interim beliefs are unique to the information on which they are conditioned. We distinguish two cases – where agents' types (payoffs and signals) belong to finite sets and where those are from intervals. The proof techniques differ across two cases.

For the case with finite types the proof hinges upon the result that in a space of all finite dimensional matrices, matrices having a full rank, i.e. having its columns and rows linearly independent, are generic. Agents belief types, conditional on their payoffs and signals are just rows in a matrix where columns are parametrized by all possible constellation of payoffs and signals of other agents. This matrix is proportional to the matrix of a prior distribution of types and so if the matrix of posterior beliefs does not have linearly dependent rows, nor the matrix of prior beliefs does have them.

For the case where agents types are from intervals the above technique could not be applied anymore. Instead we employ the result from the differential geometry – the Whitney Embedding Theorem. The Whitney Embedding Theorem states that in the space of all (continuously differentiable) mappings from a low dimensional space into some space of a higher dimension (which should be no less than twice as large than the support space) injective mappings constitute an open and dense set, i.e. they are generic. Because each agent's beliefs about others types is a projection from a space of a dimension equal to the number of his types, i.e. signals into a space which has dimensionality of the sum of dimensions of the types of all other agents, provided that these dimensions satisfy the Theorem requirement, we obtain the result. That is generic interim beliefs that an agent could have about types of all others are injective mappings. As for each dimension interim beliefs are linear in the prior, it means that generic priors induce beliefs which are fully informative about the signals on which they are conditioned.

Chapter III

In this chapter we demonstrate that in general a set of social choice functions which are implementable in a given set of types from the payoff-based universal type space is a strict subset of the set of social choice functions implementable on explicit models of information having the same payoffs and beliefs hierarchies as in that subset of types of the universal type space. Our proof is constructive – we describe an environment and a social choice function where this result holds.

The existing game theoretic and mechanism design literature has widely adopted the payoff-based universal type space construction because any given model of private information could be represented by some set of types in the payoff-based universal type space. However it could happen that a given subset of types from the payoff-based universal type space may be represented by several different strategic models. And so restricting attention to the payoff-based hierarchies of beliefs is without loss of generality only if the predictions of behaviour of types from the universal type space do not vary along the models that may be used to represent that given set of types. Ely and Peski (2006) and Dekel, Fudenberg, and Morris (2007) have demonstrated in a game-theoretic setting that predictions on a set of rationalizable strategies depend on the details of the model that is used to construct belief hierarchies about payoffs.

In our paper we verify the sensitivity of the mechanism design analysis to details of a particular specification of a type space. We show that indeed if some social choice function is not implementable on a given subset of types from the payoff-based universal type space, it does not mean that it is not implementable on other type spaces, accounting explicitly for correlated information, even though agents have the same payoff-based hierarchies of beliefs as in that subset of types from the universal type space.

The intuition behind this result is due to the fact that there may happen jamming of information on the way of transition from a specific model of information, i.e. a model which is explicit about agents payoffs and signals, to a set of types from the payoff-based universal type space where only belief hierarchies about the cross sectional distribution of payoffs are specified. By modeling only "terminal" hierarchies of beliefs about payoffs and not looking at the details of a model that gives rise to these hierarchies we naturally

may lose at least two types of information – information about agents conditional beliefs about others signals and values of signals themselves. Whereas conditioning agents' allocations on reports about this type of information may allow to relax incentive compatibility constraints and thus to achieve a broader set of allocations.

We conclude this paper with a discussion on when actually restricting attention to payoff-based hierarchies of beliefs does not entail a loss of generality. Obviously it is true when by knowing agents belief hierarchies about payoffs we have as much information as if we had known their payoffs, signals and beliefs about those. We discuss when this is likely to happen and it seems that it is likely to be true under the assumption of common priors and it is quite unlikely to be true when we cannot assume that agents priors are known, i.e. they should also constitute their private information from the very beginning.

Chapter 1

Endogenous Trade Enforcement Institutions

Abstract[1] A large population of agents heterogenous in strength endowments match one-shot to trade with each other. Two competing contract enforcement institutions are available to the population: a costless but asymmetric that imposes a punishment for cheating on weak agents but not on strong ones and a symmetric institution that imposes punishment for cheating independently of agents identities but which requires a collective investment into its capacity up-front. The paper shows: (i) Sometimes strong agents receive rents under the asymmetric system, but sometimes these are the weak who benefit from it; (ii) There exists an equilibrium where all agents, independently of their strength, pay to the fixed cost of the symmetric system provided it excludes non-contributors from enforcement of their contracts with contributors; (iii) The rate of cheating is always positive in contracts subject to asymmetric punishment and it is zero in contracts subject to impartial punishment.

[1] I am specifically grateful to Martin Hellwig for all patient and rich in ideas discussions of this research. I also would like to thank for helpful comments Avinash Dixit, Bruno Jullien, Elena Panova, Emanuele Tarantino, Eugenia Winschel. Questions from the audiences of ESNIE 2006, EEA 2007, ASSET 2007, AEA 2008 meetings and Universities of Mannheim and Toulouse were helpful to shape ideas better.

1.1 Introduction

Most of exchanges, while being mutually beneficial, are vulnerable to opportunistic behaviour of the trading agents themselves. When damaging behaviour from an encounter is to be expected, agents may decide not to start any trade at all at the first place. Hence, unless some punishment for misbehaviour is available to agents, the level of economic activity is likely to be inefficiently low.

When agents are able to have repeated economic interactions, in most cases reputational concerns help to enforce mutually beneficial behaviour within a trade. However when the size of a market grows, the reputation-based governance usually does not work well any longer. Instead, a rule-based third party governance is to be introduced (Dixit (2003b)). But in a large economy an efficient third party enforcement institution is likely to require an investment into its capacity up-front (Li (2000)). One should cover "efficiency wage" of enforcement agents, expenses on data-bases on previous cheating, investment into design of the efficient and up-to-date laws and codes, etc. Moreover as these features are likely to be complementary, failure to cover a part of these costs may result in dysfunctionalities, in other words all expenses sum up into a fixed cost of the efficient contract enforcement system.

This paper studies how an efficient contract enforcement system[2], which requires a fixed cost[3] to be covered, could collect these resources when it competes with an alternative, second-best contract enforcement systems. The efficient system is able to enforce trade agreements flawlessly (it will be called "the symmetric punishment system", or SPS, for short). Whereas under governance of the second best system some agents are more likely to obtain justice than others if they have a higher resources endowment. In a match of two agents such that agent i is "stronger", i.e. it has a higher endowment than agent j, agent i is able to punish dishonest behaviour of j. While if i cheats on j, j having less resources than i is not able to get

[2] In this paper I am concerned only with contract rights, i.e. I do not consider issues of thieft or extortion (i.e. violation of property rights).

[3] This cost is necessary to weaken all possible informational and cognitive constraints observable in reality. Paying higher "efficiency wage" to enforcers to make them immune to bribes or hiring enough of investigators to find for sure the true state of things would always improve the quality of enforcement.

compensation from i. In other words, j is "weak". The relative advantage of this system is that it functions at zero costs to agents[4]. It is called "the asymmetric punishment system" (APS).

Then the precise question of this paper is:

- Does there exist a contribution mechanism which would allow the symmetric punishment system to collect resources to its fixed cost given that each agent could enforce his contracts alternatively through the costless but asymmetric system?

But to answer this question, one should understand several preliminary issues: When the symmetric punishment system is not available, who benefits from contracting under the APS? How valuations for the SPS and the APS are distributed in the economy depending on gains from trade, gains from cheating and the distribution of endowments?

Answering to preliminary questions and looking closer at the functioning of the second best system may be of its own interest, as in the reality one observes such type of enforcement systems quite often. Many developing countries have rather asymmetric official legal systems. For example, in India only relatively resourceful agents are able to obtain justice (*The Economist*, December 26, 2006 provides examples of this). The developed countries do not necessarily have the first best system either. The importance of resources for performance within the American legal system is discussed, for example, in Galanter (1978)[5]. He theorized on different dimensions of strength bringing advantages to "the haves" over "the have-nots" within the American legal process (legal experience, financial resources and political power affect positively and non-negligibly the probability of winning the case[6]).

[4] The zero marginal cost of the APS can easily be justified if one accepts that the sector contains at least two service providers and they compete in prices. Introduction of any non-zero fixed cost for the APS does not provide qualitative changes, as what matters for the results are the relative fixed costs of two punishment systems.

[5] For another example of the importance of resources in the Italian courts see Enriques (2002).

[6] His analysis has provoked a number of subsequent studies aiming to prove or disprove, in a quantified way, the role of the resource factor in the probability of winning in courts. Lempert (1999) reviews many such studies and makes a sharper statement that it is political power and financial means that are more relevant for litigants' performance in courts.

Failure to be impartial happens to private enforcers as well. Several case studies by Hill (2003) and Hill (2006) of Japanese Yakuza demonstrate that often the "winning" party of a trade dispute is the one who is able to hire the strongest gang who arguably charges a higher price for its services (Yakuza services are sought to enforce bankruptcy law, to recollect debts in general or to enforce anti-competitive agreements among colluding companies). The study of the Italian Mafia by Gambetta (1997) (theorized later by Dixit (2003a)) makes a similar case that incentives for impartiality of Mafia enforcers are costly.

It is surprising is that despite of the apparent inefficiency of the partial systems the demand for their services does not seem to vanish completely. Inefficient legal systems are still in place and impersonal markets under those systems do not disappear. Similarly, demand for services of Yakuza is non-negligible even by the least empowered agents (Hill (2006)). While answering to the preliminary questions outlined above this paper highlights some of the reasons for this phenomenon.

There are several remarks on modelling. I study in a detailed way only the demand side of the market for enforcement services and do not consider micro-foundations of the enforcement system, e.g. incentives of enforcers. Similarly to Dixit (2003a), a transaction where each agent has an opportunity to profit at the expense of his trading partner is modelled by a two-sided prisoner's dilemma game. That is, agents could exchange goods or services, or produce something together, but receiving a positive benefit from a transaction by both agents is conditional on both agents behaving cooperatively[7].

As "experience" or "technicity of issues" are matters that could be overcome via the employment of top lawyers and technical experts. Sheehan and Songer (1992) quantitatively prove the presence of advantages of "the haves" in the United States Courts of Appeal. They show that the Federal Government is likely to win against local governments, big firms perform better than small firms, any government performs better than a firm, any individual is less likely to win than any firm, and any minority group individual or a person from the bottom of the income distribution is less likely to win than the government, a firm and any other non-minority or a relatively well-to-do individual.

[7] There are many possible examples of opportunism within bilateral contracting: a buyer cannot observe immediately the quality of the good proposed by the seller, and if the latter provides him with a lemon he cannot reject it at the moment of the exchange (and by doing this provide correct incentives to the seller ex ante); the buyer himself could renege on the payment if, for example, it is to be done in a while. Essentially, for any transaction nowadays if there is a contract, it means there is a scope for one-sided or

If both agents misbehave at the same time, the value of the exchange is the lowest.

The analysis of the preliminary questions gives following results. When applied asymmetrically punishment is of intermediate severity, any APS is beneficial *to every agent,* even the weakest one, as compared to the world with no punishment at all. In this case everyone participates in the market and applies for enforcement services to the APS. This system insures a strictly positive rate of equilibrium honest behaviour but it is never able to eliminate cheating fully by varying the level of imposed punishment. The reason for this is that contracting on the market can breakdown completely if the applied punishment is too low or too high, as stakes of cheating become in both cases important, agents prefer to cheat and so expected gain from trade on the market is lower than from not trading at all. By contrast, in the system with symmetrically imposed punishment, the rate of cheating under a wide range of the parameters is zero. In this case everyone prefers to enter the market and trade rather than stay outside. But two systems cannot be Pareto ranking between themselves at the intermediate values of punishment. As if punishment is moderately high, the strong agents obtain rents from trade under asymmetric governance relative to the symmetric one and prefer this to be the only mode for everyone. If punishment is moderately low, these are the weak who obtain rents under the APS as compared to the SPS. These results are independent of the shape of a distribution of endowments in the economy.

Now some remarks should be made before giving answers to the main question of the paper. One could recognize that the SPS possesses properties of an excludable public good: it is non-rival (in equilibrium it is only a threat as no one cheats, hence no one uses its services, although it should be still a credible option) and it is technically possible to exclude agents from enforcement of their contracts. Hence the problem of symmetric punishment system is the problem of a monopolistic provider of an excludable public good with one caveat that in case of the punishment system the "exclusion" is naturally two-dimensional. Each agent could be excluded from both *ability to punish* and from *ability to be punished*. So exclusion could be "partial" – when agents are excluded only from a possibility to punish their encounters; "full" – when agents are excluded from both, a possibility to punish their

mutual misbehaviour that this contract aims to prevent.

encounters and possibility of being punished by them; and finally what I call "quasi-none" – when both agents are allowed to punish each other, provided that at least one of them is admitted to the symmetric system.

The main result for the contribution game is that when the applied exclusion rule is sufficiently severe (i.e. it is either "partial" or "full", but not "quasi-none"), there exists an equilibrium in which *all agents*, even the strong, who prefer asymmetric governance to be the only mode available to everyone, subscribe for the services of the symmetric punishment system. That is the SPS is able to crowd the APS fully out of the market. The intuition is when all other agents are under protection of the SPS the strong agents can no longer expect any rents to their strength if the exclusion is partial. In case of full exclusion no one would wish to trade with the strong altogether, as they are excluded from ability to punish and ability to be punished and so the only equilibrium behaviour within contract is mutual cheating. That is why, unless the strong submit themselves under the order of the SPS as well, their perspectives are rather poor[8]. Whereas by paying to the SPS the strong agents commit themselves to be subject to punishment, to avoid mutual cheating and so to become attractive to others to do trade with them.

However the full subscription is not the only possible outcome, as the analysis demonstrates, a decentralized competition between two system could bring equilibria where the social welfare could be quite low because a costly SPS attracts only a fraction of agents who over-pay for its services whereas another fraction of agents remains under the APS and so two systems inefficiently coexist. This happens because agents valuations for each of systems are interdependent and the more sever is the exclusion rule that the symmetric system applies the higher is interdependence.

The related literature belongs to two different streams. First, it is research on institutions governing economic transactions, especially Dixit (2004). However Dixit (2004) does not cover explicitly the issue of competing third-party enforcement institutions one of which has to collect resources for its up-front fixed cost to provide a quality service. Another stream of related

[8] This result is reminiscent of Hobbes (1651) and Rousseau (1762). Individuals are willing to conclude a "social contract" with a third-party on their protection against others' cheating and in return a third party is allowed to take them liable for their own misbehaviour with respect to the others.

literature corresponds to the optimal design of public good provision mechanisms in a large population of agents. Specifically Hellwig (2003), Hellwig (2007), Mailath and Postlewaite (1990) and Norman (2004) provide a related analysis. Given that the type of the problem studied here is a particular one, namely exclusion from punishment is inherently two-dimensional, it makes the results available in the literature for public goods provision not to be directly applicable.

The paper is structured as follows. The next section studies individual behaviour and resulting payoffs under two punishment systems. Section 3 gives the results of the subscription game proposed by the SPS and discusses welfare implications. Section 4 contains some extensions, in particular, when the APS has a monopolistic positive pricing for its services. Section 5 concludes.

1.2 Transactions under Different Governance Modes

1.2.1 Bilateral transactions without any governance

An economy consists of a continuum of agents of mass 1. Agents match in pairs one-shot to trade with each other on some markets and as it was justified already in the introduction, each game, played bilaterally, has the strategic properties of the prisoner's dilemma game.

The next matrix presents a simple symmetric version of this strategic situation. H stands for playing "honest" and C for "cheating" behaviour:

	H	C
H	$1, 1$	$1-\nu, 1+\nu$
C	$1+\nu, 1-\nu$	$0, 0$

(1.1)

Here ν is the stake of cheating and it is such that $\nu > 1$. The only equilibrium strategy here is to misbehave for both agents.

1.2.2 Punishment

Under *asymmetric governance* agents are heterogenous in ability to punish dishonest behaviour of their encounters. This is captured by the parameter θ which is distributed in the population according to a continuous *cdf* $F(\theta)$ with a support $\left[\underline{\underline{\theta}}; \overline{\overline{\theta}}\right]$, $\overline{\overline{\theta}} - \underline{\underline{\theta}} > 0$. If an agent i matches with an agent j such

that the realization of types is $\theta_i > \theta_j$, an agent i is called the *strong (*he has $\theta = \overline{\theta}$ within the match*)*, he is able to punish for sure the deviation of j, if it occurs. Agent j is then the *weak* (he has $\theta = \underline{\theta}$ within the match) and if the strong agent cheats him, by contrast, he is not able to obtain a compensation.

The payoff matrix of the transaction is as follows, the row agent is the strong and the column agent is the weak:

$strong \backslash weak$	H	C
H	$1, 1$	$1-\nu+\rho, 1+\nu-\rho$
C	$1+\nu, 1-\nu$	$0, 0$

. (1.2)

The parameter ρ reflects the effective level of punishment[9] imposed on a cheating player[10]. This parameter may be though about as an expected punishment, i.e., it may be a probabilistically imposed punishment, where only with some probability π $(0 \leq \pi \leq 1)$ some fine R is imposed. Probability π would reflect the general quality of asymmetric punishment system whereas R could reflect a statutory amount of punishment. Thus one has $\rho = \pi R$, though given that π and R are not to be taken as endogenous choices without any loss the analysis will be given directly for ρ and it will be referred to both as punishment and efficiency of enforcement system.

Note that neither player can be punished if both agents have deviated, i.e. if the state $\{C, C\}$ has occurred. This assumption can be justified by resource constraint of agents (e.g. when nothing is produced together, nothing can be taken from each other).

[9] I incorporate directly the punishment into the payoff matrix, although there is certainly some multistage negotiation/litigation process behind. Once there is a complete information about the individual behaviour within a transaction, taking the reduced form of a dynamic bargaining model is w.l.o.g.

[10] This parameter may consist of several parts in reality, for example it may be a probabilistically imposed punishment, where only with some probability π $(0 \leq \pi \leq 1)$ a fine R is imposed. Probability π may reflect the general quality of asymmetric punishment system whereas R could reflect, for example, discretion at which the punishment is imposed (as R can be higher or lower than the cheating stake ν). Then ρ can be thought about as $\rho = \pi R$. However, under assumption of complete information among all agents about types and given that π and R are not going to be endogenized in this model, it is also without any loss that one can analyze directly for ρ.

Both agents observe each other's type at the moment of match and prior to the choice of their optimal strategies within a transaction.

Hence, potentially, every agent can be in one of two possible games: where he is strong and where he is weak. The ex ante probability (prior to the realization of matching uncertainty) to be in either game is coming from agents' type θ. The higher is θ_i, the more agent i is likely to be in the game where he is strong as the probability to be of a higher type than agent j is given by $F(\theta) = F_j(\theta_i) = \Pr(\theta_j < \theta_i) = \int_{\underline{\theta}}^{\theta_i} dF_j(\theta)$.

I assume that the service of asymmetrically imposed punishment does not cost anything to any agent[11]. I discuss in the section 4 the results when there is a monopolistic provider of the asymmetric punishment services who makes take-it-or-leave-it pricing offers.

As for the *symmetric governance system*, by definition, it is able to punish cheating of any agent, independently of his whatsoever identity. The payoffs matrix for a match is then:

$strong \backslash weak$	H	C
H	$1, 1$	$1-\nu+\rho, 1+\nu-\rho$
C	$1+\nu-\rho, 1-\nu+\rho$	$0, 0$

(1.3)

Again, as in the previous case, I assume that if both agents cheat on each other there is no way to punish both of them simultaneously. The fixed cost of this system is equal to K.

In the following section I give results on the equilibrium strategies chosen by agents under each governance mode depending on their type θ.

1.2.3 Individual Behaviour under Two Punishment Modes

Equilibrium Payoffs from Transactions under Asymmetric Punishment

Suppose only the system with asymmetrically imposed punishment is available to everyone. Agent i plays H with probability $\zeta_i(H)$ and C with the complementary probability $\zeta_i(C) = 1 - \zeta_i(H)$. Exploring the matrix (2) one

[11] The reason of why I assume away all cost considerations is to distill out the net impact of *asymmetrically* imposed punishment on agents utility.

can see there are three possible types of equilibria, depending on the values of the parameters:

Case : $\nu < \rho$

Under such constellation, there is no pure strategy equilibrium in the game specified in (1.2). The equilibrium mixed strategies are:

$$\zeta_i(\{H\}, \underline{\theta}) = \frac{1-\nu+\rho}{1+\rho}; \zeta_i(\{C\}, \underline{\theta}) = \frac{\nu}{1+\rho} \tag{1.4}$$

for the behaviour of an agent i when he is weak.

And when agent i is the strong type:

$$\zeta_i(\{H\}, \overline{\theta}) = \frac{\nu-1}{\rho-1}; \zeta_i(\{C\}, \overline{\theta}) = \frac{\rho-\nu}{\rho-1}. \tag{1.5}$$

The punishment and the deviation stake have opposite effects on probability of honest behaviour from the strong and from the weak. The higher is ρ, the more honestly behaves the weak and the less honestly behaves the strong. The higher is ν the more honest behaviour comes from the strong and the less comes from the weak. As $\rho \to \nu$ an agent who is the strong behaves honestly with probability going to 1. But such ρ deteriorates the incentives of the weak agent to behave honestly.

Corresponding utilities to be the weak and to be the strong are:

$$EU^A(\underline{\theta}) = (1+\nu-\rho)\frac{\nu-1}{\rho-1} \tag{1.6}$$

$$EU^A(\overline{\theta}) = (1+\nu)\frac{1-\nu+\rho}{1+\rho} \tag{1.7}$$

Thus, any agent of type θ has an expected utility from entry in to the market, before learning the type of his match:

$$\begin{aligned} EU^A(\theta) &= F(\theta)\cdot EU^A(\overline{\theta}) + (1-F(\theta))\cdot EU^A(\underline{\theta}) \\ &= F(\theta)\cdot(EU^A(\overline{\theta}) - EU^A(\underline{\theta})) + EU^A(\underline{\theta}) \\ &= F(\theta)\cdot\tfrac{2\rho\nu(\rho-\nu)}{(\rho^2-1)} + (1+\nu-\rho)\tfrac{\nu-1}{\rho-1} \end{aligned} \tag{1.8}$$

The expected utility is strictly increasing in θ as $\rho > \nu > 1$.

Case: $\nu - 1 < \rho < \nu$

In this case the unique, pure strategies equilibrium is

$$(\zeta_i(\{H\}, \overline{\theta}), \zeta_j(\{H\}, \underline{\theta})) = (1, 0).$$

That is an agent who is the strong behaves honestly but the weak agent always cheats. It seems to be counterintuitive, but given the results above (see (1.4) and (1.5)) it is logical. The punishment becomes too low so the weak agents have incentives only to cheat, but the strong still prefer rather *to behave honestly in order to be cheated and receive as a compensation at least a fraction of the trade benefit.* This strategy yet provides him a non-zero utility ($= 1 - \nu + \rho > 0$) as compared to the strategy where they would be cheating too and everyone would obtain zero payoff.

The expected utility from entering into the market is

$$EU^A(\theta) = 1 + (1 - 2F(\theta))(\nu - \rho) \tag{1.9}$$

It is decreasing with θ.

Case: $\rho < \nu - 1$

Here, the punishment is so low that it does not preclude any misbehaviour from either agent and in particular it can no longer compensate the strong for their honest behaviour as it was in the previous case. So the unique, pure strategies equilibrium is $(\zeta_i(\{H\}, \overline{\theta}), \zeta_j(\{H\}, \underline{\theta})) = (0, 0)$. The expected utility from participating in the market is $EU^A(\theta) = 0$.

Hence one could see that the asymmetric punishment system improves upon no punishment at all as the level of honest behaviour could be positive. But the total gain of any agent i given θ_i depends on the number of those who participate in the market, as the next section demonstrates.

Entry into contracting under the APS

It is assumed that the initial decision to participate in the market is voluntary. Relatively weak agents are free not to enter into the market at all and hence not to submit themselves under unpunished acts of the strong. Or inversely, this may be the strong who would prefer not to deal with cheating weak agents and so not to start any trade activity. Consequently one should study the question how large is the demand for the APS services given the

parameters of the model. This is defined by the gains relative to staying outside of the market.

Assume that not participating at all in the market brings the utility equal to zero to any agent[12], independently of his θ. Because the expected utility from participating in the market is increasing with type θ when $\rho > \nu$, presumably the agents with low θ would be first who would prefer to take rather outside opportunity. Similarly when $\rho < \nu$, strong agents are the first who would stay outside of contracting. But it happens that non-participation by a subgroup of agents has population-wide implications and actually *no agent at the end enters the market if for a subgroup of agents it is not beneficial*. The following proposition provides with the details of this result.

