A crash course in implementation theory

Matthew O. Jackson

Humanities and Social Sciences 228-77, California Institute of Technology, Pasadena, CA 91125, USA (e-mail: jacksonm@hss.caltech.edu)

Received: 26 March 2001/Accepted: 21 May 2001

Abstract. This paper is meant to familiarize the audience with some of the fundamental results in the theory of implementation and provide a quick progression to some open questions in the literature.

1 Introduction

There are many economic, social, and political situations where individuals interact to make decisions that affect them collectively. Examples range from voting to elect representatives or choose a public policy, to trading in a market. Based on their preferences over the possible outcomes of the interaction, individuals may act strategically in order to influence the outcome to their advantage. For instance, in an election a voter might vote for his or her second ranked candidate if the voter’s favorite candidate has no chance of being elected; or in an auction a buyer may select a bid considering trade-offs between the probability of winning the auction and the price to be paid. The specific design of the institution through which individuals interact, for instance the rules of the election or auction, can have a profound impact on the strategic behavior of the members of the society and on the outcomes of the process. Implementation theory is a study of the relationship between the structure of the institution through which individuals interact and the outcome of that interaction.

This paper grew out of lectures given at the NATO Advanced Study Institute on Game Theory and Resource Allocation: The Axiomatic Approach, which took place at SUNY Stony Brook in July of 1997. I thank the organizers and especially William Thomson for organizing the institute and the participants for feedback on the lectures. I thank Salvador Barbera for detailed suggestions an earlier version of this manuscript, and Maurice Salles and an anonymous referee for helpful comments.
Game theory plays a central role in the modeling of the strategic interaction studied in implementation theory. In many applications of game theory, the game modeling interaction is taken as given and analyzed to predict the actions of individuals and the resulting outcome. In implementation theory, instead of taking the game as given, it is something to be designed. Often, one thinks of the desired outcomes as the given and analyses whether there exist game forms for which the strategic properties induce individuals to (always) choose actions that lead to the desired outcomes. An example of an implementation question is: how can we design an auction to be sure that the individual who most highly values an object is sure to be the winner of the auction?¹ In this view, implementation theory is a normative branch and game theory is a positive branch of the same tree, and implementation theory is the design or reverse engineering process associated with game theory. Of course, this view is a bit caricatured, but indicates that there is a close relationship between the tools and understandings developed in implementation theory and game theory.

To get a feeling for the type of questions that are analyzed in implementation theory let us start by looking at a classic example. Consider a society or committee holding an election to select one out of a set of candidates. Each member of the society has a preference ranking over the candidates. The society may have certain aspirations regarding which candidate should be selected as a function of the preferences of the members of the society. For instance, it may wish to avoid selecting a candidate who is Pareto dominated by another candidate (i.e., a candidate ranked lower than another candidate by all members of society). It may also wish to select a Condorcet winner (a candidate who defeats any other candidate in a pairwise comparison according to a majority of voters’ preferences) if such a candidate exists. If these were the goals of the society then the implementation question would be, “Does there exist an election procedure for which for each possible profile of preference rankings of the voters, each equilibrium outcome of the election procedure would be Pareto optimal and Condorcet consistent?”

In order to answer this question one has to make precise what an election procedure is and what equilibrium outcomes are. This is the point at which game theoretic tools are used. The election procedure is modeled as a game form or what is commonly referred to as a mechanism in the literature. It specifies a set of possible actions or messages that each member of society can use, and then the outcome (in this case the candidate elected) as a function of the actions or messages sent by the members of society. For instance one could have each member of society submit their ranking of the candidates. If there are \( m \) candidates, then one could award a candidate \( m \) points for each voter whose submitted ranking places them highest, \( m - 1 \) points for each voter’s submitted ranking places them second highest, and so on.

¹ For an analysis of this particular problem from a mechanism design point of view, see Dasgupta and Maskin (1997). I will discuss the relationship and difference between mechanism design and implementation.