WELFARE ECONOMICS AND SOCIAL CHOICE THEORY 2ND EDITION

ALLAN M. FELDMAN
ROBERTO SERRANO

Springer

WELFARE ECONOMICS AND SOCIAL CHOICE THEORY, 2ND EDITION

WELFARE ECONOMICS AND SOCIAL CHOICE THEORY, 2ND EDITION

ALLAN M. FELDMAN

and

ROBERTO SERRANO

Brown University
USA

Springer

Library of Congress Control Number: 2005935332

ISBN:10: 0-387-29367-1
ISBN-13: 978-0387-29367-7
e-ISBN-10: 0-387-29368-X
e-ISBN-13: 978-0387-29368-4

Printed on acid-free paper.

Printed in the United States of America.

9 8 7 6 5 4 3 2 1

springeronline.com

Contents

Preface

This book covers the main topics of welfare economics — general equilibrium models of exchange and production, Pareto optimality, uncertainty, externalities and public goods — and some of the major topics of social choice theory — compensation criteria, fairness, voting, Arrow's Theorem, and the theory of implementation. The underlying question is this: "Is a particular economic or voting mechanism good or bad for society?" Welfare economics is mainly about whether the market mechanism is good or bad; social choice is largely about whether voting mechanisms, or other more abstract mechanisms, can improve upon the results of the market.

This second edition updates the material of the first, written by Allan Feldman. It incorporates new sections to existing first-edition chapters, and it includes several new ones. Chapters 4, 6, 11, 15 and 16 are new, added in this edition. The first edition of the book grew out of an undergraduate welfare economics course at Brown University. The book is intended for the undergraduate student who has some prior familiarity with microeconomics. However, the book is also useful for graduate students and professionals, economists and non-economists, who want an overview of welfare and social choice results unburdened by detail and mathematical complexity.

Welfare economics and social choice both probably suffer from excessively technical treatments in professional journals and monographs. One purpose of this book is to present these fields in a way that reduces mathematical technicalities. Knowledge of calculus and linear algebra is not a prerequisite for reading the book. However, the results are presented rigorously; there are theorems and proofs (occasionally relegated to appendices), and the reader should be familiar enough with logic to know when A implies B and when it doesn't. What mathematical notation is used is defined and explained as it is introduced. Each analytical

chapter contains examples to illustrate the concepts and the theorems, and most of the chapters provide exercises for the reader.

At the end of most chapters are lists of selected references. These references sections are not meant to be scholarly bibliographies; rather, they should be used by the reader who wants to dig a little deeper into a subject, or who wants to find out a little about the source of a possibly ancient idea.

We would like to thank all past students of welfare economics and social choice for helping us distill this material. We would like to thank the editors of *Economic Inquiry*, *The American Economic Review* and *SIAM Review*, in whose journals parts of Chapters 10, 13, 14 and 15 were previously published. Professor John Weymark made helpful comments on chapter drafts. Two of our graduate students helped: Debipriya Chatterjee prepared 16 new figures, and wrote the solutions to the exercises; Rene Saran helped with the exercises. Madeline Brewster typed much of the manuscript, patiently modified it, and prepared it for publication. Brown University provided support for preparation of this book, and Roberto Serrano gratefully acknowledges research support from Fundación Banco Herrero.

ALLAN FELDMAN AND ROBERTO SERRANO

Introduction

1. The Concerns of Welfare Economics

Welfare economics is the normative branch of economics: it is concerned with *what is good and what is bad*, rather than *what is*. This distinguishes it from the positive branches of economics, such as labor economics, health economics, economic history, financial economics, development and international trade, monetary economics and macroeconomics. Each of the positive branches is largely devoted to explaining why things are the way they are: e.g., why doctors are paid more than nurses, why some countries export agricultural commodities and some export technology, and why business fluctuations occur. Many positive economists make policy prescriptions; some say we ought to have no minimum wage, or we ought to have a higher minimum wage, or we ought to have lower tariffs, or we ought to have a 3 percent annual growth in the money supply. These prescriptions are all based partly on positive economics (the prescribers have information that indicates the likely consequences of actions that might be taken) and partly on normative economics (the prescribers have some ideas about which consequences are *good* and which are *bad*).

Welfare economics is not the Boy Scout Code of Honor of economics. It is obviously more than that: it is very analytical, very deductive, very much a collection of theorems, or statements of the form "If A, then B." In fact it might be viewed as a branch of symbolic logic, and you could strip it entirely of its normative content by simply saying it is a collection of theorems — all of which are logically true, but none of which is connected to everyday affairs or to values. None of the theorems says "The United States Government ought to do X,Y, and Z." but when welfare economics is viewed this way it's not as exciting as it can be. To

see why welfare economics is concerned with what is good or what isn't good, rather than whether or not A implies B, we turn to the contents of some of those theorems.

The first basic idea we use is the notion of a competitive economy. In the United States the economy is a complicated mixture of firms and people, tens of thousands of wheat and corn farmers, dozens of oil refiners, a handful of car manufacturers, thousands of actors and actresses, hundreds of thousands of builders and real estate developers,and hundreds of millions of consumers. Some of the markets for goods and services are extremely competitive, in the sense that no one person or firm has a significant influence on prices, such as the market for soybean meal; while some of the markets are extremely concentrated, in the sense that a few firms or people have a major influence on prices, such as the markets for various prescription drugs. The question of concentration and its effects is the subject of industrial organization and is beyond the scope of this book. We will assume for simplicity in most of what we do that people and firms take prices as given. Moreover, the properties of particular markets for particular goods, special problems like the state of the crude oil market in 2005, will not concern us here. We will deal instead with a model, an abstraction, a concept of a competitive economy, which we think does relatively little violence to the reality. To be specific, a competitive economy for us is a system of profit-maximizing firms and rational, or utility-maximizing individuals. What can be said about such a system? In particular, is there anything normative to say about it? Can we say it's good or bad?

The question leads to our next basic notion, the idea of Pareto optimality. According to the dictionary, the word *optimal* means most favorable or best. But we have a special idea of optimality, which can be illustrated with two examples of economic situations that are not Pareto optimal. First, suppose that three girls can earn \$5, \$5 and \$7 respectively (that is, \$5 for girl 1, \$5 for girl 2, and \$7 for girl 3) by doing a certain chore X for their mother. Suppose that there is a different chore Y which could earn them \$6, \$6 and \$7, respectively. Also, suppose that the chores are equally onerous, and suppose that their mother doesn't care whether they do X (and she pays \$17) or they do Y (and she pays \$19). If all of this is true, we can say that X is not Pareto optimal because there exists an alternative Y which could make some people better off than X (girls 1 and 2) and no one worse off.

Second, suppose three men own three pickup trucks. Man 1 has a Ford, man 2 has a Chevy, and man 3 has a Dodge. Suppose also that man 1 prefers Fords to Chevys to Dodges; that man 2 prefers Fords to Dodges to Chevys; and man 3 prefers Fords to Chevys to Dodges. Also,

suppose they are free to swap, if they want to. Then this distribution of trucks among the three is not Pareto optimal because there is an alternative distribution, in which man 1 still has his Ford, man 2 has the Dodge, and man 3 has the Chevy, that makes some people better off (men 2 and 3) and no one worse off.

In short, a situation is not Pareto optimal if there is an alternative situation available that makes some people better off and no one worse off. A situation, a state of affairs, or an allocation in an economy *is* Pareto optimal if there is *no* alternative available that makes some people better off and no one worse off.

With the idea of a competitive economy and of Pareto optimality in hand, we can turn to the first basic theorem of welfare economics: It says that, barring information, externality and public goods problems, a competitive economy guarantees a Pareto optimal economic outcome. And the normative idea here is that a competitive economy is "good," in the sense that it produces an optimal outcome. Any economic system that didn't produce an optimal outcome would be obviously "bad," because there would exist an alternative which some people would prefer and everyone would like as well.

The reader might note that our notion of Pareto optimality is a weak notion of goodness because it is very inclusive. It is certainly the case that if something is not optimal than it ought to be overruled, but just because something is optimal doesn't mean it's really best. For instance, in our three-girl example, if they were paid \$1, \$1 and \$17 respectively for chore Y, Y might well be optimal, although awfully hard on girls 1 and 2. Or, in the three-man example, there are many distributions of the pickup trucks among the men that are optimal, such as the distribution that assigns all three pickup trucks to man 1. But this is awful for men 2 and 3. So the theorem that says a competitive economy guarantees a Pareto optimal outcome is fine so far as it goes — but it might not go far enough. There are too many Pareto optima, some of them palatable and some not.

This observation, that the market mechanisms might produce a good (Pareto optimal) result, but not the very best result, motivates the second basic theorem of welfare economics. Suppose someone or something concludes that of all the Pareto optimal distributions of goods possible in an economy, distribution X is the very best, the ideal of the optimal. The second basic theorem says that, with minor modifications involving transfers of cash among various people, the competitive mechanism can be used to reach X. That is, X can be achieved via the interplay of profit-maximizing firms and/or utility-maximizing individuals. Consequently,

it is unnecessary to have a huge bureaucracy to decide who gets what in the economy.

In short, the first theorem says that a competitive market economy produces a Pareto optimal result (and that's good), and the second says that any Pareto optimal distribution of goods can be achieved via a slightly modified competitive mechanism. These two theorems link together the competitive market mechanism and Pareto optimality. This is the essence of classical welfare economics.

We will examine other allocative mechanisms -such as raw power in the "jungle"- which produce Pareto optimal distributions. But those mechanisms, unlike the competitive market, are involuntary; they disregard individual property rights. The market mechanism is special because it produces optimal outcomes *and* it is voluntary.

2. The Concerns of Social Choice Theory

Where do we draw the line between welfare economics and social choice theory? The two basic theorems of welfare economics link competition and optimality. But neither one answers these questions: How should we choose among Pareto optimal situations? How do we distinguish among the good? Or, in general, under what circumstances is it reasonable to say that alternative A is better for society than alternative B? The most important results in welfare economics indicate that competitive market mechanisms are good in the sense that they are Pareto optimal. The most important results in social choice theory are connected with attempts to answer this general question: When is A socially better than B?

Most economists assume that people have preferences. When faced with any two alternatives A and B a person can say whether he prefers A to B, or likes them equally well. Is it conceivable that a society — a collection of individuals — has social preferences that are similar to an individual's preferences? If it makes sense to say that society does have preferences, then there is a way to judge among the many Pareto optimal allocations, to find the very best from among the good. And there is a way to decide whether alternative A is better for society than alternative B. In short, there is a way to answer the questions left unanswered by the theorems that link the competitive market mechanisms and Pareto optimality. Social choice theory is, therefore, largely about social preferences, and how they might be found.

Social preferences might be defined by a charismatic religious leader (like Ayatollah Khomeni in Iran in 1979), or a terrorist (like Osama bin Laden in 2001), or a charismatic political leader (like Adolf Hitler or Mao Tse-Tung). This is clearly an important possibility, but in this book we

are interested in the theory of democratic institutions. Social preferences might be defined by "experts," by "leaders," by "opinion makers," or by movie stars or rappers. This is again a serious possibility that we won't take very seriously. Social preferences might be determined by economic criteria, by market criteria or by things like cost-benefit analysis. This possibility we do take seriously. Perhaps A should be considered better for society than B if A represents more dollars than B. Social preferences might be determined by egalitarian criteria, by fairness criteria, or by notions like Rawls' maximin criterion. Perhaps A should be considered better for society than B if A creates more equality than B. Again, this is an important possibility, and we do analyze it here.

But we have not yet touched on the most interesting way to discover social preferences, most interesting at least for citizens of a democratic society. We have noted that the competitive market mechanisms and the connected idea of Pareto optimality tell us something about what's good and what's not, but don't distinguish very fully among the good. The major alternative to the market mechanisms, at least for us, is the political process, and the political process ultimately depends on voting mechanisms. Voting mechanisms are in fact widely used to judge among good, (or Pareto optimal) alternatives, and voting processes are extensively used to decide when alternative A is socially better than alternative B. Voting is an exceptionally important way to define or reveal social preferences, and we shall analyze it carefully here. (Incidentally, another way to distinguish between welfare economics and social choice theory is to say that for the most part the former is about market mechanisms and the latter is about voting mechanisms, while both are about what's good and what's not good.)

There are many ways to try to find social preferences — market-related procedures and voting procedures being most important for our purposes. There are many ways to approach the problem: Is A socially better than B?

What then are the basic results of social choice theory? What general statements can we make about whether alternative A is better than alternative B? Does social choice theory provide a rule that always answers that question satisfactorily?

First, we should note that there are positive results in social choice theory — results that say things like "majority voting has properties X,Y, and Z, or plurality voting has properties W, U, and V." For instance, a theorem of Duncan Black says that if people's preferences satisfy a certain requirement, then majority voting produces social preferences that are as reasonable and as rational as any person's preferences. And if this is the case, it's easy to determine (at least conceptually)

when A is better for society than B, to determine which of the Pareto optimal (or good) situations are really best for society. Majority voting becomes a source of sensible social preferences, and can be used as the ultimate test of what's better for society than what.

But it has been known, at least since the late eighteenth century, that majority voting has one very serious logical flaw. If A, B, and C are three alternatives, it is possible that majority voting will indicate that (1) A is socially better than B, (2) B is socially better than C, and (3) C is socially better than A. That is, there might be a voting paradox, a voting cycle, and the social preferences derived by majority vote might be irrational. For instance, a country's legislature might be contemplating these alternatives: (A) to increase expenditures on national defense, leaving expenditures on other budget items constant; (b) to increase expenditures on social programs, leaving expenditures on other budget items constant; or (C) to decrease expenditures on all budget items. It is quite possible that the legislature might conclude, through majority voting, that (1) it is better to increase defense expenditures than social expenditures, (2) it is better to increase social expenditures than to decrease all expenditures, and (3) it is better to decrease all expenditures than to increase defense expenditures. But this is irrational, and leaves no acceptable course of action!

Also, as we shall see, there are serious objections that can be made against other procedures for determining when A is socially better than B. For instance, many of the market-related procedures, the procedures that provide the theoretical foundation for important everyday tests like cost-benefit tests, reduce in the end to a question of Pareto optimality vs. non-Pareto optimality, to statements like "A is socially better than B if A is Pareto optimal and B isn't." This surely isn't very helpful, because the goal is to find a way to judge among the many Pareto optima.

In fact, much of the analysis of social choice theory produces negative conclusions, conclusions of the type: Procedure X for determining when A is socially better than B has such-and-such a nasty characteristic. And all of these negative conclusions are drawn together in one important negative theorem , the most important single result of social choice theory, the Impossibility Theorem of Kenneth Arrow.

What does it say? The essence of Arrow's result is that any procedure for finding social preferences — whether it is based on the market, on voting, on expert opinion, or whatever — any procedure must have some serious flaw. So the search for a procedure to find social preferences, the search for a general answer to the question "when is A better for society than B," has to be inconclusive.

Arrow's theorem is about the problem of transforming individual preferences into social preferences. It takes as given that a Government, or what we'll call a "Central Authority," can discover what the individual preferences really *are*. A whole new set of problems arise when we acknowledge that the Central Authority might receive incomplete, or misleading, or false information about those preferences. What happens if the Central Authority relies on individuals to report their preferences, and individuals deliberately report false preferences to some how manipulate the result? Or, what happens if individuals know more about the alternatives than the Central Authority does, and attempt to exploit that superior information? Or, what happens if different individuals know different things about the characteristics of goods, or about what may happen in the future, or about people's preferences, and try to use that kind of information? (When different individuals know different things, we say there is "asymmetric information," and asymmetric information destroys the links between markets and optimality.)

Questions like these will be treated at length in this book. And here we have a mixture of theoretical answers, some negative, and some positive. For example, a theorem of Allan Gibbard and Mark Satterthwaite says that if individuals report their preferences to a Central Authority, and, based on those preferences, The Authority chooses the best alternative for society, then the individuals in society will normally be tempted to lie, so as to manipulate the outcome. This is a negative conclusion, much like Arrow's result. On the other hand, a theorem of Eric Maskin shows that under certain conditions, it is possible for a Central Authority to device a mechanism, "played" by the individuals, which has the property that if everybody but person i is telling the truth about his preferences, person i will want to tell the truth also. That is, truth telling is an equilibrium of the "game," the game has no other equilibrium outcome, and it is so designed that it yields the best outcome for society.

3. Practical Concerns of Welfare Economics and Social Choice Theory

So far we've indicated that welfare economics tells what is good (the Pareto optimal situations) and what isn't (the non-optimal situations), and that, barring information, externality and public goods problems, there is a crucial link between the competitive market mechanism and Pareto optimality. We've also indicated that general answers to the question of "Is A socially better than B?" are hard to come by. These are both fairly abstract results, but there are practical welfare economics and social choice issues that we haven't touched: What happens to that crucial link between competition and optimality when there is asym-

metric information, when there are externalities, when there are public goods?

In the presence of asymmetric information, say about the quality of the good being traded, the link between market allocations and Pareto optimality is destroyed. In cases like these, prices may not be good enough signals to separate high quality from low quality, or they may provide "too much" information making insurance contracts impossible.

When one person's decision affect another person's welfare in a way that is not reflected by market prices, we have an externality. For instance, suppose person A smokes cigarettes in the same room with person B, a nonsmoker. Then A's decisions affect B's welfare, they impose a cost upon B that A does not pay. And in a case like this, the market doesn't provide the proper signals to A, and the resulting situation is not Pareto optimal. The link between the competitive market and optimality is also destroyed.

What is a public good? It is a good whose consumption is nonexclusive: when A has it, so must B. For instance, a public park is a public good: if it is there for one person to enjoy, it's there for everyone to enjoy. National defense is a public good: if the armed forces are protecting the person and property of A, then they are doing so for A's compatriot B. A police and judicial system is a public good: if there are police officers and judges and jails which protect A from mayhem, theft and fraud, then they are also protecting B from mayhem, theft and fraud. Now when there are public goods present, the private market mechanism, in which consumers are buying goods to maximize their own utilities, again doesn't provide people with the right signals. People could (and do) hire armed guards. But who would have his own court and jail system, if it weren't provided by the state? Who would have his own army? Who would buy his own Yellowstone Park? When there are public goods like these around, private markets won't accomplish the right results, and the link between the competitive market mechanism and optimality is again destroyed.

What then should be done to establish a Pareto optimal outcome when there are informational asymmetries, externalities or public goods on the scene? These are the practical problems of welfare economics.

There are several solutions explored in this book. To solve externality problems, there are Pigouvian taxes and subsidies, which internalize the externalities. For instance, if the cigarette smoker is causing \$.10 per cigarette worth of damage to others with each cigarette he smokes, tax him \$.10 per cigarette. This will give him the right signal, and optimality will be reestablished.

To solve the problem of producing and financing a public good, there are several tax expenditure systems available. In all these systems, individuals pay taxes to a government, which decides on how much of the public good ought to be produced, and pays for its production out of tax revenues. In the Wicksell-Lindahl system, a person's marginal benefit from public expenditure is linked to the tax he pays. This system leads to optimal expenditures on the public good — providing people don't try to avoid taxes by misrepresenting their desires for the public good. In the demand revealing system, people have no incentives to misrepresent their demands for the public good. We shall discuss these tax-expenditure systems at some length, as well as the Groves-Ledyard system and a majority voting tax system.

Finally, the practical difficulties caused by information voids are addressed by the theory of implementation. According to this approach, rather than relying simply on market or voting institutions, the Central Authority creatively designs economic or social mechanisms in which persons interact. If mechanisms are designed cleverly enough, information gaps and asymmetries are surmounted. Those who have information end up revealing it, because the mechanism creates incentives for them to do so.

In short, welfare economics establishes a link between competitive markets and optimality. So competitive markets are good. Social choice theory points out all the pitfalls in the answer to the general question: "When is A socially better than B?" And the practical sides of welfare economics and social choice patch up some gaps between the theory of perfect private markets, and the reality of informational, externality and public goods problems.

Chapter 1

PREFERENCES AND UTILITY

1. Fundamental Assumptions

We suppose that there is a set of states, or alternatives, or bundles of goods, or "things" in the world. At various times we'll use various symbols to denote those things, but for now, we use the letters x, y, z, ... Later on we will be more explicit about the nature of our set of things.

The first fundamental assumption that we make about people is that they know that they like: they know their preferences among the set of things. If a person is given a choice between x and y, he can say (one and only one sentence is true):

1. He prefers x to y
2. He prefers y to x
3. He is indifferent between the two.

This is the axiom of *completeness*. It seems reasonable enough.

But some objections could be made to it. For a variety of reasons, a reasonable person might not be able to choose. If you are given the choice between shooting your dog and shooting your cat, you will balk. If your don't know what x and y really are; if, for example, both are complicated machines like cars and you don't know much about them, you may be unwilling to choose. If you are used to having your choices made for you; if you are dependent on your parents, your doctor, your religious guide, your government, you may be incapable of making choices yourself. Moreover, it may be painful, time consuming, distasteful, and nerve-wracking to make choices, and we will more or less ignore these

costs of decision making. In spite of these objections, we make the assumption.

The second fundamental assumption is the axiom of *transitivity.* The assumption has four parts:

1. If a person prefers x to y and prefers y to z, then he prefers x to z.
2. If a person prefers x to y, and is indifferent between y and z, then he prefers x to z.
3. If a person is indifferent between x and y and prefers y to z, then he prefers x to z.
4. If a person is indifferent between x and y, and is indifferent between y and z, then he is indifferent between x and z.

There are several possible objections to the transitivity assumption. Parts (1), (2), and (3) may simply not be true for some people under some circumstances. It might be the case that you prefer apple to cherry pie, and cherry to peach pie, while you prefer peach to apple. In fact, experiments with real subjects sometimes do reveal intransitivities of this sort, although when they are brought to the subjects' attention, they typically change their minds. Part (4) is the least realistic, since it can be applied repeatedly to get nonsense results: Let x_1 be a cup of coffee with one grain of sugar in it; let x_2 be a cup of coffee with two grains of sugar in it; and so on. Now it's almost certainly the case that you can't taste the difference between x_k ad x_{k+1}, for any whole number k, and so you must be indifferent between them. Therefore, by repeated applications of (4), you must be indifferent between x_0 and $x_{1,000,000}$, which is probably false. The problem here is evidently the existence of psychological thresholds. It can be escaped by assuming those thresholds away, or by assuming away the existence of finely divisible states of the world.

It is possible for some purposes to do without parts (2)-(4) of the transitivity assumption, in which case we say preferences are *quasi-transitive.* And quasi-transitivity itself can be further weakened, by assuming:

> If a person prefers x_1 to x_2, and prefers x_2 to x_3, ..., and prefers x_{k-1} to x_k, then he does *not* prefer x_k to x_1.

If preferences satisfy this assumption we say they are *acyclic.* In most of what follows, however, we assume all of transitivity for individuals' preferences.

The third and last fundamental assumption is that people always choose an alternative which is preferred or indifferent to every alternative

available to them. They choose "best" alternatives for themselves. In short, they are rational.

2. Best Alternatives and Utility Functions

In the middle and late nineteenth century it was popular in some philosophical circles to assume that pleasure and pain could be numerically measured. The measurement was in terms of *utils* or *utility units*, which were considered as scientifically real as units of length, mass, or temperature. Now a unit of length is scientifically real for several reasons: first, there is a standard object which everyone (at least everyone outside the U.S.) agrees represents one unit (e.g., a platinum rod in a vault in Paris); second, there is a natural zero for length; third, units of length can be added, subtracted, and multiplied by numbers according to the rules of arithmetic, and the results make sense: 2 meters + 2 meters = 4 meters.

Some of the nineteenth century advocates of utility calculus thought utility could be standardized and measured, like length; they thought the units could be used to measure everyone's happiness; they thought there was a natural zero between pleasure and pain; and they thought units of utility could be added and subtracted in a reasonable way.