PROPOSITION 1.1 *There exist two levels of punishment* $\rho^{\min} = \nu - 1 > 0$ *and* $\rho^{\max} = \nu + 1 < \infty$ *such that:*

- if $\rho \in \left[\rho^{\min}, \rho^{\max}\right]$ *the equilibrium payoffs for all types under asymmetric punishment are strictly greater than individual payoffs for all types under no punishment;*

- if $\rho \notin \left[\rho^{\min}, \rho^{\max}\right]$ *equilibrium payoffs are zero for all types, as they are in the absence of punishment.*

The formal proof is in the Appendix. But the intuition behind is when the lowest agent loses from participating in the market, and so he does not enter, the next to him agent does not have anymore any tiny chance to be the strong, and at most he can be only the weakest one. So, being now the lowest type, he strictly prefers not to enter either. And this process of taking outside option will continue till the moment where the very last, the strongest agent of the population remains alone, and so he does not enter either. A similar unraveling happens if the punishment is too low and the strongest agent has a negative utility from participating. This result has the following implication:

COROLLARY 1.2 *There is no* ρ *that would induce simultaneously both agents to play honestly, i.e. in this system of asymmetric application of the punish-*

[12] I assume also that when agents expect an outcome $\{C, C\}$ to happen in the market (which provides with payoff zero), they prefer rather to stay at home.

ment ρ (> 0) *the equilibrium* $(\zeta_i(\{H\}, \overline{\theta}), \zeta_j(\{H\}, \underline{\theta})) = (1, 1)$ *is unachievable.*

The proof is in the Appendix.

Hence the main conclusion this section comes to is that a system with asymmetric punishment, where only some agents deviations are punished, could be still beneficial individually and collectively as compared to a world when no punishment for misbehaviour is available. But the functioning of this system is quite fragile, as too large or too small amount of punishment that it might impose could make it totally unattractive for any agent. Of course this result of full unraveling was hinging on the assumption that everyone's outside option is equal to zero. For heterogenous or type dependent outside payoffs this result would be different and partial market participation, by some groups of agents, would be yet possible.

Equilibrium Payoffs from Transactions under Symmetric Punishment

Analyzing the game in (1.3) one can see that the unique equilibrium when the punishment is imposed symmetrically is to play honest for both agents provided that $\rho > \nu$. The resulting expected utility from the matching under symmetric punishment is

$$EU^S(\theta) = 1 \tag{1.10}$$

for each θ. *Consequently it is the way and not the amount of the punishment applied that allows to achieve an equilibrium in which both agents would behave honestly.*

Note, there is a uniform gain for everyone from installation of the symmetric punishment when the only alternative regime is to trade under no punishment at all, i.e. not to trade at all.

Comparing the Payoffs under Different Punishment Modes

One can observe that switching to a regime where for all agents only the symmetric punishment system is available has distributional implications. When $\rho \notin \left[\rho^{\min}, \rho^{\max}\right]$ each agent, even the strongest one benefits from

appearance of the symmetric punishment system. When $\rho \in [\nu, \rho^{\max}]$ relatively weak agents prefer the symmetric punishment to be the only mode of governance for everyone, because . When $\rho \in \left[\rho^{\min}, \nu\right]$ these are the strong who prefer the symmetric governance to the asymmetric (this can be seen comparing (1.8) and (1.10)).

On aggregate, when $\rho \notin \left[\rho^{\min}, \rho^{\max}\right]$ and $\rho \in [\nu, \rho^{\max}]$, the SPS brings higher aggregated payoffs than the APS.

In the remaining area of the parameters two systems are equivalent from the social welfare perspective.

1.3 The Contribution Game

Agents strength types, and hence their true valuations for the SPS, are likely to be unobservable at the ex ante stage, before they match with their trading partners and when they have to subscribe either for the SPS or the APS. Thus the entire analysis of optimal pricing for the symmetric services has to take this constraint into account.

1.3.1 General Result for Provision of a Public Good under Incomplete Information[13]

A general result in quasi-linear environments about pricing of a public good under incomplete information about agents' preferences is that there is no way to construct a game within a large population of agents where in equilibrium each agent could be made paying according to his valuation.

LEMMA 1.3 *If agent's type θ is private information, in an economy with a large number of agents, there is no way to install the symmetric punishment via a contribution game with type dependent payments and achieve the first best allocation.*

[13] Because utility is transferable across agents and installation of a system imposing symmetric punishment increases the aggregate welfare when cost of the SPS is not too high under complete information about agents' types there exists a vector of side-transfers among agents such that those who lose from introduction of the symmetric punishment system are compensated. A social contract having this outcome could be implemented via some multilateral bargaining procedure. But such payment scheme is obviously not viable when information about every individual's type θ is unobservable at the moment when agents contribute to the fixed cost.

Proof. See proof to the proposition 2.3 in Hellwig (2007). ■

The lemma implies that there is no way to learn individual information of agents by proposing a menu of incentive compatible, individually rational and hence type-dependent contributions bound to differentiated probabilities of provision of a public good.

The intuition for this result is that in a large economy, no agent is pivotal. It is known that in a quasi-linear[14] environment the unique mechanism that could collect type-dependent, incentive compatible contributions is Vickrey-Clarke-Groves mechanism (VCG)[15]. It imposes a tax on each agent for his report about his type. The tax is constructed in a way that it makes an agent i pay to the rest of the group an amount of money equal to the sum of changes in their utilities, with changes coming from the fact that there is a change in the outcome once an agent i has reported his preferences. This makes an agent i internalize fully the impact of his report on the rest of the group and hence to be truthful. However such payment scheme in general is not individually rational. Moreover as the number of agents grows, the probability that any agent's report about his preferences over a public good, however big or however small it is, would have any impact on the probability of provision of a public good is next to zero. Hence, the tax cannot vary much either, nor can it, as a result, provide with the incentives to tell the truth.

For a problem with non-excludable public goods it means that the probability of their voluntary provision in a large population is zero[16]. For a problem with excludable public goods it means that the only feasible contribution is a type-independent constant payment taking the form of an admission ticket. Hence individual exclusion should be applied, even if it is inefficient given that the public good is on place[17].

[14] And so it is in a linear environement, as here.

[15] The uniqueness of VCG for quasilinear environments was established by Green and Laffont (1979).

[16] More details on this result can be found in Mailath and Postlewaite (1990).

[17] There is an additional remark that should be made about feasibility of a voluntary contribution scheme. Here I assume the cost of a public good is actually proportional to the size of population (which is an infinite number of agents of mass 1). Otherwise there would be a scope for an incentive compatible provision of a public good via a voluntary contributions scheme. As it is shown in Hellwig (2003), in a large population of agents, if the cost of a public good is negligible compared to the size of the economy, provided that the lowest possible valuation is at least just above zero, the public good can be provided

That is why I in the following section I shall proceed for the analysis of optimal *uniform* pricing of the services of the symmetric punishment system which is, as justified in introduction, an excludable public good.

1.3.2 An Admission Fee Mechanism for Provision of the Symmetric Punishment System

The admission fee mechanism is a direct mechanism. It stipulates the probability of provision of a public good, individual payments and probability of individual admission to the public good given a profile of equilibrium reports about valuation types (see for example Hellwig (2007) or Norman (2004)).

By contrast to a usual admission mechanism where admission probability is one dimensional (stipulating probability of being allowed to enjoy a public good) here one needs to stipulate a two-dimensional admission probability because by the very nature of the punishment system agents can be excluded via two channels. An agent can be excluded from a possibility *to punish* his encounter (if the latter cheats him) and he can be excluded from a possibility *to be punished* by an encounter whom he has cheated.

The formal definition of the admission fee mechanism for provision of the symmetric punishment system is:

DEFINITION 1.4 *The admission fee mechanism to collect resources to cover the cost of the SPS K is $\Gamma = (\Theta, g(\cdot))$ where $\Theta = \prod_{i\in I}\Theta_i$ is a message space of all agents and $g(\cdot) = (y, t, \alpha)$ is the vector of outcome functions:*

$y : \Theta \to [0,1]$ *probability of emergence of the symmetric punishment system given the profile of agents messages;*

$\mathbf{t} : \Theta \to \Re_+$ *vector of individual contribution functions;*

$\mathbf{e} : \Theta \to \{n, p, f\}$ *exclusion rule (qualitative variable);*

$\boldsymbol{\alpha} : \Theta \to [0,1] \times [0,1]$ *vector of individual admission probabilities for a given exclusion rule.*

Hence instead of stipulating a single valued probability of admission to the symmetric punishment for given agent's report, the mechanism should

with a non-zero probability. The intuition for this is that in such environment a per capita contribution rate goes to zero and so agents can contribute indifferently the requested ε (equal to the lowest possbile valuation).

stipulate the probability of admission to both possibilities, i.e. to punish and to be punished. As the result, an agent i given his report about his type $\widetilde{\theta}_i$ could be excluded in three possible ways:

The full exclusion happens when agent is admitted to neither option (i.e. he cannot punish and cannot be punished). In more details:

(1) when i is admitted to the symmetric punishment and he meets j who is not admitted to the symmetric punishment, neither agent can punish misbehaviour of his encounter;

(2) when i is admitted to the symmetric punishment and he meets j who is admitted as well, either agent is punished in case he misbehaves;

(3) when i is not admitted to the symmetric punishment and he meets j who is admitted, neither agent is punished for any misbehaviour;

(4) when i is not admitted to the symmetric punishment and he meets j who is not admitted to the symmetric punishment, both are to deal within either asymmetric punishment system or none.

By contrast, **the partial exclusion** is applied when agent can punish the encounter's misbehaviour but cannot be punished for his own misbehaviour towards an encounter. In details:

(1) when i is admitted to the symmetric punishment and he meets j who is not admitted to the symmetric punishment, i can punish misbehaviour of j but agent j cannot punish misbehaviour of i if it occurs;

(2) when i is admitted to the symmetric punishment and he meets j who is admitted as well, any agent is punished if he misbehaves;

(3) when i is not admitted to the symmetric punishment and he meets j who is admitted, i is punished for his misbehaviour, but agent j, if he misbehaves towards i is not punished;

(4) when i is not admitted to the symmetric punishment and he meets j who is not admitted to the symmetric punishment, both agents are to deal within either asymmetric punishment system or none.

The **"quasi-none" exclusion rule**: where an agent who is not admitted to the symmetric punishment can still punish and be punished by his encounter if the latter is admitted:

(1) *when i is admitted to the symmetric punishment and he meets j who is not admitted to the symmetric punishment, i can punish misbehaviour of*

j and j can punish misbehaviour of i if it occurs;

(2) when i is admitted to the symmetric punishment and he meets j who is admitted as well, both agents are punished in case when misbehaviour of either occurs;

(3) when i is not admitted to the symmetric punishment and he meets j who is admitted, either agent is punished in case he misbehaves;

(4) when i is not admitted to the symmetric punishment and he meets j who is not admitted to the symmetric punishment, both are to deal within either asymmetric punishment system or none.

As the individual types are unobservable, by the revelation principle, an allocation prescribed by the mechanism should be incentive compatible. Denote each agent i's reporting strategy as $\widetilde{\theta}_i : \Theta \to \Theta$, if agent reports truthfully I denote his chosen report as θ_i^* and if he mispresents as θ_i'. with An allocation $(y(\widetilde{\theta}), \alpha(\widetilde{\theta}), t(\widetilde{\theta})$ achievable in the Bayesian Nash Equilibrium of the direct mechanism $\Gamma(\cdot)$ is incentive compatible if for each agent i the following holds[18]:

$$EU_i^P(\theta_i^*, \theta_{-i}^*, y(\cdot), \alpha(\cdot); \theta_i) - t^*(\theta_i^*, \theta_{-i}^*, y(\cdot), \alpha(\cdot); \theta_i) \quad (1.11)$$

$$\geq \quad EU^P(\theta_i', \theta_{-i}^*, y(\cdot), \alpha(\cdot); \theta_i) - t^*(\theta_i', \theta_{-i}^*, y(\cdot), \alpha(\cdot); \theta_i) \quad (1.12)$$

for a profile of truthful reports of $-i : \theta_{-i}^*$, types $\theta_i, \theta_i \neq \theta_i'$ and θ_{-i}; and under the realized punishment $P \in \{0, A, S\}$.

Given that each agent is free not to enter into contracting at all under any governance mode, an allocation induced by the provision mechanism is required to satisfy a contracting participation constraint:

$$EU_i^P(\theta_i, \theta_{-i}^*, y(\cdot), \alpha(\cdot); \theta_i) - t(\theta_i, \theta_{-i}^*, y(\cdot), \alpha(\cdot); \theta_i) \geq 0 \quad (1.13)$$

under the realized punishment $P \in \{0, A, S\}$.

Finally, the allocations should be compatible with the following budget constraint:

[18] Thus, for a given exclusion rule, the admission probabilities vector for each individual i contains two values $\alpha_i = (\alpha_i^p(\cdot), \alpha_i^{bp}(\cdot))$ with α_i^p being the probability that an agent i can punish an agent j and α_i^{bp} is the probability that an agent i can be punished by an agent j. For a match of two agents obviously the following equalities should hold $\alpha_i^p = \alpha_j^{bp}$ and $\alpha_i^{bp} = \alpha_j^p$.

$$\int_{\Theta^t=\{\theta:t(\theta)>0\}} t(\theta)dF(\theta) \geq K \tag{1.14}$$

The entire timing concludes the description of the contribution game:

1. Nature distributes θ among agents, every agent observes his θ_i;
2. Agents decide on their most preferred punishment mode out of $P \in \{0, A, S\}$ given K, θ, ρ and ν and expected payoffs from entry into the market under two different punishment modes;
3. A contribution mechanism to install the symmetric punishment stipulating a set of admissible strategies and corresponding outcomes is proposed[19];
4. Each agent i accepts participation in the contribution mechanism or not;
5. Those who accept play it, available punishment systems become known, agents make their transaction under the most preferred available punishment mode and obtain the final payoffs.

1.3.3 Highest Revenue Equilibrium of the Admission Fee Mechanism

In the following, I search for the optimal price $t^*(\theta)$ depending on the parameters (ρ, ν) and the exclusion rule such that it results in the highest collectible amount of resources for the SPS. As the cost K is exogenous, this highest revenue gives the boundary on the exogenous K above which it is not feasible to install the SPS. I evaluate how many agents subscribe for the SPS in the equilibrium with the highest revenue.

I provide the results of the admission fee mechanism for two different strategic situations. First, agents decide on contribution when the outside opportunity is to trade under the APS. In the second they consider contribution strategies when the outside option is no-punishment regime.

[19] The mechanism designer in this case may be the agents themselves.

There are some simplifying shortcuts in the analysis. In the following, I will restrict analysis only to the constellation $\nu < \rho$, i.e. where the asymmetric system is sufficiently efficient and able to withdraw at least the net gain from cheating ν. The valuation for the regime where everyone trades under symmetric punishment[20] as compared to the regime where only the asymmetric system is available, is monotonic in agents strength type when $\nu < \rho < \nu + 1$ (it is decreasing in θ); when $\rho > \nu + 1$ it is uniform in θ. Then, given this, one could simplify the analysis of the optimal mechanism as it would be natural to search for a threshold agent $\widehat{\theta}$ such that only agents below $\widehat{\theta}$ are admitted to the symmetric punishment and are charged a user fee $t^* = \frac{K}{\int_{\underline{\theta}}^{\widehat{\theta}} dF(\theta)}$.

In addition, I introduce another piece of notation: $F(\widehat{\widetilde{\theta}})$. This term denotes the total number of agents who claim themselves in equilibrium to be of type $\theta < \widehat{\theta}$ for an announced threshold of the mechanism $\widehat{\theta}$. This is needed because, when agents anticipate that a fraction of agents join the SPS their valuation (willingness to pay to be admitted to the SPS) changes depending on how big is the fraction of those whom they anticipate to join the SPS. That is why the number of those who announce them to be of "high valuation" ($\theta < \widehat{\theta}$) may be different from the actual cross-sectional number of agents below[21] $\widehat{\theta}$ (which is then denoted $F(\widehat{\theta})$).

Subscription when the APS is a viable outside option

The key result of the paper is presented in the following proposition:

PROPOSITION 1.5 *The highest collectible amount of resources for the SPS*

[20] To be precise on how is measured "valuation" for the symmetric punishment system $v(\theta, y) = EU^S(\theta) - EU^P(\theta)$ for $P \in \{0, A\}$, where $EU^P(\cdot)$ defines expected utility from the match under the alternative punishment system (asymmetric one or none). From inspection of (1.8) one can see that the valuation depends on whether the asymmetric punishment system is on place and how many other agents are admitted to it, as it depends on probability to be strong or weak $F(\theta)$ which in turn depends on decision of the others to be under asymmetric punishment system. For $\nu < \rho < \nu + 1$ if the asymmetric punishment system is active the higher is θ the lower would be $v(\theta)$. When there is no asymmetric punishment, individual valuation for the symmetric one is high for all agents, independently of θ and equal to 1 $v(\theta) = EU^S(\theta) - EU^P(\theta) = 1 - 0 = 1$)

[21] This implies implicitly that when agents mispresent their types they do so "continuously", i.e. there is no "holes" in an interval of types who mispresent their preferences.

and corresponding rate of subscription depending on the exclusion rule are as follows:

- $\int_{\Theta^t=\{\theta:t(\theta)>0\}} t^*(\theta)dF(\theta) = 1$ *and the rate of subsciption is 1 when the exclusion rule is full;*
- $\int_{\Theta^t=\{\theta:t(\theta)>0\}} t^*(\theta)dF(\theta) = 1 - EU^A(\underline{\theta})$ *and the rate of subscription is 1 when the exclusion rule is partial;*
- $\int_{\Theta^t=\{\theta:t(\theta)>0\}} t^*(\theta)dF(\theta) = \frac{1}{4}(1 - EU^A(\underline{\theta}))$ *and the rate of subscription is* $\frac{1}{2}$ *when the exclusion rule is quasi-no exclusion.*

Here $\int_{\Theta^t=\{\theta:t(\theta)>0\}} t^*(\theta)dF(\theta)$ denotes equilibrium aggregate amount of contributions. Feasibility of the SPS is then defined whether actual cost K is below or above this amount.

Proof. For each case I will consider incentives of one agents to subscribe for the SPS, given his type θ, when all the remaining agents subscribe (i.e. they claim valuations to be $\theta < \widehat{\theta}$). In particular I search what is the highest payment t that the remaining agent would agree to pay.

1. For the full exclusion rule this constraint is: $F(\widetilde{\widehat{\theta}}) \cdot 1 - t \geq 0$, hence when $F(\widetilde{\widehat{\theta}}) = 1$, $t^* \leq 1$. That is why $\int_{\Theta^t=\{\theta:t(\theta)>0\}} t^*(\theta)dF(\theta) = 1$.

2. For the partial exclusion rule, similarly: $F(\widetilde{\widehat{\theta}}) \cdot 1 + (1 - F(\widetilde{\widehat{\theta}}))EU^A(\overline{\theta}) - t \geq EU^A(\underline{\theta})$, and so when $F(\widetilde{\widehat{\theta}}) = 1$ any $t^* \leq 1 - EU^A(\underline{\theta})$ is individually rational. The maximum of aggregate resources then $\int_{\Theta^t=\{\theta:t(\theta)>0\}} t^*(\theta)dF(\theta) = 1 - EU^A(\underline{\theta})$.

3. For the quasi-no exclusion rule the threshold maximizing collectible resources is at the intermediate values of $F(\widetilde{\widehat{\theta}})$. Namely, an agent indifferent between joining the SPS and staying outside is defined by $1 - t = F(\widetilde{\widehat{\theta}}) + (1 - F(\widetilde{\widehat{\theta}}))EU^A(\underline{\theta})$. This in turn gives $F(\widetilde{\widehat{\theta}}) = \frac{1-EU^A(\underline{\theta})-t}{1-EU^A(\underline{\theta})}$. The total amount $t \cdot \frac{1-EU^A(\underline{\theta})-t}{1-EU^A(\underline{\theta})}$ is at maximum when $t = \frac{1-EU^A(\underline{\theta})}{2}$. Substituting into $F(\widetilde{\widehat{\theta}})$ gives $1/2$ participation rate result. ■

Hence, under two out of three exclusion rules, the admission fee game has one of equilibria where the SPS is provided with probability one and

all agents, even strong, who prefer to live in a world where only the APS is available for everyone, contribute to its cost.

The reason for this is the SPS treats those who have not contributed to its fixed costs either as "weaks" (under the partial exclusion rule) or as "nothing" (under the full exclusion rule). By definition of the SPS, those who sign up for the SPS's protection are "stronger" than agents with a very high θ but who remain under the APS. So the previously strong agents have no chances to obtain associated rents if all the others are under the SPS and this induces them to subscribe for the SPS as well. Both, the assumption of feasibility to exclude those who do not subscribe and the assumption that the SPS can always impose punishment on agents under the APS are crucial for the results.

By contrast, when the rule is quasi-no exclusion, an equilibrium with the full rate of subscription is impossible. This is because if too many agents are under the SPS, the remaining agents match them with a high probability and obtain the full protection without paying anything to the fixed cost of the SPS. This dilutes their own incentives to contribute and they prefer to free-ride.

Subscription when the APS is valueless

Recall from the section 2 that when $\rho \notin [\nu - 1, \nu + 1]$ no one wishes to trade under the APS. Hence the valuation for the SPS is uniformly equal to 1, even for the strong agents as they cannot benefit from their strength anyhow.

Consequently, within the admission fee game agents can be asked a simple message, say from $\{0, 1\}$. If agent says 0 he is excluded. If he says 1 he is admitted and when he matches those who are excluded, their trades are governed according to the exclusion rule in vigour.

The main results, depending on the exclusion rule, are:

PROPOSITION 1.6 *The highest collectible amount of resources for the SPS and corresponding rate of subscription depending on the exclusion rule are:*

- $\int_{\Theta^t=\{\theta:t(\theta)>0\}} t^*(\theta)dF(\theta) = 1$ *and the subscription rate is full when the exclusion rule is full;*
- $\int_{\Theta^t=\{\theta:t(\theta)>0\}} t^*(\theta)dF(\theta) = \frac{EU^A(\overline{\theta})^2}{4(EU^A(\overline{\theta})-EU^A(\underline{\theta})-1)}$ *and the subsciption rate is*

$\min\left\{\frac{EU^A(\bar{\theta})}{2(EU^A(\bar{\theta})-EU^A(\underline{\theta})-1)}, 1\right\}$ *if the exclusion rule is partial;*

- $\int_{\Theta^t=\{\theta:t(\theta)>0\}} t^*(\theta)dF(\theta) = \frac{1}{4}$ *and the subscription rate is* $\frac{1}{2}$ *when the exclusion rule is quasi-no exclusion.*

Proof. Suppose there is a subset of agents $(-i)$, call it S^*_{-i} of a mass $F(S^*_{-i}) \in [0,1]$, who subscribe for the services of the symmetric punishment system.

Under the full exclusion rule, agent $i's$ best reply would be to subscribe as well if $1 - t^* > 0$. Moreover as long as $t^* < 1$ subscription is a (weak) dominant strategy. Hence for any $K \leq 1$ the symmetric punishment is provided with probability one.

Under the partial exclusion rule, for a given t agent i is indifferent between subscribing and not when the fraction $F(S^*_{-i})$,defined below, subscribes[22]:

$$F(S^*_{-i}) \cdot 1 + (1 - F(S^*_{-i})) \cdot EU^A(\bar{\theta}) - t \geq F(S^*_{-i}) \cdot EU^A(\underline{\theta}) \tag{1.15}$$

This gives $F(S^*_{-i}) \leq \frac{EU^A(\bar{\theta})-t}{EU^A(\underline{\theta})+EU^A(\bar{\theta})-1}$ and the aggregate "incentive compatible" resources $t \cdot (\frac{EU^A(\bar{\theta})-t}{EU^A(\underline{\theta})+EU^A(\bar{\theta})-1})$. Maximizing this with respect to t and substituting back brings the equilibrium $F(S^*_{-i})$ at which is the collected resources are at maximum.

