Chapter 1

Introduction

The economic outcomes realized by a society are a function of the institutions put in place, the incentives they create, and the behavior of agents in the face of those incentives. In situations where a social planner or government agency has both a notion of desirable outcomes and an ability to put in place certain institutions, it is imperative that the planner understand the interactions between these three elements. Economic theory has provided a solid foundation for this understanding, but it is necessarily constrained by a need for tractability. As a consequence, economic models make various specific assumptions about both the ability of the planner to select institutions and the response of agents to the incentives those institutions create. In order to move this theoretical research toward the domain of application, it is necessary to understand the robustness and the realism of these assumptions.

In economic environments where goods can be made excludable and where the agent that consumes a good is the only one who receives benefit, it is well-known that under very general assumptions, establishing and maintaining property rights is sufficient to guarantee that the selfish private exchange of goods can lead to efficient

allocations for the entire society.[1] In such situations, the benevolent social planner need not consider complex mechanisms for the achievement of socially efficient outcomes; enforcing the rights of individuals to own and trade property is sufficient. This theoretical result is particularly appealing because its assumptions about individual behavior and information are minimal and it does not require that the planner actively engage in the reallocation of property. Although the theory still grapples with exactly how prices form and how economies adjust from their initial state to an efficient allocation, it is clear that under the simplest of behavioral assumptions, agents will still arrive at these optimal outcomes.

Unfortunately, such desirable results do not obtain when the consumption and production choices of one agent have an impact on the welfare of another. The provision of a pure public good is a particularly stark example where all agents necessarily consume a single good, so the production of one agent necessarily impacts the benefits of others. Property rights alone are then insufficient for the realization of efficient outcomes because each agent has an incentive to let the others fund a given level of the public good. Samuelson [90] formally demonstrates how individual incentives lead agents to select allocations different from the socially efficient allocation so that generically, an efficient allocation cannot be an equilibrium state of the economy. In such situations, more complex institutions are needed to realize the desired outcomes. Samuelson conjectures that *no* decentralized institution can be effective when agents have the ability to misrepresent their preferences. Fueled by debates about

[1]It will be assumed throughout that the social planner's goal is efficiency, as formalized by the notion of Pareto optimality. A similar methodology could be used to study institutions, incentives, and behavior under alternative desiderata and many of the results herein would still apply.

the viability of a socialist system, various authors have worked to identify and study decentralized institutions and to determine whether or not Samuelson's conjecture is universally true.

Early research focused on models of economic *adjustment* that are guaranteed to move the economy to a welfare-maximizing state, without concern for individuals' incentives to reveal their preferences truthfully.[2] An adjustment process is said to be Pareto satisfactory if it selects a unique Pareto optimal outcome and if every Pareto optimal outcome can be attained from some initial state. Thus, the classic welfare theorems for private goods economies prove that the perfectly competitive mechanism is Pareto satisfactory. At this point, authors were primarily concerned with the amount of information that a Pareto satisfactory mechanism must acquire from its agents in order to operate under the assumption that agents would always be willing to provide *truthfully* such information when asked of them.

As the mechanism design literature progressed, authors became increasingly concerned with the assumption of truthful revelation.[3] The additional requirement of incentive compatibility was imposed on mechanisms, stating that the vector of informational messages sent by each agent be a Nash equilibrium strategy profile of the 'message-sending game'. This led to the key negative result of the literature: there does not exist an incentive compatible, Pareto satisfactory mechanism where truthful revelation of one's preferences is an equilibrium behavior. Groves & Ledyard [43] demonstrated that this impossibility result could be circumvented by using abstract

[2]See Hurwicz [48] or Hurwicz [50] for a review of this development.

[3]This concern about incentives dates back at least as far as Samuelson [90].

message spaces (rather than asking for agents' preference profiles directly) and assuming that agents will always select Nash equilibrium messages of the one-shot game defined by the mechanism. Indeed, every Nash equilibrium message profile of this proposed mechanism maps to a Pareto optimal outcome of the economy without any *a priori* information about the characteristics of the economy. Maskin [67] then provided a general theorem characterizing the desiderata (including Pareto optimality, for example) that can successfully be implemented in this way. These results show that when the Nash equilibrium assumption is used, Samuelson's concerns about the incentives to misrepresent one's preferences can be successfully avoided.