But no one has yet succeeded in defining an objective unit of utility. Is it a level of electrical activity somewhere in the brain? Is it an index constructed from pulse, blood pressure, glandular activity data? Is it a rate of salivation, a degree of pupil dilation, or perspiration? We don't know. There is no way of comparing levels of satisfaction among different people. For that matter, there is no objective way of measuring utility at two different times for the same individual. This remains so despite the interesting developments in experimental psychology and neuroeconomics, although future research in these fields may shed important light on these issues.

But there is a subjective way: Ask him. (If you don't believe what a person says, you might choose instead to observe him. See what he chooses when he has what opportunities. If he chooses x when he might have chosen y, he reveals his preference for x.)

The problem with asking about utility is this. If you ask "How many units of happiness would you now get if I gave you a banana?" you will be laughed at. The question must be more subtly put. Ask instead, "Would you prefer a banana to an apple?" This is our fundamental question.

Asking "Would you prefer x to y" will never get you a measure of utility with well defined units, a zero, and other nice mathematical properties. But it will allow you to find alternatives that are at least as good

as all others, and, remarkably, it will allow you to construct a numerical measure to reflect tastes. The determination of best alternatives and the construction of a measure of satisfaction are both made possible by the completeness and transitivity assumptions on preferences. Therefore, the theory of preferences, with those two assumptions, is connected to, and is a generalization of, the old-fashioned nineteenth century theory of utility.

3. The Formal Model of Preferences

Before we can proceed, we need to introduce some notation. Let x and y be two alternatives. We consider a group of people who are numbered 1, 2, 3, and so on. To symbolize the preferences of the i^{th} person we write xR_iy for "i thinks x is at least as good as y"; xP_iy for "i prefers x to y"; and xI_iy for "i is indifferent between x and y."

The relation R_i should be viewed as the logical primitive, the "given." The relations P_i and I_i can be derived from R_i with these definitions:

$$xP_iy \text{ if } xR_iy \text{ and not } yR_ix$$

$$xI_iy \text{ if } xR_iy \text{ and } yR_ix$$

In words: Person i prefers x to y if he thinks x is at least as good as y but he does not think y is at least as good as x. And i is indifferent between x and y if he thinks x is at least as good as y and he thinks y is at least as good as x.

Now our fundamental axioms of completeness and transitivity are formally put this way:

> *Completeness.* For any pair of alternatives x and y, either xR_iy or yR_ix.

> *Transitivity.* For any three alternatives x,y, and z, if xR_iy and yR_iz, then xR_iz.

Notice that these definitions are in terms of the primary relation R_i, rather than in terms of the derived relations P_i and I_i. The verbal definitions in the section above were in terms of P_i and I_i. The reader can check that the verbal and the formal definitions are in fact logically equivalent. That is, if R_i is transitive in the sense that, for all x, y, and z, xR_iy and yR_iz implies xR_iz, then the following must also be true:

1. xP_iy and yP_iz implies xP_iz. (See Proposition 1 below.)

2. xP_iy and yI_iz implies xP_iz.

3. xI_iy and yP_iz implies xP_iz.

4. xI_iy and yI_iz implies xI_iz.

The less fundamental (and weaker) assumptions of quasi-transitivity and acyclicity are formally put this way:

Quasi-transitivity. For any three alternatives x, y, and z, if xP_iy and yP_iz, then xP_iz.

Acyclicity. For any list of alternatives x_1, $x_2, \ldots, x_k$, if $x_1P_ix_2$, $x_2P_ix_3, \ldots$, and $x_{k-1}P_ix_k$, then not $x_kP_ix_1$.

Let us now prove that if a preference relation R_i is transitive, it must be quasi-transitive, and if it is quasi-transitive, it must be acyclic:

Proposition 1. If R_i is transitive, then it is quasi-transitive. If R_i is quasi-transitive, then it is acyclic.

Proof. Suppose first that R_i is transitive. We want to show it is quasitransitive. Suppose xP_iy and yP_iz. We need to show xP_iz, that is, xR_iz and not zR_ix. Now xP_iy means xR_iy and not yR_ix and yP_iz means yR_iz and not zR_iy. Since xR_iy and yR_iz, xR_iz follows by R_i's transitivity. If zR_ix were also true, then we would have zR_ix, xR_iy and, by R_i's transitivity, zR_iy, which contradicts not zR_iy. Consequently, zR_ix cannot be true; that is, not zR_ix. But xR_iz and not zR_ix means xP_iz, and R_i is quasi-transitive.

Next suppose R_i is quasi-transitive. We want to show it is acyclic. Suppose $x_1P_ix_2, x_2P_ix_3 \ldots, x_{k-1}P_ix_k$. We need to show not $x_kP_ix_1$. Since $x_1P_ix_2$ and $x_2P_ix_3$, $x_1P_ix_3$ by quasi-transitivity. Similarly, since $x_1P_ix_3$ and $x_3P_ix_4$, $x_1P_ix_4$ by quasi-transitivity. Repeated applications of this argument gives $x_1P_ix_k$, and not $x_kP_ix_1$ follows immediately. Q.E.D

We have already noted that preferences can be quasi-transitive without being transitive: the grains-of-sugar-in-coffee example shows this. Preferences can also be acyclic without being quasi-transitive or transitive. Suppose someone likes apples (A) better than bananas (B), and bananas better than cherries (C), but is indifferent between apples and cherries. Then his preferences relation is AP_iB, BP_iC, and AI_iC. This doesn't violate acyclicity since there is no preferences cycle. (If CP_iA holds, there is a cycle.) But the preference relation is not quasi-transitive, since quasi-transitivity would require AP_iC.

With the necessary tools in hand, we proceed to define what is meant by an individual's "best" choices. Suppose S is some collection of alternatives. Let x be an element of S. Then x is said to be best for person i if i thinks it is at least as good as every other element of S.

Formally, i's *best set in* S or i's *choice set in* S, denoted $C(R_i, S)$, is defined as follows:

$$C(R_i, S) = \{x \text{ in } S | x \, R_i y \text{ for all } y \text{ in } S\}.$$

This is read: "$C(R_i, S)$ is the set of all x's in S, such that xR_iy for all y's in S." (Note that braces { } means "the set" and a slash | means "such that.")

Now to the next result. Proposition 2 answers the question "When can we be sure best things exist?" One answer is: Whenever a preference relation (defined on a finite set) is complete and transitive.

Proposition 2. Let S be a finite set of alternatives available to person i. Suppose R_i is complete and transitive. The $C(R_i, S)$ is nonempty. That is, best choices exist.

Proof. Choose one alternative, say x_1, from S. If it is best, we are done. If not, there is an alternative, say x_2, for which

$$x_1 R_i x_2$$

does not hold. By completeness $x_2 R_i x_1$ must hold, and therefore, by definition

$$x_2 P_i x_1.$$

If x_2 is best, we are done. If not, we can choose an x_3 such that

$$x_3 P_i x_2$$

by the same argument as above.

This process can either terminate at a best choice (in which case we are done), or it can go on indefinitely. Since S has only a finite number of elements, if the choice process goes on forever, it must repeat. Therefore, there must be a cycle:

$$x_1 P_i x_k P_i x_{k-1} P_i \ldots x_3 P_i x_2 P_i x_1.$$

Repeated applications of the transitivity assumption implies $x_k P_i x_1$. But this contradicts $x_1 P_i x_k$. Hence,the process cannot continue indefinitely and the choice set is nonempty. Q.E.D.

But Proposition 2 could clearly be strengthened by substituting the assumption of quasi-transitivity, or of acyclicity, for our fundamental assumption of transitivity, since the key to the proof is the possible existence of a cycle in the individual's preferences. In fact, the following proposition is also true. The proof is virtually the same as for Proposition 2, and is left to the reader.

Proposition 3. Let S be a finite set of alternatives available to person i. Suppose R_i is complete and acyclic. Then $C(R_i, S)$ is nonempty. That is, best choices exist.

Proposition 3 can itself be strengthened to more clearly indicate the connection between the existence of best or choice sets, and acyclicity of the preference relation. The following proposition says that when R_i is complete, best sets are always nonempty if *and only if* R_i is acyclic:

Proposition 4. Suppose R_i is complete. Then $C(R_i, S)$ is nonempty for every finite set of alternatives S available to person i, if and only if R_i is acyclic.

Proof: The "if" part of the proof follows from Proposition 3. To prove the "only if" part, we assume $C(R_i, S)$ is nonempty for every finite set of alternatives S. We want to show R_i is acyclic.

Suppose to the contrary that R_i is not acyclic. Then there exist alternatives $x_1, x_2 \ldots, x_k$ such that $x_1 P_i x_2, x_2 P_i x_3, \ldots, x_{k-1} P_i x_k$, and $x_k P_i x_1$. Let $S = \{x_1, x_2, x_3, \ldots, x_k\}$. Then $C(R_i, S)$ is empty, since every alternative in S is inferior to some other alternative in S. But this is a contradiction. Consequently R_i must be acyclic. Q.E.D.

The propositions above answer this question: Given particular assumptions about a person's preferences, can he always identify best alternatives? The next proposition answers a different question: Is there a numerical function, a utility function, which represents a person's preferences? If the answer is yes, then familiar mathematical tools can be applied to the problem of identifying best alternatives, since the search for a best alternative reduces to the problem of maximizing a utility function. If the answer is no, the use of utility functions, indifference curves, and all the other common tools of economics, is very likely illegitimate.

It turns out that the answer is yes if preferences are complete and transitive. (And in this case, acyclicity cannot substitute for transitivity.)

Proposition 5. Let S be a finite set of alternatives available to person i. Suppose R_i is complete and transitive.

Then we can assign numerical values $u_i(x), u_i(y), u_i(z)$, etc., to the alternatives in S so that

$$u_i(x) \geq u_i(y) \text{ and only if } xR_iy.$$

In other words, there is a utility function u_i, which places values on the alternatives that exactly reflect i's preferences. The proof is in the appendix to this chapter.

We should note that u_i could be transformed without altering its preference representation property. For instance, if we define $v_i = u_i + C$, where C is any constant, then $v_i(x) \geq v_i(y)$ if and only if $u_i(x) \geq u_i(y)$, if and only if xR_iy. Therefore, v_i represents R_i as well as u_i does. And if $u_i(x) \geq 0$ for all x's, u_i^2 would represent R_i as well as u_i. In fact, any transformation of u_i that does not change relative values leaves the representation property intact. These are called monotone transformations. If a utility function represents a person's preferences, any monotone transformation of that utility function is another utility function that represents the same preferences.

For this reason, u_i is called an ordinal utility function and, unlike the hypothesized utility functions of nineteenth century philosophers, it does not behave like a cardinal measure such as length: For our utility function, there exist no standard units, there are no natural zeros, and it makes no sense to add $u_i(x)$ to $u_i(y)$. Nor does it make any sense to add $u_i(x) + u_j(y)$, if u_j is another person's utility function.

What then is the use of an ordinal utility function? In fact, it transmits exactly the same information as the preference relation it represents: neither more, nor less. But a utility function allows us to analyze, in a compact and easy way, the behavior of an individual in an economic environment. It is quite correct to say that a consumer chooses a bundle of goods to maximize his utility, and the utility approach is mathematically and graphically convenient. It allows us to use the standard tools of the economist's trade.

To be able to represent preferences by means of utility functions, Proposition 5 has dealt with the case of finite sets of alternatives. However, in many applications in this book an individual will be choosing from infinite sets of alternatives. For example, a consumer will choose bundles of goods where the amount of each good is measured by a real number. For such settings, if one wishes to represent preferences by a utility function the assumption of *continuous preferences* is important. Intuitively, continuity means that the preference relation has "no jumps." Here's the definition.

Continuity. For any bundle of goods x, the upper contour set of R_i at x and the lower contour set of R_i at x are closed, i.e., they contain their boundaries. (The upper contour set of R_i at x is the set of bundles $\{y|yR_ix\}$. The lower contour set of R_i at x is the set of bundles $\{y|xR_iy\}$).

With the aid of continuity, Proposition 5 can be extended as follows:

Proposition 6. Let S be a (possibly infinite) set of bundles of goods. Suppose R_i is complete, transitive and continuous over S. Then there exists a utility function u_i defined on S which exactly reflects i's preference relation R_i.

This proposition will be used extensively in the following chapters.

4. Decisions under Uncertainty and Expected Utility

In this section we present an important special case of decision theory. It concerns problems involving *uncertainty.* Uncertainty has come to be viewed in recent decades as an important factor in many economic decisions. For example, an individual making investment decisions is uncertain about the returns he will obtain. A sports team making players' hiring decisions does not know for sure how these hires will translate into victories. The government of a country, when implementing a policy change, may not know exactly its consequences for society. For these cases and many more, the decision makers are facing a problem in which uncertainty and risk are essential components. It turns out that the theory developed for these decision problems has a very interesting mathematical structure, which we shall outline in this section.

Suppose that the set of pure alternatives (i.e., those not involving uncertainty) is $\{x_1, \ldots, x_k\}$. Each of these pure alternatives could be anything, but for simplicity and to fix ideas, let's think of each of them as a prize, a different amount of money that the individual could win. Thus, for example, the individual could end up with a prize of $x_1 = \$0$, $x_2 = \$10$ or $x_3 = \$100$.

Let $l = (q_1, \ldots, q_k)$ be a *lottery* over the pure alternatives. That is, l is a probability distribution, whereby alternative x_j occurs with probability q_j. Of course, $q_j \geq 0$ for $j = 1, \ldots, k$ and $\sum_{j=1}^{k} q_j = 1$. Continuing with the example of three monetary prizes, one could think of several lotteries: lottery $l_1 = (0.5, 0, 0.5)$ is a fair coin toss that pays \$100 if heads, and nothing if tails. Lottery $l_2 = (1/3, 1/3, 1/3)$ is a fair die toss that pays \$0 if faces 1 or 2 turn up, \$10 if 3 or 4 do, and \$100 if 5

or 6 do. Lottery $l_3 = (0, 1, 0)$ is also a lottery, but it is called a *degenerate lottery*, because it pays one of the prizes for sure (in this case, \$10).

Suppose that now the individual is asked to choose among the lotteries. Which should he choose? Note that two rational individuals may choose differently. For instance, presented with the choice between l_1 and l_3, one individual may choose l_3 because he is afraid of the high probability (one half) of getting nothing in l_1, while another person may choose l_1 because its expected prize (weighted average of prizes) is so much higher than that in l_3.

In any event, since individuals will be making decisions involving uncertainty, we model these situations as individuals choosing over the set of possible lotteries. Therefore, we assume that individuals have preferences over lotteries.

Given a set of pure alternatives $\{x_1, \ldots, x_k\}$, the set of lotteries over it is the set of all possible probability distributions. This is called the *probability simplex*:

$$\{(q_1, \ldots, q_k) | q_j \geq 0 \text{ for all } j, \sum_{j=1}^{k} q_j = 1\}.$$

The preference relation R_i over the probability simplex describes the preferences of the decision maker. The statement "$l_1 R_i l_2$" is read "lottery l_1 is at least as good as lottery l_2 according to person i." The preference relation R_i is used to define both the strict preference relation P_i and the indifference relation I_i, as before.

We shall assume that person i's preference relation R_i over the set of lotteries satisfies completeness, transitivity and continuity. Before we proceed, it is worth noting an important property of the set of lotteries: for any pair of lotteries l_1 and l_2 and any nonnegative constant α no greater than 1 ($\alpha \in [0, 1]$), the convex combination of the two lotteries, that is, $[\alpha l_1 + (1 - \alpha) l_2]$, is also a lottery. This is interpreted as first playing a lottery over lotteries, leading to l_1 with probability α and to l_2 with probability $1 - \alpha$, and then playing either l_1 or l_2, depending on which was chosen in the first stage. We refer to this property as the *linearity* of the set of lotteries.

Because of linearity, the assumption of continuity of preferences reduces to the following simple form:

> *Continuity.* For any three lotteries l_1 , l_2 and l_3, if $l_1 P_i l_2 P_i l_3$, there exists a number $\alpha \in (0, 1)$ such that $[\alpha l_1 + (1 - \alpha) l_3] I_i l_2$.

That is, if an individual has a strict ranking among three lotteries, so that he judges one "best" among the three, the second one "in the

middle" and the third one "worst," continuity of preferences means that there must be a way to combine the best and the worst lotteries to get something that is indifferent to the one that was judged in the middle. Preference jumps are excluded.

Finally, we shall require another assumption on preferences over lotteries, also driven by the linearity of this set.

> *Independence.* For any lotteries l_1, l_2 and l_3, $l_1 R_i l_2$ if and only if $[\alpha l_1 + (1-\alpha) l_3] R_i [\alpha l_2 + (1-\alpha) l_3]$ for every number $\alpha \in [0,1]$.

Although one can construct violations of the independence assumption, its content is very intuitive. Suppose an individual judges lottery l_1 at least as good as l_2. Then, this preference should persist, should be independent, of mixing these lotteries with the same third lottery: if the choices now are that: (a) with probability α lottery l_1 will be played, and lottery l_3 will happen with probability $1-\alpha$, or (b) with probability α lottery l_2 will be played, and lottery l_3 will happen with probability $1-\alpha$, the same individual should prefer (a) over (b) or be indifferent between the two. This is simply because with probability α he is facing the choice between l_1 and l_2 (and $l_1 R_i l_2$), while with the rest of probability he is offered the same thing, i.e., l_3.

These assumptions characterize the so-called *von Neumann-Morgenstern* or *expected utility* preferences. The four axioms on preferences over lotteries lead to the von Neumann-Morgenstern expected utility theorem, named after the great mathematician and physicist John von Neumann and the economist Oskar Morgenstern:

> *von Neumann-Morgenstern Expected Utility Theorem.* The preference relation R_i over lotteries satisfies completeness, transitivity, continuity and independence if and only if it can be represented by a function that has the expected utility form. That is, there exist numbers $u_1, \ldots, u_k$ such that for any pair of lotteries $l = (q_1, \ldots, q_k)$ and $l' = (q'_1, \ldots, q'_k)$, $l R_i l'$ if and only if $\sum_{j=1}^{k} q_j u_j \geq \sum_{j=1}^{k} q'_j u_j$.
>
> *Proof*: It is easy to see that, if preferences are representable by a utility function that has the expected utility form, those preferences must satisfy the four axioms required.
>
> For the other direction, we provide a graphic proof for the case of three pure alternatives x_1, x_2 and x_3, which correspond to the degenerate lotteries l_1, l_2 and l_3, respectively. We deal with the nontrivial case in which the individual has a strict preference among these three. Let's say that $l_1 P_i l_3 P_i l_2$. The probability simplex is depicted in Figure 1.1.

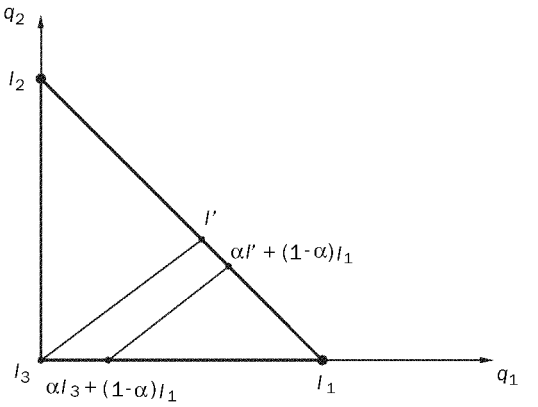


Figure 1.1.

A point in this triangle represents a lottery over the three pure alternatives (which are the degenerate lotteries l_1, l_2 and l_3). Note how the coordinates (q_1, q_2) of any point, measured from the usual origin, tell us the probabilities that the given lottery assigns to the best and to the worst alternatives (obviously, the probability that this lottery assigns to the middle alternative l_3 is simply $1 - q_1 - q_2$).

Now, completeness, transitivity and continuity of R_i guarantee the existence of a utility function representing those preferences (Proposition 6). Given such a utility function u, let $u_1 = u(l_1)$, $u_2 = u(l_2)$ and $u_3 = u(l_3)$, with $u_1 > u_3 > u_2$. What we shall show now is that this function is linear in probabilities: for any lottery $l = (q_1, q_2, q_3)$, the utility of lottery l is $u(l) = q_1 u_1 + q_2 u_2 + q_3 u_3$.

Since $l_1 P_i l_3 P_i l_2$, by continuity, there exists $\alpha \in (0, 1)$ such that $l' = [\alpha l_1 + (1-\alpha) l_2]$ is indifferent to l_3, i.e., $l' I_i l_3$, which implies that these two lotteries, l' and l_3, lie on the same *indifference curve* (a locus of points among which the individual is indifferent). Furthermore, by independence, one has that for any $\alpha \in [0, 1]$:

$$l_3 = [\alpha l_3 + (1-\alpha) l_3] I_i [\alpha l' + (1-\alpha) l_3],$$

which implies that the indifference curve passing through l' and l_3 is a straight line (recall that the locus of points that are convex combinations of two extreme points is the straight line segment connecting them). See Figure 1.1.

Finally, also from independence, since $l' I_i l_3$, one also has that for any $\alpha \in [0, 1]$, $[\alpha l' + (1-\alpha) l_1] I_i [\alpha l_3 + (1-\alpha) l_1]$, and applying the previous step, we construct a new indifference curve for each value of α that

is parallel to the one through l' and l_3. Next, taking combinations of l' and l_3 with l_2, one concludes that the indifference map is one of parallel straight lines. This corresponds to a function that is linear in probabilities. See Figure 1.1 again. Q.E.D.

Thus, in the problems involving uncertainty that we shall cover, we shall assume that agents have von Neumann-Morgenstern or expected utility preferences.

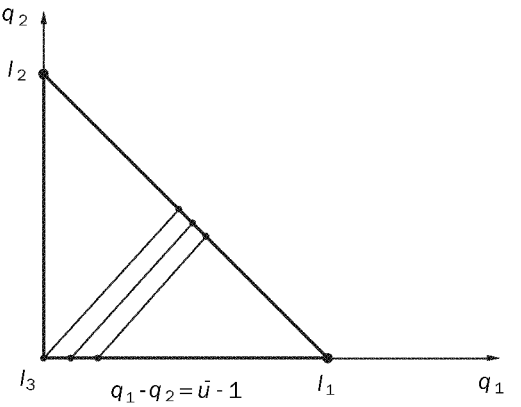


Figure 1.2.

As an illustration, Figures 1.2 and 1.3 depict two different preferences over the probability simplex, where the three degenerate lotteries are l_1, l_2, and l_3. In Figure 1.2, let the corresponding utilities $u_i(l_1) = 2$, $u_i(l_3) = 1$ and $u_i(l_2) = 0$ according to preferences R_i. For these preferences, the indifference curve of level $\bar{u}$ is the locus of points in the simplex whose equation is $2q_1 + (1 - q_1 - q_2) = \bar{u}$ or $q_1 - q_2 = \bar{u} - 1$. Not surprisingly, the top ranked point in the simplex is the degenerate lottery l_1, while the worst lottery is l_2. Figure 1.3 shows an indifference map with different expected utility preferences over lotteries. In it, $u_i'(l_1) = 4$, $u_i'(l_3) = 3$ and $u_i'(l_2) = 0$, and we call these preferences R_i'. For them, the indifference curve of level $\bar{u}$ has the equation $4q_1 + 3(1 - q_1 - q_2) = \bar{u}$ or $q_1 - 3q_2 = \bar{u} - 3$.