Under the quasi-no exclusion rule, an agent indifferent between two regimes is $1 - t \geq F(S^*_{-i}) \cdot 1$. Proceeding as in the case with the partial exclusion one obtains the results. ■

Compared to the previous subsection, under the partial exclusion rule there may be no longer the full rate of participation as agents have incentives to free-ride on those who have subscribed for the symmetric punishment. By contrast under the quasi-no exclusion rule agents have less incentives to free-ride compared to the previous section, i.e. a collectable revenue here is higher. Finally, the full exclusion guarantees the first best efficiency result

[22] One should not be surprised by seeing utilites of the strong and the weak. The partial exclusion by SPS has exactly the same strategic impact on agent's behaviour in mixed matches (i.e. of those who are admitted and those who are not) as APS has for strongs and weaks.

because joining the SPS is a dominant strategy for any agent under this rule.

The most important conclusion of this section is that by excluding from protection (to different degrees) the SPS could induce even the strongest agent to contribute to its cost. This is because in order to obtain a gain from economic activity both agents should play honest and for any constellation of parameters for this both agents should have behind a threat of punishment.

1.3.4 Boundaries of Efficiency of the Subscription Game and other Possible Equilibria

In this subsection I describe the sets of possible equilibria when agents subscribe in a decentralized manner for the symmetric punishment. And because different equilibria bring about different individual and aggregate welfare, for each type of exclusion rule I define the lowest social welfare.

The results of this section are given in order to discuss possible reasons why in reality two punishment systems may coexist and to stress the underlying coordination issue. As it is known from the literature on unique implementation (e.g. Repullo (1992), Palfrey (1992)) undesirable equilibria usually may be rather easily dispensed with through some extended indirect mechanisms (though those may be by quite unnatural). So the primary goal of this section is only to demonstrate that a market for enforcement services does not have a built-in mechanism to regulate itself for efficient pricing and quality of services. A SPS which has a too high start-up cost could yet arise and make everyone pay to its fixed cost. Similarly, there are also situations in which a SPS which has a small fixed cost could yet fail to attract enough of demand and even if there is a huge social welfare gain from a switch to regime where only the symmetric system is available for everyone this is not guaranteed to happen in equilibrium.

Then, the first observation is that sometimes it is preferable not to have the SPS when its cost is too high relative to the social gain when it is available.

PROPOSITION 1.7 *There exists a threshold level of the cost $\overline{K}(\rho, \nu)$ such that if the actual cost of the SPS is above $\overline{K}(\rho, \nu)$, the aggregate welfare is higher when this system is not in place.*

The formal proof of this in the appendix. The next observation is straightforward from the previous proposition.

PROPOSITION 1.8 *In any constellation of* (ρ, ν)*, as long as* $K < \overline{K}$ *the social welfare is lower when two punishment modes, the asymmetric and symmetric one, coexist in equilibrium as compared to the social welfare when only the symmetric punishment regime is available.*

The proof is rather trivial and omitted (a sum of payoffs from APS and SPS (i.e. when some agents are under the SPS and some under the APS) is inferior to the aggregate of payoffs when everyone is under SPS).

Given these two propositions I shall check if there are equilibria where the cost of the SPS is covered even if it is too high and whether there exist equilibria where two punishment system coexist despite of apparent inefficiency of such outcome.

The set of equilibria under the partial exclusion rule:

It happens that under the partial exclusion rule, for each stipulated admission threshold $\widehat{\theta}^*$ there could arise a continuum of subscription equilibria. The main reason for this is that the subscription strategies are strategic complements. So naturally of this there are multiple equilibria and actual outcome depends on beliefs agents share about the subscription behaviour in the population.

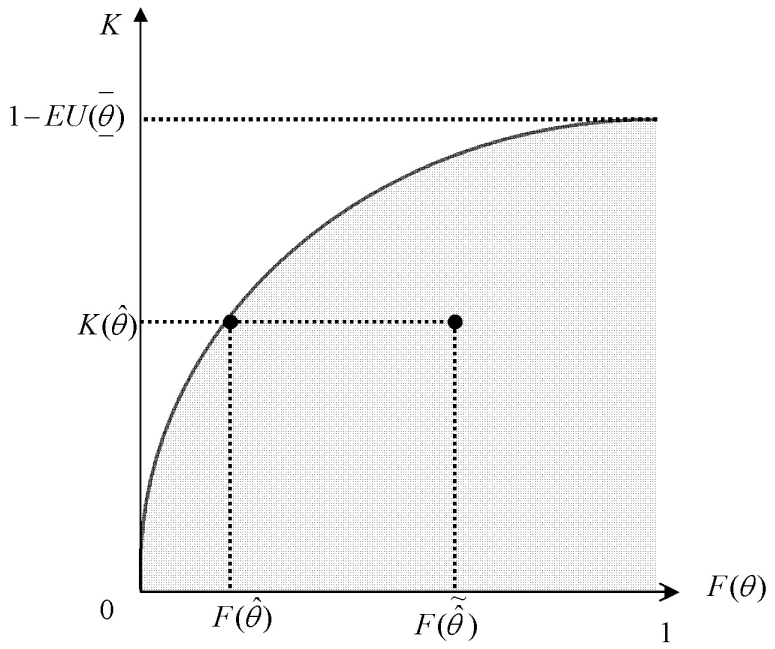
PROPOSITION 1.9 *In the admission fee game based on the partial exclusion rule, for a given* $K \in (0, 1 - EU(\underline{\theta}))$ *each threshold* $\widehat{\theta}$ *posted by the SPS induces three possible types of equilibria:*

(1) a separating equilibrium where for a given threshold $\widehat{\theta}$ *agents report their strength types truthfully* θ *and hence all agents* $\theta < \widehat{\theta}$ *are admitted to the symmetric punishment and all agents* $\theta > \widehat{\theta}$ *remain under asymmetric;*

(2) a continuum of semi-pooling equilibria where for a given $\widehat{\theta}$*, the actual rate of subscription is defined by some agent* $\widetilde{\widehat{\theta}} \in \left[\widehat{\theta}, \overline{\theta}\right]$ *such that all agents with* $\theta \in \left[\widehat{\theta}, \widetilde{\widehat{\theta}}\right]$ *claim to be* $\theta < \widehat{\theta}$ *so are admitted to the SPS; the remaining enforce their contracts under the APS;*

(3) two fully pooling equilibria where all agents report either their types to be $\theta > \widehat{\theta}$*, i.e. the subscription rate is zero, or* $\theta < \widehat{\theta}$*, i.e. everyone subscribes for the SPS.*

The results of the proposition are illustrated in the following Figure:



The grey area on the picture is for the set of semi-pooling subscription equilibria for a given level of the cost K, $y^* = 1$; the bold concave line gives the separating equilibrium for a given cost K, $y^* = 1$, in the white area $y^* = 0$ (as the number of those who subscribe and pay the corresponding t is not enough to cover the cost K)

This picture essentially shows the following. Take any cost $K < 1 - EU^A(\underline{\theta})$ and stipulate an incentive compatible admission threshold $\widehat{\theta}^*$, then the number of agents who in equilibrium could claim high valuation ($\theta < \widehat{\theta}^*$) ranges from $F(\widetilde{\widehat{\theta}}) = F(\widehat{\theta}^*)$ till $F(\widetilde{\widehat{\theta}}) = 1$, i.e. anything in between could be an equilibrium. If all agents believe that everyone believes that a fraction $F(\widehat{\widetilde{\theta}}) > F(\widehat{\theta}^*)$ would claim high valuation and subscribe for symmetric governance, for each single agent with $\widehat{\widetilde{\theta}} > \theta > \widehat{\theta}^*$ it is profitable to subscribe for the symmetric punishment as well (provided $F(\widehat{\widetilde{\theta}}) \cdot t(\widehat{\widetilde{\theta}}) \leq K$, i.e. the symmetric punishment system is commonly expected to be in place). The

opposite holds as well, if an agent from $\left[\widehat{\theta}^*, \widehat{\widehat{\theta}}\right]$ believes that everyone else in $\left[\widehat{\theta}^*, \widehat{\widehat{\theta}}\right]$ remains to trade under the APS, for him it is optimal not to join the SPS either.

This strategic pattern leads to the following evaluation of the inefficiency:

PROPOSITION 1.10 *There exists a level of cost $\widehat{K}^{PE}$ with $\widehat{K}^{PE} > \overline{K}$, such that under the partial exclusion rule, for $K \in \left[\overline{K}, \widehat{K}^{PE}\right]$ the symmetric punishment is installed, even though it is welfare decreasing.*

The proof is rather straightforward. Consider an equilibrium where the entire population subscribes (for a given cost K inefficiency even higher when a fraction of agents remain under the APS while the costly SPS is on place). In equilibrium where everyone subscribes agents are insensitive to the actual level of contribution t^* provided $t^* \leq 1 - EU^A(\underline{\theta})$ as everyone's behaviour depends only on beliefs. Then, even if the actual cost is $K = 1 - EU^A(\underline{\theta}) \equiv \widehat{K}^{PE}$ agent i subscribes and pays $t^* = 1 - EU^A(\underline{\theta})$ if he believes all others are under protection of the SPS. The highest potential efficiency loss due to over-provision is then equal to $\widehat{K}^{PE} - \overline{K} = \frac{\nu\rho(\rho-\nu)}{\rho^2-1} > 0$ as $\rho > \nu$. This number is equal to the aggregated surplus from trade that the SPS could extract from agents under the partial exclusion rule.

To summarize, the decentralized play of the contribution game may result in two types of inefficiencies: 1. The SPS and APS may coexist in equilibrium, i.e. different groups of agents use different systems and 2. The SPS may be able to collect money for its cost K even when it is too costly to have it or, in contrast, it may not able to do it even when it is efficient to have this system.

Equilibria under full exclusion rule:

The full exclusion rule as well has a continuum of subscription equilibria.

PROPOSITION 1.11 *In the admission fee game based on the full exclusion rule, for a given $K \in (0,1)$, each threshold $\widehat{\theta}$ induces three possible types of equilibria:*

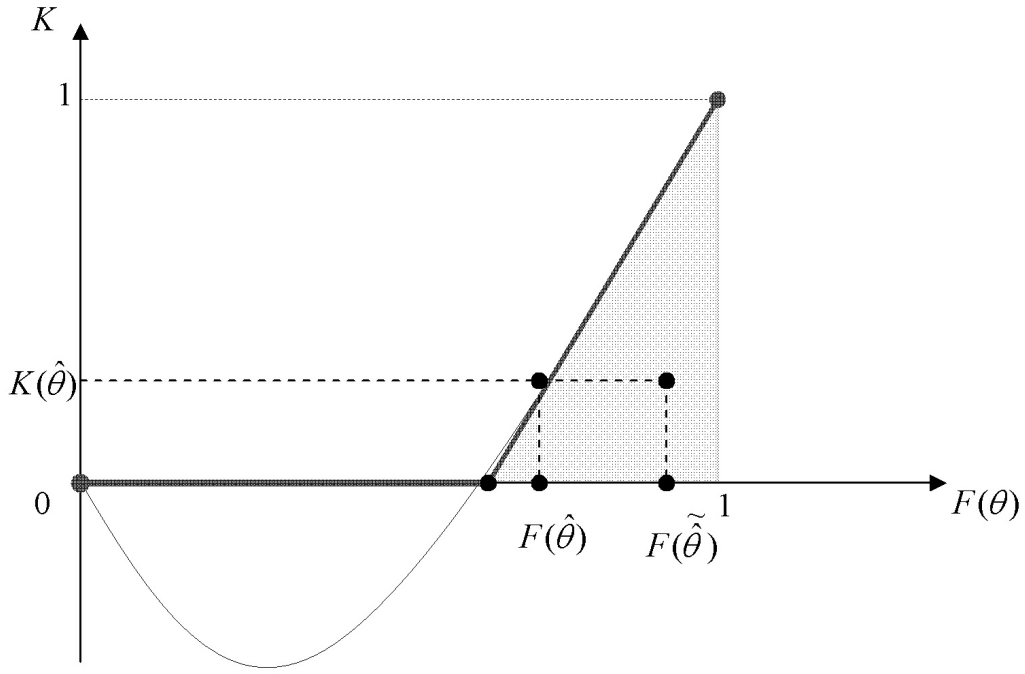
(1) a separating equilibrium where for a given threshold $\widehat{\theta}$ agents report truthfully their type θ, all agents $\theta < \widehat{\theta}$ are admitted to the symmetric punishment and all agents $\theta > \widehat{\theta}$ remain under asymmetric;

(2) a continuum of semi-pooling equilibria where for a given $\widehat{\theta}$*, the actual rate of subscription is defined by some* $\widetilde{\widehat{\theta}} \in \left[\widehat{\theta}, \overline{\theta}\right]$ *such that all agents with* $\theta \in \left[\widehat{\theta}, \widetilde{\widehat{\theta}}\right]$ *claim to be* $\theta < \widehat{\theta}$ *so are admitted to the SPS; the remaining enforce their contracts under the APS;*

(3) two fully pooling equilibria where all agents report either their types to be $\theta > \widehat{\theta}$*, i.e. the subscription rate is zero, or* $\theta < \widehat{\theta}$*, i.e. everyone subscribes for the SPS.*

The proof is in the Appendix. Note that the previous section has described exactly one of two equilibria from (3) type.

The sets of equilibria under full exclusion rule for each K are depicted on the following figure:



The bold line defines the separating equilibrium for each given K. The grey area corresponds to the set of semi-pooling equilibria (defined by $F(\widetilde{\widehat{\theta}})$)

for each level of the cost K, $F_{\min}(\widehat{\widehat{\theta}})$ is the minimal critical mass of agents to whom it is profitable to subscribe (given agents believe that at least fraction $F_{\min}(\widehat{\widehat{\theta}})$ subscribe).

Contrary to the previous case, with the partial exclusion rule, each i's beliefs about the subsciption behaviour should be sufficiently optimistic for him to join the SPS. That is each agent should believe that there is at least some minimal number ("a critical mass ") of agents who are under protection by the SPS. The reason for this threshold is as follows. Under the full exclusion rule, if there are too few people who are under the SPS, it is with a high probability that an agent who joins the SPS meets the one who is under the APS. In this case, according to the full exclusion rule on neither agent is imposed any punishment. Consequently the only payoffs that both agents can expect are zero. Anticipating this, if an agent believes that only few of others subscribe, he prefers to remain under the APS which gives him a higher expected utility. The opposite holds as well. If many agents join the SPS, for the remaining agents the utility from joining it higher than not doing it because with a very high probability agents outside of the SPS will meet the insiders and will not trade with them because of absence of any punishment. The highest collectible amount of resources is the higher under the full exclusion rule ($= 1$) as compared to the partial exclusion rule ($= 1 - EU^A(\underline{\theta})$).

One can see that the coexistence of two modes in equilibrium is possible under the full exclusion rule too. And similarly to the case with the partial exclusion rule there exist equilibria when agents pay too much to the SPS.

PROPOSITION 1.12 *There exists a level of cost $\widehat{K}^{FE}$ with $\widehat{K}^{FE} > \overline{K}$, such that under the partial exclusion rule, for $K \in \left[\overline{K}, \widehat{K}^{FE}\right]$ the symmetric punishment is installed, even though it is welfare decreasing.*

The logic of the proof is similar to the one in the proposition 1.10. Informally, the highest individually rational payment in the equilibrium where all agents subscribe for the symmetric punishment system services is $t^* = 1$. Hence if the actual cost is $K = 1 \equiv \widehat{K}^{FE}$ the symmetric punishment is provided, even if it is inefficient (from the proposition 1.7 it follows that $\widehat{K}^{FE} > \overline{K}$). The departure from efficiency is up to $\widehat{K}^{FE} - \overline{K} = 1 - \frac{(\rho-\nu)\nu}{(\rho^2-1)} > 0$.

Finally as $\widehat{K}^{FE} > \widehat{K}^{PE}$ one concludes that the full exclusion provides

with a larger highest amount of resources that could be extracted inefficiently from the agents by the SPS, i.e. the risk of over-provision of the SPS is higher under the full exclusion rule.

Results for the quasi-no exclusion rule:

By contrast to the previous two exclusion rules, agents subscription strategies are strategic substitutes, each agent would like to subscribe for the SPS if his encounter does not do it and not to subscribe if he expects his match to do it. As a result, as the following proposition puts it, there are no equilibria where all agents would contribute to the costs of the SPS.

PROPOSITION 1.13 *In the admission fee game based on the quasi-no exclusion rule, for each K such that $K \leq \frac{1}{4}(1 - EU^A(\underline{\theta}))$ there exist two associated thresholds $\widehat{\theta}$ resulting in truthful revelation. However both thresholds imply coexistence of the SPS and the APS in equilibrium. When $K > \frac{1}{4}(1 - EU^A(\underline{\theta}))$, the only equilibrium is all agents report $\theta > \widehat{\theta}$ for any $\widehat{\theta}$ and so $y^* = 0$.*

The proof is in the Appendix.

The intuition why there are only two possible equilibria is that under quasi-no exclusion rule each agent's willingness to join the SPS is decreasing in the number of agents whom he believe to be under the SPS. This is because the higher is share of agents who are under the SPS the higher is the probability for i to match with them and obtain for free the utility equal to 1. As encounter's subscription to protection extends the full protection to agent i, agents are willing to free-ride on others subscriptions. At the same time, when others do not subscribe each agent prefers to join the SPS and obtain rather a sure utility of 1. These gives rise to only two possible subscription equilibria: with low rate of participation (and high individually rational t) and with high rate of the participation (and reduced t), and there are no equilibria with extreme participation rates as it was with other exclusion rules.

As the result, the quasi-no exclusion rule has opposite welfare implications. Namely there is in general undercollection of resources for the cost of the SPS.

PROPOSITION 1.14 *Under the quasi-no-exclusion rule there exist a level of*

the cost $K \in \left[\frac{1}{4}(1 - EU^A(\underline{\theta})), \overline{K}\right]$*, such that it is valuable to introduce the SPS but the subscription fee game fails to do it.*

The proof is omitted, it is straightforward from the proof of the previous proposition.

Conclusions to the admission fee game with the quasi-no exclusion rule are 1. There is likely to be under-provision the SPS, i.e. agents fail to collect resources for the SPS even for intermediate levels of the cost K; 2. There *always* exists a positive fraction of agents who remain under the APS even if the SPS is in place.

1.4 Some Extensions and Discussions

1.4.1 Results when there is a monopolistic provider of asymmetric punishment

Everywhere I assumed that the marginal costs of using the APS were zero due to competition among providers. In this section I assume that there is a monopolistic provider who can make take-it-or-leave-it offers for the price of his asymmetric services while he does not change the quality. The results change a bit.

Assume that the price that a monopolistic provider charges is ϕ. Then the payoff matrix from trading can be modified as follows:

$strong \backslash weak$	H	C
H	$1, 1$	$1 - \nu + \rho - \phi, 1 + \nu - \rho$
C	$1 + \nu, 1 - \nu$	$0, 0$

(1.16)

i.e. agents pay a monetary fee ("the individual marginal cost of asymmetric punishment") only when the punishment ρ is imposed after equilibrium plays and it is only when a weak agent cheats on the strong one. This would affect the equilibrium strategies in the following way (under assumption that $\rho > \nu > 1$):

$$\zeta_i(\{H\}, \underline{\theta}) = \frac{1 - \nu + \rho - \phi}{1 + \rho - \phi}; \zeta_i(\{C\}, \underline{\theta}) = \frac{\nu}{1 + \rho - \phi} \tag{1.17}$$

$$\zeta_i(\{H\}, \overline{\theta}) = \frac{\nu - 1}{\rho - 1}; \zeta_i(\{C\}, \overline{\theta}) = \frac{\rho - \nu}{\rho - 1}. \tag{1.18}$$

As the result the expected payoff from contracting under asymmetric punishment would be

$$EU^A(\theta) = F(\theta)\frac{2\rho\nu(\rho-\nu)+\nu\phi(\nu+1-2\rho)}{(\rho-1)(1+\rho-\phi)} + \frac{(1+\nu-\rho)(\nu-1)}{\rho-1} \quad (1.19)$$

It can be shown that the result of full or zero participation rate in trade under asymmetrically imposed punishment carries over into the case with non-zero marginal costs. However there appears a condition on the maximal price that asymmetric enforcer can collect from agents (given that their outside option is not to trade at all). In particular, it should hold simultaneously that $\rho < \nu + 1$ and $\phi < \rho + 1$ and $\phi < \frac{2\rho(\rho-\nu)}{2\rho-\nu-1}$.

When $\rho \in [\nu, \nu+1]$ the third condition implies the second. Hence there is an upper boundary on the price that a monopolistic provider can charge $\overline{\phi} \equiv \frac{2\rho(\rho-\nu)}{2\rho-\nu-1}$.

In this case he extracts the entire surplus of the strong agent and the utilities of any strong and any weak agent would be equal, i.e.

$$EU^A(\overline{\theta}) = EU^A(\underline{\theta}) = \frac{(1+\nu-\rho)(\nu-1)}{\rho-1}. \quad (1.20)$$

Thus, at the ex ante stage, all agents expect to have the same level of utility, independently of the type θ, $EU^A(\theta) = \frac{(1+\nu-\rho)(\nu-1)}{\rho-1}$. As the result all agents would value equally the symmetric punishment system.

The highest efficient level of the cost K of the symmetric punishment system would be $\overline{K} = 1 - EU^A(\underline{\theta}) = \nu\frac{\rho-\nu}{\rho-1}$. It can be seen immediately that the partial exclusion rule is the most (second-best) optimal exclusion rule. As the admission fee mechanism would always result in the full subscription if the admission fee is $t^* \leq 1 - EU^A(\underline{\theta})$. Moreover to subscribe is a dominant strategy for any agent, i.e. he would subscribe for the symmetric punishment independently of his type θ and of strategies of the others. Summarizing all this into Proposition:

PROPOSITION 1.15 *When there is a monopolistic provider of the asymmetric punishment services, all agents have the same valuation for the SPS. Subscribing for the SPS when it applies the partial exclusion rule and posts the fee $t^* \leq 1 - EU^A(\underline{\theta})$ is a dominant strategy. Hence the subscription fee game will always have a socially efficient outcome.*

Thus, this result has a repercussion with the results of the section 3.3.2.: provided that the APS is relatively inefficient compared to the SPS (it imposes too high prices or too high or too low punishment), the coordination issue vanishes and it becomes a dominant strategy for everyone to subscribe for the SPS. In this case for the SPS "to win" the entire market for enforcement services, given its cost, it should apply partial exclusion (when the APS is monopolistic service) or full exclusion (when the APS is too inefficient and impose either too high or too low punishment).

One could argue that in order to avoid such outcome, the APS may find it beneficial to reduce its price and to leave some rents from trade to agents so that the SPS does attract so easily everyone under its protection. A full analysis of simultaneous competition in prices and qualities of two systems is beyond the scope of the current paper and left for future research.

1.4.2 Discussion of implications for the legal system

One could derive some implications on the optimal choice of the exclusion rules applied by an official legal system in order to win over the shadow sector of an economy (assuming that the latter enforces its contract through some APS). The optimal choice is likely to depend on how large is the black market.

First of all, the admission fee could be thought about as a tax that an entrepreneur has to pay for being a legal entity. By becoming legal, he would be entitled to enforce contracts via the (efficient) official court system. Then, if the shadow sector of an economy is large for an official legal system in order to attract the whole population under its protection, it is better to start with the quasi-no exclusion rule. In this case the "weakest" agents are likely to be the first to join, as it is a dominant strategy for them. Afterwards, once the weakest are "legalized", it is efficient to start to apply partial exclusion rule, in order to attract intermediate "strength" types. Given that previously weak agents are under protection by the official system which applies partial exclusion rule, the intermediate types are now "the weakest" ones in the population. Hence their valuation of the protection by the official system increases and they join the legal sector as well. Finally, when these types are also attracted and the legal sector becomes of sufficiently large size, it would be beneficial to use the full exclusion, as even the "strongest" types

would find it profitable to pay taxes and trade under the efficient official system as turf of weak agents on which they were getting rents previously is no longer around.

The exclusion rules are arguably defined by different rules within a legal system. One could speculate that:

- the full exclusion rule would be representative for a legal system that considers any contract made with non-legal entity (e.g. who do not pay taxes) to be void and not subject to any consideration.

- the partial exclusion rule would be representative for a legal system which imposes a very high fine on those who are caught to be on the black market and not being paying taxes.

  Hence in the case where it is a legal entrepreneur who cheats the one who is on the black market, the latter, being afraid of fines, simply will not apply to the official system because of fear to be punished himself by the system (and so he is excluded); at the same time, when it is the black market entrepreneur who misbehaves towards the legalized one, the legal system could pursue the black market entrepreneur and so he gets punished;

- the quasi-no exclusion rule would represent a legal system in which an agent who is a legal entity is both under full protection and under full liability for his behaviour within a contract, even with respect to those who are on the black market. For the full liability of agents from the legal sector to be effective one needs that the agents who are on the black markets do not fear to apply for justice. This arguably could be insured by various leniency programs for those who revealed themselves to be previously in the black market.