The first step in taking these theoretical solutions to real world problems is determining the conditions under which agents will behave according to the Nash equilibrium prediction. If a mechanism is proposed whose Nash equilibrium outcomes are guaranteed to be Pareto optimal, it is of use in real world settings only if Nash equilibrium is an accurate predictor of behavior. Previous experimental research on behavior in games clearly indicates that the Nash equilibrium concept is highly predictive in some situations and terribly inaccurate in others. It is generally true that in market-like trading interactions, the selfish utility maximizing model performs quite well, while in situations where one's strategies directly and obviously affect others' payoffs, behavior often deviates from the selfish prediction in a way consistent with models of fairness, inequality aversion, or reputation building. It is therefore difficult to extrapolate from these observations a prediction about behavior in mechanisms for the provision of a public good, and it certainly appears plausible that the existence of

a public good in the economy will lead players to deviate from the Nash prediction.

A second necessary extension of the theoretical work is to consider situations where the mechanism is repeated periodically. Institutions are often long-lived and agents' behavior may not be identical through time. Players may learn more about the preferences of others. They may attempt to build reputations. They might use past information to predict what others will do in an upcoming iteration. They may expect that others will use current announcements to shape their play in later periods and will thus be motivated to deviate in the present to positively affect the future. Repetition of a one-shot game opens the door to a much larger set of equilibrium predictions and off-equilibrium motivations, raising questions about the validity of the original solution in such a setting. It may be that repetition will 'undo' certain mechanisms by generating unexpected inefficiencies through dynamic behavior, while improving the outcomes of some inefficient mechanisms through the selection of repeated game equilibria that Pareto dominate the one-shot prediction.

These two concerns are summarized by two empirical questions: Will agents play the Nash equilibrium prediction, and will repetition of the mechanism alter behavior? Chapters 2 and 4 experimentally consider these two questions in the domains of public goods provision and contracting under moral hazard, respectively. The first of these chapters identifies a particular 'best response' learning model of behavior that is fairly consistent with actual play of a repeated mechanism. This model not only provides insight about which particular mechanisms will converge to their equilibrium points, but also provides sufficient conditions for a mechanism to guarantee that equilibrium

outcomes will eventually obtain. The latter chapter examines a situation in which repeated game behavior may improve the outcomes of a mechanism through rational reputation-building in the presence of stereotyping behavior. Specifically, employees may select levels of effort that are socially optimal for long periods of time because they believe their employer will erroneously categorize them (and, through stereotyping, their co-workers) as 'reciprocal' players who irrationally respond to higher wages with increased effort. The employer, believing that its workers may be reciprocal, will attempt to motivate these workers with higher pay and in fact, will receive higher effort in response for all but the final period, even if all of the workers are in fact selfish-minded individuals. Thus, in an institution where selfish behavior is generally thought to cause highly inefficient outcomes, the repeated nature of the interaction can lead to significant welfare gains for all parties involved.

Finally, the standard model of mechanism design assumes that the social planner needs only to identify desirable allocations in order for such outcomes to be realized. This paradigm implicitly assumes that the planner is equipped with some ability to enact the chosen allocations through a credible and undesirable action, should some agents not comply with the allocation. This action may take the form of explicit penalties for deviating, such as fines or imprisonment, or even the seizure of the agents' endowments and the forceful reallocation of assets. It may also take the form of an alternative allocation that any deviating agent would prefer less than the one suggested. In the example of providing a public good, it may be credible for the planner to provide no public good if even one agent deviates from the chosen

allocation. Here, any allocation preferred by every agent to their initial endowment can be credibly enforced by this possibility. On the other extreme, if the planner has no credible outside option, agents can freely ignore the suggested allocation and voluntarily provide any level of the public good they prefer – an outcome that is generically different from optimality.

Chapter 3 examines an intermediate case where the central planner cannot credibly commit to cancel all production of the public good if a single agent deviates, but does have the ability to decline the contribution of an individual if that contribution is not consistent with the chosen allocation. In other words, if the central planner requests a certain amount of money from an agent to be put into production of the public good, and that agent responds by sending a different amount of money, the planner can credibly 'tear up the check' and produce only the level of public good that can be achieved with all others' contributions. In such a situation, the planner is unable to guarantee that there exists an allocation that is mutually agreeable by all agents, and therefore, must expect that there will be situations in which one or more agents will have an incentive to not comply with the proposed allocation. Even worse, it is shown that as an economy becomes large, it is guaranteed that at least one individual will always prefer noncompliance. Thus, the ability to identify desirable allocations through a mechanism is hardly sufficient for those allocations to obtain when the planner has this level of enforcement available.