Note that, despite the fact that u_i' is a *monotone transformation* of u_i, both utility functions do not represent the same preferences over lotteries. This is true because, to preserve the expected utility feature, preferences can be represented only by functions that are *positive affine transformations* of one another. That is, if u_i and u_i' are two expected utility functions representing the same preferences over lotteries,

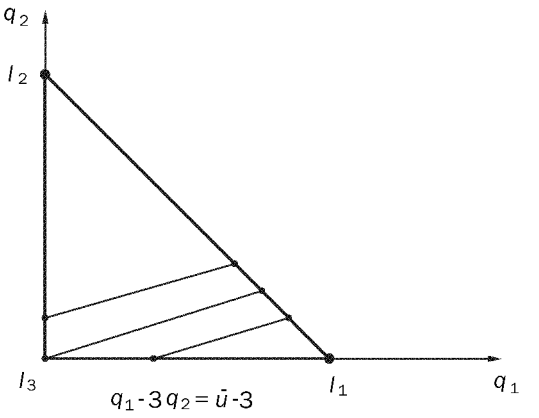


Figure 1.3.

there must exist a positive constant α and another constant β such that $u_i'(l_j) = \alpha u_i(l_j) + \beta$ for each degenerate lottery l_j.

To see that the preferences depicted in Figures 1.2 and 1.3 differ, note that the indifference curves have different slopes, and so the indifference maps are not the same. More clearly, let's exhibit two lotteries l_1 and l_2 such that l_1 is preferred to l_2 according to preferences R_i ($l_1 P_i l_2$), while l_2 is preferred to l_1 according to R_i' ($l_2 P_i' l_1$). Such lotteries could be, for example, $l_1 = (1/3, 1/3, 1/3)$ and $l_2 = (0.1, 0.7, 0.2)$. Indeed, for this pair of lotteries, $u_i(l_1) = 1 > 0.9 = u_i(l_2)$, but $u_i'(l_1) = 7/3 < 2.5 = u_i'(l_2)$.

5. Introduction to Social Preferences

Interest in quasi-transitivity and acyclicity arises largely from the analysis of social preferences, rather than of individual preferences. It is hard to imagine, for instance, that a person could have preferences which are acyclic but not quasi-transitive. But society's preferences are not, as we shall explain at length in later chapters, nearly so sensible as a person's.

A few examples will clarify the idea of social preferences, and the possibilities of nontransitivities for them. Suppose a group is making choices between alternatives, by using some voting rule. If x defeats y in a vote, let us say x is socially preferred to y, which we now write xPy. If x and y tie, let us say x and y are socially indifferent, which we now write xIy. If x is socially preferred to y or socially indifferent to y, we now write xRy. Where we had R_i, P_i, and I_i for individual i's preference, strict preference, and indifference relations, we now have

R, P, and I for society's preference, strict preference, and indifference relations.

Let us be more specific about the voting rules. Assume for simplicity that there are only three people in the group that is making the choices, and assume there are only three alternatives, x, y, and z.

Our first example is an instance of Condorcet's voting paradox, to which we shall return in Chapter 9 below. The voting rule is simple majority rule: a vote is taken between a pair of alternatives, and if alternative A gets more votes than alternative B, then A wins. Suppose the individuals' preferences are as follows: Person 1 prefers x to y to z. Person 2 prefers y to z to x. Person 3 prefers z to x to y. Each individual has sensible transitive preferences, but they evidently disagree on the relative merits of the three alternatives. We can indicate these preferences diagrammatically by listing the alternatives from top to bottom in the order of each person's preferences:

$\underline{1}$	$\underline{2}$	$\underline{3}$
x	y	z
y	z	x
z	x	y.

Consider a vote between x and y. Evidently, if the individuals vote according to their preferences, which we assume they do, person 1 votes for x; person 2 votes for y; and person 3 votes for x. Consequently, xPy. Next, consider a vote between y and z. Now person 1 votes for y; person 2 votes for y; and person 3 votes for z. Consequently, yPz. Finally, consider a vote between x and z. Now person 1 votes for x; person 2 votes for z; and person 3 votes for z. Consequently, zPx. We have a cycle here, since xPy, yPz, and zPx. These social preferences are not even acyclic.

The moral is social preferences might be very odd indeed — they need not share the sensible rational qualities of individual preferences. What about best sets in this example? We do have $C(R, \{x, y\}) = \{x\}$: x is best if the choice is limited to x and y. Similarly, $C(R, \{y, z\}) = \{y\}$, and $C(R, \{x, z\}) = \{z\}$. But R has a cycle. So Proposition 4 warns us that there is some set of available alternatives S for which $C(R, S)$ is empty. And, in fact, $C(R, \{x, y, z\})$ is empty: if all three alternatives are available, none is best according to majority rule. Each alternative is worse than one of the others.

Now we turn to a slightly different example. Suppose the people, alternatives, and preferences are as above, but the majority rule mechanism is modified as follows: A vote is taken between a pair of alternatives,

and if alternative A gets more votes than alternative B, then A wins — unless person 1 prefers B to A. If 1 prefers B to A, and A wins a majority over B, then A and B are declared tied, or socially indifferent. We call this rule simple majority rule with a vetoer. Person 1 has a veto, in the sense that he can prevent any alternative from actually beating another alternative he prefers. What are the voting results for this rule? Consider a vote between x and y. Alternative x gets two votes to one for y, and person 1, who prefers x anyway, does not exercise his veto. Consequently, xPy. Next, consider a vote between y and z. Alternative y gets two votes to one for z, and person 1 again does not exercise his veto. Consequently, yPz. Finally, consider a vote between x and z. Alternative z gets two votes to one for x, but now person 1 does exercise his veto, since he prefers x to z. Consequently, xIz. In sum, xPy, yPz and xIz. These social preferences are acyclic, although they are not quasi-transitive. Since they are acyclic, Proposition 4 tells us that best sets are always nonempty. In fact, $C(R\{x, y, z\}) = \{x\}$ in this case; the alternative x is socially best. (It is no accident, of course, that x is also person 1's favorite.)

For the third example, we again continue with the people, alternatives and preferences above, but majority rule is now discarded. The new rule is an oligarchy of persons 1 and 2, and it works like this: A is socially preferred to B if and only if both persons 1 and 2 prefer A to B. Otherwise, A and B are socially indifferent. Now consider a "vote" between x and y. Person 1 prefers x to y, but 2 prefers y to x. Consequently, xIy. Next, consider a vote between y and z. Person 1 prefers y to z and person 2 prefers y to z. Consequently, yPz. Finally, consider a vote between x and z. Person 1 prefers x to z but person 2 prefers z to x. Consequently, xIz. In sum, xIy, yPz, and xIz. Here there are no cycles, so the social preference relation is acyclic. Moreover, the definition of quasi-transitivity is (vacuously) satisfied. (It would not be satisfied if xPy and yPz, and xIz, as in the former example.) But the social preference relation is not transitive, because transitivity requires that if xIy and yPz, then xPz must follow. So this is an example of a quasi-transitive, but not transitive, social preference relation. Note that $C(R, \{x, y, z\}) = \{x, y\}$, the favorite alternatives of the two oligarchs.

The next examples are not hypothetical as the three preceding ones. They were first discussed, in the 1970s, by Donald Brown:

We now consider two voting rules used by the United Nations Security Council. The first was in force prior to August 31, 1965. At that time there were five permanent and six nonpermanent members of the Security Council. To be passed, a motion needed seven affirmative votes, and the concurrence of all five permanent members. That is, each perma-

nent member had to vote aye on a motion, or to abstain, or that motion would be defeated. Each permanent member had a veto. Now assuming that each nation's Ambassador had transitive (i.e., sensible) preferences, the procedure could not cycle. To see this, suppose there were a series of motions, or amendments to motions, or amendments to amendments, such that x_1 defeated x_2, x_2 defeated x_3, x_3 defeated $x_4, \ldots,$ and x_{k-1} defeated x_k. Since x_1 defeated x_2, x_1 got seven affirmative votes from the eleven members of the Council. Consequently, one of the permanent members must have voted affirmatively for x_1 over x_2. Say the United States voted affirmatively for x_1. Then the United States presumably preferred x_1 to x_2. Now x_2 was passed over x_3. Consequently, x_2 had seven affirmative votes over x_3, and the concurrence of all five permanent members. That means every permanent member either preferred x_2 to x_3, or was indifferent between the two In particular, the United States either preferred x_2 to x_3, or was indifferent between the two. Similar reasoning shows the United States either preferred x_n to x_{n+1}, or was indifferent between the two, for $n = 3, 4, \ldots, k-1$. Consequently, by repeated applications of transitivity, the United States preferred x_1 to x_k. Therefore, the United States would have used its veto power to prevent x_k's winning over x_1: so x_k could not possibly defeat x_1. A cycle could not occur: the voting rule was acyclic. From Proposition 4 we know that no matter what set of alternatives was available, the voting procedure would sensibly identify at least one best alternative.

The second United Nations Security Council voting rule was put in force on September 1, 1965. At that time, the nonpermanent membership of the Council was increased from six to ten. The permanent membership remained at five. To be passed, a motion now needs nine affirmative votes, and the concurrence of all five permanent members. (This rule remains in effect in 2005.) This procedure can cycle. To see this, we construct an example. There are ten alternatives, labeled $x_1, x_2, \ldots, x_{10}$. Assume for the sake of argument that the five permanent members are all indifferent about all these alternatives: None feels strongly enough about any of the alternatives to veto it. Assume that the preferences of the nonpermanent members are as follows: (Under member 1, we list the alternatives, from top to bottom, in that Ambassador's order of preference; similarly for 2, 3, and so on.)

The table is formidable, but the analysis is perfectly simple: Consider a vote between x_1 and x_2. Everyone except the Ambassador from Country 10 prefers x_1 to x_2. (The permanent members are indifferent.) Consequently, x_1 defeats x_2. Consider a vote between x_2 and x_3. Everyone except the Ambassador from Country 9 prefers x_2 to x_3. (The permanent members are indifferent.) Consequently, x_2 defeats x_3. Sim-

1	2	3	4	5	6	7	8	9	10
x_1	x_{10}	x_9	x_8	x_7	x_6	x_5	x_4	x_3	x_2
x_2	x_1	x_{10}	x_9	x_8	x_7	x_6	x_5	x_4	x_3
x_3	x_2	x_1	x_{10}	x_9	x_8	x_7	x_6	x_5	x_4
x_4	x_3	x_2	x_1	x_{10}	x_9	x_8	x_7	x_6	x_5
x_5	x_4	x_3	x_2	x_1	x_{10}	x_9	x_8	x_7	x_6
x_6	x_5	x_4	x_3	x_2	x_1	x_{10}	x_9	x_8	x_7
x_7	x_6	x_5	x_4	x_3	x_2	x_1	x_{10}	x_9	x_8
x_8	x_7	x_6	x_5	x_4	x_3	x_2	x_1	x_{10}	x_9
x_9	x_8	x_7	x_6	x_5	x_4	x_3	x_2	x_1	x_{10}
x_{10}	x_9	x_8	x_7	x_6	x_5	x_4	x_3	x_2	x_1

ilarly, x_3 defeats x_4, x_4 defeats x_5, and so on, until x_9 defeats x_{10}. Now consider a vote between x_1 and x_{10}. Everyone except the Ambassador from Country 1 prefers x_{10} to x_1. (The permanent members are indifferent.) Consequently, x_{10} defeats x_1, and there is a voting cycle!

To briefly summarize the observations of this section, the question of transitivity for a preference ordering, which hardly arises for an individual's ordering, does arise with a vengeance for a social preference ordering. In our discussion of individuals, where it is comfortable to assume completeness and transitivity for preferences, we shall largely use Proposition 5 and the utility functions that proposition guarantees exist. But when we return to social preferences, we shall have to return to the concepts of this chapter, and pay careful attention to ideas like completeness, transitivity, and transitivity's weaker cousins, quasi-transitivity and acyclicity.

6. Exercises

1 Show that if a preference relation R_i is transitive in the sense that xR_iy and yR_iz implies xR_iz for all x, y, and z, then (i) xP_iy and yI_iz implies xP_iz, and (ii) xI_iy and yI_iz implies xI_iz.

2 Hockey team A defeats hockey team B. Hockey team B defeats hockey team C. Hockey team A ties hockey team C.

(a) Is this preference order complete? Is it transitive? Quasi-transitive? Acyclic?

(b) Can you identify a best hockey team?

(c) Can you construct a "quality" function u for hockey teams, with the property that $u(x) > u(y)$ if and only if x defeats y? Show with numbers why you can or cannot do this.

(d) Can you construct a pseudo quality function v for hockey teams, which only satisfies this property: if x defeats y then $v(x) > v(y)$?

3 Show that if preferences over lotteries satisfy independence, then for all lotteries l_1, l_2 and l_3, one has that $l_1 I_i l_2$ if and only if $[\alpha l_1 + (1 - \alpha)l_3] I_i [\alpha l_2 + (1 - \alpha)l_3]$ for every $\alpha \in [0, 1]$.

4 Show that if preferences over lotteries are representable by an expected utility function, they must satisfy completeness, transitivity, continuity and independence.

5 Suppose a committee has five rational members, and, for motion x to defeat motion y, x needs four affirmative votes out of the five.

(a) Show that if there are five alternatives available, there can be a voting cycle.

(b) Show that if there are only four alternatives available, there cannot be a voting cycle.

7. Appendix

Proof of Proposition 5. For notational convenience in this proof, we will drop the subscript i wherever it appears.

Suppose S is finite and R is complete and transitive. We want to show that there exists a utility function u such that

$u(x) \geq u(y)$ if and only if xRy.

First, we subdivide S into "indifference classes."

Let $C_1 = C(R, S)$. C_1 is nonempty by Proposition 2.

The alternatives in S which are not in C_1 we call $S - C_1$.

Let $C_2 = C(R, S - C_1)$. C_2 is nonempty by Proposition 2.

The alternatives in S which are not in C_1 or in C_2 we call $S - C_1 - C_2$.

Let $C_3 = C(R, S - C_1 - C_2)$. C_3 is nonempty by Proposition 2.

We continue in this fashion until we have exhausted S. This we must be able to do because S is finite. Let C_h be the last class so constructed.

$$\text{Now define } u(x) = \begin{cases} h \text{ if } x \text{ is in } C_1 \\ h - 1 \text{ if } x \text{ is in } C_2 \\ \cdot \\ \cdot \\ \cdot \\ 1 \text{ if } x \text{ is in } C_h \end{cases}$$

Next we show that $u(x) \geq u(y)$ implies xRy. Suppose $u(x) \geq u(y)$. Then x is in the same class as y, or in a class constructed before the class containing y. Let C_k be the class containing x. Then x is in $C(R, S - C_1 - C_2 - \ldots - C_{k-1})$ while y is in $S - C_1 - C_2 - \ldots - C_{k-1}$. Therefore, xRy.

Finally, we will establish that xRy implies $u(x) \geq u(y)$. We will argue that $u(x) < u(y)$ implies not xRy. Suppose $u(x) < u(y)$. Let C_k be the indifference class containing x, and C_j be the indifference class containing y.

Since $u(x) < u(y)$, x's class C_k was constructed after y's class C_j. Therefore, y is in $C(R, S - C_1 - \ldots - C_{j-1})$, x is in $S - C_1 - \ldots - C_{j-1}$, but x is not in $C(R, S - C_1 - \ldots - C_{j-1})$. Therefore, yRx and there is some alternative z in $S - C_1 - \ldots - C_{j-1}$ such that yRz (because y is in the best set $C(R, S - C_1 - \ldots - C_{j-1})$) but not xRz (because x is not). By completeness, if not xRz, then zPx.

Now by transitivity, if yRz and zPx, then yPx. Hence, not xRy, which is what we wanted to establish. Q.E.D.

8. Selected References

(Items marked with an asterisk (*) are mathematically difficult.)

1. K. Arrow, *Social Choice and Individual Values*, 2nd Edition, John Wiley and Sons, Inc., New York, 1963, Chapter II.

   This is an easy to read chapter of the classic monograph by Kenneth Arrow. It has short but useful observations on older literature. Arrow's notation and formalization of preferences and best or choice sets are the ones followed in this book.

2. D.J. Brown, "Aggregation of Preferences," *Quarterly Journal of Economics*, V. 89, 1975, pp. 456-469.

   This relatively nontechnical piece by Donald Brown is meant to introduce the nonspecialist to modern variants of Arrow's Impossibility Theorem. In these variants *oligarchies* and what Brown calls *collegial polities* take the place of dictators in Arrow's original theorem. Our example above of a majority rule mechanism with a vetoer is a Brown collegial polity. Our observations about the Security Council of the United Nations are taken from this Brown article.

*3. G. Debreu, *Theory of Value*, John Wiley and Sons, Inc., New York, 1959, Chapter 4.

   Chapter 4 of Gerard Debreu's classic monograph has a rigorous proof for existence of continuous utility functions. The mathematics is rather sophisticated.

*4. R.D. Luce, "Semiorders and a Theory of Utility Discrimination," *Econometrica*, V. 24, 1956, pp. 178-191.

This article deals with individual preferences that are complete, but not transitive. Instead of transitivity, quasi-transitivity, or acyclicity, Luce assumes the following intuitive property: It is possible to "string out all the elements of S [the set of available alternatives] on a line in such a fashion that an indifference interval never spans a preference interval." For such preferences, there is a theorem similar to, but slightly different than our Proposition 5.

5. A.K. Sen, *Collective Choice and Social Welfare*, Holden-Day, Inc., San Francisco, 1970, Chapter 1*.

Amartya Sen's book provides an extremely clear treatment of most topics in social choice. Chapter 1*, on preferences relations, is formal but not difficult. Sen's Lemma 1*1 is the original version of our Proposition 4 above.

6. J. von Neumann and O. Morgenstern *Theory of Games and Economic Behavior*, Princeton University Press, 1st edition, 1944.

In its first part, this fundamental book provides the model of decision making under uncertainty, and derives expected utility from assumptions on preferences over lotteries. The rest of the book studies problems involving more than one individual, in which the decisions made by each one influence the others. These problems are called games. Although one could cite some previous contributions, this book constitutes the formal birth of game theory.

Chapter 2

BARTER EXCHANGE

1. Introduction

In the previous chapter we analyzed preferences and developed, in Propositions 5 and 6, the connections between preferences and utility functions. In that chapter the set of alternatives was abstract; it might have been interpreted as virtually anything: The alternative might have been consumption bundles, political candidates, meals in a restaurant, careers. In fact, the alternatives might have been potential dates, or potential spouses. People have analyzed the system of choice of spouses in terms of completeness and transitivity of preferences, and so on.

In this chapter and several that follow, we shall focus on a particular set of alternatives, the alternatives of primary concern to a microeconomist. A microeconomist analyzes the economic behavior of consumers and producers of goods and services. Now a consumer is someone who buys, or gets in trade, quantities of goods, or things, like food, housing, cars, clothes, medical care, haircuts, and so on. He acquires these things because he wants them, because he likes them, because he needs them, or, in our language, because he prefers them to what he gives up in exchange for them, or because they give him a higher level of utility than the things he gives up in exchange. If A swaps B a 1999 Toyota and a dog for a 2000 Nissan, A does it because he prefers the 2000 Nissan to the 1999 Toyota/dog combination, or because the Nissan gives him more utility. B does it because the 1999 Toyota/dog combination gives him more utility.

This chapter will be about such swapping, or barter exchange. In order to analyze barter exchange, we will construct a model in which a group of people exchange bundles of goods among themselves. The

story goes roughly like this: Each trader starts with some given initial bundle of goods. A new distribution of goods, or allocation, is proposed by someone. The traders discuss the proposal among themselves. If some group of traders decides that it can do better on its own, with its given initial resources, it objects to the proposal, and the proposal is rescinded or blocked. Then a new proposal is made by someone. The process continues until an allocation is found to which no group of traders objects. The traders then accept that new distribution of goods.

This is basically the model of barter exchange developed by Francis Y. Edgeworth in 1881. It will hopefully provide us with certain insights about how an economic system of exchange works, or ought to work, and what a system of exchange might accomplish for the people who comprise it.

2. Allocations

The set of alternatives is now a set of distributions or allocations of goods in an economy.

Let's suppose that there are n people, numbered $1, 2, \ldots, n$. Usually a person is indexed with the letter i. Often we'll let $n = 2$, in which case we are talking about what happens when there are only two people (like Adam and Eve exchanging fruit in the Garden of Eden).

We assume there are m different goods. Typically we index a good with the letter j, so the goods are numbered $j = 1, 2, \ldots, m$. This is the entire list of goods. In some context, there might be only one ($m = 1$) or two ($m = 2$) goods. In others, there might be tens of thousands of goods.

A good is something that a consumer might want to consume. It might be apples, or fruit. But if one good is apples, another good cannot be fruit without producing boundless confusion. The goods are distinct, non-overlapping. Housing might obviously be a good. Happiness probably isn't a good, since you can't easily buy it or trade it. We don't like to think of a consumer going to a store to buy happiness. But services can be goods in our list. A shoeshine is a plausible good; a visit to a dentist is another.

We will assume for mathematical simplicity that every good is perfectly homogeneous and perfectly divisible, like water, gasoline, or natural gas. Obviously pianos aren't divisible. But the divisibility assumption is very convenient in economic analysis since it allows us to use continuity arguments, and, anyway, it does not produce misleading conclusions. That is, the results we derive using divisibility are, in the main, true when indivisibility is explicitly allowed. Besides, indivisibilities are sometimes a consequence of the units in which we measure consump-

tion. For example, instead of counting pianos or cars, we could count the hours a week that a consumer likes to play the piano or drive a car.

Now let's turn to some more notation. We let x_{ij} be person i's quantity of good j. The first subscript identifies the person, the second identifies the good. If good 3 is bananas, then x_{13} is the number of bananas for person 1. Obviously, x_{ij} must always be nonnegative: a person cannot possess a negative quality of a good.

We let x_i be person i's bundle or vector of goods. Thus $x_i = (x_{i1}, x_{i2}, \ldots, x_{im})$; so x_i shows m things: i's quantity of good 1, i's quantity of good 2, i's quantity of good 3, ..., i's quantity of good m. for example, if $m = 3$, good 1 is apples, good 2 is peaches, and good 3 is bananas, and if $x_1 = (8, 2, 0)$, then person 1 has 8 apples, 2 peaches, and no bananas.

Next, we define $x = (x_1, x_2, \ldots, x_n)$. Now x shows person 1's bundle of goods, person 2's bundle of goods, ..., person n's bundle of goods. It is a list of bundles of goods. To say the same thing, it is a vector of vectors.

In the theory of exchange there is no production; what goods are available are there in the beginning. We let ω_{ij} be person i's starting or initial quantity of good j. similarly, ω_i is his initial bundle, and ω is the initial list of bundles of goods. The total of good j available must be

$$\omega_{1j} + \omega_{2j} + \omega_{3j} + \ldots + \omega_{nj} = \sum_{i=1}^{n} \omega_{ij}.$$

The symbol

$$\text{“}\sum_{i=1}^{n} \omega_{ij}\text{”}$$

is short-hand for “summation of the ω_{ij}'s, where i ranges from 1 to n.”

An *allocation* is a list of bundles of goods x with totals of goods consistent with the totals initially available. This means

$$\sum_{i=1}^{n} x_{ij} \text{ must equal } \sum_{i=1}^{n} \omega_{ij}$$

for every j, that is, every good. We call the set of allocations a, and a is formally written as follows:

$$a = \{x | x_{ij} \geq \text{ for all } i, j \text{ and } \sum_{i=1}^{n} x_{ij} = \sum_{i=1}^{n} \omega_{ij} \text{ for all } j\}.$$

The set of alternatives in the theory of exchange is a.