1.4.3 Further Research

Further research should be devoted to understand the structure of the fixed cost K. One way to impose some structure on K would be to assume that K is a function of ρ which itself is a function of π, where π is as it was discussed in the beginning is the probability that the punishment is imposed.

Another possible way would be explicit modelling of incentives of agents who provide punishment[23]. Additionally, if the SPS is the official legal system, the design of up-to-date codes is likely to be costly too. Gennaioli and Shleifer (2007) argues for costly codes. The common law system produces for free, by precedent, efficient rules and codes, (even when judges are biased, all inefficiencies are averaged away). By contrast, the centralized civil law system, at least in theory, has to put efforts to design the codes explicitly in advance. So in addition to incentives one may model explicitly the process of learning of which laws are efficient.

Obviously, studying explicitly the details of the cost K may allow to consider other and potentially more optimal mechanisms to collect resources for the SPS.

1.5 Conclusions

In the paper I have studied whether an institution which is costly but inducing within a contract mutually efficient behaviour could collect resources to cover its up-front costs.

The main results of the paper are:

- The asymmetric punishment system alone could bring to every agent, even the weakest one a positive level of utility. Moreover the rate of honest behaviour is higher as compared to a world with no punishment at all. The higher is the rate of the effective punishment (or the higher is the efficiency of the APS) the higher is the rate of honest behaviour in the population. However payoffs of the weakest agents worsen proportionally as the level of punishment increases and because of agents' freedom not to trade at all, the punishment applied asymmetrically cannot be increased indefinitely. As the result of existence

[23] The fixed cost assumption can be endogenised through an ex post bargaining game between strong agents and enforcers. This game has an equilibrium, such that each agent who is strong within a match *ex post* is able to convince the enforcing agent to rule in his favour the contract dispute, whoever was wrong, at no cost to him, if there are competing enforcers.As the result, at ex ante stage, in order to counteract the aggregate advantages of all agents who are strong at the ex post stage, the population of agents have to pledge to the enforcement agents the whole surplus available to the strong agents ex post. This means that the cost of the symmetric punishment system is actually equal to $K = \frac{1}{2}(EU^A(\overline{\theta}) - 1)$. Whether $K \lesseqgtr \overline{K}$ depends on the parameters ρ and ν.

of such upper boundary there is always a non-zero level of cheating in the equilibrium under the APS.

- If the level of punishment is larger than the cheating stake, relatively strong agents prefer the world where only asymmetric punishment is available to everyone. Under such parameters constellation, however, the symmetric punishment system brings higher aggregated payoffs than the asymmetric one.

- The contribution game based on exclusion from punishment of non-payers, in general, is successful in covering costs of the SPS. Even the strong agents may pay to the cost of the SPS when too many of weak agents get protected by the SPS.

- This also has a drawback for decentralized provision of the symmetric punishment – agents entire surplus from trade could be extracted by threatening them with exclusion from contract enforcement services. As a result there may be an overprovision of the symmetric punishment system, i.e. it may be installed even when it has a socially inefficient, high cost.

- The more severe is the exclusion rule the larger is amount of resources that could be collected from agents. The quasi-no exclusion rule may be unable to collect the necessary amount even if it is efficient to have the SPS.

- Because under partial and full exclusion rules contribution strategies are strategic complements there exist equilibria where relatively strong agents remain under the APS and relatively weak join the SPS. In other words, two modes could inefficiently coexist.

- When there is a monopolistic provider of the asymmetric punishment who charges a fee and extracts the entire ex post surplus of the strong agents, the unique dominant strategy equilibrium is where all agents subscribe for the SPS. But, the exclusion rule yet should be either partial or full and not the mild rule of quasi-no exclusion.

1.6 Appendix: Omitted Proofs

Proof to the Proposition 1.1
Proof.

1. Case $\rho > \nu$: The expected utility from entry into the market is increasing monotonically in type. Hence I will check incentives of the lowest entering type (it is unique). For any (ρ, ν) define this type θ^e, and he is found from the identity $EU^A(\theta^e) \equiv 0$. Namely

$$F(\theta^e) = \frac{(\rho+1)(\nu-1)(\rho-\nu-1)}{2\nu\rho(\rho-\nu)}.$$

One can see that the participation is full, as r.h.s is below zero, if $\rho < \nu + 1 \equiv \rho^{\max}$. Now I show that for $\rho > \rho^{\max}$ there is *no* any agent who would enter.

Suppose $\rho > \rho^{\max}$. Then there is a fraction of agents, namely $F(\theta^e)$ who take the outside option. The remaining fraction $(1 - F(\theta^e))$ reassess the utility from participating in the market given (ρ, ν). The expected utility of an agent $\theta \in \left[\theta^e, \overline{\theta}\right]$ is:

$$EU^{A\prime}(\theta) = (F(\theta) - F(\theta^e))EU(\overline{\theta}) + (1 - F(\theta))EU^A(\underline{\theta}).$$

I search again for an agent such that $EU^{A\prime}(\theta) \equiv 0$, call him $\theta^{e\prime}$:

$$F(\theta^{e\prime}) = F(\theta^e)\frac{(1+\nu)(1-\nu+\rho)(\rho-1)}{2\rho\nu(\rho-\nu)} - \frac{(1+\nu-\rho)(\nu-1)(\rho+1)}{2\rho\nu(\rho-\nu)}$$

Now it can be shown that $\rho > \rho^{\max}$ is sufficient condition for $F(\theta^{e\prime}) > F(\theta^e)$ to happen. I.e. if ρ is too high a new fraction of agents would prefer to take the outside option. This argument is repeated for any new θ^e, hence no agent enters.

2. Case $\rho < \nu$: From (1.9) it follows that entry type is defined by (again the monotonicity of EU implies that there is the unique cut-off):

$$F(\theta^e) = \frac{1+\nu-\rho}{2(\nu-\rho)} \tag{1.21}$$

It holds that $F(\theta^e) < 1$ if $\rho < \nu - 1 \equiv \rho^{\min}$ i.e. participation is partial (for $F(\theta^e) \leq 1$ it is enough to find under which condition $\frac{1+\nu-\rho}{2(\nu-\rho)} \leq 1$).

If ρ is too low it is not profitable to enter for a fraction of agents of *high* type ($\theta \in \left[\theta^e, \overline{\overline{\theta}}\right]$). It can be shown similarly to the case with $\rho > \nu$ that full unraveling occurs, but starting with the strongest type.

■

Proof to the Corollary 1.2

Proof. In the subgame of trading under the asymmetric punishment system there are following equilibria depending on (ρ, ν).

1. if $\rho < \nu - 1$, the unique pure NE is $(\zeta_i(\{H\}, \overline{\theta}), \zeta_j(\{H\}, \underline{\theta})) = (0, 0)$, the rate of participation in the market is zero;
2. if $\rho \in (\nu - 1, \nu)$ there is full participation with the unique equilibrium $(\zeta_i(\{H\}, \overline{\theta}), \zeta_j(\{H\}, \underline{\theta})) = (1, 0)$;
3. if $\rho \in (\nu, \nu + 1)$ there is full participation, no pure strategy NE, the mixed strategies are defined in (1.4),(1.5);
4. if $\rho > \nu + 1$ the rate of participation is zero, there is a NE in mixed strategies as in (1.4),(1.5)

■

Proof to the Proposition (1.7):

Proof. Consider the following constellations in (ρ, ν) plane

1. $\nu < \rho < \nu + 1$

   The aggregate interim social welfare under asymmetric punishment is (for any $F(\theta)$) $SW^A = \frac{\rho(\rho-\nu)+\nu^2-1}{(\rho^2-1)} > 0$. The aggregate social welfare under symmetric punishment is $SW^A = 1$. When $\nu < \rho < \nu + 1$ it holds that $\frac{\rho(\rho-\nu)+\nu^2-1}{(\rho^2-1)} < 1$. Hence the highest cost $\overline{K}$ is defined from $SW^S - K \geq SW^A$. It is equal $\overline{K} = \frac{(\rho-\nu)\nu}{(\rho^2-1)}$, $0 < \frac{(\rho-\nu)\nu}{(\rho^2-1)} < 1$.

2. $\nu - 1 < \rho < \nu$

   Here, under asymmetric punishment, the interim social welfare is equal to 1 too. It is never efficient to install the symmetric punishment system from interim perspective[24]. (But ex post efficient threshold of $\overline{K}$ in this case is $\overline{K} = \frac{1}{2}$).

[24] But recall that the lower ρ is, the higher is the equilibirum rate of cheating behaviour.

3. $\rho \notin [\nu - 1, \nu + 1]$

   The asymmetric punishment is not active. Hence the social welfare under absent symmetric punishment is $SW^0 = 0$. The threshold level is then $\overline{K} = SW^S - SW^0 = 1$ (both interim and ex post).

■

Proof to the Proposition 1.9:

Proof. The multiplicity of equilibria is due to the fact that agents beliefs about the actual rate of participation determine their own subscription strategies. Depending on what is the common belief about the subscription rate, for a given $\widehat{\theta}$, there may be different actual subscription rates, including the rate of subscription meant by the mechanism, when $\widehat{\theta}$ was chosen.

Firstly I shall prove the possibility of the truthful equilibrium. Suppose all agents report their types truthfully. Then the optimal $\widehat{\theta}$ for a given level of the cost K is defined by IC condition

$$1 \cdot F(\widehat{\theta}) + (1 - F(\widehat{\theta})) \cdot EU(\overline{\theta}) - t \geq EU(\underline{\theta}). \tag{1.22}$$

merged to the budget balance requirement $t \cdot F(\widehat{\theta}) = K$:

$$F(\widehat{\theta})^2(1 - EU(\overline{\theta})) + F(\widehat{\theta}))(EU(\overline{\theta}) - EU(\underline{\theta})) = K. \tag{1.23}$$

It is easy to check that indeed agents with $\theta > \widehat{\theta}$ state truthfully their types and remain under the APS.

The second equilibrium comes from the observation that agents beliefs are free and so if every agent expects (by whichever reason) that there is a subgroup of agents, say an interval $\left[\widehat{\theta}, \widetilde{\widehat{\theta}}\right]$ who report their types to be $\theta < \widehat{\theta}$, any isolated agent with $\theta \in \left[\widehat{\theta}, \widetilde{\widehat{\theta}}\right]$ is better off from reporting his type to be $\theta < \widehat{\theta}$ rather than saying the truth, as:

$1 \cdot F(\widetilde{\widehat{\theta}})^2 + F(\widetilde{\widehat{\theta}}) \cdot (1 - F(\widetilde{\widehat{\theta}})) \cdot EU(\overline{\theta}) - K > F(\widetilde{\widehat{\theta}}) EU(\underline{\theta})$ when $F(\widetilde{\widehat{\theta}}) > F(\widehat{\theta})$ and keeping the cost K constant[25]. Consequently even totally uninformative equilibrium with $F(\widetilde{\widehat{\theta}}) = 1$ can occur for any given $\widehat{\theta}$.

[25] This is because the LHS $F(\widehat{\theta})^2(1 - EU(\overline{\theta})) + F(\widehat{\theta})(EU(\overline{\theta}) - EU(\underline{\theta})) \equiv K$ is increasing in $F(\widehat{\theta})$. Hence when K remains fixed but $F(\widehat{\theta})$ increases, the LHC of the participation constraint remains valid for a given $F(\widehat{\theta})$.

The third equilibrium arises when i believes all other agents report a low valuation for the symmetric punishment, it is not installed and hence i does not strictly gain from reporting any positive valuation (= self-fulfilling). ■

Proof to the Proposition 1.11

Proof. The proof is very similar to the case with partial exclusion rule. The truthful equilibrium is defined by equation (merging IR and BB requirement)

$$(1 - EU(\overline{\theta})) \cdot F(\widehat{\theta})^2 + EU(\underline{\theta}) \cdot F(\widehat{\theta}) - K = 0. \tag{1.24}$$

The only difference is that for y to be positive, for a relatively high K, there should exist a common belief $F(\widehat{\theta})$ that satisfies this equation.

The proof of untruthful equilibria, i.e. that it may happen $F(\widehat{\widetilde{\theta}}) \neq F(\widehat{\theta})$ quite the same in the case with partial exclusion. If all agent follow some untruthful equilibrium strategy, every remaining agent has incentives to play it as well, due to complementarity of participation decision. ■

Proof to the Proposition 1.13:

Proof. An agent indifferent between joining the SPS and remaining under the APS is defined by the IC condition $1 - t = F(\widehat{\theta}) + (1 - F(\widehat{\theta}))EU^A(\underline{\theta})$.

Merging this with the budget balance requirement defines

$$F(\widehat{\theta})(1 - EU^A(\underline{\theta})) - F(\widehat{\theta})^2(1 - EU^A(\underline{\theta})) = K \tag{1.25}$$

The LHS is a concave function with a maximum at $F(\widehat{\theta}) = 1/2$. From here follows that the maximum collectible amount is $\frac{1}{4}(1 - EU^A(\underline{\theta}))$ and that there are two possible participation equilibria for $K \leq \frac{1}{4}(1 - EU^A(\underline{\theta}))$. Namely, (1.25) has two solutions for $F(\widehat{\theta})$; both belong to $(0, 1)$ and so define two incentive compatible participation thresholds. By proposition 1.8 I chose the highest (SW is increasing in the number of agents who are under the SPS).

When $K > \frac{1}{4}(1 - EU^A(\underline{\theta}))$ there is no such IC payment that once aggregate would match this level of the cost. ■

Chapter 2

On Uniqueness of Payoffs to Beliefs

Abstract[1] Neeman (2004) and Heifetz and Neeman (2006) have shown that, in models with correlated values, full surplus extraction is only possible if agents' payoffs can be inferred from their beliefs about other agents. For models with exogenous payoff and belief functions defined on an abstract type space, Heifetz and Neeman (2006) show that this so-called BDP property ("beliefs determine preferences") is non-robust in a measure-theoretic sense. By contrast, we show that BDP is generic in a topological sense if beliefs are formed by conditioning on available information and the set of objects about which agents form beliefs is sufficiently rich.

2.1 Introduction

The seminal work of Crémer and McLean (1988) (in what follows CM) has shown that, in common-prior models with incomplete information, correlations of agents types can be used to reduce information rents. In particular, in an auction with two or more potential buyers, all surplus can be extracted

[1] This chapter is the joint paper of Martin Hellwig and me. The original title of the paper is: "Payoffs Can be Inferred From Beliefs, Generically, When Beliefs are Conditioned on Information". We would like to thank for helpful discussions Felix Bierbrauer, Jacques Crémer, Bruno Jullien, Benny Moldovanu and Stephen Morris.

if the buyers have correlated private values. This work has inspired a number of other papers, in particular, McAfee and Reny (1992) extending the CM result to auctions with a continuum of types. Most recently, Kosenok and Severinov (2008) provide a set of necessary and sufficient conditions for an arbitrary allocation of surplus among types to be incentive-compatible.

The analysis of CM has, however, been challenged by Neeman (2004) and Heifetz and Neeman (2006). They show that CM's conclusions concerning full surplus extraction depend on a fairly restrictive assumption, which is implicit in CM's modelling of incomplete information. They refer to this assumption as the BDP condition: "beliefs determine preferences". Under this condition, an agent's payoff can be precisely inferred from his beliefs, or, equivalently, in any two states of the world in which an agent has different payoffs, he must also have different beliefs about the rest of the world.

The analysis of Neeman (2004) and Heifetz and Neeman (2006) greatly improves our understanding of the CM result: Differences in beliefs induce differences in attitudes towards bets or, more generally, state-contingent payment schemes. These differences in attitudes can be used to extract rents, i.e., to make an agent surrender the surplus he gets even though this surplus depends on the state of the world and the mechanism designer has no way to observe the state or the surplus directly. However, this is only possible if differences in payoffs across states of the world are aligned with differences in beliefs. If there are two states of the world where an agent has different payoffs and the same beliefs about the rest of the world, there is no way to prevent the agent from earning an information rent. Thus, Heifetz and Neeman (2006) show that full surplus extraction in an auction is impossible if BDP fails to hold. For a public-good provision problem with participation constraints, Neeman (2004) shows that, as in the independent-private-values analysis of Mailath and Postlewaite (1990), in an environment with many participants, feasible provision levels are close to zero if BDP is violated in such a way that for each agent and each state of the world in which this agent gets a positive payoff from the public good, there is another state of the world in which the agent has the same beliefs, but a zero payoff from the enjoyment of the public good.

At this point, the question is what to make of the BDP property in models with common priors. Neeman (2004) and Heifetz and Neeman (2006) indicate that this property is very restrictive. In Heifetz and Neeman (2006),

a geometric and a measure-theoretic version of this statement are stated and proved; a topological version is discussed, but, for lack of a clear view as to the appropriate topology on the space of incomplete-information models, such a result is not established. In a companion paper, we show that non-BDP models with common priors are indeed generic if we map any incomplete-information model into the associated universal-type-space formulation à la Mertens and Zamir (1985) and we endow the set of universal-type-space formulations with the natural topology as suggested by Mertens and Zamir (1985).[2][3]

Statements about genericity, robustness, or non-robustness of a certain property always depend on the space to which one is referring. Most of the literature, including Crémer and McLean (1988), Neeman (2004), and Heifetz and Neeman (2006), avoid working with a universal type space, which would be clumsy and cumbersome. In Crémer and McLean (1988), the type space is specified so that people observe their payoffs, and their beliefs correspond to regular conditional distributions given the payoffs that they have observed. In this setting, genericity of BDP follows from the assumption that the set of payoff types is finite and, for a generic set of priors, the map from payoff types to conditional distributions is one-to-one. By contrast, in Neeman (2004) and Heifetz and Neeman (2006), payoffs and beliefs are both determined by exogenously given functions of some abstract "type" variables.

Our approach lies between Crémer and McLean (1988) and Neeman (2004) or Heifetz and Neeman (2006). Like Crémer and McLean (1988), we treat beliefs as being endogenous, rather than part of the specification of a "type". In contrast to Crémer and McLean (1988), however, we allow for the possibility that beliefs may reflect the influence of variables other than payoffs. Specifically, we assume that agents form beliefs by observing

[2] I.e., the weak topology on the space of priors that is induced by the product topology on the universal type space when the parameters of the underlying game are topogized so that payoff functions are continuous and the beliefs at each level of the hiearchy are given the weak topology.

[3] However, whereas this genericity result implies that any BDP model can be approximated by a sequence of non-BDP models, we have a continuity result showing, as one moves along this sequences, the maximum level of surplus extraction in the non-BDP models converges to the maximum level of surplus extraction in the BDP model, i.e., to full surplus extraction.

certain information variables and conditioning on these observations. The information variables include the agents' own payoff parameters. However, the influence of payoff parameters on beliefs may be confounded by other information variables. The question then is to what extent the payoff parameters, or any of the other variables on which people condition, can be recovered from their beliefs.

To get an idea for what we are after, consider an environment with two agents in which the preferences of agent i depend on a parameter θ_i, which is taken to be private information. Suppose also that each agent receives a signal s_i about the other agent's preference parameter. If the signal s_2 that agent 2 receives is informative about the preference parameter θ_1 of agent 1, these two variables must be correlated. This implies that the agent 1's expectation about the signal s_2 of agent 2 must depend on his payoff parameter θ_1. If this dependence is one-to-one, e.g., monotonic, the BDP property arises as a matter of course. By looking at agent 1's beliefs about the signal that agent 2 has received about agent 1's payoff, one can infer agent 1's own payoff parameter. By contrast, θ_1 cannot be inferred from agent 1's belief about s_2 if, in addition to his own payoff parameter, agent 1 also observes a signal s_1 about s_2; in this case, the effects of θ_1 and s_1 on agent $1's$ belief about s_2 cannot be disentangled. However, the inference from agent $1's$ beliefs to his payoff type should be possible if the set of variables about which agent 1 forms beliefs is sufficiently large and, as a source of information, none of the variables on which he conditions is redundant.

The formal analysis below makes this intuition precise. Following Aumann (1987), we assume that all heterogeneity in beliefs across agents is due to a heterogeneity of information. Agents have a common prior on the state of the world. Given this prior, their beliefs are determined by updating after the observation of additional signals; their own preference parameters are among the signals on which they condition. To the extent that signals differ from agent to agent, the beliefs to which they give rise are naturally heterogeneous.

In such a setting, agents form beliefs about the other agents' signals as well as their payoffs. If the set of parameters of other agents about which each agent forms his beliefs is sufficiently rich, these beliefs will tend to reflect all the information on which they are based, including the agent's

own payoff parameters. In this case, the BDP property arises as a generic property of common priors.

We prove two versions of this claim. The first version concerns models with finitely many types for each agent. We show that, in such models, if the number of possible types is the same for all agents, the BDP property holds for an open and dense set of priors. With finitely many types for each agent, there are finitely many states of the world, a prior is just a finite-dimensional vector, and the terms open and dense need no further explanation.

The second version of our claim concerns models with a continuum of types. Here, we treat the type t_i of any agent i as an n_i-dimensional vector, an element of some compact set $T_i \subset \mathbb{R}^{n_i}$. We restrict our attention to common priors with densities that belong to the space $C^1(\mathbb{R}^N, \mathbb{R})$ of continuously differentiable functions from $\mathbb{R}^N$ into $\mathbb{R}$ where $N = \sum_i n_i$. If we endow the set of such priors with the topology that is induced by the strong topology on the space $C^1(\mathbb{R}^N, \mathbb{R})$ of density functions, we find that, if n_i is the same for all agents and if there are at least four agents, then, for an open and dense set of priors, the BDP property holds for all agents. More generally, the BDP property is generic if, for each agent i, we have $2n_i + 1 \leq N_{-i} := \sum_{j \neq i} n_j$.

The result is a consequence of Whitney's Embedding Theorem[4]. The vector $t_{-i} = (t_j)_{j \neq i}$ of the other agents' types, about which agent i forms his beliefs, belongs to the space $\mathbb{R}^{N_{-i}}$. For any value t_i of the agent's own type, therefore, the conditional expectation $\bar{t}_{-i}(t_i)$ of the other agents' types that is induced by t_i is an N_{-i}-dimensional vector. For priors with densities belonging to $C^1(\mathbb{R}^N, \mathbb{R})$, the map $t_i \to \bar{t}_{-i}(t_i)$ is a continuously differentiable function from $\mathbb{R}^{n_i}$ into $\mathbb{R}^{N_{-i}}$. Whitney's Embedding Theorem implies that, if $2n_i + 1 \leq N_{-i}$, then any such function can be approximated by a sequence of embeddings. We show that, if we start from a given C^1 density on $\mathbb{R}^N$, then the embeddings approximating the conditional-expectations function $\mathbb{R}^N$ are actually the conditional-expectations functions for t_{-i} given t_i under suitable perturbations of the prior density. Given that, by definition, embeddings are injective, under these perturbed priors, it is always possible to recover the type t_i of agent i from the agent's conditional expectation $\bar{t}_{-i}$ about the

[4] In economics, Whitney's Embedding Theorem has previously played an important role in the literature on generic existence of completely revealing rational expectations equilibria, in particular, Allen (1981).

other agents' types. In particular, it is possible to recover the agent's payoff parameters, which are just one part of his type t_i.

These genericity results are at odds with the results of Neeman (2004), and Heifetz and Neeman (2006). The difference is due to our treating beliefs as resulting from conditioning on information and taking account of the fact that payoff parameters are part of the information about the information of the others on which agents condition.

Our results are also at odds with the claims of Barelli (2009). Relying on an approach similar to that of Heifetz and Neeman (2006), but using a topological rather than measure-theoretic notion of genericity, he asserts that BDP fails to hold for an open and dense set of models. There are some difficulties with his analysis, however. He begins by introducing an abstract type space Θ, together with pairs of mappings, one for each agent, that specify payoffs and beliefs for each type. Then he defines a model as a pair Y, μ such that (i) $Y \subset \Theta$ is minimal with respect to the requirement that for any agent and any type of agent, beliefs assign probability one to the set Y and (ii) μ is a prior on Y that is consistent with the specified beliefs as posteriors at the given types. In this construction, Y is endogenous to the specification of payoff and belief functions. In the subsequent analysis of genericity, however, Y is treated as exogenous, e.g. in considering the approximation of a model with a continuum of types by a sequence of models with finitely many types. Moreover, in these approximations, no account is given of the requisite adaptations of belief functions: Presumably, in a model with finitely many types, beliefs will be probability measures on finite sets, rather than the continuum. Given these lacunae, we find it difficult to interpret his results. In any case, like Heifetz and Neeman (2006), Barelli (2009) does not model the notion that beliefs result from conditioning on information and that an agent's payoff parameters are part of the information on which he conditions his beliefs.

In the following, Section 2 lays out the basic framework of our analysis. Section 3 introduces the BDP property and gives a few examples in order to build some intuition. Section 4 formulates and proves our genericity results.

2.2 The Basic Framework

Whereas our analysis concerns abstract issues of incomplete-information modelling, we find it convenient to base our presentation on a concrete model of social choice. In this model, there are $I \geq 2$ agents, indexed by $i = 1, .., I$. There is a single private good and a social decision variable g. Each agent i has a quasi-linear utility function

$$u_i(g, \theta_i, m_i) = v(g, \theta_i) + m_i, \tag{2.1}$$

where m_i is the amount of private-good consumption, and θ_i is a payoff parameter. The allocation problem is to choose g and $m_1, ...m_I$. subject to the feasibility constraint

$$K(g) + \sum_{i=1}^{I} m_i \leq Y; \tag{2.2}$$

here $K(g) \geq 0$ is a resource cost and Y is a an exogenously given measure of aggregate resource availability.

The choice that is taken will typically depend on $\theta_1, ..., \theta_I$. However, we assume that, for each i, the payoff parameter θ_i is private information of agent i. To implement a *social choice function*

$$(\theta_1, ..., \theta_I) \rightarrow (g(\theta_1, ..., \theta_I), m_1(\theta_1, ..., \theta_I), m_I(\theta_1, ..., \theta_I)), \tag{2.3}$$

one must have a way of extracting the relevant information about $\theta_1, ..., \theta_I$ from the different participants.

To model information, we assume that there is some underlying space Ω of possible states of the world, and that all beliefs are derived from a common prior F on the space Ω. Agent i's payoff parameter θ_i is given by a function

$$\omega \rightarrow \theta_i(\omega); \tag{2.4}$$

in addition to this payoff parameter, the agent also observes a signal s_i, which is given by a function

$$\omega \rightarrow s_i(\omega); \tag{2.5}$$

The functions $\theta_i(\cdot), s_i(\cdot)$ take values in given sets Θ_i, S_i, respectively.

In this setting, we may think of the agent's *type* as a pair

$$t_i = (\theta_i, s_i), \tag{2.6}$$

an element of the type space

$$T_i = \Theta_i \times S_i. \tag{2.7}$$

In our analysis, the underlying probability space Ω matters only to the extent that it affects the types $t_1, ..., t_I$ of agents $1, ..., I$. there is therefore no loss of generality in assuming that

$$\Omega = \prod_{i=1}^{I} T_i \tag{2.8}$$

and that the maps $\theta_i(\cdot)$ and $s_i(\cdot)$ are simply the projections from Ω to Θ_i and S_i, respectively.

Given the information that $\theta_i(\cdot)$ and $s_i(\cdot)$ take the values θ_i and s_i, agent i updates his expectations, replacing the prior F by the conditional distribution on Ω that is induced by this information. Denote this conditional distribution as $B_i(\theta_i, s_i)$ and consider the induced marginal distributions on T_i and on $T_{-i} := \prod_{\substack{j=1 \\ j \neq i}}^{I} T_j$. Because the agent knows his own type, the marginal distribution on T_i is simply the degenerate distribution that assigns all mass to the observed $t_i = (\theta_i, s_i)$. The marginal distribution on T_{-i} represents the agent's conditional beliefs about the other agents's types; we denote this distribution as $b_i(\theta_i, s_i)$.

2.3 The BDP Property

2.3.1 Definition

The distribution $b_i(\theta_i, s_i)$, i.e. the conditional distribution on T_{-i} that is induced by the prior F and the information that $\theta_i(\cdot)$ and $s_i(\cdot)$ take the values θ_i and s_i, corresponds to what is usually referred to as the agent's *belief type*. As discussed in Neeman (2004) and Heifetz and Neeman (2006), a key question is to what extent the payoff type can be inferred from the observation of the agent's belief type $b_i(\theta_i, s_i)$. Given that we have identified Ω

with the product $\prod_{i=1}^{I} T_i$ and the functions $\theta_i(\cdot)$ and $s_i(\cdot)$ with the projections to Θ_i and S_i, this is a question about the prior F. The following definition defines the BDP property as a property of the prior F.

DEFINITION 2.1 *A prior F on the product $\prod_{i=1}^{I} T_i$ exhibits the **BDP property** if, for $i = 1, ..., I$ and F-almost all $(\theta_i, s_i) \in T_i$,*

$$\{(\theta_i', s_i') \in T_i | b_i(\theta_i', s_i') = b_i(\theta_i, s_i)\} \subset \{\theta_i\} \times S_i. \tag{2.9}$$

According to this definition, any two types $t_i = (\theta_i, s_i)$ and $t_i' = (\theta_i', s_i')$ that induce the same beliefs $b_i(\theta_i', s_i') = b_i(\theta_i, s_i)$ must also involve the same payoff types $\theta_i' = \theta_i$. As discussed in the introduction, this condition plays a key role in the analysis of surplus extraction in models with correlated values.

2.3.2 Examples

We illustrate the BDP property by means of a few examples. We begin with the example given in the introduction.

EXAMPLE 2.2 *Let $I = 2$, $\Theta_1 = S_2 = \mathbb{R}$, $\Theta_2 = S_1 = \{0\}$, and suppose that F is a multivariate normal distribution. Then*

$$E[s_2|\theta_1, s_1] = \frac{cov(s_2, \theta_1)}{var\theta_1}(\theta_1 - E\theta_1) + Es_2. \tag{2.10}$$

If $cov(s_2, \theta_1) \neq 0$, i.e., if the signal s_2 contains any information about θ_1, one can infer θ_1 from the belief variable $E[s_2|\theta_1, s_1]$ and the parameters $E\theta_1, Es_2, cov(s_2, \theta_1), var\theta_1$ of the prior F. Thus, for any prior on T that is multivariate normal, the BDP property holds unless the signal s_2 is uncorrelated with the payoff type θ_1. Within the set of models covered by Example 2.2, BDP is generic.

In Example 2.2, agent 2 receives a signal about agent 1's payoff type. Knowing this, agent 1 treats his own payoff type as a signal about agent 2's signal. Therefore, his belief about agent 2's signal varies with his payoff type. The relation is monotonic, and his payoff type can be inferred from his belief type.

EXAMPLE 2.3 *Let $I = 2$, $\Theta_1 = S_1 = S_2 = \mathbb{R}$, $\Theta_2 = \{0\}$, and suppose that F is a multivariate normal distribution. Then*

$$E[s_2|\theta_1, s_1] = \alpha_\theta(\theta_1 - E\theta_1) + \alpha_s(s_1 - Es_1) + Es_2, \tag{2.11}$$

where

$$\begin{pmatrix} \alpha_\theta & \alpha_s \end{pmatrix} = \begin{pmatrix} cov(s_2, \theta_1) & cov(s_2, s_1) \end{pmatrix} \begin{pmatrix} var\theta_1 & cov(\theta_1, s_1) \\ cov(\theta_1, s_1) & vars_1 \end{pmatrix}^{-1}. \tag{2.12}$$

In this case, as in Example 2.2, agent 1's belief about agent 2's signal is affected by agent 1's payoff type unless $cov(s_2, \theta_1) = 0$. However, if agent 1's own signal is also correlated with s_2, it is not possible to infer θ_1 from the belief variable $E[s_2|\theta_1, s_1]$ and the parameters of the prior F. For such an inference, one would also have to know the realization of agent 1's own signal. In this setting, for any prior on T that is multivariate normal, even if $cov(s_2, \theta_1) \neq 0$, the BDP property fails to hold except in the special case where $cov(s_2, s_1) = 0$.

EXAMPLE 2.4 *Let $I = 2$, $\Theta_1 = S_1 = \mathbb{R}, \Theta_2 = \{0\}$, $S_2 = \mathbb{R}^2$, and suppose that F is a multivariate normal distribution. Then*

$$\begin{pmatrix} E[s_2^1|\theta_1, s_1] \\ E[s_2^2|\theta_1, s_1] \end{pmatrix} = A \begin{pmatrix} \theta_1 - E\theta_1 \\ s_1 - Es_1 \end{pmatrix} + \begin{pmatrix} Es_2^1 \\ Es_2^2 \end{pmatrix}, \tag{2.13}$$

where $A = \Sigma_{21}\Sigma_{11}^{-1}$ and Σ_{21}, Σ_{11} are submatrices of the variance-covariance matrix Σ of $s_2^1, s_2^2, \theta_1, s_1$ partitioned as

$$\Sigma = \begin{pmatrix} \Sigma_{22} & \Sigma_{21} \\ \Sigma_{12} & \Sigma_{11} \end{pmatrix} \tag{2.14}$$

so as to reflect the distinction between the variables that agent 1 observes and what he does not observe. In this specification, both θ_1 and s_1 can be inferred from the conditional expectations $E[s_2^1|\theta_1, s_1]$ and $E[s_2^2|\theta_1, s_1]$ whenever the matrix A is invertible, i.e., whenever the matrix Σ_{21} in the partition (2.14) is nonsingular. Given that the set of nonsingular two-by-two matrices is open and dense in the set of all two-by-two matrices, the BDP property is generic within the set of models covered by Example 2.4.

In Examples 2.3 and 2.4 the robustness of nonrobustness of the BDP property depends on whether or not the dimension of the set of objects about which agent 1 forms his beliefs is large enough to permit a precise inference about the different variables on which he conditions. If it is large enough to permit a disentangling of the different information variables on which he relies, then, in particular, one can infer his payoff type from his beliefs.

Whereas Example 2.3 involves a failure of BDP due to a confounding of influences of different information variables, the following Example 2.5 shows that BDP will also fail if, for given parameters of the prior F, the map from payoff types to conditional expectations is not monotonic (not one-to-one). Subsequently, Example 2.6 will show that this problem is likely to disappear if there are more variables about which to form expectations so that the vector of conditional expectations has a sufficiently high dimension.

EXAMPLE 2.5 *Let $I = 2$, $\Theta_1 = S_2 = \mathbb{R}$, $\Theta_2 = S_1 = \{0\}$, and suppose that F is the distribution that is generated when*

$$s_2 = (\theta_1)^2 + \varepsilon, \tag{2.15}$$

where θ_1 and ε are independent normal random variables. In this case, the conditional distribution of s_2 given θ_1 is normal with mean

$$E[s_2|\theta_1, s_1] = (\theta_1^2 - E\theta_1^2) + Es_2 \tag{2.16}$$

and variance $Var\varepsilon$. From the belief $E[s_2|\theta_1, s_1]$, one can infer θ_1^2, but one cannot tell whether is the positive of the negative solution of the equation

$$\theta_1^2 = E[s_2|\theta_1, s_1] - Es_2 + E\theta_1^2. \tag{2.17}$$

The BDP property fails to hold.

EXAMPLE 2.6 *Let $I = 2$, $\Theta_1 = \mathbb{R}$, $\Theta_2 = S_1 = \{0\}$, $S_2 = \mathbb{R}^2$, and suppose that F is the distribution that is generated when*

$$s_2^1 = (\theta_1)^2 + \varepsilon, \tag{2.18}$$

$$s_2^2 = A\theta_1 + B(\theta_1)^2 + \eta \tag{2.19}$$

where θ_1,ε, and η are independent normal random variables. In this case, the conditional distribution of s_2^1 and s_2^2 given θ_1 is normal with means

$$E[s_2^1|\theta_1, s_1] = (\theta_1^2 - E\theta_1^2) + Es_2^1, \tag{2.20}$$

$$E[s_2^2|\theta_1, s_1] = A(\theta_1 - E\theta_1) + B(\theta_1^2 - E\theta_1^2) + Es_2^2 \tag{2.21}$$

and variance-covariance matrix $\begin{pmatrix} Var\varepsilon & 0 \\ 0 & Var\eta \end{pmatrix}$. *If $A \neq 0$, then from (2.20) and (2.21), one obtains*

$$\theta_1 = E\theta_1 + \frac{1}{A}\left[E[s_2^2|\theta_1, s_1] - Es_2^2 - B(E[s_2^1|\theta_1, s_1] - Es_2^1\right], \tag{2.22}$$

which shows that θ_1 can be inferred by looking at $E[s_2^1|\theta_1, s_1]$ and $E[s_2^2|\theta_1, s_1]$ jointly. By looking at the two belief variables together, one overcomes the difficulty that neither belief variable alone is injective in θ_1. The BDP property holds unless $A = 0$.

2.4 Genericity Results

2.4.1 BDP with Finite Type Sets

Turning from these examples to a more general analysis, we first consider the case where type sets are finite. If each type set T_i is finite, with n_i elements, the state space T is also finite, with $N = \prod_{i=1}^{I} n_i$ elements, and the prior F is represented by a vector $\Pi \in \mathbb{R}^N$, such that $\sum_{k=1}^{N} \Pi_k = 1$. The set of such vectors is endowed with the usual (Euclidean) topology.

PROPOSITION 2.7 *Assume that, for each i, T_i is a finite set with n_i distinct elements. For any i, let*

$$N_{-i} := \prod_{\substack{j=1 \\ j\neq i}}^{I} n_j \tag{2.23}$$

be the cardinality of the set $T_{-i} := \prod_{\substack{j=1 \\ j\neq i}}^{I} T_j$. If $n_i \leq N_{-i}$ for all i, then BDP holds for an open and dense set of priors on T.

Proof. For any i, a vector Π of probabilities on T can be written in matrix form as $\Pi(i) = (\pi_{t_i t_{-i}})_{t_i t_{-i}}$ where the different rows refer to different types t_i of agent i and different columns refer to the different elements t_{-i} of T_{-i}. For each t_i, the conditional belief $b_i(t_i)$ given t_i of agent i about the other agents' types is represented by a vector of conditional probabilities on T_{-i}. Bayes' Law implies that, under the prior Π, this vector is proportional to the vector $(\pi_{t_i t_{-i}})$; one can write:

$$b_i(t_i) = \lambda(t_i) \times (\pi_{t_i t_{-i}})_{t_{-i}}, \tag{2.24}$$

where

$$\lambda(t_i) = \frac{1}{\sum_{t_{-i} \in T_{-i}} \pi_{t_i t_{-i}}} \tag{2.25}$$

is chosen to ensure that the entries in (2.24) sum to one. (2.24) and (2.25) imply that, if the rows $(\pi_{t_i t_{-i}})_{t_{-i}}$, , $t_i \in T_i$, of the matrix $\Pi(i)$ are linearly independent, then so are the belief vectors $b_i(t_i), t_i \in T_i$. This implies, in particular, that the belief vectors $b_i(t_i), t_i \in T_i$, are all distinct and the function $t_i \to b_i(t_i)$ is invertible, i.e. one can infer the type t_i of agent i from his belief vector. Given that $t_i = (\theta_i, s_i)$, this means, in particular, that one can infer θ_i from $b_i(t_i)$.

By Lemma 2.9, the assumption that $n_i \leq N_{-i}$ implies that the set of $n_i \times N_{-i}$ matrices with linearly independent rows is an open and dense subset $\mathcal{P}_i$ of $\mathbb{R}^{n_i} \times \mathbb{R}^{N_{-i}} = \mathbb{R}^N$. Because I is finite, the intersection $\mathcal{P} = \cap_{i=1}^I \mathcal{P}_i$ of these sets for different i is still open and dense. If the prior Π belongs to $\mathcal{P}$, then, for any i, belief vectors $b_i(t_i), t_i \in T_i$, are all distinct and one can infer the type t_i of agent i from his belief vector; in this case, BDP is satisfied. ■

Proposition 2.7 echoes similar statements in Heifetz and Neeman (2006) and Barelli (2009). The assumption that $n_i \leq N_{-i}$ for all i ensures that the spaces of things about which agents from their beliefs are always richer than the spaces of variables on which they condition. Moreover, finiteness ensures that, generically, the map from types to beliefs is one-to-one.

2.4.2 BDP with a Continuum of Types

We next allow for a continuum of types of each agent. We assume that, for each i, there is a positive integer n_i such that the set T_i of types of agent

i is a subset of $\mathbb{R}^{n_i}$. The products $\prod_j T_j$ and $\prod_{j\neq i} T_j$ are subsets of $\mathbb{R}^N$ and $\mathbb{R}^{N_{-i}}$, respectively, where $N = \sum_{j=1}^I n_j$ and $N_{-i} := N - n_i$.

We restrict attention to priors ν with compact supports and with continuously differentiable densities, i.e., with densities f_ν that are C^1 functions on $\mathbb{R}^N$. The space of such priors is endowed with the topology that is induced by convergence of supports in the Hausdorff metric and by strong (uniform) C^1 convergence of the densities. Thus, a sequence $\{v^k\}$ of priors with supports $T^{\nu^k} \subset \mathbb{R}^N$ and densities $f_{\nu^k} : \mathbb{R}^N \to \mathbb{R}_+$ converges to a limit ν with support T^ν and density f_ν if and only if the Hausdorff distance between T^{ν^k} and T^ν converges to zero and f_{ν^k} converges to f_ν, C^1-uniformly.

The beliefs $b_i(t_i)$ of any agent i are now elements of the space of measures on $\mathbb{R}^{N_{-i}}$. By the same logic as before, the BDP property holds if, for any i, the function $t_i \to b_i(t_i)$ is injective.

PROPOSITION 2.8 *If $2n_i + 1 \leq N_{-i}$ for all i, then BDP holds on an open and dense subset of the set of priors with compact support and continuously differentiable densities, endowed with the topology of Hausdorff convergence of supports and strong C^1 convergence of densities.*

Proof. We will show that, for any i, there is an open and dense set $\mathcal{P}_i$ of priors with compact supports and continuously differentiable densities such that BDP holds for agent i if the prior belongs to $\mathcal{P}_i$. As in the proof of Proposition 2.7, the desired result then follows from the observation that the finite intersection $\mathcal{P} = \cap_{i=1}^I \mathcal{P}_i$ of these open and dense sets for different i is itself open and dense.

Fix any i. For any ν with compact support T^ν and continuously differentiable density f_ν, let

$$f_\nu^i(\cdot) = \int_{T^\nu_{-i}} f_\nu(\cdot, t_{-i}) dt_{-i}$$

be the density of the marginal distribution for t_i that is induced by ν. Because f_ν is nonzero only on the compact set T^ν and because f_ν is continuously differentiable, the values of f_ν and its derivatives are bounded. By Lebesgue's bounded-convergence theorem, it follows that f_ν^i is continuously differentiable. Moreover, the values of f_ν^i are nonzero only on the projection

T_i^ν of T^ν to the space of agent i's types.[5].

Because f_ν^i is continuously differentiable, the interior $\mathring{T}_i^\nu$ of the set T_i^ν is nonempty. Moreover, for any $t_i \in \mathring{T}_i^\nu$, we have $f_\nu^i(t_i) > 0$. For such t_i, the posterior belief $b_i(t_i, \nu)$ is a measure on $\mathbb{R}^{N_{-i}}$ that has a density $\beta_i(\cdot|t_i, \nu)$ satisfying

$$\beta_i(t_{-i}|t_i, \nu) = \frac{f_\nu(t_i, t_{-i})}{f_\nu^i(t_i)} \tag{2.26}$$

for all $t_{-i} \in \mathbb{R}^{N_{-i}}$. The conditional expectation of t_{-i} given $t_i \in \mathring{T}_i^\nu$ can then be written as

$$\bar{t}_{-i}^\nu(t_i) := \int_{\mathbb{R}^{N_{-i}}} t_{-i}\beta_i(t_{-i}|t_i, \nu)dt_{-i} = \frac{\int_{\mathbb{R}^{N_{-i}}} t_{-i} f_\nu(t_i, t_{-i})dt_{-i}}{f_\nu^i(t_i)}. \tag{2.27}$$

Observe that $f_\nu(t_i, t_{-i}) = 0$ unless the pair (t_i, t_{-i}) belongs to the compact set T^ν. Therefore, the integrand in (2.27) is zero unless t_{-i} belongs to the projection T_{-i}^ν of T^ν to the space of types of agents other than i. By the continuity of the projection mapping, compactness of T^ν implies compactness of T_{-i}^ν. Therefore, the conditional expectation (2.27) is well defined.

Because the densities f_ν and f_ν^i are continuously differentiable, the map is $t_i \to \bar{t}_{-i}^\nu(t_i)$ belongs to the space $C^1(\mathring{T}_i^\nu, \mathbb{R}^{N_{-i}})$ of continuously differentiable functions from $\mathring{T}_i^\nu$ into $\mathbb{R}^{N_{-i}}$. If this map is an injection, then so must be the map $t_i \to f_\nu(t_i, \cdot)$ from $\mathring{T}_i^\nu$ into $C^1(\mathbb{R}^{N_{-i}}, \mathbb{R})$. Hence, it suffices to show that, for an open and dense set of priors ν, the map $t_i \to \bar{t}_{-i}^\nu(t_i)$ from $\mathring{T}_i^\nu$ into $\mathbb{R}^{N_{-i}}$ is an injection.

Given the assumption that $2n_i + 1 \leq N_{-i}$, Whitney's Embedding Theorem (Mas-Colell (1985), p. 37) implies that, for any ν and any $\varepsilon > 0$ there exists an embedding $t_i \to \hat{t}_{-i}^\varepsilon(t_i)$ such that

$$\left\|\hat{t}_{-i}^\varepsilon(t_i) - \bar{t}_{-i}^\nu(t_i)\right\| < \varepsilon \tag{2.28}$$

for all t_i. Given this embedding, we consider the function f_ν^ε such that, for $t_i \in \mathring{T}_i^\nu$ and $t_{-i} \in \mathbb{R}^{N_{-i}}$,

$$f_\nu^\varepsilon(t_i, t_{-i}) = f_\nu\left(t_i, t_{-i} - \hat{t}_{-i}^\varepsilon(t_i) + \bar{t}_{-i}^\nu(t_i)\right) \tag{2.29}$$

and, for $t_i \in \mathbb{R}^{n_i} \backslash \mathring{T}_i^\nu$ and $t_{-i} \in \mathbb{R}^{N_{-i}}$,

$$f_\nu^\varepsilon(t_i, t_{-i}) = 0. \tag{2.30}$$

[5] We allow for the possibility that T^ν is a proper subset of the product $\prod T_i^\nu$.

Using the change of variables

$$\hat{t}_{-i} = t_{-i} - \left(\hat{t}^{\varepsilon}_{-i}(t_i) - \bar{t}^{\nu}_{-i}(t_i)\right), \tag{2.31}$$

we compute

$$\int_{\mathbb{R}^{N_{-i}}} f^{\varepsilon}_{\nu}(t_i, t_{-i}) dt_{-i} = \int_{\mathbb{R}^{N_{-i}}} f_{\nu}(t_i, \hat{t}_{-i}) d\hat{t}_{-i} = f^{i}_{\nu}(t_i) > 0 \tag{2.32}$$

for any $t_i \in \mathring{T}^{\nu}_i$ and

$$\int_{\mathbb{R}^N} f^{\varepsilon}_{\nu}(t_i, t_{-i}) dt_{-i} dt_i = \int_{\mathring{T}^{\nu}_i} \int_{\mathbb{R}^{N_{-i}}} f^{\varepsilon}_{\nu}(t_i, t_{-i}) dt_{-i} dt_i = \int_{\mathring{T}^{\nu}_i} f^{i}_{\nu}(t_i) dt_i = 1, \tag{2.33}$$

which shows that the function f^{ε}_{ν} is actually the density $f_{\nu^{\varepsilon}}$ of a probability measure ν^{ε}.

We claim that f^{ε}_{ν} is also continuously differentiable. To simplify the notation, write $\hat{t}^{\varepsilon}_{-i}(t_i) = \bar{t}^{\nu}_{-i}(t_i) = 0$ for $t_i \in \mathbb{R}^{n_i} \backslash T^{\nu}_i$ and combine (2.29) and (2.30) into the single equation

$$f^{\varepsilon}_{\nu}(t_i, t_{-i}) = f_{\nu}\left(t_i, t_{-i} - \hat{t}^{\varepsilon}_{-i}(t_i) + \bar{t}^{\nu}_{-i}(t_i)\right). \tag{2.34}$$

From (2.34), one immediately sees that f^{ε}_{ν} is continuously differentiable at any point (t_i, t_{-i}) at which $\hat{t}^{\varepsilon}_{-i}(\cdot)$ and $\bar{t}^{\nu}_{-i}(\cdot)$ are continuously differentiable; in particular, f^{ε}_{ν} is continuously differentiable at any point (t_i, t_{-i}) with $t_i \notin (T^{\nu}_i \backslash \mathring{T}^{\nu}_i)$. Moreover, (2.30) yields

$$Df^{\varepsilon}_{\nu}(t_i, t_{-i}) = 0 \tag{2.35}$$

for any for $t_i \in \mathbb{R}^{n_i} \backslash T^{\nu}_i$ and $t_{-i} \in \mathbb{R}^{N_{-i}}$. Continuity of Df^{ε}_{ν} requires that (2.35) also holds $t_i \in (T^{\nu}_i \backslash \mathring{T}^{\nu}_i)$ and that $\left\| Df^{\varepsilon}_{\nu}(t'_i, t'_{-i}) \right\|$ is uniformly close to zero if t'_i is close to some $t_i \in (T^{\nu}_i \backslash \mathring{T}^{\nu}_i)$.