One of the simplest examples of an exchange economy involves only two people and one good; so $n = 2$ and $m = 1$. Let the total quantity

of the good initially available, $\omega_{11} + \omega_{21}$, be equal to 1. Then a is the set of all pairs (x_{11}, x_{21}) such that $x_{11}, x_{21} \geq 0$ and $x_{11} + x_{21} = 1$. The set of allocations in this economy can be easily diagrammed. To picture a, draw a line segment one unit long. Choose a point x on the line segment, and let the distance from the lefthand end of the line to x represent person 1's quantity of the good x_{11}, and let the distance from the righthand end of the line x represent person 2's quantity of the good x_{21}. Now x_{11}, $x_{21} \geq 0$, and $x_{11} + x_{21} = 1$, so every such point represents an allocation, and, conversely, every allocation can be represented by such a point, or division of the line.

3. The Edgeworth Box Diagram

The most useful example of an exchange economy is one in which there are two people and two goods. This economy's set of allocations can be illustrated in an Edgeworth box diagram, which is constructed in the following manner:

Again suppose that the total quantity of each good available is one unit. Instead of a line segment one unit long, we draw a square, each of whose sides is one unit long. Good 1 will be measured horizontally, and good 2 will be measured vertically. A division of the horizontal side(s) of the box represents a division of the first good between the two people, that is the quantities x_{11} and x_{21}. A division of the vertical side(s) of the box represents a division of the second good between the two people, that is, the quantities x_{12} and x_{22}. Dividing the horizontal and vertical sides is equivalent to choosing a single point in the box, and therefore, any point in the Edgeworth box diagram represents an allocation, and conversely, any allocation is represented by a point in the Edgeworth box.

In Figure 2.1, person 1's bundle x_1 can be read off by measuring x's coordinates from the origin in the lower lefthand corner, and person 2's bundle x_2 can be read off by measuring x's coordinates from the upper righthand corner.

The next step is to represent the preferences of the two people. We will always assume that person i has a utility function $u_i(x)$, defined for any allocation x. A locus of allocations for which the function u_i is constant is called an *indifference curve*. This is analogous to an elevation line on topographical map, or an isobar on a weather map. In Figure 2.2 we have drawn a few indifference curves for person 1 (the solid lines) and for person 2 (the dashed lines). It is typically assumed that people's utilities rise as the quantities of goods they possess increase, so indifference curves further from person 1's origin (the lower lefthand corner) represent higher levels of u_1; and 2's indifference curves further from his

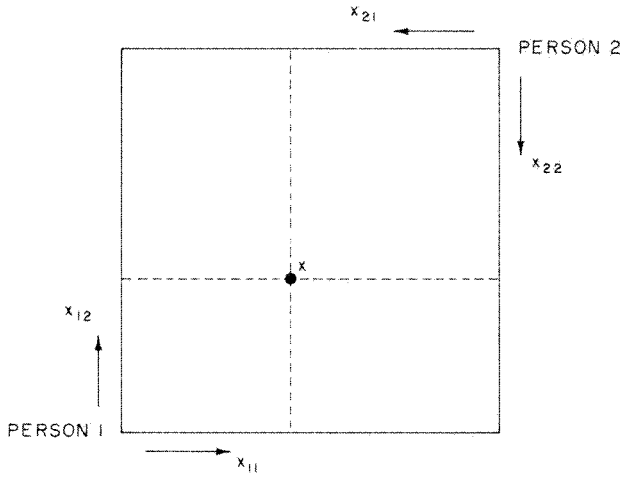


Figure 2.1.

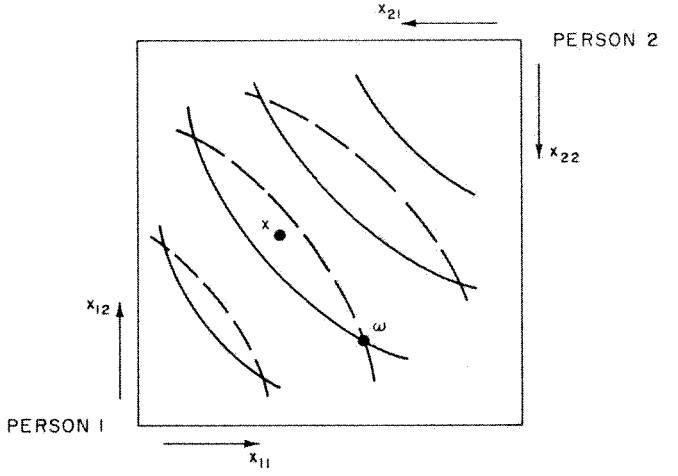


Figure 2.2.

origin (the upper righthand corner) represent higher levels of u_2. The point ω represents the initial allocation.

4. Pareto Optimal Allocations and the Core

Some allocations are unambiguously inferior to others. For example, in Figure 2.2 x is preferred by both people to the initial allocation ω.

Moreover, x is feasible for them: 1 and 2 can join together, trade, and reach x — given that they have ω in the beginning. If an arbitrary group of people can join together and improve upon a proposed allocation, with their initial resources, we say they block the proposal. This notion is formalized as follows:

First, individual i is said to be *self-interested* if u_i is a function of x_i alone: $u_i(x) = u_i(x_i)$. This means i is neither helped nor harmed by any other person's consumption of any good. Now let us assume that all individuals are self-interested.

Consider a group of traders, called S. S might be one person, it might be several persons, or it might be all. Let $\{s_i\}_{i \text{ in } S}$ represent a set of bundles of goods, one bundle for each member of S. We say that $\{s_i\}_{i \text{ in } S}$ is *feasible* for S if

$$\sum_{i \text{ in } S} s_{ij} = \sum_{i \text{ in } S} \omega_{ij} \text{ for every good } j.$$

The symbol

$$\text{“} \sum_{i \text{ in } S} s_{ij} \text{”}$$

is shorthand for “summation of the s_{ij}'s over the i's who belong to S.” A set of bundles is feasible for S if the totals of the goods in those bundles are consistent with the totals the members of S initially possess. If x is a proposed allocation and S is a coalition, we will say that S can *block* x, (or S can *block x from ω*) if there is a feasible set of bundles $\{s_i\}_{i \text{ in } S}$, such that:

$$u_i(s_i) \geq u_i(x_i) \text{ for all } i \text{ in } S$$
$$u_i(s_i) > u_i(x_i) \text{ for at least one } i \text{ in } S$$

The *core* (or the *core from ω*) is that set of allocations which cannot be blocked (from ω) by any coalition. Note that the definition of the core depends on our assumption of self-interestedness.

Now let us temporarily drop the assumption of self-interestedness. Consider a proposed allocation x, and, instead of an arbitrary coalition, the whole group of people in the economy. The whole group objects to x if there is a feasible alternative which is unambiguously better. A set of bundles $y = (y_1, \ldots, y_n)$ is feasible for the whole group if it is an allocation. We therefore make the following definition: An allocation x is not Pareto optimal if there is another allocation y such that

$$u_i(y) \geq u_i(x) \text{ for all } i = 1, 2, \ldots, n$$
$$u_i(y) > u_i(x) \text{ for at least one } i.$$

If there is no such alternative, x is a *Pareto optimal*, or *efficient* allocation.

When the self-interestedness assumption holds and the core is therefore well defined, Pareto optimality is implied by inclusion in the core: An allocation is in the core if no group, including the whole set of traders, blocks it. An allocation is Pareto optimal if the whole set of traders does not block it. In what follows we will assume self-interestedness unless we say otherwise.

Let's pause to consider the significance of the three crucial ideas here: blocking, the core, and Pareto optimality. All have been defined in the context of a simple exchange-economy model. But the ideas are clearly generalizable. The notion of Pareto optimality is especially ubiquitous. It can be used as a test of the adequacy of any arrangement that is meant to satisfy people's preferences, economic or not. For suppose a situation — economic, political, or whatever — has the property that it can be changed so as to make everyone as well off ($u_i(y) \geq u_i(x)$ for all i) and some people better off ($u_i(y) > u_i(x)$ for at least one i). Then the situation has the potential for unambiguous improvement. Since it has the potential for unambiguous improvement, it is unambiguously imperfect. The ideas of blocking and the core are most easily applied to an exchange economy with self-interested traders; so suppose an allocation has the property that some group of people can unambiguously improve upon it. Then the allocation is clearly objectionable for that group, and is therefore unambiguously imperfect. On the other hand, if an allocation is in the core, then no group can raise a clear objection to it.

Of course, a group's blocking or objecting to a proposed allocation depends on that group's initial resources. Consequently, the core depends on the initial distribution of goods or allocation ω. If ω makes A rich and B poor, that fact is reflected in the core. Whether or not a group can clearly object to a proposed allocation depends on what the group initially has.

The set of Pareto optimal allocations, on the other hand, does not depend on the initial distribution of goods among people, although it obviously does depend on the initial totals of the various goods.

Let's consider an illustration in an Edgeworth box diagram. When there are just two traders and two goods, and both traders have convex-shaped smooth indifference curves like the curves in Figure 2.3, the Pareto optimal allocations are the points, like x, y, and z, at which indifference curves of the two people are tangent. At these points it is impossible to make one party better off without hurting the other. Point x, however, is blocked by person 1, since $u_1(\omega_1) > u_1(x_1)$. Similarly, z

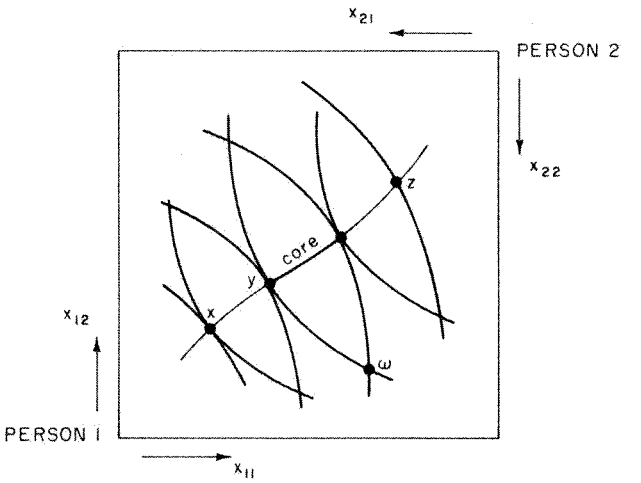


Figure 2.3.

is blocked by 2. The core is the locus of Pareto optimal points, such as y, lying on or within the lens shaped area bounded by the indifference curves passing through ω.

5. Algebraic Examples

It will be useful at this point to work through a few algebraic examples, to see how to calculate the Pareto optimal and core allocations in an exchange economy.

As a first example, suppose there are two people and two goods and suppose the people have these utility functions

$$u_1 = x_{11}x_{12}$$
$$u_2 = x_{21} + 2x_{22}.$$

In other words, person 1's utility level is the product of the quantities of the two goods he has, person 2's utility level is equal to the amount of good 1 he has plus twice the amount of good 2 he has. Let the initial allocation be $\omega_1 = (1/2, 1/2), \omega_2 = (1/2, 1/2)$. Each starts with $1/2$ unit of each good.

In this case, person 1's indifference curves are hyperbolic (because $x_{11}x_{12} =$ constant is the formula for a hyperbola) and person 2's indifference curves are straight lines. In order to proceed, we need to find expressions for the marginal rates of substitution or the absolute values of the slopes of the two people's indifference curves.

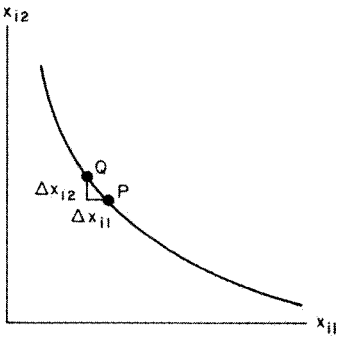


Figure 2.4.

An arbitrary indifference curve is drawn in Figure 2.4. The absolute value of the slope of the indifference curve at the point P is $|\Delta x_{i2}/\Delta x_{i1}|$. (The reader familiar with calculus can substitute d's for Δ's everywhere.) Now $|\Delta x_{i2}/\Delta_{i1}|$ is person i's *marginal rate of substitution* of good 2 for good 1, which we can abbreviate MRS for person i. To calculate an MRS it is most convenient to consider the marginal utilities of goods 1 and 2 for person i. The *marginal utility* of good 1 for i, which we abbreviate MU of good 1 for person i, is the rate at which his utility changes as his consumption of good 1 changes, or $\Delta u_i/\Delta x_{i1}$. The marginal utility of good 2 for i, or MU of good 2 for person i, is defined analogously. Now if person i loses Δx_{i1} units of good 1 his loss of utility is $\Delta u_i = \Delta x_{i1} \cdot$ MU of good 1 for i, by the definition of marginal utility. Similarly, if he acquires Δx_{i2} units of good 2, his gain in utility is $\Delta u_i = \Delta x_{i2} \cdot$ MU of good 2 for i, again by definition. If person i starts at the point P in Figure 2.4 and goes to the point Q, he loses Δx_{i1} units of good 1 and gains Δx_{i2} units of good 2. But he ends up on the same indifference curve, so the net change in his utility is zero. Consequently, the loss in utility equals the gain in utility, in absolute value, or

$$|\Delta x_{i1} \cdot \text{ MU of good 1 for } i| = |\Delta x_{i2} \cdot \text{MU of good 2 for } i|,$$

from which it follows that

$$\text{MRS for person } i = \left|\frac{\Delta x_{i2}}{\Delta x_{i1}}\right| = \frac{\text{MU of good 1 for } i}{\text{MU of good 2 for } i}.$$

We can use this formula in our example.

Person 1's utility function is $u_1 = x_{11}x_{12}$. The marginal utility of good 1 for him is the rate at which $x_{11}x_{12}$ grows as his consumption

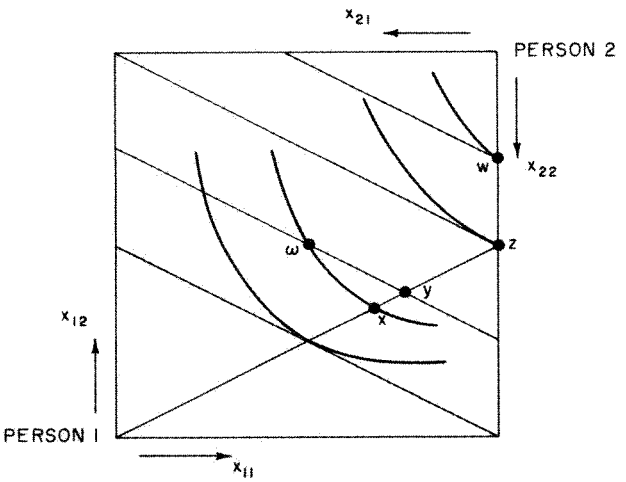


Figure 2.5.

of good 1, or x_{11}, grows. Now if x_{11} rises by Δx_{11} units, then u_1 rises by $\Delta x_{11} \cdot x_{12}$ units. Consequently, MU of good 2 for person 1 $= x_{12}$. Similarly, MU of good 2 for person 1 $= x_{11}$. Thus MRS for person 1 $= x_{12}/x_{11}$.

Person 2's utility function is $u_2 = x_{21} + 2x_{22}$. The marginal utility of good 1 for him is the rate at which $x_{21} + 2x_{22}$ grows as his consumption of good 1, or x_{21}, grows. Now if x_{21} rises by Δx_{21} units, then u_2 rises by $\Delta x_{21} \cdot 1$ units. Consequently, MU of good 1 for person 2 $= 1$. Similarly, MU of good 2 for person 2 $= 2$. Thus MRS for person 2 $= 1/2$. (Note that there are no approximation complexities in this example, that is , no Δx_{ij} terms in the expressions for marginal utilities. If such terms appear, it is necessary to take the limits of the expressions involved as the Δx_{ij}'s approach zero.)

We are ready at this point to illustrate our little exchange economy in an Edgeworth box diagram, in Figure 2.5. The box is one unit on each side. We have sketched in four (hyperbolic) indifference curves for person 1, and four (straight line) indifference curves (with |slope| $= 1/2$) for person 2. The two individuals's indifference curves are tangent where their slopes are equal, or where

MRS person 1 = MRS person 2.

This gives

$$\frac{x_{12}}{x_{11}} = \frac{1}{2}, \text{ or } x_{12} = \frac{1}{2}x_{11}.$$

Graphically, this is the straight line from person 1's origin to the point z. This straight line segment gives part of the set of Pareto optimal allocations, since any move from a point on it (like x or y) must make someone worse off.

But there are Pareto optimal allocations other than these tangency points. Consider, for instance, the point w. In order to make person 1 better off, starting at w, we would have to move above the hyperbolic indifference curve for person 1 going to w. This would make person 2 worse off. That is, there is no way to make everyone as well off and at least one better off. In fact, all the points on the righthand side of the box above point z are nontangency Pareto optimal allocations.

Where are the core allocations in this example? Any core allocation must also be Pareto optimal, so the core allocations lie somewhere on the lines we've already identified as Pareto optimal. But a core allocation must not be blocked by person 1, or by person 2. Person 1 would block any allocation that gives him less utility than ω_1, or any allocation below his hyperbolic indifference curve through ω. Similarly, person 2 would block any allocation that gives him less utility than ω_2, or any allocation below (with respect to his origin) his straight line indifference curves through ω. Consequently, the core is the locus of points on the straight line between points x and y, including the endpoints x and y.

In this example, then, we would expect barter exchange between persons 1 and 2 to move the economy from ω to a point on the line segment from x to y.

Our next example has three people and two goods. We suppose the three traders have the following utility functions and initial bundles:

$$\begin{array}{lll} \text{Person 1:} & u_1(x_1) = x_{11}x_{12} & \omega_1 = (1, 9) \\ \text{Person 2:} & u_2(x_2) = x_{21}x_{22} & \omega_2 = (5, 5) \\ \text{Person 3:} & u_3(x_3) = x_{31}x_{32} & \omega_3 = (9, 1) \end{array}$$

Instead of explicitly calculating the Pareto optimal and core allocations, let us simply consider one proposed allocation. Let x be given by:

$$\begin{aligned} x_1 &= (3, 3) \\ x_2 &= (5, 5) \\ x_3 &= (7, 7) \end{aligned}$$

Clearly, no single individual will want to block x, because $u_i(x_i)$ is greater than or equal to $u_i(\omega_i)$ in every case. Moreover, the allocation x is Pareto optimal: If you draw Edgeworth boxes for all the pairs of traders, you will find each pair's indifference curves just touch at the points which correspond to x. Yet x will be blocked. This is so because the move from ω to x is a bad deal for 1 and 2. If they acted on their own, and ignored 3, they could rearrange their resources as follows:

$$\begin{aligned} &\text{For 1:} \quad s_1 = (2, 5) \\ &\text{For 2:} \quad s_2 = (4, 9) \end{aligned}$$

This is a feasible set of bundles for them because

$$\begin{aligned} 2 + 4 &= 1 + 5 \, (= \omega_{11} + \omega_{21}) \\ 5 + 9 &= 9 + 5 \, (= \omega_{12} + \omega_{22}). \end{aligned}$$

Moreover, both traders would be better off with (s_1, s_2) than they would under the proposed allocation x, since $u_1(s_1) = 10 > 9 = u_1(x_1)$, and $u_2(s_2) = 36 > 25 = u_2(x_2)$. In short, we would not expect barter exchange to lead this economy to the allocation x.

Our final example has two parts. Consider first a two-person, two-good economy in which the traders have utility functions

$$\begin{aligned} u_1(x_1) &= x_{11}x_{12} \\ u_2(x_2) &= x_{21}x_{22} \end{aligned}$$

and in which the initial bundles are $\omega_1 = (9, 1)$, $\omega_2 = (1, 9)$. An Edgeworth box diagram for this economy would be ten units on each side, the initial allocation would be close to the lower righthand corner, and both people would have nice hyperbolic indifference curves. Reasoning like that used in the first example shows person 1's MU of good $1 = x_{12}$, person 1's MU of good $2 = x_{11}$, and consequently, the MRS for person $1 = x_{12}/x_{11}$. Similarly, MRS for person $2 = x_{22}/x_{21}$. Now consider the particular allocation x given by:

$$\begin{aligned} x_1 &= (3, 3) \\ x_2 &= (7, 7) \end{aligned}$$

At this allocation, MRS for person $1 = 3/3 = 1$; MRS for person $2 = 7/7 = 1$; so the indifference curves of the two are tangent, and it follows that x is a Pareto optimal allocation. Moreover, each person likes x at least as much as he likes ω:

$$\begin{aligned} u_1(x_1) &= 3 \cdot 3 = 9 \geq u_1(\omega_1) = 9 \cdot 1 = 9 \\ u_2(x_2) &= 7 \cdot 7 = 49 \geq u_2(\omega_2) = 1 \cdot 9 = 9 \end{aligned}$$

Therefore, neither individual would block x. (Obviously person 1 is getting the short end of the stick here, but he cannot block unless he is actually worse of than he is initially.) Since x is Pareto optimal and since x would not be blocked by either person, x is in the core of this economy.

Next let's "replicate" the economy by creating an identical twin for each of our traders. Person 1's twin is person 3. Person 2's twin is person 4. The utility functions for the two newcomers are:

$$u_3(x_3) = x_{31}x_{32}$$
$$u_4(x_4) = x_{41}x_{42}$$

The initial bundles are $\omega_3 = (9, 1)$, and $\omega_4 = (1, 9)$. Now consider the particular allocations x given by:

$$x_1 = (3, 3)$$
$$x_2 = (7, 7)$$
$$x_3 = (3, 3)$$
$$x_4 = (7, 7)$$

The new twinned x allocation is clearly analogous to the old one. Now the old x allocation was in the core. What of the new one?

Suppose x is proposed, or on the table, and persons 1, 2, and 3 get together. Under x, their utilities are $u_1(x_1) = 9$, $u_2(x_2) = 49$ and $u_3(x_3) = 9$. The total quantities of the two goods they are endowed with are given by sums of their initial bundles $\omega_1 = (9, 1)$, $\omega_2 = (1, 9)$ and $\omega_3 = (9, 1)$. Consequently, they start with 19 units of good 1 and 11 units of good 2. Consider this set of bundles:

$$s_1 = (4, 3)$$
$$s_2 = (11, 5)$$
$$s_3 = (4, 3)$$

Note that the sums of the goods in the s_i bundles are nineteen units of good 1 and 11 units of good 2. Moreover, $u_1(s_1) = 12 > u_1(x_1) = 9$; $u_2(s_2) = 55 > u_2(x_2) = 49$, and $u_3(s_3) = 12 > u_3(x_3) = 9$. Therefore, the group of people $S = \{1, 2, 3\}$ blocks x. The x of the replicated economy is not in the core.

The implication of this example is that as an economy gets large (at least via replication), the core in some way shrinks. This shrinkage was first analyzed in 1881 by Edgeworth. He reasoned that a large core is associated with a small number of traders; when there are few people there is lots of room for bargaining. But a small core is associated with a large number of traders; when there are lots of people there are well established market prices, and there is little room for bargaining.

Edgeworth's intuition was confirmed in the 1960's by Gerard Debreu, Herbert Scarf and Robert Aumann, who proved under certain assumptions that as the number of traders gets very large the core shrinks to the set of competitive equilibrium allocations.

6. Final Notes on the Core: The Number of Coalitions

In a move to the core every group of traders, no matter how small, middling, or large, must be satisfied — satisfied in the sense that it could not withdraw from the negotiations and do better on its own. Moving to the core is perfect free trading and perfect free trading is moving to the core. Of course, the core depends on the initial allocation, while the set of Pareto optimal allocations does not. But given the initial allocation, the core is a good set of outcomes for society.

Unfortunately, we have up till now overlooked a potentially disastrous problem. When there are two traders, it is computationally simple to decide whether or not a particular allocation is in the core. One must answer yes or no to these questions: Would trader 1 block it? Would trader 2 block it? Would traders 1 and 2 acting together block it? When there are three traders it is still not too difficult to decide whether or not a particular allocation is in the core. One must answer yes or no to each of these questions: Would trader 1 block it? Would trader 2 block it? Would trader 3 block it? Would traders 1 and 2 acting together block it? Would traders 1 and 3 acting together block it? Would traders 2 and 3 acting together block it? Would traders 1, 2, and 3 acting together block it?