Using (2.32), for any $t_i \in \mathring{T}^{\nu}_i$, one also computes

$$\begin{aligned}
\bar{t}^{\nu^{\varepsilon}}(t_i) &= \frac{\int_{\mathbb{R}^{N_{-i}}} t_{-i} f^{\varepsilon}_{\nu}(t_i, t_{-i}) dt_{-i}}{\int_{\mathbb{R}^{N_{-i}}} f^{\varepsilon}_{\nu}(t_i, t_{-i}) dt_{-i}} \\
&= \frac{\int_{\mathbb{R}^{N_{-i}}} \left(\hat{t}_{-i} + \left(\hat{t}^{\varepsilon}_{-i}(t_i) - \bar{t}^{\nu}_{-i}(t_i)\right)\right) f_{\nu}(t_i, \hat{t}_{-i}) d\hat{t}_{-i}}{\int_{\mathbb{R}^{N_{-i}}} f_{\nu}(t_i, \hat{t}_{-i}) d\hat{t}_{-i}} \\
&= \frac{\int_{\mathbb{R}^{N_{-i}}} \hat{t}_{-i} f_{\nu}(t_i, \hat{t}_{-i}) d\hat{t}_{-i}}{\int_{\mathbb{R}^{N_{-i}}} f_{\nu}(t_i, \hat{t}_{-i}) d\hat{t}_{-i}} + \hat{t}^{\varepsilon}_{-i}(t_i) - \bar{t}^{\nu}_{-i}(t_i) \\
&= \hat{t}^{\varepsilon}_{-i}(t_i).
\end{aligned}$$

Because the map $t_i \to \hat{t}^{\varepsilon}_{-i}(t_i)$ is an embedding, it follows that the map $t_i \to f^{\varepsilon}_{\nu}(t_i, \cdot)$ from $\mathring{T}^{\nu}_i$ into $C^1(\mathbb{R}^{N-i}, \mathbb{R})$ is an injection. The measure ν^{ε} that has the density $f_{\nu^{\varepsilon}} = f^{\varepsilon}_{\nu}$ must therefore have the BDP property for agent i.

By (2.29) and (2.30), the support $T^{\nu^{\varepsilon}}$ of the measure ν^{ε} is the closure of the set of pairs (t_i, t_{-i}) such that $(t_i, t_{-i} - \left(\hat{t}^{\varepsilon}_{-i}(t_i) - \bar{t}^{\nu}_{-i}(t_i)\right) \in T^{\nu}$. By (2.28), it follows that the Hausdorff distance between the sets $T^{\nu^{\varepsilon}}$ and T^{ν} is no greater than ε. As ε goes to zero, the sets $T^{\nu^{\varepsilon}}$ converge to T^{ν}.

Because $\left\|\hat{t}^{\varepsilon}_{-i}(t_i) - \bar{t}^{\nu}_{-i}(t_i)\right\| < \varepsilon$ for all $t_i \in \mathring{T}^{\nu}_i$, one also has

$$|f^{\varepsilon}_{\nu}(t_i, t_{-i}) - f_{\nu}(t_i, t_{-i})| \leq \left\|D_{t_{-i}} f_{\nu}\right\| \varepsilon$$

for all $t_i \in \mathring{T}^{\nu}_i$ and $t_{-i} \in \mathbb{R}^{N-i}$. For $t_i \in \mathbb{R}^{n_i} \backslash \mathring{T}^{\nu}_i$ and $t_{-i} \in \mathbb{R}^{N-i}$

$$|f^{\varepsilon}_{\nu}(t_i, t_{-i}) \ - f_{\nu}(t_i, t_{-i})| = 0$$

by construction. Given that the derivatives of f_{ν} are continuous, a similar argument shows that, for any $\delta > 0$, there exists $\varepsilon(\delta)$ such that, if $\varepsilon \in (0, \varepsilon(\delta))$, one has

$$\|Df^{\varepsilon}_{\nu}(t_i, t_{-i}) - Df_{\nu}(t_i, t_{-i})\| < \delta$$

for all $t_i \in \mathring{T}^{\nu}_i$ and $t_{-i} \in \mathbb{R}^{N-i}$. For $t_i \in \mathbb{R}^{n_i} \backslash \mathring{T}^{\nu}_i$ and $t_{-i} \in \mathbb{R}^{N-i}$

$$\|Df^{\varepsilon}_{\nu}(t_i, t_{-i}) - Df_{\nu}(t_i, t_{-i})\| = 0.$$

Hence, as $\varepsilon \to 0$, the priors ν^{ε}, which satisfy BDP converge to the given prior ν.

This shows that the set of priors satisfying BDP for agent i is dense.

To show that the set of such priors is also open, we note that, if a sequence of priors $\{\nu^{\varepsilon}\}$ converges to a prior ν, then the conditional-expectations functions $t_i \to \bar{t}^{\varepsilon}_{-i}(t_i)$ that are induced by the measures ν^{ε} converge in $C^1(\mathring{T}_i, \mathbb{R}^{N-i})$ to the conditional-expectations function $t_i \to \bar{t}_{-i}(t_i)$ that is induced by ν. If the function $t_i \to \bar{t}_{-i}(t_i)$ is an embedding then so, for any sufficiently small $\varepsilon > 0$ must be the function $t_i \to \bar{t}^{\varepsilon}_{-i}(t_i)$. This follows because embeddings are open and dense in $C^1(\mathring{T}_i, \mathbb{R}^{N-i})$. ■

2.5 Concluding Remarks

We conclude the paper with two remarks. First, the assumption of a common prior is strictly speaking not necessary for our analysis. If different agents

have different priors, our analysis will still go through, provided be allow for the space of other agents' types to be rich enough so that, generically, each agent's beliefs make active use of all his information, including the information about his own payoff.

Second, we conjecture that the dimensionality assumption in Proposition 2.8 is not necessary for the conclusion. In the continuous-type model, each agent's belief about the other agents is a probability measure on a set with the cardinality of the continuum. The space of such probability measures itself is an infinite-dimensional space. In principle, therefore, it should be possible to use something like the Whitney Embedding Theorem to show that, generically, the maps $t_i \to b_i(t_i)$ are injective. The problem is that these mappings do not directly lend themselves to a differential-topology argument. Any detour, e.g., by looking at the integrals with respect to the measures $b_i(t_i)$ of the elements $g_1, g_2, ...$ of a separating sequence of functions, raises questions about the measures that would generate the approximating embeddings. As yet, we have not been able to resolve these questions. Therefore, we have imposed the dimensionality assumption that $2n_i + 1 \leq N_{-i}$ for all i, i.e., that the dimension of any one agent's type space be less than one half the sum of the dimensions of all the other agents' type spaces.

2.6 Appendix: Omitted Proofs

LEMMA 2.9 *For any m and n, the set $\mathcal{A}^{m,n}$ of $m \times n$ matrices that have rank $\min(m,n)$ is an open and dense subset of $\mathbb{R}^{mn}$ (in the usual topology).*

Proof. It suffices to prove the result for the case $m = n$. For any $m \times m$ matrix A, let $D_m(A)$ be the determinant. The map $A \to D_m(A)$ is a continuous function from $\mathbb{R}^{m^2}$ into $\mathbb{R}$. For any $A \in \mathcal{A}^m$, one has $D_m(A) \neq 0$ by continuity of $D_m(\cdot)$, there exists an open neighbourhood $\mathcal{B}(A)$ of A such that $D_m(A') \neq 0$ for all $A' \in \mathcal{B}(A)$ and, hence, $\mathcal{B}(A) \subset \mathcal{A}$. It follows that

$$\cup_{A \in \mathcal{A}} \mathcal{B}(A) \subset \mathcal{A}^m.$$

Since, trivially, one also has

$$\mathcal{A}^m \subset \cup_{A \in \mathcal{A}} \mathcal{B}(A),$$

it follows that

$$\mathcal{A}^m = \cup_{A \in \mathcal{A}} \mathcal{B}(A).$$

As a union of open sets, $\mathcal{A}^m$ itself is open.

To prove that $\mathcal{A}^m$ is dense in $\mathbb{R}^{m^2}$, an induction argument is applied. For $m = 1$, $\mathcal{A}^1 = \{A \in \mathbb{R} | A \neq 0\}$, so obviously any $A \in \mathbb{R}$ can be approximated by a sequence in $\mathcal{A}^1$.

If $m > 1$, suppose that the claim is true for $m-1$ and note that, for any $A \in \mathbb{R}^{m^2}$, one has

$$D_m(A) = \sum_{i=1}^{m} a_{1i}(-1)^{i+1} D_{m-1}(A_{1i}),$$

where, for $i = 1, ..., m$, A_{1i} is the $(m-1) \times (m-1)$ matrix that is obtained from A by eliminating the first row and the i-th column. By the induction hypothesis, there is a sequence of matrices $\{A_{11}^k\}$ converging to A_{11} such that $D_{m-1}(A_{11}^k) \neq 0$ for all k. For any k and any $\hat{a}_{11}$, let $\hat{A}^k(\hat{a}_{11})$ be the matrix that is obtained from A if a_{11} is replaced by $\hat{a}_{11}$ and the submatrix A_{11} is replaced by A_{11}^k. Then

$$D_m(\hat{A}^k(\hat{a}_{11})) = \hat{a}_{11} D_{m-1}(A_{11}^k) + \sum_{i=2}^{m} a_{1i}(-1)^{i+1} D_{m-1}(A_{1i}^k),$$

where, for $i = 2, ...m$, A_{1i}^k is the $(m-1)\times(m-1)$ matrix that is obtained from $\hat{A}^k(\hat{a}_{11})$ by eliminating the first row and the i-th column; these submatrices are obviously independent of $\hat{a}_{11}$.

Consider the sequence $\{\hat{A}^k(a_{11})\}$ that is obtained by setting $\hat{a}_{11} = a_{11}$ so that the first diagonal element of the matrix is not changed at all. By construction, this sequence converges to A. If it has a subsequence $\{k^r\}$ such that $\lim_{r\to\infty} k^r = \infty$ and $D_m(\hat{A}^{k^r}(a_{11})) \neq 0$ for all r, this subsequence provides the desired approximation of A by elements of $\mathcal{A}^m$. Alternatively, suppose that $D_m(\hat{A}^k(a_{11})) = 0$ for all but finitely many k. In this case, consider the sequence $\{\hat{A}^k(a_{11}+\frac{1}{k})\}$ that is obtained by setting $\hat{a}_{11} = a_{11}+\frac{1}{k}$ for $k = 1, 2, ...$ By construction, this sequence also converges to A. For any k, one has

$$D_m(\hat{A}^k(a_{11} + \frac{1}{k})) = \frac{1}{k} D_{m-1}(A_{11}^k) + D_m(\hat{A}^k(a_{11})).$$

Because $D_m(\hat{A}^k(a_{11})) = 0$ for all but finitely many k, it follows that

$$D_m(\hat{A}^k(a_{11} + \frac{1}{k})) = \frac{1}{k} D_{m-1}(A_{11}^k) \neq 0$$

and, hence, $\hat{A}^k(a_{11} + \frac{1}{k}) \in \mathcal{A}^m$ for all but finitely many k. ■

Chapter 3

Details Behind Belief Hierarchies Matter

Abstract[1] In the existing mechanism design literature different types, i.e. types with different sets of information, are usually treated as "redundant" if they have the same beliefs hierarchies about others payoffs. Here we model strategically relevant signals as an explicit part of agents' types and show that there exist social choice functions which are not implementable on a payoff-based universal type space but which are implementable on a set of "redundant" payoff-and-signals-based type spaces which map into that payoff-based universal type space.

3.1 Introduction

Several recent papers in the mechanism design have sought implementation results for the case where agents, in addition to having private information about their payoff characteristics, know privately their beliefs about others' potential payoff characteristics and others' beliefs (e.g. Neeman (2004), Heifetz and Neeman (2006), Bergemann and Morris (2005)).

[1]This chapter is the joint paper of Martin Hellwig and me. We would like to thank Felix Bierbrauer, Jacques Crémer, Bruno Jullien, Stephen Morris, Tymofiy Mylovanov and Michael Vogt for helpful discussions of ideas at different stages of this project.

Agents' private information about their beliefs has been shown, in general, to lead to further allocative inefficiencies (Fang and Morris (2006), Skrzypacz and Feinberg (2005), Neeman (2004)). Specifically, Neeman (2004) and Heifetz and Neeman (2006) have demonstrated that the result of Crémer and McLean (1988) fails to hold once agents' beliefs about others' types become imperfectly correlated to their own payoffs, i.e. once the so-called "belief-determine-preferences" (BDP) condition is violated. The BDP condition requires that any interim belief type that an agent could have is associated with at most one possible payoff type. It is the key condition for the lottery mechanism in Crémer and McLean (1988) to work.

In order to model agents' heterogeneity in beliefs, this literature recurs to the construction of the universal type space of Mertens and Zamir (1985) and Brandenburger and Dekel (1993) (who, in turn, have formalized the idea of Harsanyi (1967/68)). In this construction one specifies some underlying space of strategic uncertainty Ω relevant for agents payoffs and a description of agents' hierarchies of beliefs about it, i.e. agents beliefs about payoff relevant uncertainty, agents beliefs about others' beliefs about payoff relevant uncertainty, etc. A collection of all hierarchies of beliefs together with the set Ω is called "Ω-based universal type space". It is "universal" because by its construction it contains any coherent hierarchy of beliefs about Ω that could exist and each hierarchy has a unique representation in it. This, in turn, justifies overwhelming use of this construction in the game-theoretic and mechanism design literature – it is a "terminal" space of types for all models of information and so, it is argued, all results could be expressed in terms of types in the Ω-based universal type space.

In this type space, different models of information are regarded as "redundant" if they map into the same Ω-based hierarchies of beliefs. In other words, any sequence of information types which all have the same hierarchies of beliefs about Ω is not treated as a sequence of separate types even if they are different in the way that these hierarchies have been obtained (i.e. in how agents have updated some prior given their information – their signals).

Ely and Peski (2006) and Dekel, Fudenberg, and Morris (2007) have demonstrated, as a part of motivation of their analysis, sensitivity of the game theoretic predictions to the details of a type space used to model a given set of hierarchies. Different types having the same hierarchies of

beliefs about the payoff uncertainty may have different sets of rationalizable strategies. Given this, the question arises whether the mechanism design problems are also sensitive to the specification of a type space that gives rise to a given set of hierarchies of beliefs about the underlying uncertainty, e.g., payoff characteristics in the universal type space. In other words, which of the predictions obtained in the mechanism design literature with payoff-based hierarchies of beliefs are robust to the details of type spaces that are used to model those hierarchies?

To start with, one should mention that the existing mechanism design literature has been working with the following version of the universal type space. First, the fundamental uncertainty about which hierarchies of beliefs are formed consists only of agents' payoff types (as it is what the mechanism designer ultimately cares about). Second, agents' interim types are constructed by specifying a set of "abstract types" and two mappings from the set of abstract types into a set of payoffs and a set of feasible beliefs types. Each belief type is a probability measure on other agents' abstract types and thus on their payoffs and their beliefs about everyone else's abstract type, i.e. on the entire hierarchy of beliefs about payoff parameters. In these constructions an agent's belief types, i.e. probability measure over types of others could be arbitrary and just appended to the set of possible payoffs rather than derived from a prior conditionally on the information.

We show that as a result of such modelling approach where different models of agents' information are unified into one model of the universal type space, the mechanism designer, by disregarding some dimensions of agents' information, may lose the ability to condition agents' allocations on reports on these dimensions whereas such conditioning may allow to relax the incentive compatibility constraints. The reason of such loss lies in both above mentioned issues. Because of the second reason, agents' informative signals are not treated explicitly and so there is no scope for conditioning on them. The information contained in the signals may be useful, for example, for a scoring mechanism, where agent i is scored depending on how his report about his payoff type matches $-i$'s reports of signals about i's payoff. As it is known in the literature, such scoring could induce agents to reveal their types truthfully, e.g. because they are informationally small in the sense of McLean and Postlewaite (2002). And because of the first issue, even if agents informative signals are treated explicitly, the mechanism designer yet may

lose information about beliefs about signals if he restricts attention to payoff based hierarchies of beliefs. For example, the lottery mechanism of Crémer and McLean (1988) may work well if i's beliefs about others' signals are informative about i's payoff type, whereas his beliefs about others payoffs are not.

In this paper we deal in details exactly with this first issue[2]. We demonstrate that there exist social choice functions which are implementable on a type space where "redundant types" are considered as separate types may not be implementable on a type space where details of these "redundant types" are ignored. To formalize this idea we construct first an extended type space – a type space with payoff-and-signals-based hierarchies of beliefs. Then we provide an example how such payoff-and-signals-based hierarchies of beliefs could be mapped into payoff-based hierarchies of beliefs only and which information could be lost on the way. Finally we give an example of social choice function which is implementable on type spaces with payoff-and-signals-based hierarchies of beliefs and which is not implementable on a corresponding representation of a type space with payoff-based hierarchies of beliefs only.

Our results imply that impossibility results obtained on a given type space with payoff-based hierarchy of beliefs are non-robust. They are valid only for that type space but not for "redundant" type spaces which have the same payoffs-based belief hierarchies but where agents differ in their information. And so there could be an arbitrary number of models of information with the same payoffs and hierarchies of beliefs about the payoffs and for which there exist mechanisms achieving efficiency by exploiting heterogeneity in information, i.e. in signals.

In this paper we do not require that agents share a common prior at the ex ante stage, they could have any type of beliefs, e.g. also private priors about others prior beliefs about the joint realization of payoffs and signals. However, in the last section we discuss under which conditions restricting attention to payoff-based hierarchies of beliefs is without loss of generality

[2] Concerning the issue of exogeneous structure on belief mappings, in Gizatulina and Hellwig (2009) we have demonstrated that when agents form their hierarchies of beliefs about others' *payoffs and signals* and when those are sequences of conditional beliefs derived from a prior according to the Bayes rule, these hierarchies enjoy generically the BDP property, i.e. they are unique to a given constellation of a payoff characteristic and signals. Hence generically it is possible to achieve various efficiency results.

and it seems that the common prior plays a role in the sense that under common priors two approaches are more likely to be equivalent than when agents priors are heterogeneous and privately known.

In the following section we describe the environment. Then we show a usual approach in the existing literature, i.e. construction of type spaces with payoff based hierarchies of beliefs. Then we construct a payoff and signals based universal type space where agents' informative signals are a part of explicit description of agents' types. In the subsequent section we show that the set of social choice functions implementable on a type space with payoff and signals based beliefs hierarchies is strictly larger than the set of social choice functions implementable on type spaces with payoff-based beliefs hierarchies into which an initial type space with payoff and signals-based belief hierarchies is mapped. Finally, we provide the above mentioned discussion and give conclusions.

3.2 The Environment

There is a finite set I of agents indexed by $i = 1, ..., I$. Each agent has a payoff type $\theta_i \in \Theta_i$. Denote as usual the set of payoff types in the economy by $\Theta = \times_i \Theta_i$ and its element θ.

There is a set of social outcomes G, with elements g. Each agent has a utility function $u_i : G \times \Theta_i \times \mathbb{R} \to \mathbb{R}$, giving him some degree of utility as a function of a social alternative g, his payoff type θ_i and a potential monetary transfer from $\mathbb{R}$. That is we restrict attention to private values environments – agents utility does not depend as such on Θ_{-i}. A social choice function is a mapping $F : \Theta \to G$, so the choice of some alternative g depends only on the vector θ and if the mechanism designer knows θ, he would like the outcome to be $F(\theta)$.

3.3 Type Spaces

3.3.1 Payoffs-Based Universal Type Space

Most of the mechanism design literature working with type spaces allowing for privately know belief hierarchies (e.g. Neeman (2004), Bergemann and Morris (2005), Heifetz and Neeman (2006), etc.) employ a construction

suggested by Harsanyi (1967-68), which was formalized by Mertens and Zamir (1985) and Brandenburger and Dekel (1993).

In this construction each agent's information, relevant for his behaviour in a given strategic situation, is contained in his "type". Specifically, each agent's type describes his payoff relevant characteristic as part of the description of the state of nature and also his beliefs about others payoffs characteristics and others' beliefs. It has been shown that all this information could be succinctly represented by a triple $(Y_i, \widehat{\theta}_i, \widehat{\pi}_i)$ where Y_i is agent i's "type space" containing a set of possible types y_i, $\widehat{\pi}_i : Y_i \to \Delta(Y_{-i})$ is a mapping defining agent i's probabilistic beliefs about others' types and $\widehat{\theta}_i : Y_i \to \Theta_i$ is a mapping defining agent i's payoff type. Thus a given realization of $y_i \in Y_i$ defines uniquely some (θ_i, π_i).

From a given belief type π_i, defined on $\Delta(Y_{-i})$ it is possible to unfold agent's entire hierarchy of beliefs, i.e. his beliefs about others' payoff types[3], his beliefs about other agents' beliefs about everyone's payoffs and so on. In other words each $y_i = (\theta_i, \pi_i)$ could be equivalently represented as an infinite dimensional vector

$$y_i = (y_i^0, y_i^1, y_i^2, ...) \tag{3.1}$$

where

$y_i^0 = \theta_i \in \Theta_i$, i.e. $Y_i^0 \equiv \Theta_i$ (the zeroth level is agent' payoff);

$y_i^1 \in Y_i^1 \equiv \Delta(Y_{-i}^0)$ (agent i's beliefs about others' payoffs);

$y_i^2 \in Y_i^2 \equiv \Delta(Y_{-i}^1)$ (agent i's beliefs about others beliefs about everyone's payoffs);

....

$y_i^k \in Y_i^k \equiv \Delta(Y_{-i}^{k-1})$, etc.

The result of Mertens and Zamir (1985) and Brandenburger and Dekel (1993) was that any strategic situation, describable by some explicit hierarchy as in (3.1) has a unique representation as $y_i = (\widehat{\theta}_i(y_i), \widehat{\pi}_i(y_i))$ provided each order of beliefs $y_i^k = \Delta(Y_{-i}^{k-1})$ satisfies the following coherency condition

$$marg_{Y_{-i}^{k-2}} y_i^k = y_i^{k-1}$$

where $marg_{Y_{-i}^{k-2}}$ is the marginal on the space Y_{-i}^{k-2}. Any two types who differ in their information but who has the same belief hierarchy about others

[3]For an analysis from an interim perspective, it is usually assumed that each agent knows his own payoff type and own beliefs structure.

payoffs and others beliefs are called redundant. A type space containing all types with hierarchies of beliefs about Θ which are non-redundant, i.e. which do not have already a copy of themselves in this type space was named the Θ-based universal type space.

Note that agents may possess a common prior, i.e. there exist a $\Delta_i(Y) = \Delta_j(Y) = \Delta(Y)$ such that conditionally on observing y_i each agent's belief type is defined by

$$\pi_i(y_{-i}|y_i) = \frac{\Delta(y_{-i}, y_i)}{\sum_{y_{-i} \in Y_{-i}} \Delta(y_{-i}, y_i)}$$

Alternatively agents may have heterogeneous though commonly known priors about Y, then the above formula should be substituted by an appropriate version, accounting for each Δ_i and Δ_{-i} (as other agents' interim beliefs types would depend on Δ_{-i}).

Finally, at the ex ante stage each agent may be uncertain about the priors of the others. Then the underlying uncertainty consists of a cross sectional distribution of payoffs, signals and prior beliefs from which other agents derive their beliefs hierarchies. In other words, the prior Δ_i is a probability measure over (Δ_{-i}, Y).

A collection (Δ, Y) with $Y = \times_i Y_i$ and $\Delta = \times_i \Delta_i$ defines a type space from the prior perspective.