But what if there are, say fifty people? Is it computationally possible to decide whether or not an allocation x is in the core? Is it possibly in our life-times to examine each possible group or coalition in the set of fifty people, to determine whether that group would block? Unfortunately, it is not. When there are fifty people, there are $2^{50} - 1$ different groups that might block. And $2^{50} - 1$ is approximately equal to 1,125,900,000,000,000.

When there are reasonably large numbers of people in the economy do there exist ways to find allocations in the core? Or is the idea of the core generally useless because the number of coalitions is astronomical? We'll see in the next chapter that we can in fact reach core allocations.

7. Exercises

1. In an economy with two people and two goods, suppose

$$\begin{aligned} u_1(x_1) &= 3x_{11} + 2x_{12} \\ u_2(x_2) &= x_{21}x_{22} \\ \omega_1 &= (10, 9) \\ \omega_2 &= (0, 1) \end{aligned}$$

a. Carefully draw an Edgeworth box diagram to represent the economy. Include a few indifference curves and the point ω.

b. Solve for the Pareto optimal allocations. Illustrate them graphically. Illustrate the core graphically.

2. Consider an economy with three people, who have the following utility functions and initial bundles:

$$\begin{array}{ll} u_1(x_1) = 3x_{11} + 2x_{12} + x_{13} & \omega_1 = (0,1,0) \\ u_2(x_2) = 2x_{21} + x_{22} + 3x_{23} & \omega_2 = (1,0,0) \\ u_3(x_3) = x_{31} + 3x_{32} + 2x_{33} & \omega_3 = (0,0,1) \end{array}$$

a. Could any pair of people block ω?

b. Could any single person block ω?

c. Is ω Pareto optimal?

3. There are three people and two goods in an economy. The utility functions and initial bundles are given below:

$$\begin{array}{ll} u_1 = x_{11}x_{12} & \omega_1 = (0,8) \\ u_2 = 2x_{21} + x_{22} & \omega_2 = (2,2) \\ u_3 = x_{31}x_{32} + x_{32} & \omega_3 = (8,0) \end{array}$$

Suppose it is suggested that the three traders move to the allocation x given by $x_1 = (1,2), x_2 = (8,4), x_3 = (1,4)$.

a. Calculate the marginal rates of substitution for the three at the allocation x. (Note: If the MRS's are all equal, then x will be Pareto optimal.)

b. Show that x makes no one worse off than the original allocation.

c. Show that x is not in the core.

8. Selected References

(Items marked with an asterisk (*) are mathematically difficult.)

*1. R. J. Aumann, "Markets with a Continuum of Traders," *Econometrica* V. 32, pp. 39-50, 1964.

This article introduces a model of a continuum of traders as a useful way to understand economies in which each agent has negligible influence in the aggregate outcome. In these economies, Aumann establishes that the core coincides with the set of market equilibrium allocations.

*2. G. Debreu and H. Scarf, "A Limit Theorem on the Core of an Economy," *International Economic Review*, V. 4, 1963, pp. 235-246.

This was the first rigorous treatment of the Edgeworth conjecture that the core shrinks to the set of competitive equilibrium allocations as the economy grows large. Debreu and Scarf formalized the replication argument and proved Edgeworth's conjecture.

3. F.Y. Edgeworth, *Mathematical Psychics*, London, Kegan Paul, 1881, reprinted in 2003, P. Newman (ed.) *F. Y. Edgeworth's Mathematical Psychics and Further Papers on Political Economy*, Oxford University Press. See especially pages 16-42.

In addition to laying the foundation for the Edgeworth box diagram (which he did not actually use in the form later popularized), Edgeworth develops the idea of the core in this readable book. He also develops the notion of a "recontracting process," which works as follows. An allocation is proposed by someone. Then each group of traders or coalition in the economy meets, and decides whether or not the allocation is acceptable. If not, some coalition blocks it, and another proposal is put forward. If no coalition blocks, the proposal is in the core, and the process terminates.

*4. A. M. Feldman, "Recontracting Stability," *Econometrica*, V. 42, 1974, pp. 35-44.

*5. J. Green, "The Stability of Edgeworth's Recontracting Process," *Econometrica*, V. 42, 1974, pp. 21-34.

This paper has a proof of stability for a recontracting process. That is, Green provides a theorem that says under certain assumptions a recontracting process will be dynamically stable. It will eventually reach the core and stop. Also see reference 4 above.

*6. H. Scarf, "The Core of an N-Person Game," *Econometrica*, V. 35, 1967 pp. 60-69.

It is possible to construct examples of economies in which the core is empty, and any allocation is blocked. For such examples, perfect bargaining might go on until all the participants die of exhaustion. But these examples are special. In this paper Scarf shows that the crucial ingredient is convex indifference curves — with a convexity assumption core allocations do in fact exist. Economists normally assume that indifference curves are convex, so Scarf's result is reassuring.

*7. R. Serrano and O. Volij, "Mistakes in Cooperation: the Stochastic Stability of Edgeworth's Recontracting," Working Paper 2003-23, Department of Economics, Brown University, 2003.

This paper studies a recontracting process similar to those in references 4 and 5, but in which agents make mistakes with small probability. The tools of stochastic dynamical systems –stochastic stability– are employed to uncover the allocations that are visited by the process a positive proportion of time in the long run. The results relate to the core and the strong core of the economy.

Chapter 3

WELFARE PROPERTIES OF MARKET EXCHANGE

1. Introduction

In chapter 2 we analyzed barter exchange — that is, exchange without prices. In the Edgeworth model of that chapter, the traders start at some given initial allocation, and a proposal is made for a new allocation. Each group of traders decides whether or not it can, by itself, better the proposal. If no group can better it, the proposal is in the core, and is accepted.

In this chapter, we analyze a more familiar type of economic exchange — exchange governed by prices. However, the model remains rather abstract, since there are no stores, cash registers, checks, or dollar bills in our description of price exchange.

The story goes roughly like this. Each trader starts with some given initial bundle of goods. But now there is a market at which people buy and sell according to given prices. Someone in the market announces a list of prices. Each trader decides, on the basis of those prices, how much of each good he wants to buy and sell. The things he buys are always paid for with the things he sells (since there are no bank accounts, no loans, no accumulations of cash). Consequently the value of what he wants to buy (at the announced market prices) must not exceed the value of what he wants to sell (at the announced market prices). Or, equivalently, the value of the bundle of goods he wants to consume must be less than or equal to the value of the bundle of goods he starts with.

But what people can consume is limited by what actually exists in the exchange economy. If we start with ten units of food distributed between two people, and there is no production, then the total amount of food that those two can consume is, in the end, ten units. That is,

for each good, total supply must equal total demand, for there to be an equilibrium. The first announced list of prices might not yield an equilibrium. If it does not, there is no actual exchange made, and the person who announces the prices changes them. Eventually a list of prices is found which does have the equilibrium property that supply equals demand for each good. When this list is found, the transactions are actually made, and the equilibrium allocation of goods is established.

This is essentially the model of price-governed exchange first analyzed by Leon Walras in 1874. The subtle part of the Walrasian model is its characterization of the behavior of the person who announces prices; fortunately, we won't dwell on that. What is important for us in the model is the characterization of the equilibrium in a competitive price-governed economy: In a Walrasian equilibrium, each person is buying the best bundle (for him) that he can afford, and all the individual decisions are consistent, in the sense that total demand equals total supply for each good. In spite of its level of abstraction, this is a good description of a smoothly running economic machine. What we can learn about the Walrasian model therefore should give us some insight into real competitive economies.

2. The Two-Person, Two-Goods Model

The simplest price-governed exchange model is a two-person, two-goods model. Suppose persons 1 and 2 have self-interested utility functions u_1 and u_2, respectively, and initial bundles ω_1 and ω_2, respectively.

Now suppose 1 and 2 act as if prices are given. What does this mean? It means, for 1, that he ignores 2, and behaves as if the only thing governing his behavior is the pair (or vector) of prices (p_1, p_2). How does (p_1, p_2) govern his behavior? Well suppose he has zero apples, worth \$.10 each, and ten grapefruits, worth \$.25 each. In all probability he will want to eat some apples as well as some grapefruits, so he will have to sell some of his grapefruits, to buy the apples he wants. If he decides he wants to consume A apples and G grapefruits, the value of the bundle (A, G) must be less than or equal to the value of the bundle he starts with, namely $(0, 10)$, otherwise, he can't afford (A, G). This means that (A, G) must satisfy

$$\$.10 \cdot A + \$.25 \cdot G \leq \$.10 \cdot 0 + \$.25 \cdot 10 = \$2.50.$$

In general, in a price-governed exchange world, person 1 chooses a bundle x_1 to maximize $u_1(x_1)$ subject to the constraint

$$p_1 x_{11} + p_2 x_{12} \leq p_1 \omega_{11} + p_2 \omega_{12}.$$

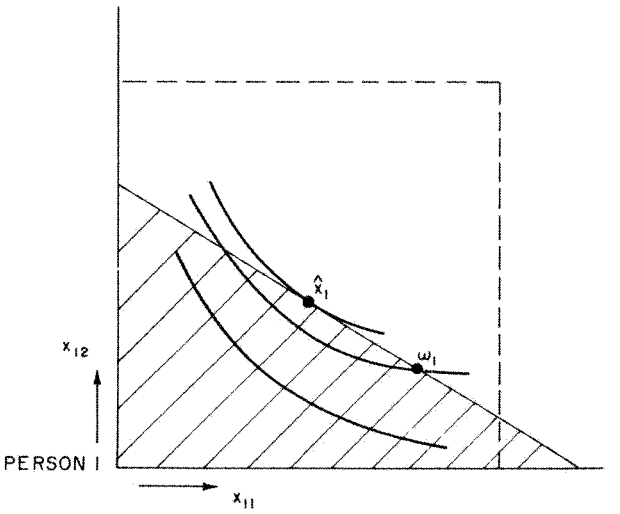


Figure 3.1.

Figure 3.1 represents his problem. The dashed lines show the outside boundaries of the Edgeworth box diagram, which 1 is now ignoring, because he believes that the only thing which constrains him is the set of prices (p_1, p_2). We have drawn a line, called a *budget* or *price line*, through ω_1 to represent the set of points which satisfy

$$p_1 x_{11} + p_2 x_{12} = p_1 \omega_{11} + p_2 \omega_{12}.$$

The absolute value of the slope of this line is p_1/p_2. Every bundle (x_{11}, x_{12}) which lies on or below the line satisfies

$$p_1 x_{11} + p_2 x_{12} \leq p_1 \omega_{11} + p_2 \omega_{12}.$$

We have shaded in the set of bundles of goods which lie in this region.

Given his budget line, which depends only on ω_1 and (p_1, p_2), person 1 chooses a best bundle of goods. If he has indifference curves like the ones drawn, he will choose the bundle $\hat{x}_1$, where one of his indifference curves just touches, or is tangent to, his budget line.

If person 1's preferences are unusual, $\hat{x}_1$ might not lie on the budget line. For example, if he is indifferent about all bundles, he might choose a bundle below the line. In what follows however, we typically assume that people generally prefer larger quantities of goods to smaller, and this assumption will guarantee that the chosen bundle is on the budget line. Also observe that $\hat{x}_1$ might fall outside of the Edgeworth box, outside of the dashed line.

Now we put person 2 explicitly back into the picture. Let us assume for simplicity that the total quantity of each good is 1 unit, so the Edgeworth box diagram must be 1 unit on each side.

Where is person 2's budget line? Formally it is given by

$$p_1 x_{21} + p_2 x_{22} = p_1 \omega_{21} + p_2 \omega_{22}.$$

If we substitute 1's quantities for 2's in this equation, according to the rules

$$x_{1j} + x_{2j} = \omega_{1j} + \omega_{2j} = 1 \text{ for } j = 1, 2$$

we get

$$p_1(1 - x_{11}) + p_2(1 - x_{12}) = p_1(1 - \omega_{11}) + p_2(1 - \omega_{12})$$

or

$$p_1 x_{11} - p_1 + p_2 x_{12} - p_2 = p_1 \omega_{11} - p_1 + p_2 \omega_{12} - p_2$$

or

$$p_1 x_{11} + p_2 x_{12} = p_1 \omega_{11} + p_2 \omega_{12}$$

which is exactly 1's budget line equation. Therefore, 2's budget line coincides with 1's inside the Edgeworth box. Outside the box the two budget lines are cut off differently, since the goods axes for the two traders are different.

Person 2 can afford any bundle on or below (with respect to his origin) his budget line; he can consume any bundle in the shaded area of Figure 3.2. The figure includes a few of 2's indifference curves. Evidently, given the prices p_1 and p_2, person 2 will choose the bundle $\hat{x}_2$.

The next step is to combine the two diagrams into one. We will now drop the extensions of the budget line outside the boundaries of the box, as they play no important role in what follows. In Figure 3.3 we show the budget line through the initial allocation ω, the two crucial indifference curves of persons 1 and 2 that are tangent to the budget line, and the desired bundles of persons 1 and 2, $\hat{x}_1$ and $\hat{x}_2$. Note that we have indicated the quantities $\hat{x}_{11}$, $\hat{x}_{12}$, $\hat{x}_{21}$ and $\hat{x}_{22}$ on the sides of the box.

As the picture is drawn, the desired quantities $\hat{x}_{11}$, $\hat{x}_{12}$, $\hat{x}_{21}$ and $\hat{x}_{22}$ are inconsistent. This is so because $\hat{x}_{11} + \hat{x}_{21} > \omega_{11} + \omega_{21} = 1$, the length of the box; and $\hat{x}_{12} + \hat{x}_{22} < \omega_{12} + \omega_{22} = 1$, the height of the box. The two people want to consume more of good 1 than exists. And this is impossible.

We call $\hat{x}_{11} + \hat{x}_{21}$ the (total) demand for good 1; $\hat{x}_{12} + \hat{x}_{22}$ the (total) demand for good 2; $\omega_{11} + \omega_{21}$ the (total) supply of good 1; and $\omega_{12} + \omega_{22}$ the (total) supply of good 2. In this example there is excess demand for good 1, and excess supply of good 2.

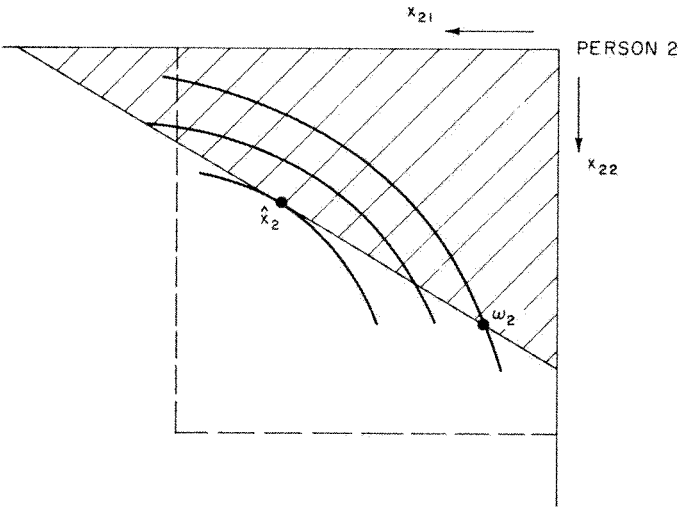


Figure 3.2.

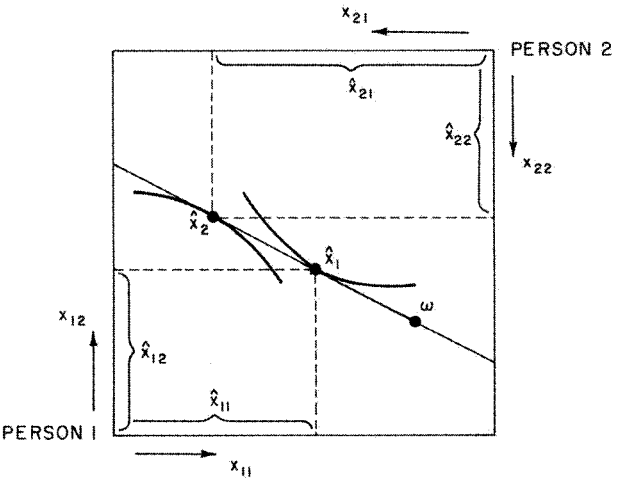


Figure 3.3.

If the desires of the two traders are not consistent, it is impossible for them to trade from ω to $(\hat{x}_1, \hat{x}_2)$. Therefore, something must give. Either one or both will be frustrated, or the prices will change.

In the everyday world, excess demand means shortages. When there are shortages, if prices are unregulated, they go up. For if there are shortages, there are frustrated consumers who offer to pay a little more

to get the quantities they want, and there are sellers who realize that they can boost prices and still sell all they want to sell. Similarly, in the everyday world excess supply means surpluses. When there are surpluses, prices tend to go down. (This phenomenon might be masked in an economy in which there is persistent inflation which hides price declines. But even if all absolute prices are rising, the goods for which there are surpluses will fall in price relative to other goods. And, only relative prices matter in our exchange economy model, since doubling all prices has no real effect on any consumer's budget equation.) When there are surpluses, sellers have unplanned and unwanted inventories, so they have "special sales." Buyers see extra stocks of merchandise, so they try to bargain with sellers, again pushing prices down.

Let's now incorporate these natural price movements in our analysis. Good 1 is in excess demand, and good 2 is in excess supply. Therefore, p_1 will tend to rise, and p_2 will tend to drop, or, at any rate, p_1 will tend to rise relative to p_2. Therefore, p_1/p_2 will go up, and the budget line will get steeper. This shift will continue until supply equals demand for both goods, that is, until the desired bundles $\hat{x}_1$ and $\hat{x}_2$ of the two traders coincide.

An allocation $\hat{x} = (\hat{x}_1, \hat{x}_2)$, where the traders are consistently maximizing their respective utilities, subject to their budget constraints, is called a competitive equilibrium allocation. It is competitive because both traders are acting as price takers, which is presumably how people act in large competitive markets with lots of traders, and it is an equilibrium because their consumption plans are consistent; there is no excess supply or excess demand for either good; and there is no reason for prices to change further. A competitive equilibrium allocation is illustrated in Figure 3.4.

The figure suggests a very important result. Since the two traders' indifference curves are tangent to the same budget line at $\hat{x}$, they are also tangent to each other. Consequently, $\hat{x}$ is Pareto optimal. It is also clear that each person likes $\hat{x}$ at least as much (in fact more) than he likes ω. Consequently, $\hat{x}$ is in the core.

3. Competitive Equilibrium in an Exchange Economy: Formal Preliminaries

We formally define a competitive equilibrium for an economy of self-interested traders as follows. Suppose there are n people and m goods.

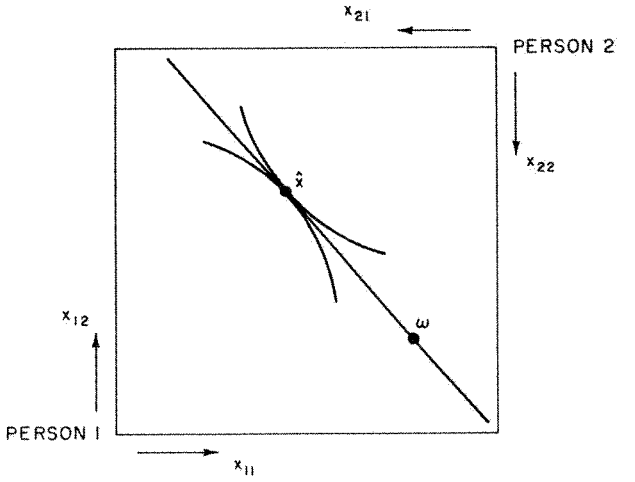


Figure 3.4.

The allocation $\hat{x}$, along with the price vector $p = (p_1, p_2, \ldots, p_m)$, form a *competitive equilibrium* if, for all i,

$\hat{x}_i$ maximizes $u_i(x_i)$ subject to the budget constraint

$$p_1 x_{i1} + p_2 x_{i2} + \ldots + p_m x_{im} \leq p_1 \omega_{i1} + p_2 \omega_{i2} + \ldots + p_m \omega_{im}.$$

Let's make three remarks about the definition. First, $p_1 x_{i1} + p_2 x_{i2} + \ldots + p_m x_{im}$ is the value of the bundle x_i, while $p_1 \omega_{i1} + p_2 \omega_{i2} + \ldots + p_m \omega_{im}$ is the value of the bundle ω_i. The budget inequality then says that, when valued at the prices given by p, the value of what person i consumes must be less than or equal to the value of what he starts with. Second, since $\hat{x}$ is an allocation, we must by definition have supply equal to demand for every good:

$$\sum_{i=1}^{n} \hat{x}_{ij} = \sum_{i=1}^{n} \omega_{ij}, \text{ for all goods } j.$$

Third, a competitive equilibrium depends on ω, the initial allocation. If you change ω, you change the competitive equilibria.

At this point we can introduce an important notational simplification. We define $p \cdot x_i$, the *dot product* of the vector p and the vector x_i as follows:

$$p \cdot x_i = \sum_{j=1}^{m} p_j x_{ij} = p_1 x_{i1} + p_2 x_{i2} + \ldots + p_m x_{im}.$$

Thus $p \cdot x_i$ is the value of the bundle x_i, given the price vector p. Similarly, $p \cdot \omega_i$ is the value of the initial bundle ω_i, given the price vector p, and person i's budget constraint can be compactly rewritten

$$p \cdot x_i \leq p \cdot \omega_i.$$

Now for the theorems below we need a formalization of the idea that a person prefers more goods to less. Suppose person i is self-interested, so his utility depends only on his own consumption. Let x_i be any bundle of goods for him. If i would prefer to x_i any bundle y_i that includes more of every good, then i's utility function is said to be monotonic. That is, u_i is *monotonic* if $y_{ij} > x_{ij}$ for all j implies $u_i(y_i) > u_i(x_i)$.

For one of the theorems below we need two simple preliminary observations, in addition to the definitions and notation developed so far. The first observation says if i prefers some other bundle to the bundle he chooses to buy, that other bundle must be too expensive for him: it must be more valuable than the bundle he starts with. This makes splendid sense, for if the preferred bundle weren't too expensive for him, he would buy it, since his goal, after all, is to maximize his utility. The second observation says that if i likes some other bundle at least as much as the bundle he chooses to buy, that other bundle can't be less valuable than the bundle he starts with. This depends on the prior assumption of monotonicity, but it too makes fine sense. For if the other bundle were actually less valuable than the bundle he starts with, he could afford a bundle that is slightly more expensive than the other bundle, one that contains slightly more of every good. He would prefer this third bundle to the bundle he chooses to buy. So he would buy it instead, since his goal is, again, to maximize his utility.

Let's turn to the formal statements of these observations.

Observation 1. Let $(\hat{x}, p)$ be a competitive equilibrium. If $u_i(y_i) > u_i(\hat{x}_i)$ for some bundles y_i, then

$$p \cdot y_i > p \cdot \omega_i.$$

Proof. If this were not the case, then $\hat{x}_i$ would not be the bundle that maximizes the utility function u_i subject to i's budget constraint, which would contradict the assumption that $(\hat{x}, p)$ is a competitive equilibrium. Q.E.D.