3.3.2 Extended Interim Type Space

Here instead of allowing agents to have at the interim stage beliefs only about payoffs and others beliefs, we allow that each agent's type is characterized in addition by a finite vector of signals which he believes to be strategically relevant as they maybe correlated to others' payoffs, signals and belief types. Each agent's belief type describes then his beliefs about others' payoff, signal and beliefs types, this is so-called "payoffs and signals"-based, i.e. $(\Theta \times S)$-based universal type space.

The notation is standard, the set of possible payoffs of an agent i is Θ_i with $\Theta = \times_i \Theta_i$, the set of signals is $\mathbf{S}_i$ with $\mathbf{S} = \times_i \mathbf{S}_i$ and we introduce $T_i = \Theta_i \times \mathbf{S}_i$, with $T = \times_i T_i$. The realization of agent's type is the following infinite dimensional vector:

$$t_i = (t_i^0, t_i^1, t_i^2, ...) \tag{3.2}$$

with

$$\begin{aligned} t_i^0 &\in T_i^0 \equiv \Theta_i \times \mathbf{S}_i \\ t_i^1 &\in T_i^1 \equiv T_i^0 \times \Delta_i(T_{-i}^0) \\ t_i^2 &\in T_i^2 \equiv T_i^0 \times \Delta_i(T_{-i}^1) \end{aligned}$$

and so on.

Similarly to the previous section the beliefs over the orders are required to satisfy the coherency condition. Given this any agent type could be represented as a triple

$$t_i = (\theta(t_i), s(t_i), \pi(t_i)).$$

We do not impose any requirement on the process that is behind a given set of interim types, e.g. there does not have to be a common prior. Instead for example at the ex ante stage each agent may possess a prior Δ_i over $(\Delta_{-i} \times \Theta \times S)$, i.e. describing his prior beliefs about other agents' priors and about a process distributing θ and s across agents.

This construction subsumes, e.g., the construction of Δ-hierarchies of Ely and Peski (2006) if we interpret their "types" as payoff relevant characteristics and informative signals.

Finally, as in the preceding section, this construction is consistent with three possible types of priors: a common prior, heterogeneous but commonly known and privately known priors.

3.3.3 Motivating Example

In this section we provide an example which is special cases of the construction in the section 2.2 allowing for commonly known, heterogeneous prior. This example shows how a mapping of $(\Theta \times S)$-based model into Θ-based universal type space fails to satisfy the BDP property, whereas within $(\Theta \times S)$-based model agents beliefs about others signals in the extended model are fully informative about his type.

Assume there are two agents, $I = 2$. Each agent's type is two dimensional and consists of his payoff and his signal about other agent's payoff: $\theta_i \in \Theta_i$ and $s_i \in S_i$. We assume that at the ex ante stage two agents have

heterogeneous prior about the joint realization of payoffs and signals. Agent i's prior is

$$\Delta_i \begin{pmatrix} \theta_i \\ s_i \\ \theta_j \\ s_j \end{pmatrix} = \mathcal{N}\left(\begin{pmatrix} \mu \\ 0 \\ \mu \\ 0 \end{pmatrix}, \begin{pmatrix} 1/\alpha & 0 & 0 & \rho \\ 0 & 1/\alpha & 0 & 0 \\ 0 & 0 & 1/\alpha & 0 \\ \rho & 0 & 0 & 1/\alpha \end{pmatrix} \right)$$

with $\rho > 0$. Whereas agent j' prior is

$$\Delta_j \begin{pmatrix} \theta_i \\ s_i \\ \theta_j \\ s_j \end{pmatrix} = \mathcal{N}\left(\begin{pmatrix} \mu \\ 0 \\ \mu \\ 0 \end{pmatrix}, \begin{pmatrix} 1/\alpha & 0 & 0 & 0 \\ 0 & 1/\alpha & \rho & 0 \\ 0 & \rho & 1/\alpha & 0 \\ 0 & 0 & 0 & 1/\alpha \end{pmatrix} \right).$$

Agents agree on priors' means, variance and that $cov_{\Delta_i}(\theta_i, s_i) = cov_{\Delta_i}(\theta_j, s_j) = cov_{\Delta_j}(\theta_i, s_i) = cov_{\Delta_j}(\theta_j, s_j) = 0$. But each agent believes that other agent's signal about his payoff is informative, i.e. $cov_{\Delta_i}(\theta_i, s_j) = \rho$ (and similarly for $cov_{\Delta_j}(\theta_j, s_i) = \rho$) whereas each of them believes that his own signal about other agent's payoff is fully noisy, i.e. $cov_{\Delta_i}(\theta_j, s_i) = cov_{\Delta_j}(\theta_i, s_j) = 0$.

Now let us consider each agent's hierarchy of beliefs in the extended universal type space and in the standard one.

The $(\Theta \times S)$-based hierarchy of beliefs looks as follows:

$k = 0 : \ t_i^0 = (\theta_i, s_i)$

$k = 1 : \ t_i^1 = \Delta_i(t_j^0)$ conditional on (θ_i, s_i) each agent believes that the other agent's payoff and signals are distributed normally with the following conditional means:

$$\begin{aligned} E_i\left[\theta_j | \theta_i, s_i\right] &= E_i\left[\theta_j | s_i\right] = \frac{Cov_{\Delta_i}(s_i, \theta_j)}{Var(s_i)}(s_i - E_i\left[s_i\right]) + E_i\left[\theta_j\right] \\ &= E_i\left[\theta_j\right] \\ &= \mu \end{aligned}$$

and

$$\begin{aligned} E_i\left[s_j|\theta_i,s_i\right] &= E_i\left[s_j|\theta_i\right] = \frac{Cov_{\Delta_i}(s_j,\theta_i)}{Var(\theta_i)}(\theta_i - E_i\left[\theta_i\right]) + E_i\left[s_j\right] \\ &= \rho\alpha(\theta_i - \mu) \end{aligned}$$

$k = 2 : t_i^2 = \Delta_i(t_j^1)$: agent i's beliefs about agent j's belief about i's payoffs and signals are his beliefs about $E_j\left[\theta_i|\theta_j,s_j\right]$ and $E_j\left[s_i|\theta_j,s_j\right]$ respectively.

$$\begin{aligned} E_i\left[E_j\left[\theta_i|\theta_j,s_j\right]|\theta_i,s_i\right] &= E_i\left[\frac{Cov_{\Delta_j}(s_j,\theta_i)}{Var(s_j)}(s_j - E_j\left[s_j\right]) + E_j\left[\theta_i\right]|\theta_i,s_i\right] \\ &= E_i\left[E_j\left[\theta_i\right]|\theta_i,s_i\right] \\ &= \mu \end{aligned}$$

$$\begin{aligned} E_i\left[E_j\left[s_i|\theta_j,s_j\right]|\theta_i,s_i\right] &= E_i\left[\frac{Cov_{\Delta_j}(s_i,\theta_j)}{Var(\theta_j)}(\theta_j - E_j\left[\theta_j\right]) + E_j\left[s_i\right]|\theta_i,s_i\right] \\ &= E_i\left[\rho\alpha(\theta_j - \mu)|\theta_i,s_i\right] \\ &= E_i\left[\theta_j|\theta_i,s_i\right] - \mu\rho\alpha \\ &= \mu(1-\rho a) \end{aligned}$$

and so on for $t_i^3, t_i^4, ...$

Now we shall construct a Θ-based hierarchy of beliefs of this model.

$y_i^0 = \theta_i$

$y_i^1 = \Delta_i(y_j^0)$ which, again, conditional on (θ_i, s_i) each agent believes that the other agent's payoff is distributed normally with the following conditional mean:

$$\begin{aligned} E_i\left[\theta_j|\theta_i,s_i\right] &= E_i\left[\theta_j|s_i\right] = \frac{Cov_{\Delta_i}(s_i,\theta_j)}{Var(s_i)}(s_i - E_i\left[s_i\right]) + E_i\left[\theta_j\right] \\ &= E_i\left[\theta_j\right] \\ &= \mu \end{aligned}$$

$y_i^2 = \Delta_i(y_j^1)$ as i knows that $E_j[\theta_i|\theta_j, s_j] = \mu$ his beliefs about j's belief about i's payoff are degenerate – he puts probability 1 on j believing that θ_i is distributed normally with a mean μ (and corresponding variance $1/\alpha$ defined by the priors).

All Θ-based higher order beliefs $y_i^3, y_i^4, ...$ are similarly degenerate.

Consequently, with such priors and information structure, each agent's Θ-based beliefs hierarchy (in the form $y_i = (\theta(y_i), \pi(y_i))$ is non-BDP, but their $(\Theta \times S)$-based beliefs hierarchies, i.e. their types $t_i = (\theta(t_i), s(t_i), \pi(t_i))$ are BDP.

3.4 Results

The main result of the paper is

THEOREM 3.1 *The set of $F(\Theta)$ implementable on a Θ-based interim type space induced by some type space with $(\Theta \times S)$-based hierarchies of beliefs is a strict subset of $F(\Theta)$ implementable on the initial $(\Theta \times S)$-based type space.*

First of all, the proof that any SCF which is implementable on a Θ-based representation of some $(\Theta \times S)$-based type space is also implementable on the initial $(\Theta \times S)$-based type space could left as a remark. This result is a straightforward consequence of the revelation principle.

Consider a mechanism in which all agents are asked to submit their Θ-based hierarchies of beliefs. By hypothesis, there exist incentive compatible decision rule $g(\theta_i(y_i), \theta_{-i}(y_{-i}))$ and payments $\{t_i(y_i, y_{-i})\}_{i \in I}$. Now suppose that agents are asked to submit their $(\Theta \times S)$-based hierarchies of beliefs. Because the designer has no less informations as in the case where agents are asked to submit their Θ-based hierarchies of beliefs, he could design the allocations as if agents had submitted Θ-based hierarchies of beliefs and achieve efficiency.

We shall proceed to the second part of the proof. We show that there exist functions $F(\Theta)$ and $(\Theta \times S)$-based type spaces such that when only a reduced type space, with Θ-based hierarchies of beliefs is considered $F(\Theta)$

is not implementable, whereas it is implementable on the initial type space with $(\Theta \times S)$-based hierarchies.

The proof is constructive – we describe an environment (utilities and types) and a social choice function which is implementable if the designer could ask agents to report their payoffs, signals and agents beliefs about others' payoffs, signals and others' beliefs, but which is not implementable if the only information that the designer could have is a space of (reported) payoffs and beliefs about others' payoffs induced by payoff and signals based space. This construction is based on the fact that agents' Θ-based beliefs are non-informative about his payoff type, i.e. they fail the BDP condition, whereas his $(\Theta \times S)$-beliefs satisfy the BDP property, hence we could employ the mechanism as in Crémer and McLean (1988) to achieve efficiency.

3.4.1 Proof of the Theorem: Part 2

Utilities, SCF and Type Space

Recall, there is a finite number of agents, indexed by $i = 1, .., I$. Agents have to make a decision whether to change a status quo allocation. There is a social alternative g (in our case the level of a public good) taking a value in a set G of possible values. Agents have to aggregate their preferences whether to switch to an alternative allocation (i.e. how much of a public good to have).

Each agent's utility is quasi-linear:

$$u_i(\theta_i, g, m_i) = \theta_i g + m_i \tag{3.3}$$

where m_i is utility from money, θ_i is preferences parameter and $v_i(\theta_i, g)$ is utility from the public good when it is at some level g (throughout the paper only the private values case is considered so $v_i(\cdot)$ depends only on θ_i).

A social choice rule $F : \Theta \to G$ maps a profile of payoff types $(\Theta = \times\Theta_i)$ into a social alternative (the level of public good). In this paper it is assumed that the objective is to maximize the aggregate welfare for each realization of the vector θ, i.e. the social choice rule is ex post efficient.

The cost of a public good is constant per capita and equals κ with $I \cdot \kappa = K(I)$. In this case an efficient social choice rule, based on the goal of maximization of social welfare, prescribes provision of a public good whenever $\sum_i \theta_i \geq K(I)$.

We shall restrict attention to a belief-closed subset of $(\Theta \times \mathbf{S})$-based universal type space satisfying three assumptions, labeled F, R and I. The first assumption says that agents payoffs are from finite set whereas agents signals are allowed to belong to an interval in $\mathbb{R}^{S_i}$.

Assumption F: *For each i the set Θ_i is finite and the set $\mathbf{S}_i \in \mathbb{R}^{S_i}$*

We also assume that it is common knowledge that each agent's belief type could be fully described by a vector of parameters, taking values in a compact interval in $\mathbb{R}^M$:

Assumption R[4]**:** *Any interim belief type $\pi(t_{-i}|t_i)$ is uniquely describable by a vector $\boldsymbol{\rho}_i \in \mathbb{R}^M$*

$$\boldsymbol{\rho}_i = (\rho_i^1, .., \rho_i^M) \tag{3.4}$$

and it is continuous in $(\rho_i^1, .., \rho_i^M)$ i.e. for each m there is a sequence of $\rho_i^m(k)$ such that

$$\lim_{k\to\infty} \rho_i^m(k) \to \rho_i^m$$

and

$$\lim_{k\to\infty} \int_{t_{-i}\in T_{-i}} f(.)d\pi_k(t_{-i}|t_i) \to \int_{t_{-i}\in T_{-i}} f(.)d\pi(t_{-i}|t_i)$$

for all bounded and continuous functions $f : T_{-i} \to \mathbb{R}$.

For example a vector $(\rho_i^1, .., \rho_i^M)$ could be a vector of M moments defining the probability measure over T_{-i} (note that multidimensionality of T_{-i} is not an issue – a vector $(\rho_i^1, .., \rho_i^M)$ is just a convention for corresponding parameters of the p.d.f. which may be matrices as well). Then, by the continuity theorem for moment generating functions we know that if $X_1, X_2, ...$ is a sequence of random variables with corresponding (presumably existing) moment generating functions $\varphi_1(.), \varphi_2(.), ...$ then $X_k \to_d X$ (in distribution) if and only if $\varphi_n(.) \to \varphi(.)$ for each moment.

Because by construction $(\theta(t_i), s(t_i), \pi(t_i))$ is homeomorphic to $(t_i^0, t_i^1, t_i^2, ...)$, the continuity of $(\theta(t_i), s(t_i), \pi(t_i))$ in $(\rho_i^1, .., \rho_i^M)$ implies the continuity of $(t_i^0, t_i^1, t_i^2, ...)$ in $(\rho_i^1, .., \rho_i^M)$ (for each order t_i^k).

[4] This quite simplfying assumption is made in order to avoid dealing with any topological notions of closeness of infinite belief hierarchies.

This means that, in addition to $(t_i^0, t_i^1, t_i^2, ...)$ and $(\theta(t_i), \pi(t_i))$ the third representation of each agent's type, given the assumed restriction, is possible:

$$t_i = (\theta_i, s_i, \rho_i^1, .., \rho_i^M) \tag{3.5}$$

where $(\rho_i^1, .., \rho_i^M)(t_i)$ defines belief type $\pi_i(t_i)$ of an agent i when he is type t_i, from which in turn a sequence as in (3.2) could be unfold.

Given the above condition the dimensionality of agent type T_i is labelled by $n_i = M + S_i + 1$. We will refer to a member of n_i by in.

The last assumption that $T = \times T_i$ should be satisfying is the following (where $\mathbf{s}_{-i-j}^i$ is the vector of signals of all agents but j about i's type):

Assumption I:

- *There exist no functions $q(\theta_{-i})$ and $q(\rho_{-i})$ such that either $E_i\left[q(\theta_{-i})|t_i\right]$ or $E_i\left[q(\rho_{-i})|t_i\right]$ would be an injective mapping from $(\theta_i, \rho_i^1, .., \rho_i^M)$ into $\mathbb{R}$.*

- *But there exists a function $q(s_{-i-j}^{in})$ and agents j and k such that for each t_i $(i \neq j \neq k)$ and for each dimension $in \in n_i$ it holds:*

  A. $E_i\left[q(\mathbf{s}_{-i-j}^{in})|t_i\right]$ is an injective mapping from $(\theta_i, \rho_i^1, .., \rho_i^M)$ into $\mathbb{R}$;

  B.

$$\sum_{in \in n_i} E_i\left[\left(q(\mathbf{s}_{-i-j}^{in}) - E_i\left[q(\mathbf{s}_{-i-j}^{in})|t_i\right]\right)^2 |t_i\right] \tag{*}$$
$$\leq \sum_{kn \in n_k} E_i\left[\left(q(\mathbf{s}_{-k-i}^{kn}) - E_k\left[q(\mathbf{s}_{-k-i}^{kn})|t_i\right]\right)^2 |t_i\right]$$

We discuss in details the consistence of the first bullet point with the part A of the second bullet point in the next subsection, in the Proposition 3.2. The last term in the 2nd bullet point requires that agent i's sum of conditional variances in the value of $q(.)$ as a function of signals about his type is weakly smaller than the sum of conditional variances of $q(.)$ as function of signals about some agent k. This condition is satisfied for example when the variance in signals is independent of types and the same across the agents.

Result

The first proposition asserts that the Assumption I has more than an empty set of type spaces, i.e. it is internally consistent.

PROPOSITION 3.2 *There exists type spaces* (Δ, S, Θ) *which satisfy the Assumption I.*

Proof. Example 1 is one such type space. Recall that there are two agents each having one signal, so we let $q(s^{in}_{-i-j}) = s_j$ and the same for signal about agent j. As it was shown, agent i's expectation about agent j's signal about his payoff is a 1:1 function of his payoff type i.e. $E_i\left[q(s^{in}_{-i-j})|t_i\right] = E_i\left[s_j|t_i\right] = \theta_i$, idem for agent j. Then take $q(\theta_{-i}) = \theta_j$, we have $E_i\left[\theta_j|t_i\right] = \mu$ and obviously there is no any function which would be dependent only on θ_j which would provide with more information about t_i. Finally as all ρ^M_j and ρ^M_i are commonly known and independent of the realization of θ_i we have the result. ■

In the remaining we show that asking agents for their $(\Delta_i, \widehat{\theta}(t_i), \widehat{s}(t_i))$ allows us to achieve efficiency under assumptions R, I and F.

PROPOSITION 3.3 *For each interim type space* $(T_i, \widehat{\theta}(t_i), \widehat{s}(t_i), \widehat{\pi}(t_i))$ *satisfying the assumptions R,I and F there exists a scoring mechanism with the truthful Bayesian Nash Equilibrium. The allocations achievable in the truthful equilibrium of the scoring mechanism are ex post efficient, satisfy the ex post budget balance condition and interim individual rationality constraints.*

Perfect working of the scoring mechanism that we construct is insured by the assumptions I and F. By the assumption I if we know agents beliefs about others signals we could learn his payoff and his belief type, so we could construct a vector of payments conditional on others' reports of their signals about i such that each i by revealing his expectation about others signals reveals uniquely his payoff and belief type. The assumption F allows to disentangle each different payoff type by putting a finite weight on the score for agent's report of his payoff type.

In the remaining part of this section we provide a constructive proof of the theorem. Agent i's reporting strategy is $\widetilde{t}_i : T_i \to T_i$. When an agent

makes a truthful announcement his strategy is denoted $\widetilde{t}_i(t_i) = t_i$ and when he mispresents on either dimension of t_i we use $\widetilde{t}_i(t_i) = t'_i$.

DEFINITION 3.4 (SCORING MECHANISM) *The scoring mechanism is the following three stage game:*

1. *The designer proposes an allocation rule (as a function of agents' reports about their types):*

   - *the individual payment function* $m_i : \widetilde{T} \to \mathbb{R}$ *with:*
     $m_i^*(\widetilde{t}) = \phi(\theta_i(\widetilde{t}_i), \widetilde{t}_{-i}) + Q_i(\widetilde{t}_{-j}) - Q_k(\widetilde{t}_{-i})$ *if* $g = 1$ *and*
     $m_i^*(\widetilde{t}) = Q_i(\widetilde{t}_{-j}) - Q_k(\widetilde{t}_{-i})$ *if* $g = 0$;
   - *the public good provision rule* $g : \widetilde{T} \to \{0, 1\}$ *with:*
     $g^* = 1$ *if* $\sum_i \theta_i(\widetilde{t}_i) \geq K(I)$ *and*
     $g^* = 0$ *otherwise;*
   - *the ex post budget balance condition:*
     $\sum_i m_i^*(\widetilde{t}) = K(I)$ *if* $g^* = 1$ *and*
     $\sum_i m_i^*(\widetilde{t}) = 0$ *otherwise.*

2. *Agents vote to play the scoring mechanism given the above allocation rule. If there is an unanimous vote to play the scoring game, each agent reports* $\widetilde{t}_i$*. Otherwise the allocation is:* $g^* = 0$, $m_i^*(\widetilde{t}) = 0$

3. *Given the behaviour at the stage 2 individual allocations specified by the mechanism are obtained.*

In the payment function $m_i^*(\widetilde{t})$, the term $Q_i(\widetilde{t}_{-j})$ is the score each i is assigned conditionally on his report and reports of all other agents but j about the distribution of types in the economy. This score is a weighted average of the scores on each dimension $in \in \left\{\theta_i, \boldsymbol{\rho}_i, \mathbf{s}_i^{\rho}, \mathbf{s}_i^{\theta}\right\}$

$$Q_i(\widetilde{t}_{-j}) = \sum_{in} w_n Q_{i(-j)}^{in} \tag{3.6}$$

where w_n is a weight given to a score $Q_{i(-j)}^{n}$ from a dimension in. The applied scoring rule is the quadratic score

$$Q_{i(-j)}^{in} = (\widehat{t}_{in}^{-i-j} - \widetilde{t}_{in})^2 \tag{3.7}$$

where $\widehat{t}_{in}^{-i-j}$ is an estimate of i's type in dimension in by all the remaining agents but agent j, i.e. by $-i-j$. We set the estimator $\widehat{t}_{in}^{-i-j}$ to be $\widehat{t}_{in}^{-i-j} = q(s_{-i-j}^{in})$, i.e. it is a function of signals reported by $-i-j$ about i's type on some dimension in satisfying the assumption $I.2$. The term $Q_k(\widetilde{t}_{-i})$ is a similar score assigned to an agent k which is independent of the reports of agent i. The term $\phi(\theta_i(\widetilde{t}_i), \widetilde{t}_{-i})$ in $m_i^*(\widetilde{t})$ is a function that defines the way the ex post surplus from the provision of a public good is redistributed among the agents given all announcements. Because the mechanism is efficient we have

$$\phi(\theta_i(\widetilde{t}_i), \widetilde{t}_{-i}) \leq \theta_i(\widetilde{t}_i)$$

i.e. agents never contribute to the public good more than their valuation.

The next proposition provides details on incentive compatibility of the scoring game.

PROPOSITION 3.5 (INCENTIVE COMPATIBILITY) *For each i there exist a vector of weights $\mathbf{w}_{n_i}$ such that, if agent i conjectures that other agents $(-i)$ report truthfully t_{-i} reporting truthfully t_i constitutes the best reply strategy for agent i.*

Proof. Each agent's expected utility given his strategy and given that he expect truthful strategies from the others (i.e. $\widetilde{t}_{-i} = t_{-i}$) is

$$\begin{aligned} EU_i(\widetilde{t}_i, t_{-i}; t_i) &= E_i\left[g(\widetilde{\theta}_i, \theta_{-i})|t_i\right]\theta_i - E_i\left[m_i(\widetilde{t}_i, t_{-i})|t_i\right] = \\ &= E_i\left[g(\widetilde{\theta}_i, \theta_{-i})|t_i\right](\theta_i - \phi(\widetilde{\theta}_i, t_{-i})) - \\ &\quad -E_i\left[Q_i(\widetilde{t}_i, t_{-i-j})|t_i\right] + E_i\left[Q_k(t_k, t_{-k-i})|t_i\right] \end{aligned}$$

Here $\widetilde{\theta}$ is a short-cut for $\theta_i(\widetilde{t}_i(t_i))$.

The term $E_i\left[g(\widetilde{\theta}_i, \theta_{-i})|t_i\right](\theta_i - \phi(\widetilde{\theta}_i, t_{-i}))$ arises because by rules of the mechanism each agent pays $\phi(\widetilde{\theta}_i, t_{-i})$ ex post only if the public good is provided, it happens with a probability[5] $E_i\left[g(\widetilde{\theta}_i, \theta_{-i})|t_i\right]$.