Observation 2.. Let $(\hat{x}, p)$ be a competitive equilibrium. Assume i has a monotonic utility function. If $u_i(y_i) \geq u_i(\hat{x}_i)$ for some bundle y_i then

$$p \cdot y_i \geq p \cdot \omega_i.$$

Proof. Suppose to the contrary that $p \cdot y_i < p \cdot \omega_i$. Define a bundle z_i by adding a small quantity ϵ of every good to the bundle y_i:

$$z_{ij} = y_{ij} + \epsilon \text{ for all } j.$$

Choose ϵ small enough so that

$$\begin{aligned} p \cdot z_i &= p_1 y_{i1} + p_1 \epsilon + p_2 y_{i2} + p_2 \epsilon + \ldots + p_m y_{im} + p_m \epsilon \\ &= p \cdot y_i + \epsilon(p_1 + p_2 + \ldots p_m) \\ &\leq p \cdot \omega_i. \end{aligned}$$

This can be done since, by assumption, $p \cdot y_i < p \cdot \omega_i$. By the monotonicity assumption, person i prefers the bundle z_i to the bundle y_i. That is,

$$u_i(z_i) > u_i(y_i) \geq u_i(\hat{x}_i).$$

Therefore, $u_i(z_i) > u_i(\hat{x}_i)$, while $p \cdot z_i \leq p \cdot \omega_i$, which again contradicts the assumption that $(\hat{x}, p)$ is a competitive equilibrium.

Q.E.D.

This completes our formal preliminaries to the first main result.

4. The First Fundamental Theorem of Welfare Economics

Since around the time of Adam Smith (*Wealth of Nations*, 1776), economists have argued that the price system has a special virtue that other allocative systems might not share. The price system induces self-interested individuals, independently maximizing their private well being, to bring the economy to a socially optimal state. The invisible hand of competition transforms private greed into public welfare. Why should this be the case? In the complicated real world, people see prices and adjust their behavior accordingly, in order to maximize private welfare or private profit. High prices steer people away from wasteful uses of resources and technology, low prices attract them to effective production processes, effective technologies, and effective ways to satisfy wants. With prices giving the signals, wants are fulfilled in an economical way.

Moreover, the system works automatically. When supply is greater than demand, prices fall; when demand is greater than supply, prices rise. There is no need for an expensive or cumbersome centralized bureaucracy to tell us what to do; the hand of competition will lead in the right direction, and without force or coercion. Respecting the individual's property rights, the market system is based on the voluntary participation of people. Each person's wants and the resources that he has to satisfy them determine his final position in the system.

That's not all, though. Smith famously argued that the market system maximizes wealth. But 19th and 20th century economists, including Francis Ysidro Edgeworth, Martin Shubik, Herbert Scarf Gerard Debreu and Robert Aumann, proved that the market does more: it produces an outcome that is in the core. As we will see, a market equilibrium has a special quality, that no coalition of traders, large or small, can object to it.

In the simple world of exchange, the market mechanism automatically distributes goods among people in an optimal way. The person who likes coffee will end up with a lot of coffee, while the one who likes tea will end up with a lot of tea. The person who likes bread will have bread, while the one who likes potatoes will have potatoes. Each consumer will get the bundle of goods he likes best (given his budget constraint), and all those diverse desires will be invisibly reconciled through the adjustment of prices.

Let us be more precise. First, for us, a socially optimal state in the exchange economy model is an allocation that is in the core — one with the property that no coalition of traders, large or small, could better itself by an internal redistribution of its own resources. Second, the outcome of a competitive system or the market mechanism is simply a competitive equilibrium: An allocation $\hat{x}$ and a price vector p with the property that, given his initial holdings ω_i and the prices p, each and every trader is maximizing his own (self-interested) utility function. So the fundamental theorem we are after is: "A competitive equilibrium allocation is in the core."

In the last chapter, we saw that it is generally computationally impossible to determine whether or not an allocation is in the core via an examination of all possible blocking coalitions. There are just too many possible coalitions or groups of traders. But the fundamental theorem points the way to arriving at a core allocation: Use the competitive mechanism; use the free market.

Now we can analyze the theorem to see whether or not it is right. Very few meaningful propositions are always true, and the one at hand is not one of the few. In fact, we can construct a clear example, with two self-interested traders and two goods, where the proposition is false.

Consider a two-person, two-goods economy, in which person 1 has the following odd utility function:

$$u_1(x_1) = \begin{cases} 1 \text{ when } x_{11} + x_{12} < 1 \\ \\ x_{11} + x_{12} \text{ when } x_{11} + x_{12} \geq 1. \end{cases}$$

Note that this utility function is not monotonic in the region where $x_{11}+x_{12} < 1$. That is, person 1 is indifferent among all bundles satisfying $x_{11} + x_{12} < 1$; if you start at one such bundle, and give him just a little more of both goods (so that $x_{11} + x_{12} < 1$ remains true), then he is no better off. He is said to have a "fat" indifference curve in this region. Let person 2's utility function be

$$u_2(x_2) = x_{21}x_{22}.$$

Suppose the initial allocation is $\omega_1 = (1, 0), \omega_2 = (0, 1)$.

Now let $p = (1, 1)$, let $\hat{x}_1 = (1/2, 1/2)$, and $\hat{x}_2 = (1/2, 1/2)$. Obviously $\hat{x}$ is an allocation. The totals of the two goods are 1 and 1, as they must be. With the prices $p = (1, 1)$ and the initial bundle $\omega_1 = (1, 0)$, person 1 can afford any bundle that costs no more than $p_1\omega_{11} + p_2\omega_{12} = 1$. That is, he can afford any bundle in his fat region of indifference, as well as any bundle on the line $x_{11} + x_{12} = 1$. But all these bundles give him 1 unit of utility. Consequently, $\hat{x}_1$ maximizes person 1's utility subject to his budget constraint.

With the prices $p = (1, 1)$ and the initial bundle $\omega_2 = (0, 1)$, person 2 can afford any bundle satisfying

$$p_1x_{21} + p_2x_{22} \leq p_1\omega_{21} + p_2\omega_{22} = 1.$$

He finds the best such bundle using the tangency condition:

$$\text{MRS for person 2} = \frac{p_1}{p_2}$$

or

$$\frac{x_{22}}{x_{21}} = \frac{p_1}{p_2} = 1.$$

The bundle $\hat{x}_2 = (1/2, 1/2)$ solves person 2's problem; so $\hat{x}_2$ maximizes person 2's utility subject to his budget constraint.

In short, $\hat{x}$ is a competitive equilibrium allocation. But it is not Pareto optimal, and therefore it is not in the core. There are allocations that make person 2 better off and person 1 no worse off. For instance, let $\hat{y}_1 = (0, 0)$ and $\hat{y}_2 = (1, 1)$. Then $u_1(\hat{y}_1) = 1 \geq u_1(\hat{x}_1) = 1$, while $u_2(\hat{y}_2) = 1 > u_2(\hat{x}_2) = 1/4$.

The source of the difficulty in this example is the first person's fat indifference curve. But this fat indifference curve is really quite bizarre. People with fat indifference curves in the everyday world are the ones who literally throw their money away. Lots of us claim that we know someone else who throws money away, but we deny that we do it ourselves! (Putting money in a bank for future use is not throwing it away, nor is donating it to a worthwhile charity.) In short, fat indifference

curves can be comfortably assumed away. They are in fact ruled out by the assumption of monotonic preferences.

How do things change when monotonicity is assumed? It turns out that in our exchange model, with self-interested monotonic utility functions, the competitive mechanism automatically distributes goods among people in an optimal way. A competitive equilibrium allocation is in the core, and the market does achieve a socially desirable state of affairs.

Let's now turn to a formal statement and proof of this most basic theorem of welfare economics.

First Fundamental Theorem of Welfare Economics. If all traders have monotonic self-interested utility functions, and if $(\hat{x}, p)$ is a competitive equilibrium, then $\hat{x}$ is in the core (and is, therefore, Pareto optimal as well).

Proof. Suppose $(\hat{x}, p)$ is a competitive equilibrium.

Suppose, contrary to the theorem, that $\hat{x}$ is not in the core.

Then some coalition can block $\hat{x}$ from ω. Let us say S can block x, and let us say it can do so with the bundles $\{s_i\}_{i \text{ in } S}$. This means that

$$\sum_{i \text{ in } S} s_{ij} = \sum_{i \text{ in } S} \omega_{ij} \text{ for all goods } j \qquad (i)$$

$$u_i(s_i) \geq u_i(\hat{x}_i) \text{ for all } i \text{ in } S \qquad (ii)$$

$$u_i(s_i) > u_i(\hat{x}_i) \text{ for at least one } i \text{ in } S \qquad (iii)$$

Combining (ii) with Observation 2, we have

$$p \cdot s_i \geq p \cdot \omega_i \text{ for all traders } i \text{ in } S.$$

Combining (iii) with Observation 2, we have

$$p \cdot s_i > p \cdot \omega_i \text{ for at least one } i \text{ in } S.$$

Now let us add these inequalities over all the traders in S, to get

$$\sum_{i \text{ in } S} p \cdot s_i > \sum_{i \text{ in } S} p \cdot \omega_i.$$

This inequality can be rewritten

$$\sum_{i \text{ in } S} (p_1 s_{i1} + p_2 s_{i2} + \ldots p_m s_{im})$$

$$> \sum_{i \text{ in } s} (p_1 \omega_{i1} + p_2 \omega_{i2} + \ldots p_m \omega_{im})$$

Rearranging, we have

$$p_1 \sum_{i \text{ in } S} s_{i1} + p_2 \sum_{i \text{ in } S} s_{i2} + \ldots + p_m \sum_{i \text{ in} S} s_{im}$$

$$> p_1 \sum_{i \text{ in } S} \omega_{i1} + p_2 \sum_{i \text{ in } S} \omega_{i2} + \ldots + p_m \sum_{i \text{ in } S} \omega_{im}$$

or

$$p_1 \left[\sum_{i \text{ in } S} s_{i1} - \sum_{i \text{ in } S} \omega_{i1} \right] + p_2 \left[\sum_{i \text{ in } S} s_{i2} - \sum_{i \text{ in } S} \omega_{i2} \right]$$

$$+ \ldots + p_m \left[\sum_{i \text{ in } S} s_{im} - \sum_{i \text{ in } S} \omega_{im} \right] > 0$$

But by (i), every term in brackets is zero. Therefore, the assumption that $\hat{x}$ is not in the core leads to a contradiction, which proves the theorem. Q.E.D.

5. The Second Fundamental Theorem of Welfare Economics

The first fundamental theorem says that a competitive equilibrium allocation is in the core, and is Pareto optimal. The competitive market, in other words, brings about a distribution of goods that is desirable in the sense that no group of traders could do better on its own. Moreover, it brings about this desirable distribution automatically: prices tend to rise in response to excess demand and tend to fall in response to excess supply; the prices adjust by themselves to solve the distribution problem.

However, the ideal distribution of goods brought about by the competitive mechanism depends heavily on the initial allocation. That is, the competitive allocation and the core are determined by the initial allocation ω as well as by preferences, and if, for example, the initial allocation is very unequal, so will be the competitive allocation.

This is an important objection to complete reliance on the competitive market: it might produce great inequalities. What does this mean in a real economy? An economy with production as well as exchange has people selling (or renting) their assets to firms, as well as buying and consuming goods and services. People sell their labor, or rent their capital goods. Some people have lots of capital to rent, and some people have very valuable labor to sell. Usually we think of industrialists, doctors and lawyers in this regard, but T.V. personalities, baseball players, and movie stars are better examples. And some people have no capital to lend or rent, and very little valuable labor to sell. Some people have

few talents, few skills, and maybe not even much muscle power. There are haves, and there are have nots. And the have nots might be have nots through no fault of their own. They might be disabled, afflicted by disease, or just very unlucky. The free market mechanism will produce a distribution of goods that gives Rolls-Royces and homes in Palm Springs to the baseball stars. The distribution will give Fords and suburban tract homes to many of us. But it will give poverty and hunger to the have nots. And the result will very likely be Pareto optimal and in the core.

We can illustrate the inequality that might result from a competitive equilibrium allocation in an Edgeworth box diagram.

In Figure 3.5, ω is an initial allocation that strongly favors person 1, and $\hat{x}$ is the competitive equilibrium based on it. A Pareto optimal allocation like y might seem preferable to $\hat{x}$ on equity grounds. So the question arises: Can the competitive mechanism with some modifications be used to move the economy to an alternative Pareto optimal allocation like y, even given the initial distribution ω?

Do we really need to ask this question? Examination of Figure 3.5 seems to indicate an obvious solution to the inequality problem. Simply make person 1 give to person 2 $\omega_{11} - y_{11}$ units of good 1, and $\omega_{12} - y_{12}$ units of good 2. End of discussion. (Person 1 doesn't want to do this, of course, but reducing inequality almost always means causing someone to do something he doesn't want to do.)

This obvious solution to the inequality problem involves the direct transfer of quantities of various goods from one person to another. Why can't this solution be used in general for cases more complicated than the one drawn in the figure? It cannot be used for the same reason that a proposed allocation cannot be checked to see whether or not it is in the core by examining all possible blocking coalitions. That is, when there are many people and many goods the direct transfer solution is computationally impossible. It is just too burdensome to work.

Consider the economy of the United States, with its nearly 300 million consumers. How many different goods do we have? Obviously, the answer to this question depends on the level of aggregation we use: we might say food is a good, rather than apples, tomatoes, bread, etc. But if we talk about reasonable levels of aggregation, there are surely hundreds of things we want to identify as distinct goods. And if we walk into a large department store, we can count tens of thousands of different items; in a large supermarket there are probably tens of thousands of different food, grocery, and household goods available. So on a finely disaggregated level, there are hundreds of thousands of different goods being produced and consumed in the United States. Now imagine the

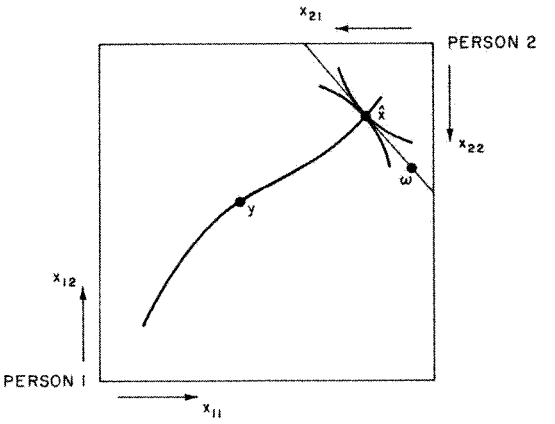


Figure 3.5.

problem of directly transferring either hundreds, or hundreds of thousands, of goods among hundreds of millions of people. Could a Central Authority, say a branch of the U.S. Government, effect such a transfer in a reasonable way?

To answer that question we can look at an effort that took place in the 70's by the U.S. Government to partially direct the reallocation of one good: gasoline. The United States Department of Energy employed some 20 thousand people (of whom, in truth, only a fraction worked on the chore of gas allocation). It attempted, in the spring and summer of 1979, to direct the distribution of gasoline in the United States, on a regional basis. That is, it attempted to dictate how much each state should get and how the gas should be distributed between urban and rural areas within each state. It did not attempt to decide how much each driver should get. Now the Energy Department allocated gasoline by, in effect, short circuiting the market mechanism. Gasoline sellers were not allowed to adjust their prices freely, and they were not allowed to decide by themselves where to sell what they want to sell. Their actions were governed by Department regulations, which were literally thousands of pages in length.

Did the Energy Department, with its thousands of employees and thousands of pages of regulations, succeed in distributing gasoline to U.S. consumers in a reasonable way? The answer to the question is a rather clear No. With the suppression of the price mechanisms a new rationing device appeared: the gas station line. Consumers spent, in the aggregate, millions of man hours waiting in lines. They burned, in the aggregate, millions of gallons of gasoline simply looking for open

gas stations or waiting on lines. They ran into each other's cars while jockeying for positions on lines. They actually shot and killed each other over gallons of gas. This was a distribution system that wasted time, gasoline, people's nerves, and even occasionally lives. The attempt to partially direct the distribution of one good was a dramatic failure.

Now, imagine the complexities of having the Central Authority effect a transfer of hundreds or thousands of goods among 300 million people. If an initial allocation is unequal or inequitable in a large economy, like the U.S. economy, and if a more equitable allocation is sought on equity grounds, a movement from the initial allocation to the more equitable allocations might theoretically be brought about by a centralized and purely nonmarket mechanism. The Central Authority might theoretically inform everyone of the precise quantities of the thousands of goods to which he is entitled. But the costs, the wastes, and the information requisites of such nonmarket reallocations are enormous.

Back then, to the question: Can the automatic, decentralized, competitive market mechanism with some modifications be used to move an economy to a more equitable Pareto optimal allocation? In terms of Figure 3.5, can a modified market mechanism be used to get the economy from ω to a point like y?

The answer is generally yes, and the modified mechanism works like this. Instead of transferring quantities of hundreds of thousands of goods among millions of people, the Central Authority transfers cash. That is, generalized purchasing power, or money, is taken from some people, and given to others. After people's bank accounts have been lightened, or enhanced, as the case may be, they are left to their own devices, and the market proceeds to work as usual. Prices adjust to automatically equate supply and demand in each market, and a new, more equitable competitive equilibrium allocation comes about. The new allocation is Pareto optimal. But it is probably not in the core for the original, pretransfer allocation.

The cash transfer system has important advantages over the direct transfer of goods system. First, it is not computationally staggering. A single human mind can grasp the idea that everyone should have a minimum (cash) income of X dollars. A single human mind cannot devise a good distribution of tens of thousands of goods among hundreds of millions of people. Second, it preserves people's freedom of choice. Adjusting a person's bank book and then letting him do his shopping is less onerous than adjusting his consumption bundle directly. And third, adjusting bank balances will not create the disastrous and wasteful market problems that direct transfers have created in, for instance, the market for gasoline.

Now let's concentrate on the meaning of the cash-transfer system in the context of our exchange economy model. This might be confusing, since we have said there is no cash, no money, in the exchange model: goods trade for goods. That position must be modified. To picture what is going on here, imagine that everyone has an account with a central bank. The bank lists, in person i's account, all the goods he has. Initially the accounts lists ω_i. Now suppose there is some list of prices for the goods, or price vector p. At the bottom of person i's account book, the bank evaluates i's goods. Initially, this value is $p_1\omega_{i1}+p_2\omega_{i2}+\ldots+p_m\omega_{im}$, or $p \cdot \omega_i$. If there are no cash transfers and if i wants to trade his initial bundle ω_i for a consumption bundle x_i, i must clear it through the bank, which approves the transaction if

$$p \cdot x_i \leq p \cdot \omega_i$$

that is, if $p \cdot x_i$ is less than or equal to the value of i's account. What we have described so far is just a variation of our basic exchange model, made a little complex by the existence of the bank, whose sole function is to keep an eye on people's budget constraints.

When there are cash transfers, the bank is instructed by the Authority running the system to add an amount T_i to person i's bottom line amount $p \cdot \omega_i$. The number T_i could be positive or negative. The right-hand side of i's budget inequality becomes $p \cdot \omega_i + T_i$. And now if i wants to trade his initial bundle ω_i plus his transfer T_i for the consumption bundle x_i, he must clear it through the bank, which approves the transaction if and only if

$$p \cdot x_i \leq p \cdot \omega_i + T_i.$$

What does the Central Authority do to effect a more equitable allocation of goods? It assigns positive T_i's to those people who are judged too poor, and negative T_i's to those who are judged too rich. In other words, it transfers bank balances from the rich to the poor. Once the balances are transferred, the individuals buy and sell as usual and the market mechanism is allowed to work by itself.

We have asserted that a modified mechanism can be devised to get the economy to an equitable allocation, like y in Figure 3.5. At this point we must indicate the formal nature of the problem, and be more precise about the assertion. Mathematically, the problem is this. Suppose we are given an initial allocation ω, and a (desired) Pareto optimal allocation y. Does there exist a vector of bank balance transfers $(T_1, T_2, \ldots, T_n)$ and a price vector $p = (p_1, \ldots, p_m)$, such that, for every person i, y_i maximizes u_i subject to $p \cdot y_i \leq p \cdot \omega_i + T_i$?

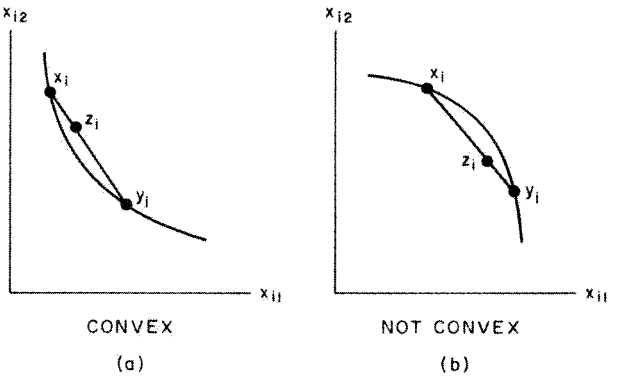


Figure 3.6.

The affirmative answer depends on three assumptions about the utility functions: (1) self-interestedness and (2) monotonicity, both of which have already been defined, and (3) convex indifference curves. Formally, we say u_{i1} has *convex indifference curves* if the following is true: If $u_i(x_i) \geq u_i(y_i)$ for two different bundles x_i and y_i, then $u_i(z_i) \geq u_i(x_i)$ for any bundle z_i that lies on the straight-line segment connecting x_i and y_i. the idea of convexity for indifference curves is illustrated in the figure below. Figure 3.6a shows indifference curves that are convex; while 3.6b shows indifference curves that aren't. Note that the figure illustrates the $u_i(x_i) = u_i(y_i)$ case, and that it assumes monotonicity.

The formal assertion that a modified competitive mechanism can be used to get the economy to almost any desired Pareto optimal allocation is called the Second Fundamental Theorem of Welfare Economics. The proof of this important result is considerably more complex than the proof of the First Fundamental Theorem, so we shall not include it here.

Instead we only state the theorem for the general n-person, m-good case:

> *Second Fundamental Theorem of Welfare Economics.* Suppose all traders have self-interested, monotonic utility functions, and convex indifference curves. Let y be any Pareto optimal allocation that assigns positive quantities of every good to every trader: $y_{ij} > 0$ for all i and j.
>
> Then there exists a vector of bank balance transfers $(T_1, T_2 \ldots, T_n)$ and a price vector $p = (p_1, \ldots, p_m)$ such that y and p are a competitive equilibrium given the transfers. That is, for all i, y_i maximizes u_i subject to $p \cdot y_i \leq p \cdot \omega_i + T_i$.

Let's make some observations about the theorem at this point.

First, the assumption that $y_{ij} > 0$ for all i and j, which seems quite restrictive, is made largely for mathematical convenience, and can be greatly weakened without affecting the conclusion of the theorem.

Second, the T_i's must sum to zero: All transfers to people must be financed by taxing other people. This is so because monotonicity ensures that people will want to spend to the limits of their budgets; therefore, for all i,

$$p \cdot y_i = p \cdot \omega_i + T_i.$$

Adding over all the i and rearranging terms gives

$$\sum_{j=1}^{m} p_j \sum_{i=1}^{n} y_{ij} = \sum_{j=1}^{m} p_j \sum_{i=1}^{n} \omega_{ij} + \sum_{i=1}^{n} T_i$$

or

$$\sum_{i=1}^{n} T_i = \sum_{j=1}^{m} p_j \left[\sum_{i=1}^{n} y_{ij} - \sum_{i=1}^{n} \omega_{ij} \right] = \sum_{j=1}^{m} p_j [0] = 0.$$

There is no way to induce the desired change by simply printing money and giving it to the poor. Some people must be taxed so that others can be subsidized.

Third, the theorem does not indicate how y ought to be chosen! We have argued that y is a more equitable allocation than the original ω, or the competitive allocation that would arise without transfers. But the decision to pick a particular y is made, somehow, by some people, and the process they use to make that decision is left in the dark at this point. In subsequent chapters we will discuss at some length the problem of choosing y.

Fourth, the Second Fundamental Theorem says that the competitive mechanism, with modifications, is even more useful, and more robust, than the First Fundamental Theorem indicates. Even if unmodified competition brings about distributions of goods that are inequitable, the price mechanism with modifications can be used to bring about almost any equitable and optimal allocation.