By construction of the mechanism each agent's report on a dimension in' does not affect his score on the dimension in'', so these could be treated

[5] As usual in the literature, we use interchangely the expected level of the public good and expected probability of provision of a public good in a unit of 1.

separately. First we study agents incentives when reporting θ_i. Consider any deviation $\theta_i' \neq \theta_i$. The existence of a weight w^θ, inducing truthful reporting of θ_i, follows from the IC constraint

$$\begin{aligned} & E_i\left[g(\widetilde{\theta}_i, \theta_{-i})|t_i\right](\theta_i - \phi(\widetilde{\theta}_i, t_{-i})) - w^\theta E_i\left[Q_i(\widetilde{\theta}_i, s^{\theta_i}_{-i-j})|t_i\right] \\ \geq & E_i\left[g(\widetilde{\theta}_i', \theta_{-i})|t_i\right](\theta_i - \phi(\widetilde{\theta}_i', t_{-i})) - w^\theta E_i\left[Q_i(\widetilde{\theta}_i', s^{\theta_i}_{-i-j})|t_i\right] \end{aligned}$$

(we omit $E_i\left[Q_k(\widetilde{t}_k, t_{-k-i})|t_i\right]$ as it is independent on the report of i and so does not affect his incentives). Rearranging this brings us a condition on w^θ :

$$w^\theta \geq \frac{E_i\left[g(\widetilde{\theta}_i', \theta_{-i})|t_i\right](\theta_i - \phi(\widetilde{\theta}_i', t_{-i})) - E_i\left[g(\widetilde{\theta}_i, \theta_{-i})|t_i\right](\theta_i - \phi(\widetilde{\theta}_i, t_{-i}))}{E_i\left[Q_i(\widetilde{\theta}_i', s^{\theta_i}_{-i-j})|t_i\right] - E_i\left[Q_i(\widetilde{\theta}_i, s^{\theta_i}_{-i-j})|t_i\right]} \tag{3.8}$$

with $w^\theta = \max_{t_i, t_i', t_{-i} \in \mathbf{T}} w^\theta(t_i, t_i')$.

A finite weight w^θ exists if for any possible $\widetilde{\theta}_i$ and $\widetilde{\theta}_i'$ this expression is finite. In turn it means that the nominator should be finite and the denominator strictly greater than zero. This requirement is obviously satisfied for the nominator, as all variables are bounded. The strict positivity of the denominator is due to the following. By reporting any $\widetilde{\theta}_i$ agent expects his score to be

$$\begin{aligned} E_i\left[Q_i^\theta(\widetilde{\theta}_i, s^{\theta_i}_{-i-j})|t_i\right] &= E_i\left[(q(s^{\theta_i}_{-i-j}) - \widetilde{\theta}_i)^2|t_i\right] \\ &= E_i\left[(q(s^{\theta_i}_{-i-j}) - E_i\left[q(s^{\theta_i}_{-i-j})\right] + E_i\left[q(s^{\theta_i}_{-i-j})\right] - \widetilde{\theta}_i)^2|t_i\right] \\ &= E_i\left[((q(s^{\theta_i}_{-i-j}) - E_i\left[q(s^{\theta_i}_{-i-j})\right])^2 + (E_i\left[q(s^{\theta_i}_{-i-j})\right] - \widetilde{\theta}_i)^2|t_i\right] \end{aligned}$$

He expects it to be minimized when $\widetilde{\theta}_i^* = E_i\left[q(s^{\theta_i}_{-i-j})|t_i\right]$ (from which the value of θ_i could be uniquely found by the Assumption $I.2$.) If agent mispresents his type, i.e. he reports $\widetilde{\theta}_i^* \neq E_i\left[q(s^{\theta_i}_{-i-j})|t_i\right]$ the second term is strictly positive in expectation, hence we have $E_i\left[Q_i^\theta(\widetilde{\theta}_i', s^{\theta_i}_{-i-j})|t_i\right] > E_i\left[Q_i^\theta(\widetilde{\theta}_i, s^{\theta_i}_{-i-j})|t_i\right]$.

Extraction of agent's belief type is straightforward. Given that i expects that $s^{\rho_i^m}_{-i-j}$ are reported truthfully, each i while maximizing his expected

utility is minimizing his expected payment as it is the only part of the utility affected by his report of ρ_i^m. Hence he minimizes:

$$E_i\left[Q_{i\rho_i^m}|t_i\right] = E_i\left[(\widehat{t}_{i\rho_i^m}^{-i-j} - \widetilde{t}_i^{\rho_i^m})^2|t_i\right] \tag{3.9}$$

Rewriting this again as in the above case:

$$E_i\left[Q_{i\rho_i^m}\right] = E_i\left[(\widehat{t}_{i\rho_i^m}^{-i-j} - E_i\left[\widehat{t}_{i\rho_i^m}^{-i-j}\right])^2 + (E_i\left[\widehat{t}_{i\rho_i^m}^{-i-j}\right] - \widetilde{t}_i^{\rho_i^m})^2|t_i\right] \tag{3.10}$$

one observes that $E_i\left[Q_{i\rho_i^m}\right]$ is minimized when

$$\widetilde{t}_i^{\rho_i^m *} = E_i\left[\widehat{t}_{i\rho_i^m}^{-i-j}|t_i\right] = E_i\left[q(\mathbf{s}_{-i-j}^{\rho_i^m})|t_i\right] \tag{3.11}$$

As $E_i\left[q(\mathbf{s}_{-i-j}^{\rho_i^m})|t_i\right]$ is an injective mapping from (θ_i, ρ_i^m) one learns ρ_i^m for any $m \in M$.

The incentive compatibility for revealing signals is straightforward, it follows from the weak indifference argument. No misreporting of s_i^{jn}($j \neq k \neq i$) could improve allocation of i, it could only worsen it by increasing the expected score. This is because apart of $E_i\left[Q_{s_i^{jn}}\right]$ a report of s_i^{jn} does not affect the level of a public good $g(\theta)$ nor it changes $E_i\left[Q_k(\widetilde{t}_{-i})\right]$. This means i could not sponge off into his pocket proceeds from an increased score of any other agent as the ex post payment of Q_j for some $j \neq k$ is paid to agent l and i in turn receives proceeds from Q_k which are independent of any report of i by construction. Thus we could always find an ε-reward scheme, such that if i reports truthfully his signals he gets ε (conditionally on some finite score) and he gets 0 otherwise. ■

Proposition 3.6 (Interim Individual Rationality) *Given that the decision rule is efficient, any i, when votes to play the scoring game, expects the interim individual rationality constraint*

$$E_i\left[\theta_i g(t^*) - m_i(t^*)|t_i\right] \geq 0.$$

to be satisfied in the truthful equilibrium of the scoring game.

Proof. Recall that the expected utility in the truthful equilibrium is

$$\begin{aligned} E_i\left[\theta_i g(t^*) - m_i(t^*)|t_i\right] &= E_i\left[g(\theta_i^*, \theta_{-i}^*)|t_i\right](\theta_i - \phi(\theta_i^*)) \\ &\quad - E_i\left[Q_i(t_i^*, t_{-i-j}^*)|t_i\right] + E_i\left[Q_k(t_k^*, t_{-k-i}^*)|t_i\right] \end{aligned}$$

Given the condition $(*)$ in the Assumption I we have that in the truthful equilibrium

$$E_i\left[Q_i(t_i^*, t_{-i-j}^*)|t_i\right] \leq E_i\left[Q_k(t_k^*, t_{-k-i}^*)|t_i\right]$$

Given that $\theta_i - \phi(\theta_i) > 0$ for any θ_i in the truthful equilibrium (by the efficiency of provision decision) we have that

$$E_i\left[g(t_i^*, t_{-i}^*)|t_i\right](\theta_i(t_i^*) - \phi(\theta(t_i^*))) \geq 0$$

Hence is the result. ■

PROPOSITION 3.7 (BUDGET BALANCE) *Any equilibrium of the scoring game satisfies the ex post budget balance condition.*

Proof. We balance the budget as follows. If ex post the public good is not provided, no payments are made to the designer, but agent i pays to j the ex post value of $Q_i(t_{-j}^*)$ and receives from agent k his corresponding $Q_k(t_{-i}^*)$. With $I > 2$ we could always settle such transfers.

If $g^* = 1$ and so $m_i^*(t)$ are positive, find some $\phi(.)$ such that $\sum_i \phi(\theta(t_i^*)) = K$. The side transfers among agents are as in the case with $g^* = 0$. ■

Three proofs imply the Theorem.

There is one additional remark. To demonstrate the possibility result we have assumed that the set Θ is finite. Allowing for Θ to be an interval does not change much the qualitative part of our conclusions – each agent's surplus could be extracted up to an arbitrary ε amount, similarly to the result of McAfee and Reny (1992). It could be demonstrated as follows. Divide Θ into K intervals denoted $\left\{\widehat{\theta}_1, .., \widehat{\theta}_k, .., \widehat{\theta}_K\right\}$ with $\widehat{\theta}_k = \left[\underline{\theta}_k, \overline{\theta}_k\right)$ and $\overline{\theta}_{k-1} = \underline{\theta}_k$, such that $\overline{\theta}_k - \underline{\theta}_k \leq \varepsilon$. Each agent, instead of reporting θ_i is allowed to report only the interval $\widehat{\theta}_k$ such that $\theta_i \in \widehat{\theta}_k$, that is we "discretize" the reports[6].

Run the scoring mechanism as above with the only modification of the agents scoring rule as

$$\begin{aligned} Q_i^{\theta}(\widetilde{\theta}_i, q(\widetilde{s}_{-i-j}^{\theta_i})) &= 0 \text{ if } q(\widetilde{s}_{-i-j}^{\theta_i}) \in \widehat{\theta}_k \text{ and } \widetilde{\theta}_i \in \widehat{\theta}_k \\ Q_i^{\theta}(\widetilde{\theta}_i, q(\widetilde{s}_{-i-j}^{\theta_i})) &> 0 \text{ if } q(\widetilde{s}_{-i-j}^{\theta_i}) \in \widehat{\theta}_k \text{ and } \widetilde{\theta}_i \notin \widehat{\theta}_k \end{aligned}$$

[6] A similar discretizing technique was employed in Miller, Pratt, Zeckhauser, and Johnson (2007).

Next define a system of weights $\mathbf{w}$ as in the case with finite Θ, fully similarly to (3.8) inducing agents to report truthfully the interval $\widehat{\theta}_k$ where their valuations fall into. Such finite weights exists again because the denominator is strictly positive, as if $\widetilde{\theta}_i = \widetilde{\theta}_i' \notin \widehat{\theta}_k$ $E_i\left[Q_i^\theta(\widetilde{\theta}_i', q(\widetilde{s}^{\theta_i}_{-i-j}))|t_i\right] > 0$ and $E_i\left[Q_i^\theta(\widetilde{\theta}_i', q(\widetilde{s}^{\theta_i}_{-i-j}))|t_i\right] = 0$ if $\widetilde{\theta}_i \in \widehat{\theta}_k$. Each agent contributes only $\phi(\underline{\theta}_k)$ which, by choosing an appropriate constellation of $\mathbf{w}$ and K could be made to be arbitrary close to $\phi(\theta_i^*)$. The individual rationality and the budget balance conditions follow immediately, similarly to the case with finite Θ.

Thus, the richness of the payoff type space as such does not preclude us to extract almost the entire agents' surplus.

3.5 Discussion: Further Research

3.5.1 On equivalence of two approaches

The question that arises from the result of this paper is then the following: under which conditions, asking agents to report only their Θ-based hierarchies of beliefs is without loss of generality?

There are two possible questions to consider:

1. When by knowing only Θ-based hierarchies of beliefs of agents implementation results are the same as if we knew agents $(\Theta \times S)$-based hierarchies of beliefs?

2. When by knowing only Θ-based hierarchies of beliefs of agents could we recover precisely what would be their $(\Theta \times S)$-based hierarchies of beliefs?

For the first question it is obvious that when it is impossible to implement a given social choice function on a type space with $(\Theta \times S)$-based hierarchies of beliefs it is also impossible to implement it on a Θ-based hierarchies. Also, as a part of our theorem shows, when it is possible to implement a SCF on Θ-based hierarchies of beliefs it is also possible to do it on $(\Theta \times S)$-based hierarchies of beliefs.

The answer to the second question is two-fold. It is impossible to recover agents $(\Theta \times S)$-hierarchy of beliefs from Θ-based one if:

1. (a) Θ-based hierarchy of beliefs is independent of values of s_i (as it was in our motivating example);
   (b) Θ-based hierarchy is dependent on s_i but from Θ-based beliefs we cannot solve for the values of s_i.

The question a. in the essence is just a special case of the second one, but for clarity we shall discuss them rather separately. In the following we consider potential severity of both issues in the models with a common prior, with heterogeneous but known priors and finally where agents have heterogenous and privately known prior beliefs.

Common prior case Under a common prior the first issue of independence does not arise, as agents agree on whether signals bear any information or not, so if signals are non-informative the Θ-based hierarchy of beliefs would not depend on them, whereas the $(\Theta \times S)$-hierarchy of beliefs would depend on them but would not provide any more information about agents payoffs than the Θ-based hierarchy of beliefs.

However under common priors the second issue yet may be present. Gizatulina and Hellwig (2009) have demonstrated that generically if the mechanism designer knows agents' $(\Theta \times S)$-hierarchies of beliefs he could learn the specific values of their payoffs and signals. However if the designer knows only the Θ-based hierarchies of agents beliefs, the answer to the question whether information about the signals is available, i.e. whether he could "reconstruct" $(\Theta \times S)$-hierarchies and whether he could condition the allocation on the reported signals is yet open.

Recall that when agents form their hierarchies of beliefs about others payoffs they condition the common prior on their signals. Thus any $y^1 = \Delta_i(\theta_{-i})$ is a conditional probability, i.e. it could be rewritten as $y_i^1 = \Delta_i(\theta_{-i}|\theta_i, s_i)$. Similarly for each agent $j \neq i$, $y_j^1 = \Delta_j(\theta_{-j}|\theta_j, s_j)$. For the second order we have $y_i^2 = \Delta_i(y_{-i}^1) = \Delta_i(\Delta_{-i}(\theta_{-(-i)}|\theta_{-i}, s_{-i})|\theta_i, s_i)$. So under common priors either y_i^1 or y_i^2 should be dependent on agent i's *entire* profile of signals about others' payoffs and others' signals as each signal that i has is informative either about others' payoffs (and so it enters into his first order beliefs) or about others signals (then it enters then into his beliefs about others' beliefs, i.e. his second order beliefs).

This means that we could learn agents signals from Θ-based beliefs if

we could "invert" it. To do this we may, for example, construct a system of equations of conditional moments of $y_i^1, y_i^2, ..$ (which, arguably, depend on unknown (θ_i, s_i)) and solve this system for values of (θ_i, s_i). If it is possible to accomplish, the two approaches are equivalent and so there is no loss in generality in considering only payoff based hierarchies of beliefs. Arguably, this possibility depends on the details of the information process and it is an open question how often one could do it, i.e. how generic is equivalence of two approaches under the common prior assumption.

Heterogenous known priors When agents do not share a common prior though their priors are known it may happen, as in our motivating example, that no one's first or second order beliefs about others' payoffs depend on agents signals as no one believes signals to be informative. Hence we cannot recover information about signals from Θ-based beliefs alone, so two approaches are not equivalent. Also, regarding the issue in b. even if agents' first or second order beliefs are yet dependent on signals, again as in the case with a common prior the equivalence of two approaches depends on the existence of solutions to the system of moments equations.

Private heterogenous prior beliefs Finally, when agents prior beliefs are also their information, two approaches are even less likely to be equivalent. As in the case with commonly known heterogenous priors, agents beliefs about everyone's payoffs may be well independent from their beliefs about everyone's signals. The second problem – possibility to solve for (θ_i, s_i) aggravates. If agents prior beliefs are not commonly known (e.g. if we would like to avoid any parametrization of the ex ante stage) and if the payoff and signals are from compact intervals (rather than discrete sets), in order to reconstruct $(\Theta \times S)$-based hierarchy of beliefs from the Θ-based one we have to solve for (θ_i, s_i) and $\Delta_i(\Delta_{-i} \times \Theta \times S)$. Hence when Δ_i is allowed to be arbitrary we need to solve for an infinity of values[7]. Even if on the other hand, Θ-based hierarchy of beliefs provides us with an infinity

[7] But if we would like to restrict our analysis to a class of priors, which are fully paramterizable by some finite vector of paramters, as in our construction of priors descirbable by $\rho_i = (\rho_i^1, .., \rho_i^M)$, the problem of finding the values of (θ_i, s_i) is easened. Simply from the Θ-based infinite hierarchy of beliefs the designer has to solve for the vector $(\theta_i, s_i, \rho_i^1, .., \rho_i^M)$.

of observations on (θ_i, s_i) and $\Delta_i(\Delta_i \times \Theta \times S)$ it is not trivial to say how an infinite dimensional system of equation will behave.

Full force exploration of the question when two approaches are equivalent is left for our further research.

3.5.2 On "minimal type spaces"

Another question that arises from the results of this paper is what is the "minimal" type space capturing all necessary and sufficient details of agents' information that may be employed for the analysis of the mechanism design problems?

Ely and Peski (2006) have proved that to find type-space invariant predictions about rationalizable strategies in a given game it is necessary and sufficient to specify agents conditional hierarchies of beliefs, i.e. $\Delta(\Theta)$-based universal type space. But, as they themselves say this, such specification is working only for that solution concept and changing the concept would entail necessity to change the minimal description of agents information (see also their discussion of the danger that too many strategically redundant types who behave similarly but who has different belief hierarchies maybe generated, p.23, footnote 6).

What would be the minimal universal type space for the mechanism design problems is an open question. In our paper suggests that specifying $(\Theta \times S)$-based hierarchies changes predictions from the case where only Θ-based UTS is considered. However we do not prove that this $(\Theta \times S)$-based specification is either necessary or sufficient to obtain type space invariant prediction. Though we suspect that under the common prior assumption it would be both necessary and sufficient description of the environment.

3.6 Concluding Remarks

Any impossibility result obtained on a type space with Θ-based hierarchies of beliefs is tight, it is valid only for that specific environment and there may exists a large set of "redundant" type spaces, having the same payoffs and belief hierarchies about payoffs and beliefs about payoffs etc. but where yet agents differ in informative signals on which they condition their beliefs about others payoffs and others beliefs, and where the mechanism designer

could achieve implementation by conditioning on reported signals.

The issue of robust implementation notwithstanding, our result questions to which extent asking agents to report only their payoff and beliefs types from the universal type space in a direct revelation mechanisms is an appropriate technique to be used in environments where agents possess nontrivial privately known beliefs.

Also, among other things, our result extends the set of type spaces on which mechanisms exploiting the BDP property work. Specifically, compared to Crémer and McLean (1988) and McAfee and Reny (1992) we allow agents not to possess a common prior and their beliefs about others payoffs types do not have to be informative about their own payoff types.

Bibliography

Aumann, R. (1987): "Correlated equilibrium as an expression of Bayesian raionality," *Econometrica*, 55, 1–18.

Barelli, P. (2009): "On the genericity of full surplus extraction in mechanism design," *Journal of Economic Theory*, 144, 1320 – 1333.

Bergemann, D., and S. Morris (2005): "Robust mechanism design," *Econometrica*, 73, 1521–1534.

Brandenburger, A., and E. Dekel (1993): "Hierarchies of Beliefs and Common Knowledge," *Journal of Economic Theory*, 59, 189–198.

Crémer, J., and R. McLean (1988): "Full extraction of the surplus in Bayesian and dominant strategy auctions," *Econometrica*, 56, 1247–1257.

Dekel, E., D. Fudenberg, and S. Morris (2007): "Interim Correlated Raionalizability," *Theoretical Economics*, 2, 15–40.

Dixit, A. (2003a): "On Modes of Economic Governance," *Econometrica*, 71, 449–481.

——— (2003b): "Trade expansion and contract enforcement," *Journal of Political Economy*, 111, 1293–1317.

——— (2004): *Lawlessness and Economics: Alternatives Modes of Governance*. Princeton University Press, Princeton.

Ely, J., and M. Peski (2006): "Hierarchies of beliefs and interim rationalizability," *Theoretical Economics*, 1, 19–65.

Enriques, L. (2002): "Do Corporate Law Judges Matter? Some Evidence from Milan," *European Business Organization Law Review*, 3, 765–821.

FANG, H., AND S. MORRIS (2006): "Multidimensional private value auctions," *Journal of Economic Theory*, 126, 1–30.

GALANTER, M. (1978): "Why the 'Haves' Come Out Ahead: Speculations on the Limits of Legal Change," *Law and Society Review*, 9, 95–160.

GENNAIOLI, N., AND A. SHLEIFER (2007): "The Evolution of Common Law," *Journal of Political Economy*, 115, 43–68.

GIZATULINA, A., AND M. HELLWIG (2009): "Payoffs could be inferred from beliefs, generically, when beliefs are conditioned on information," Working paper, Max-Planck Institute for Research on Collective Goods.

HARSANYI, J. (1967/68): "Games with Incomplete Information Played by Bayesian Agents," *Management Science*, 14, 159–182, 320–334, 486–502.

HEIFETZ, A., AND Z. NEEMAN (2006): "On the generic (im)possibility of full surplus extraction in mechanism design," *Econometrica*, 117, 213–233.

HELLWIG, M. (2003): "Public-Good Provision with Many Participants," *Review of Economic Studies*, 70, 589–614.

——— (2007): "The Provision and Pricing of Excludable Public Goods: Ramsey-Boiteux Pricing versus Bundling," *Journal of Public Economics*, 91, 511–540.

HILL, P. (2003): "Heisei Yakuza: Burst Bubble and Botaiho," *Social Science Japan Jounral*, 6, 1–18.

——— (2006): "The Japanese Mafia, Take Two: Postscript to the Paperback Edition," *Working Paper, Department of Sociology, University of Oxford*.

HOBBES, T. (1651): *Leviathan*. available at http://oregonstate.edu/instruct/phl302/texts/hobbes.

KOSENOK, G., AND S. SEVERINOV (2008): "Individually Rational, Budget-Balanced Mechanisms and Allocation of Surplus," *Journal of Economic Theory*, 140, 126–161.

LEMPERT, R. (1999): "A classic at 25: Reflections on Galanter's "Haves" article and works it has inspired," *Law and Society Review*, 38.

Li, J. (2000): "The benefits and costs of relation-based governance: and explanation of the East-Asian miracle and crisis," *Working Paper City University of Hong Kong*.

Mailath, G., and A. Postlewaite (1990): "Asymmetric information bargaining problems with many agents," *Review of Economic Studies*, 57, 351–367.

Mas-Colell, A. (1985): *The Theory of General Economic Equilibrium: A Differentiable Approach, Econometric Society Monograph 9*. Cambridge University Press, Cambridge.

McAfee, P., and P. Reny (1992): "Correlated information and mechnanism design," *Econometrica*, 60, 395–421.

McLean, R., and A. Postlewaite (2002): "Informational size and incentive compatibility," *Econometrica*, 70, 2421–2453.

Mertens, J., and S. Zamir (1985): "Formulation of Bayesian analysis for games with incomplete information," *International Journal of Game Theory*, 10, 619–632.

Miller, N., J. Pratt, R. Zeckhauser, and S. Johnson (2007): "Mechanism design with multidimensional, continuous types and interdependent valuations," *Journal of Economic Theory*, 136, 476–496.

Neeman, Z. (2004): "The relevance of private information in mechanism design," *Journal of Economic Theory*, 117, 55–77.

Norman, P. (2004): "Efficient Mechanisms for Public Goods with Use Exclusions," *Review of Economic Studies*, 71, 1163–88.

Rousseau, J.-J. (1762): *Du Contrat Social ou Principes du Droit Politique*. available at http://fr.wikisource.org/wiki.

Sheehan, R., and D. Songer (1992): "Who Wins on Appeal? Upperdogs and Underdogs in the United States Courts of Appeals," *American Journal of Political Science*, 36, 235–258.

Skrzypacz, A., and Y. Feinberg (2005): "Uncertainty about uncertainty and delay in bargaining," *Econometrica*, 73.

Appendix A

Appendix

A.1 Ehrenwörtliche Erklärung

Hiermit bestätige ich ehrenwörtlich, dass ich diese Dissertationsschrift selbstständig eingefertigt habe und die benutzten Hilfsmittel vollständig und deutlich angegeben habe.

Alia Gizatulina

Bonn, den 7 Mai, 2009

A.2 Vita

Alia Gizatulina

Born: *in 1979, in Bishkek, Kyrgyzstan*

Nationality: *Kyrgyzstan*

Education:

1996 – 2001: Kyrgyz-Russian Slavic University, Bishkek, Kyrgyzstan (MA in Economics)

2001 – 2003: TSE, University of Toulouse, Toulouse, France (DEA "Economie Mathématique et Econométrie" and courses of DEEQA)

since 2003: CDSE, University of Mannheim, Mannheim, Germany (Doctoral studies in Economics)