Fifth and finally, although the second fundamental theorem is an important mathematical result, one should also be aware of its limitations in terms of its practical application. In the real world, the Central Authority in charge of bringing about a desired Pareto efficient allocation y will often not have the necessary information to even determine whether or not y is Pareto optimal whether it is or is not depends on the true preferences of the agents, who may have an incentive to misrepresent them before the Authority.) Thus, the task is hard because one is effec-

tively trying to shoot at a "moving target." We will address these issues in the last chapters of this book on the theory of implementation.

6. An Algebraic Example

Consider first a consumer, person 1, with a linear utility function, $u_1 = x_{11} + 2x_{12}$, and an initial bundle $\omega_1 = (1, 1/2)$. If he is faced with prices $p = (p_1, p_2)$, he will want to buy the best bundle he can afford, that is, the best bundle x_1 satisfying

$$p_1 x_{11} + p_2 x_{12} \leq p_1 \cdot 1 + p_2 \cdot \frac{1}{2}.$$

His indifference curves are straight lines, with slope in absolute value equal to

$$\text{MRS person } 1 = \frac{\text{MU of good } 1}{\text{MU of good } 2} = \frac{1}{2}.$$

Figure 3.7a illustrates indifference curves and a budget line for person 1. The dark line is 1's budget line; the absolute value of its slope, which evidently exceeds $1/2$, is equal to p_1/p_2. Given this p_1/p_2, person 1 wants to consume the bundle $\hat{x}_1$. For a different p_1/p_2, he wants to consume a different bundle: As p_1/p_2 rises, the budget line pivots around ω_1 and gets steeper, and the desired bundle $\hat{x}_1$ climbs up the vertical axis. As p_1/p_2 drops, the budget line gets flatter, and the desired bundle $\hat{x}_1$ climbs down the vertical axis, until $p_1/p_2 = 1/2$. When $p_1/p_2 = 1/2$, the budget line coincides with 1's indifference curve through ω, and every bundle on that indifference curve maximizes 1's utility subject to his budget constraint. When $p_1/p_2 < 1/2$, 1's desired bundle $\hat{x}_1$ moves to the horizontal axis, and as the budget lines get flatter, the desired bundle $\hat{x}_1$ moves out the horizontal axis. The locus of all desired bundles, for all possible prices, is the dashed line in Figure 3.7b. This is the path that $\hat{x}_1$ traces as p_1/p_2 goes from zero to infinity, and it is called person 1's *offer curve*. Figure 3.7b also includes the budget line that appears in 3.7a. Note that when we have the offer curve and the budget line we can read off the bundle $\hat{x}_1$ that 1 wants to consume. This is the point (other than ω_1) where the offer curve and the budget line intersect.

Now suppose we have another consumer, say person 2, with the utility function

$$u_2 = x_{21} x_{22}$$

and the initial bundle $\omega_2 = (0, 1/2)$. If person 2 is faced with prices $p = (p_1, p_2)$, he will want to buy the best bundle he can afford, that is, the best bundle x_2 satisfying

$$p_1 x_{21} + p_2 x_{22} \leq p_1 \cdot 0 + p_2 \cdot \frac{1}{2}.$$

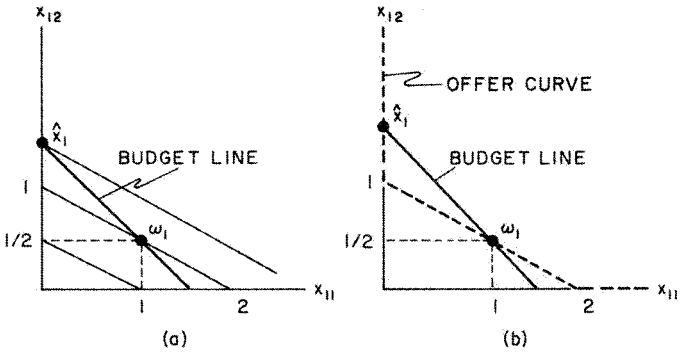


Figure 3.7.

Since his utility function is monotonic he will always want to consume at a point on, rather than below, his budget line; so the bundle he buys will satisfy

$$p_1 x_{21} + p_2 x_{22} = p_1 \cdot 0 + p_2 \cdot \frac{1}{2}.$$

Person 2's indifference curves are hyperbolic, and the absolute value of the slope of an indifference curve for him is given by

$$\text{MRS person 2} = \frac{\text{MU of good 1}}{\text{MU of good 2}} = \frac{x_{22}}{x_{21}}.$$

With these indifference curves, his utility maximizing bundles will always be at points of tangency of his indifference curves and budget lines. That is, for the bundles person 2 wants to consume we will have

$$\text{MRS person 2} = \frac{x_{22}}{x_{21}} = \frac{p_1}{p_2}.$$

Consequently, person 2's choices are governed by two equations

$$p_1 x_{21} + p_2 x_{22} = \frac{1}{2} p_2$$

and

$$p_2 x_{22} = p_1 x_{21}.$$

The solution to this pair of equations is

$$x_{21} = \frac{1}{4}\frac{p_2}{p_1} \text{ and } x_{22} = \frac{1}{4}.$$

This is the algebraic form of person 2's offer curve.

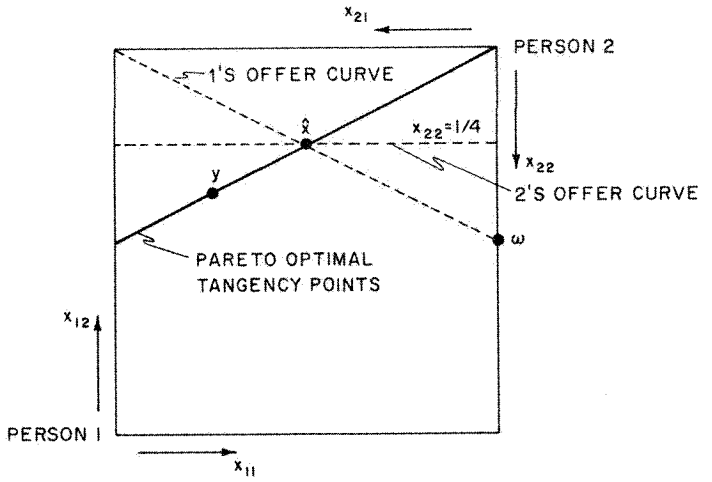


Figure 3.8.

Now consider an economy made up of persons 1 and 2. We have pictured it in an Edgeworth box diagram in Figure 3.8. To minimize complexity, all the figure shows is the offer curve of person 1 (taken from Figure 3.7b), the offer curve of person 2 (the dashed line where $x_{22} = 1/4$), and the locus of Pareto optimal points where the two individuals' indifference curves are tangent. (This is determined by setting MRS for person 1 = MRS for person 2, or $x_{22}/x_{21} = 1/2$.) The initial allocation is the point ω.

The two offer curves intersect at the point $\hat{x}$, or $\hat{x}_1 = (1/2, 3/4)$, $\hat{x}_2 = (1/2, 1/4)$. This means that for the appropriate prices person 1 wants to consume $\hat{x}_1$ and person 2 wants to consume $\hat{x}_2$. What are the prices? The budget line needed to get the two to $\hat{x}$ goes through ω and $\hat{x}$; so its slope in absolute value is 1/2. Consequently, $p_1/p_2 = 1/2$ is required, and, since one of the prices can be chosen arbitrarily, the price vector $p = (p_1, p_2) = (1, 2)$ works fine. In short, $p = (1, 2)$ and $\hat{x}_1 = (1/2, 3/4)$, $\hat{x}_2 = (1/2, 1/4)$ is a competitive equilibrium based on ω.

The First Fundamental Theorem says that $\hat{x}$ is Pareto optimal, and in the core. The figure shows it is Pareto optimal because it lies on the locus of tangencies, and it's in the core because it is Pareto optimal and it makes each individual at least as well off as ω.

Now we turn to the Second Fundamental Theorem. Suppose $\hat{x}$ is judged objectionable because it gives too much to person 1. Consider

the Pareto optimal allocation y given by $y_1 = (1/4, 5/8), y_2 = (3/4, 3/8)$. Suppose we want to get to y from ω via a modified competitive mechanism. What bank balance transfers T_1 and T_2 are required?

The first thing to notice about this example is that person 1, whose indifference curves are straight lines with slope $1/2$, in absolute value, will only choose y_1 to maximize his utility if his budget line also has slope $1/2$ in absolute value. Consequently, the two people will end up at y only if $p_1/p_2 = 1/2$. Again, one price can be chosen arbitrarily, so assume $p = (p_1, p_2) = (1, 2)$ is the competitive equilibrium price vector.

Next note that person 1 will choose point y_1 only if y_1 satisfies the equation

$$p \cdot y_1 = p \cdot \omega_1 + T_1.$$

Similarly, y_2 must satisfy the equation

$$p \cdot y_2 = p \cdot \omega_2 + T_2.$$

Substituting (1, 2) for p, and the given values for y_1, ω_1, y_2 and ω_2, we find that

$$T_1 = -\frac{1}{2} \text{ and } T_2 = +\frac{1}{2}.$$

(Obviously, these magnitudes depended on our setting $p_1 = 1$. If we let $p_1 = 2$, then p_2, T_1 and T_2 would have to be doubled.)

The reader should check that when $p = (1, 2)$, $T_1 = -1/2$ and $T_2 = +1/2$, person 1 will actually maximize his utility at y_1, and person 2 will actually maximize his utility at y_2.

7. Exercises

1. Let $u_1 = 3x_{11} + x_{12}$ and

   $$u_2 = x_{21}x_{22}.$$

   Let the initial allocation be $\omega_1 = (2, 1)$, $\omega_2 = (1, 2)$. Solve for the competitive equilibrium.

2. An economy is made up of two individuals and two goods. Their utility functions are:

   $$u_1 = x_{11} + x_{12}$$
   $$u_2 = 5x_{21}x_{22}.$$

   Their initial endowments are:

   $$\omega_1 = (100, 0)$$
   $$\omega_2 = (0, 50).$$

Let the price of the second good, p_2, be 1.

a. Find 1's offer curve. (Hint: Do it graphically, and pay particular attention to what happens when $p_1 = 1$.)

b. Graph a representative indifference curve for person 2. Is it symmetric around the line $x_{21} = x_{22}$?

c. Find a competitive equilibrium.

3. Consider an economy with two people who have the utility functions and initial endowments given below:

$$\begin{aligned} u_1 &= 2x_{11} + x_{12} \quad & \omega_1 &= \left(\frac{1}{2}, \frac{1}{2}\right) \\ u_2 &= x_{21} + 2x_{22} \quad & \omega_2 &= \left(\frac{1}{2}, \frac{1}{2}\right). \end{aligned}$$

a. Solve for a competitive equilibrium.

b. Show that MRS person 1 = MRS person 2 doesn't hold at the competitive equilibrium.

c. Is the competitive equilibrium allocation Pareto optimal? Why?

4. Again consider the two-person, two goods economy given by

$$\begin{aligned} u_1 &= 2x_{11} + x_{12} \quad & \omega_1 &= \left(\frac{1}{2}, \frac{1}{2}\right) \\ u_2 &= x_{21} + 2x_{22} \quad & \omega_2 &= \left(\frac{1}{2}, \frac{1}{2}\right). \end{aligned}$$

Suppose someone decides that $y_1 = (3/4, 0)$, $y_2 = (1/4, 1)$ is the best allocation. Show how y could be achieved via the modified competitive mechanism, with the appropriate bank balance transfers.

5. Consider the economy given by

$$\begin{aligned} u_1 &= x_{11}x_{12} \quad & \omega_1 &= (1, 0) \\ u_2 &= 2x_{21} + x_{22} \quad & \omega_2 &= (0, 1). \end{aligned}$$

a. Solve for the set of Pareto optimal allocations. Where is the core?

b. Draw an Edgeworth box diagram to represent the economy.

c. Solve for the competitive equilibrium.

d. Calculate bank balance transfers T_1 and T_2, and prices p_1 and p_2, such that the equilibrium of the modified competitive mechanism is $y_1 = (1/4, 1/2)$, $y_2 = (3/4, 1/2)$.

8. Selected References

(Items marked with an asterisk (*) are mathematically difficult.)

*1. K.J. Arrow, "An Extension of the Basic Theorems of Classical Welfare Economics," *Second Berkeley Symposium on Mathematical Statistics and Probability* (J. Neyman, ed.), University of California Press, 1951, pp. 507-532.

In this article Arrow proves the fundamental theorems in a model with exchange and production. There are three main contributions. 1. Arrow formulates the theorems in precise mathematical terms. 2. He uses the theory of convex sets rather than the calculus arguments of earlier authors. 3. Using this powerful mathematical tool, he relaxes the unrealistic assumption made by earlier writers that all quantities of goods produced and consumed by all people are strictly greater than zero.

*2. R. J. Aumann, "Markets with a Continuum of Traders," *Econometrica* V. 32, pp. 39-50, 1964.

See it as reference 1 in Chapter 2.

*3. G. Debreu, *Theory of Value*, John Wiley and Sons, Inc., New York, 1959, Chapters 5 and 6.

Chapter 5 provides an elegant proof of a theorem about the Walrasian model of an economy with exchange and production: Debreu proves that under reasonable assumptions, a competitive equilibrium in fact exists. Consequently, our theorems about the optimality of the competitive equilibrium aren't vacuous!

Chapter 6 proves rigorous proofs for general versions of the two fundamental theorems of welfare economics.

*4 G. Debreu and H. Scarf, "A Limit Theorem on the Core of an Economy," *International Economic Review*, V. 4, 1963, pp. 235-246.

This is the first formal presentation of the so-called core convergence theorem. See Chapter 2, references 1 and 2.

5. O. Lange, "The Foundations of Welfare Economics," *Econometrica*, V. 10, 1942, pp., 215-228.

Lange derives conditions for maximizing person i's utility subject to the constraint that all other people's utilities be held constant, that is, conditions for a Pareto optimum. He uses calculus with Lagrange multipliers to do this. The equations he derives "contain *in nuce* most theorems of welfare economics, e.g., all the propositions in Pigou's

Economics of Welfare." However, he does not explicitly present the fundamental theorems.

6. A.P. Lerner, "The Concept of Monopoly and the Measurement of Monopoly Power," *Review of Economic Studies*, V. 1, 1934, pp. 157-175.

   Although this article is principally about monopoly, it contains a clear statement of a weak version of the first fundamental theorem, with an intuitive, nonrigorous argument. A competitive situation in an economy is Pareto optimal because utility maximizing individuals set marginal rates of substitution equal to price ratios, and profit-maximizing competitive firms set prices equal to marginal costs. Consequently, there is no possibility of further trade between individuals, or of further arrangements between individuals and firms, that would make some people better off and no one worse off. Lerner's version of the theorem does not mention the core.

7. P.K. Newman, *The Theory of Exchange*, Prentice-Hall, Inc., Englewood Cliffs, 1964, Chapters 3, 4, 5.

   Newman gives a detailed exposition of exchange theory. He includes excellent notes on the literature.

8. A.C. Pigou, *The Economics of Welfare*, MacMillan and Co., London, 1920. Part II, Chapters I, II, and III.

   Pigou provides the 1920 version of the first fundamental theorem. In this version, the "free play of self-interest" generally leads to the maximization of the "national dividend," which is roughly analogous to the modern concept of gross national product. So competition maximizes a measure of total output. Pigou is not particularly concerned with Pareto optimality or the core as outcomes of competition, since he feels the national dividend is a reasonable measure of economic welfare.

9. J. Quirk and R. Saposnik, *Introduction to General Equilibrium Theory and Welfare Economics*, McGraw-Hill, New York, 1968, Chapter 4, especially 4-5.

   This text includes a good exposition of the fundamental theorems, as well as material on the existence of competitive equilibria.

10. J.A. Schumpeter, *History of Economic Analysis*, Edited by E.B. Schumpeter, Oxford University Press, 1954.

Part II, Chapter 3 provides an uncomplimentary but useful guide to "Adam Smith and the *Wealth of Nations*." According to Schumpeter, Smith was heavy on examples and weak on analysis. Smith's espousal of laissez faire policies might follow from the first fundamental theorem, but Smith had no clear view of the theorem.

Part IV, Chapter 7 provides a section on "The Walrasian Theory of General Equilibrium." This is excellent even though its completion was interrupted by Schumpeter's death.

11. M. Shubik, "Edgeworth Market Games," in *Contributions to the Theory of Games IV*, R. D. Luce and A. W. Tucker (eds.), Princeton University Press, 1959.

    This article provides the first modern presentation of Edgeworth's conjecture regarding the connection between the core and the market equilibrium allocations.

12. G.J. Stigler, *Production and Distribution Theories*, The MacMillan Company, New York, 1941.

    Stigler devotes Chapter IX to an explanation of the Walrasian general equilibrium system.

13. L. Walras, *Elements of Pure Economics* (Translated by William Jaffe), George Allen and Unwin, Ltd., London, 1954.

    This is a translation of *Eléments d' économie politique pure*, 1926 Edition, Parts II and III are particularly relevant.

    Walras must be given most of the credit for developing two important analytical tools that we use. (1) The general equilibrium framework which puts all consumers, all goods, and in its general form all producers, together in one closed model. In the model the consumers maximize utility subject to their budget constraints, and supply equals demand for all goods. (2) The "tatonnement" price adjustment story, in which a set of prices is called out, desired supplies and demands at those prices are collected, and the prices are adjusted in response to excess supplies and demands. When at last there is no excess supply or demand in any market, actual transactions are made.

    We should note, incidentally, that there is a real question about whether such a (hypothetical) price adjustment mechanism would be stable: Would it really eventually lead to a set of market-clearing prices? We have and will continue to ignore the stability problem in this book. See Quirk and Saposnik for references on stability.

Chapter 4

WELFARE PROPERTIES OF "JUNGLE" EXCHANGE

1. Power as a Basis for Exchange

We will now describe a model of exchange that is quite different from barter exchange or market exchange. In both barter and market transactions, exchanges are voluntary. Person i does not swap a bundle of commodities to person j unless he's better off (or no worse off) by so doing.

But sometimes exchanges are forced: You give me your wallet and your watch, or I'll kill you! This is the "law of the jungle." The more powerful takes from the less powerful.

To model this kind of situation, we will assume there is a *power relation* among individuals $\{1, 2, \ldots, n\}$. That is, there is a strict ordering of the individuals from strongest to weakest. For example, suppose that person 1 is strongest, person 2 is second strongest, and so on. Then if person i meets person j in the jungle, and $i < j$, person i can take whatever he wishes from j, whether j agrees or not.

Let's pause for a moment to consider the plausibility, or lack of plausibility, of jungle exchange. Firstly, we do not intend to claim that there is more forcible taking in tropical rain forests than in New York City. The economic jungle is not a geographic locality. Secondly, we do not intend to claim that taking things based purely on power is pervasive, or even common, in everyday economic activity. Thankfully, it is not. But, thirdly, there are examples where the model might apply: in times of war, in some despotic states, and even, occasionally, in the distribution of certain goods in modern democracies. For example, in some academic departments, offices and/or parking places may be distributed

on the basis of seniority. If professor i is senior to professor j, then i has the right to claim j's office, or his parking place.

Michele Piccione and Ariel Rubinstein, who wrote the seminal paper on this model, called the structure in which power and coercion govern the distribution of goods "the jungle." They might have as well called it "anarchy," but the important point is that there is a power ordering of the participants, and the more powerful take from the less.

Let us now consider what is taken in the jungle.

If we simply laid a power ordering over our barter and price exchange models, which involve allocations of m divisible goods among n people, and which assume monotonic self-interested utility functions, we wouldn't have much of interest. The most powerful person would just seize the goods of everyone else. This would produce a trivial and degenerate jungle equilibrium. In order to make the n-person m-goods jungle model interesting, with divisible goods, we would have to assume that individuals have satiation points, so the most powerful person could seize the consumption bundle that he most wants, but would leave something for the others to fight over.

However, we are not going to assume the usual consumption bundle of m (infinitely divisible) goods, in our jungle model.

2. A Model of Indivisible Objects

We will now assume that each person consumes *one and only one indivisible object.* (Think for example of a house.) We assume that there are n of these objects to be distributed, 1 each, to each of the n people. We will assume that there is some initial distribution of the objects.

The general economic exchange model where n people are allocated n indivisible objects and where each person only wants to consume 1 object, was first developed by Lloyd Shapley and Herbert Scarf. Shapley and Scarf proved that such economies have core allocations, that is, allocations which are unblocked by any coalitions. They also discussed related allocation problems, such as the "roommate problem" (how to pair up n people, who have preferences about *each other*, in some optimal way), and the "marriage problem" (how to pair up n men with n women, again in some optimal way.)

In this chapter, we are considering how to allocate n indivisible objects among n people. Under barter or market exchange, each person swaps his object, in a voluntary exchange, for someone else's, or he trades it in the market for a different one. But under jungle exchange, there is a power relation which permits the more powerful to take from the less powerful.

A *jungle equilibrium* is a distribution of the n objects among the n persons such that, for every i and j, if person i is more powerful than j, person i's utility from the object in his possession is greater than or equal to the utility he would get from j's object. If this is the case, the more powerful i does not bother to force an exchange on the less powerful j.

From this point onward, we will call the indivisible objects houses. As we have done before, we let u_i represent i's utility function. The houses are labeled $h_1, h_2, \ldots, h_n$, where h_i is the house that person i starts with. An allocation of houses is a permutation of the vector $(h_1, h_2, \ldots, h_n)$. The original allocation is called h. We will use g or g' to represent alternative allocations, for example $(h_2, h_1, h_3, \ldots, h_n)$, which switches houses between persons 1 and 2.

(To translate this notation back into the notation used in Chapters 2 and 3, we would proceed as follows: Call house h_1 good 1, call h_2 good 2, etc. Each person can consume a bundle x_i of goods, comprised of exactly one unit of one good, and zero of all the rest. For example, $x_i = (0, 1, 0, \ldots, 0)$ means i is consuming the house h_2. The initial allocation is the set of unit vectors: $\omega_1 = (1, 0, 0, \ldots, 0)$, $\omega_2 = (0, 1, 0, \ldots, 0)$, and so on. An allocation is any permutation of the houses, so $x = (x_1, x_2, \ldots, x_n)$ is an allocation if each x_i is a unit vector (but not necessarily the one with a 1 at the ith place), and the sum of the x_i's is $(1, 1, \ldots, 1)$. A price vector $p = (p_1, p_2, \ldots, p_n)$ attaches a price to each house: p_i is the price on h_i, the house originally owned by person i. If person i is buying or selling according to his budget constraint, the price of any house he buys would have to be less than or equal to the price of the house he is selling; and this would give $p \cdot x_i \leq p \cdot \omega_i$, as before.)

We now revert to our house allocation notation, and proceed with formal definitions.

Let g be an allocation of houses. Let S be a subset of the set of n individuals. We will say S *blocks* g if there is some redistribution of the houses that members of S started with, redistribution that we will call g', such that $u_i(g_i') \geq u_i(g_i)$ for all i in S, and $u_i(g_i') > u_i(g_i)$ for at least one i in S. If no subset of individuals blocks g, we will say g is in the *core*. (Note that the definition of the core is formally much the same as it was in Chapter 2.)

Let g be an allocation of houses. We will say g is *Pareto optimal*, or *optimal* for short, if there is no alternative allocation g' such that $u_i(g_i') \geq u_i(g_i)$ for all i, and $u_i(g_i') > u_i(g_i)$ for at least one i. That is, g is Pareto optimal if it is not blocked by the set of all individuals

$\{1, 2, \ldots, n\}$. As in Chapter 2, it is obviously the case that if g is in the core, it is also Pareto optimal.

Let g be an allocation of houses. Let $p = (p_1, p_2, \ldots, p_n)$ be a house price vector, with p_i, the price of h_i (the house originally allocated to person i). The house going to person i under allocation g, that is, g_i, is not necessarily h_i, the one going to him originally. We will write $p(g_i)$ for the price of that particular house: that is, if i gets house 5 under the allocation g, then $p(g_i) = p_5$.

We say (g, p) is a (competitive) *market equilibrium* if, for all i, g_i maximizes u_i subject to the budget constraint $p(g_i) \leq p_i$.

Finally, we call an allocation of houses g a *jungle equilibrium* if, whenever i is more powerful than j, $u_i(g_i) \geq u_i(g_j)$. For example, if the power ordering is 1 over 2, 2 over 3, etc., then g is a jungle equilibrium if, whenever $i < j$, $u_i(g_i) \geq u_i(g_j)$.

3. A 4-Person 4-Houses Example

We now turn to an example to illustrate.

We assume there are 4 people. The initial allocation of houses is $h = (h_1, h_2, h_3, h_4)$. An arbitrary allocation of houses is a permutation of h. For example, $g = (h_2, h_1, h_4, h_3)$ swaps houses between persons 1 and 2, and also between persons 3 and 4. The number of possible allocations is the number of permutations of $n = 4$ things, which equals $4 \cdot 3 \cdot 2 \cdot 1 = 24$.

We assume the 4 individuals have the following preferences: (As in Chapter 1, we list the houses under person i's number, in person i's order of preference).

Table 4.1a.

1	2	3	4
h_2	h_4	h_1	h_2
h_1	h_3	h_3	h_4
h_3	h_2	h_2	h_1
h_4	h_1	h_4	h_3

Therefore, for example, person 1 has the following relative utility levels: $u_1(h_2) > u_1(h_1) > u_1(h_3) > u_1(h_4)$; he likes the house originally allocated to 2 best, his own house second best, and so on.

The reader can see that these preferences have some interesting characteristics. For example, if we focus on persons 1, 2, and 3, and pretend person 4 and his house are not there, the preferences reduce to:

Table 4.1b.

$\underline{1}$	$\underline{2}$	$\underline{3}$
h_2	h_3	h_1
h_1	h_2	h_3
h_3	h_1	h_2

But this looks just like the Condorcet voting paradox preferences introduced in Chapter 1. In fact, there is a nice house swapping cycle here: If these 3 people get together and talk about their preferences they will realize that person 2 can transfer his house to person 1, person 3 can transfer his house to person 2, and person 1 can transfer his house to person 3. This swap would make these 3 better off. Since such a swap exists, it is clear that the original house allocation $h = (h_1, h_2, h_3, h_4)$ is *not* in the core and is *not* Pareto optimal. Interestingly, there is another swap available with these preferences. If we focus on persons 2 and 4, and pretend that persons 1 and 3, and their houses, are not there, the preferences reduce to:

Table 4.1c.

$\underline{2}$	$\underline{4}$
h_4	h_2
h_2	h_4

Now we can see another obvious swap, which would make 2 and 4 better off. This again proves $h = (h_1, h_2, h_3, h_4)$ is not in the core and is not Pareto optimal. Note also that the 3-person swap and the 2-person swap cannot *both* be done, because person 2 is in both cycles. If he swaps his house in a deal with 1 and 3, he no longer has it to swap in a deal with 4.

Now let's consider which of the 24 possible allocations are efficient. As it turns out, there are 5 Pareto optimal allocations. One is the allocation that results from the 3-way swap among persons 1, 2, and 3. This is (h_2, h_3, h_1, h_4). To get a clear picture of this allocation, it is helpful to return to the 4-person, 4-house preference picture, and indicate in bold non-italics the (h_2, h_3, h_1, h_4) allocation:

Table 4.2.

1	2	3	4
$\mathbf{h}_2$	h_4	$\mathbf{h}_1$	h_2
h_1	$\mathbf{h}_3$	h_3	$\mathbf{h}_4$
h_3	h_2	h_2	h_1
h_4	h_1	h_4	h_3

Note that with (h_2, h_3, h_1, h_4), 2 people (persons 1 and 3) are getting their favorite houses, and the remaining 2 cannot make a mutually agreeable swap.

Another optimal allocation is generated by starting with the initial house distribution, and letting 2 and 4 do their swap. This would produce (h_1, h_4, h_3, h_2). This swap would be particularly notable because 2 and 4 both would be getting their top choices. We will call such a swap a "top swap" below, and we will see how such swaps and cycles can be used to derive market equilibria.

The 5 Pareto optimal allocations in this example are as follows:

$$(h_2, h_3, h_1, h_4)$$
$$(h_2, h_4, h_1, h_3)$$
$$(h_3, h_4, h_1, h_2)$$
$$(h_2, h_4, h_3, h_1)$$
$$(h_1, h_4, h_3, h_2)$$

The interested reader can illustrate each one of these by reproducing the original set of preferences, and, as we did in Table 4.2 above, and underlining or highlighting the houses assigned to persons 1 through 4 under each of these 5 allocations.

In their seminal paper Shapley and Scarf proved that economies like this one have non-empty cores. Since a core allocation of houses also has to be Pareto optimal, the reader can simply check the 5 Pareto optimal allocations to discover which are core allocations. For instance, (h_2, h_3, h_1, h_4) is blocked by the set $S = \{2, 4\}$. These two people can swap their original houses between themselves, giving each his favorite. For person 2, $u_2(h_4) > u_2(h_3)$, and for person 4, $u_4(h_2) > u_4(h_4)$. Therefore, (h_2, h_3, h_1, h_4) is not in the core. A few minutes of examination of the Pareto optimal allocations should convince the reader that there is one and only one core allocation in this particular example. It is (h_1, h_4, h_3, h_2).

4. Finding a Market Equilibrium with Top Cycles

Now let's consider whether or not there exists a market equilibrium of houses, given our original assignment of house 1 to person 1, house 2 to person 2, and so on. Recall that a price vector $p = (p_1, p_2, p_3, p_4)$ and an allocation of houses g is a market equilibrium if, for every person i, g_i maximizes i's utility subject to his budget constraint. His budget constraint says he can afford house g_i if its price, what we call $p(g_i)$, is less than or equal to the price of the house he starts with, p_i.

We will show how to construct a market equilibrium allocation, and an equilibrium price vector. Looking back at Tables 4.1.b and 4.1.c, we recall that these are 2 potential trading cycles in our example. The subset {1,2,3} of individuals could swap houses among themselves, making all better off, and the same is true of {2,4}.

The {2,4} swap is especially interesting, because if person 2 gives h_2 to person 4, and person 4 gives h_4 to person 2, each party is receiving his *favorite* house. (This is not the case if a swap is made within the group {1,2,3}.)

If some subset of traders can execute a swap among themselves, so that each person gets his favorite house from among all the houses that are available, we call that subset a *top trading cycle*, and we call the swap a *top swap.* A look at Table 4.1.a should convince the reader that, in that example with 4 people and 4 houses in play, $\{2, 4\}$ is the top trading cycle, and the exchange between 2 and 4 is the top swap.

The idea of a top trading cycle is easily extended beyond our 4-person 4-house example. Suppose there are n people and n houses, and assume no one is even indifferent between 2 houses. Construct a table showing orders of preference similar to Table 4.1.a. Now ignore everything in the table except for the 1st row. In our example this is:

1	2	3	4
h_2	h_4	h_1	h_2

We claim that we can always pick out a top trading cycle; no matter what the preferences and what n may be. Here is how: Start with any person i. If i likes his own house best, we are done, person i himself, that is, the set $\{i\}$, is a top trading cycle. If not, i likes someone else's house best, say h_j. Start a list with i at the left, and j next. If j likes i's house best, we are done, the set $\{i, j\}$ is a top trading cycle. If not, there must be a new person, say k, such that j likes h_k best. Add k to the list, which now reads i, j, k. Continue in this fashion. Eventually, since there are only a finite number of people and houses, the list must loop back on itself, e.g., we must have a list like i, j, k, l, m, k. Once it loops back, we have a top trading cycle, e.g., k, l, m: person k likes h_l

best, person 1 likes h_m best, and person m likes h_k best. The top cycle $\{k, l, m\}$ can then execute the obvious top swap.

In our 4 person example, if we arbitrarily start with person 1, we would list our people 1,2,4,2, and $\{2, 4\}$ would be revealed as a top trading cycle.

A top trading cycle may have just one person in it; for instance, if the top row of the preference table is

$$\begin{array}{cccc} \underline{1} & \underline{2} & \underline{3} & \underline{4} \\ h_2 & h_3 & h_4 & h_4 \end{array}$$

the top trading cycle is $\{4\}$.

Also, there can be more than 1 top trading cycle; for example, if the first row is

$$\begin{array}{cccc} \underline{1} & \underline{2} & \underline{3} & \underline{4} \\ h_2 & h_1 & h_4 & h_3 \end{array}$$

there are 2 top cycles, and if each person starts with his favorite house

$$\begin{array}{cccc} \underline{1} & \underline{2} & \underline{3} & \underline{4} \\ h_1 & h_2 & h_3 & h_4 \end{array}$$

then there are 4 top cycles.

But in any case, given n persons and n houses, there must exist (at least one) top trading cycle.

Here is how to construct a competitive equilibrium. Start with all persons and all houses. Find a top trading cycle. Assign a (single) price to each house in that 1^{st} cycle, and choose a (relatively) high price. For our example, the top trading cycle is $\{2, 4\}$, and we will let $p_2 = p_4 = 3$. Next, remove persons 2 and 4 and houses h_2 and h_4 from the lists of persons and houses. Focusing on the remaining persons and houses, construct a preference table. For our example, this is

Table 4.3.

$$\begin{array}{cc} \underline{1} & \underline{3} \\ h_1 & h_1 \\ h_3 & h_3 \end{array}$$

In this remaining population and set of houses, find a top trading cycle. In our example, it is $\{1\}$. Assign a (single) price to each house in this cycle, and choose a price lower than the price chosen previously. For instance, set $p_1 = 2$. Next, remove this person and house from the lists

of persons and houses, and repeat. In the next round, choose a price lower than the previously chosen price. In our example, for instance, choose $p_3 = 1$.

Now let the above constructed prices be the market prices (e.g., set $p = (2, 3, 1, 3)$), and let traders "go to the market" with their original houses. That is, let them choose utility maximizing houses subject to their budget constraints based on these prices.

With respect to person 1, he starts with a house worth 2. Table 4.1.a shows he would most like h_2, but it costs 3, and he cannot afford it. Of the houses he can afford, he likes his original house h_1 best; he buys it. Persons 2 and 4 each start with houses worth 3. Person 2 likes h_4 best, he can afford it; he buys it. Person 4 likes h_2 best, he can afford it; he buys it. Person 3 starts with a house worth 1. He would prefer h_1, but he cannot afford it, in fact the only house he can afford is h_3. He buys it. The result is that each person is buying the house he likes best, subject to his budget constraint, and all the houses get allocated to all the people.

In short, $p = (2, 3, 1, 3)$ and the house allocation (h_1, h_4, h_3, h_2) comprise a market equilibrium.

5. Fundamental Theorems of Welfare Economics and Jungle Economics

Recall the first fundamental theorem of welfare economics from Chapter 3. It says that a competitive equilibrium allocation is in the core, and is therefore Pareto optimal. The Chapter 3 assumptions are somewhat different than the assumptions being made here, but the market-equilibrium–implies–core-implies-Pareto-optimality result survives.

For the purposes of the results to follow we are assuming that n houses are being distributed among n people, and that no person is indifferent between any pair of houses. (The assumption of no-indifference is crucial here; without it, we can construct market equilibrium allocations that are not Pareto optimal. See Roth and Postlewaite (1977).) We now have the following in the house allocation model:

> *First Fundamental Theorem of Welfare Economics.* Let (p, g) be a market equilibrium allocation of houses. Then g is in the core, and is Pareto optimal.
>
> *Proof*: Omitted.

Let us reflect for a moment on the fact that the market allocation of houses is in the core. This means that market exchange captures the essence of being voluntary, of being non-coercive. Every possible coali-

tion has to acquiesce to a proposal for it to be in the core. Given the initial allocation of houses, and given the preferences, there are many Pareto optimal allocations (5 in our numerical example), but far fewer core allocations (1 in our numerical example). The competitive allocation must not only be Pareto optimal, it must also be in the core.

At this point we can return to our thoughts about the jungle.

Are there jungle equilibria? Of course there are. To find one, first determine the power relation. In our example, suppose the power relation is 1, 2, 3, 4; meaning 1 is strongest, 2 is second strongest, and so on. Consider an allocation constructed as follows: Ask person 1 which house he wants most. The answer is h_2. Assign h_2 to person 1 and remove person 1 and h_2 from the lists of persons and houses. Next ask person 2 which house he likes most, of the set of remaining houses, $\{h_1, h_3, h_4\}$. The answer is h_4. Assign h_4 to person 2 and remove person 2 and h_4 from the lists. Next, ask person 3 which house he likes most, of the remaining houses, $\{h_1, h_3\}$. The answer is h_1. Assign it to him, and remove person 3 and h_1 from the lists. Finally, ask person 4 which house he likes most of the one left, and the answer is h_3. Assign it to him. This process produces the allocation (h_2, h_4, h_1, h_3).

This is obviously a jungle equilibrium under the assumed power relation: person 1 likes his house h_2 more than the houses of his inferiors 2, 3, and 4. Person 2 likes his house h_4 more than the houses of his inferiors 3 and 4, and so on.

This procedure can easily be followed for any power relation among the 4 individuals. The number of such power relations, like the number of house allocations, is $4 \cdot 3 \cdot 2 \cdot 1 = 24$, since there are four ways to name the most powerful, and having named the most powerful there are 3 ways to name the second most powerful, and so on.

But the jungle equilibrium outcome is the same for many power relations. The reader can check, for example, that if the power relation is 1, 2, 3, 4, the jungle equilibrium is (h_2, h_4, h_1, h_3), and if the power relation is 1, 3, 2, 4, the equilibrium is the same.

In fact, a mechanical examination of all 24 power relations reveals that there are 5 jungle equilibria, and they are exactly the 5 Pareto optimal allocations. (It turns out that, in general, one way to identify all the Pareto optimal allocations is to follow this so-called serial dictatorship procedure; see, for example, Abdulkadiroglu and Sonmez (1998).)

There is a jungle theorem for the house allocation model that corresponds to, but is weaker than, the first fundamental theorem of welfare economics. Piccone and Rubinstein established that "efficiency also holds in the jungle." Their proof is for a divisible-goods model, and is more complex than the one that follows. We now turn to our version:

First Fundamental Theorem of Jungle Economics. Let g be a jungle equilibrium allocation. Then g is Pareto optimal.

Proof: Suppose not. Then g is Pareto-dominated by another allocation g'. In a hypothetical move from g to g' some individuals would be better off and others would stay the same. An individual would keep the same utility under our assumptions if and only if he kept the same house. We will ignore those individuals. Let V be the individuals who are made better off by a hypothetical move from g to g'. They are all evidently getting different houses. But their houses are coming from within the group.

Since houses are being shifted within V itself, and since everyone in V is being made better off by the hypothetical move from g to g', there must exist a subset of V, which we will denote $\{a, b, c\}$, such that, in the move from g to g', a's house shifts to b; b's house shifts to c, and c's house shifts to a. (We are assuming a 3-person cycle to illustrate; the actual cycle has to have at least 2 people, and at most n. Our argument obviously applies to cycles of length 2 through n.) In this hypothetical 3 way switch, every person is getting a house he likes better (under $g\prime$) than the one he had (under g).

Now consider the power relation. It must make one of $\{a, b, c\}$ the most powerful of that subset. Say it is a. But in the hypothetical shift from g to g', he's getting c's house, which he likes better than the one he has. But c is a's inferior in the power relation.

This is a contradiction, because if g is a jungle equilibrium, and person c is person a's inferior in the power relation, a cannot possibly envy c's house: if he did envy it he would already have taken it under the law of the jungle. Q.E.D.

Note the difference between the market result and the jungle result. The market leads to the core, which is based on the initial allocation of houses plus voluntary trade. Since it leads to the core, it *ipso facto* also leads to a Pareto optimal outcome. The jungle also leads to a Pareto optimal outcome, but certainly not to the core, because the essence of the jungle is power, and power overrides voluntary transactions based on initial endowments.

At this point we consider whether or not there is a second fundamental theorem of jungle economics. Such a theorem would say that if g is any Pareto optimal allocation of houses, there must be a power relation such that g is a jungle equilibrium.

The interested reader can check the 5 Pareto optimal distributions of houses are also jungle equilibrium allocations for some power relation

among the individuals in our example. To see that this is in fact the case, consider (h_2, h_3, h_1, h_4), for instance. For a power relation, let person 1 be most powerful (he chooses h_2); let person 3 be second most powerful (he chooses h_1); person 4 be third most powerful (he chooses h_4, since his superior 1 has already claimed h_2); and let person 2 be at the end of the list (he takes the remaining house h_3). So if the power relation is 1, 3, 4, 2, the jungle equilibrium will be the Pareto optimal house allocation (h_2, h_3, h_1, h_4).

While a Pareto optimal allocation is not necessarily a jungle equilibrium in general models with divisible goods, it turns out that in our house allocation model, every Pareto optimal allocation must in fact be a jungle equilibrium allocation. This is our second theorem of the jungle:

Second Fundamental Theorem of Jungle Economics. Let g be a Pareto optimal allocation of houses. Then there exists a power relation such that, given that power relation, g is a jungle equilibrium allocation.

Proof: (Following Piccione and Rubinstein). Assume g is an optimal allocation. Write iTj if, under the allocation g, person i would prefer person j's house, i.e. $u_i(g_j) > u_i(g_i)$. In this case we say "i envies j." (For more on the concept of envy, see our discussion of "fairness" in Chapter 10.)

Note that the relation T cannot cycle. For instance, we cannot have iTj, jTk, and kTi. If we did, we could arrange a swap that benefits all people in the cycle. This would contradict the assumption that g is a Pareto optimal distribution of houses.

We proceed in stages. In stage 1, we separate $\{1, 2, \ldots, n\}$ into two subsets; those who are envied by someone, and those whom no one envies. Call the latter set B_1 (for bottom-set 1). We claim that B_1 is nonempty. If not, any person k would be envied by at least one other person j. So, jTk. In turn, j would be envied by another person, say i. So, iTj. Repeating this argument would eventually produce a cycle, such as iTj, jTk, and kTi, which would be a contradiction. The individuals in B_1 will be placed (in any order) at the bottom of the power relation. In stage 2, we consider $\{1, 2, \ldots, n\} \setminus B_1$ (the set of remaining persons after removing those in B_1). We separate this set into two subsets: those who are envied by someone in this set, and those whom no one in this set envies. Call this latter set B_2 (for bottom-set 2). By arguments similar to those made above for B_1, B_2 must be nonempty. The individuals in B_2 will be placed (in any order) next to the bottom of the power relation (i.e., above the B_1 people, but below everyone else).

Continue in this fashion until everyone has been separated out. Given the power relation so defined, it is clear that iTj only if j is higher in the power relation than i. It follows that g is a jungle equilibrium allocation. Q.E.D.

We will finish this section by illustrating the logic of the second fundamental theorem of jungle economics with our example, as shown in Table 4.1.a.

Consider the Pareto optimal allocation (h_2, h_4, h_3, h_1). With this allocation, persons 2 and 4 are getting their favorite houses. Therefore they envy no one. Person 3 is getting h_3, and he envies person 4, who is getting h_1. Therefore $3T4$. Person 4 envies persons 1 and 2, and so $4T1$ and $4T2$. Nobody envies 3, but each of the others is envied by someone. Therefore place 3 at the bottom of the power relation. Then, looking only at persons 1, 2, and 4, we do this again. Again, 1 and 2 are getting their favorite houses, so they envy no one, and in particular no one envies person 4. But we still have $4T1$ and $4T2$. Now place 4 next to the bottom of the power relation, above person 3, but below the remaining pair 1 and 2. Then, looking only at persons 1 and 2, repeat the process. Neither envies the other; so place them at the top, in any order.

The conclusion of all of this is that the optimal house allocation (h_2, h_4, h_3, h_1) is a jungle equilibrium from any power relation consistent with the above argument. There are two such power relations. Going as usual from left (most powerful) to right (least powerful), they are 1, 2, 4, 3, and 2, 1, 4, 3.

6. Exercises

1 For the 4 person, 4 house example in the text, show why (h_2, h_4, h_1, h_3), (h_3, h_4, h_1, h_2), and $(h_2, h_4, h_3, h_1,)$ are not in the core.

2 For the 4 person, 4 house example in the text, show that if the power relation is 1,2,3,4, the jungle equilibrium is (h_2, h_4, h_1, h_3), and if the power relation is 1,3,2,4, the jungle equilibrium is again (h_2, h_4, h_1, h_3).

What is the jungle equilibrium if the power relation is 4,3,2,1?

3 Consider the 4 person, 4 house example in the text. For each of the 5 Pareto optimal house allocations, find a power relation for which the jungle equilibrium is the given Pareto optimal allocation.

7. Selected References

(Items marked with an asterisk (*) are mathematically difficult.)

*1. A. Abdulkadiroglu and T. Sonmez, "Random Serial Dictatorship and the Core from Random Endowments in House Allocation Problems," *Econometrica* V. 66, 1998, pp. 689-701.

This is an interesting article that uses the characterization of Pareto optimal allocations in terms of outcomes of serial dictatorships to obtain its main result, which concerns the core of an economy in which endowments are allocated at random.

*2. M. Piccione and A. Rubinstein, "Equilibrium in the Jungle," Mimeo, 2005.

This is an innovative and thorough analysis of jungle economics in various alternative models, including the house allocation model.

*3. A. Roth and A. Postlewaite, "Weak versus Strong Domination in a Market for Indivisible Goods," *Journal of Mathematical Economics* V. 4, 1977, pp. 131-137.

This articles studies versions of Pareto optimality and the core based on strict improvements (if all members of the coalition improve) or weak improvements (if only some do). In particular, the authors construct examples showing violations of the conclusions of the first welfare theorem if indifference is allowed.

*4. L. Shapley and H. Scarf, "On Cores and Indivisibility," *Journal of Mathematical Economics*, V. 1, 1974, pp. 23-28.

A mathematically rigorous treatment which includes helpful non-mathematical discussions of several different exchange models with indivisibilities.

Chapter 5

ECONOMIES WITH PRODUCTION

1. Introduction

In the last three chapters we focussed on models of exchange. In those models, the quantities of the various goods are fixed; nothing is produced. In this chapter, we will focus on a model of production. This model has firms, goods that are used as inputs in the production process, and goods that are produced as outputs. But it has no consumers. If the exchange model was half the story, the production model is the other half. Near the end of the chapter we will indicate how the two models might be merged.

The typical textbook treatment of production starts with production functions, the analogs of utility functions. These functions indicate precisely what levels of output a firm can achieve with given inputs. For instance, the Cobb-Douglas production function $q = L^{2/3}K^{1/3}$ says that with 27 units of labor (L) and 8 units of capital (K), it is possible to produce $(27)^{2/3} \times (8)^{1/3} = 9 \times 2 = 18$ units of output (q). The model of production developed below, however, is slightly more abstract than the typical production function approach.

We shall characterize firms not with production functions, but with production sets. To explain this, we first need to say something about what goods are inputs and what are outputs.

Let's assume as before that there are M goods. In the production model, some goods are used to produce other goods. These are inputs. For instance, unskilled labor, farm land, iron ore, and seed corn are used to produce, respectively, lots of things, farm products, iron, and corn. Now some of these inputs are not themselves produced. Such goods are occasionally called pure inputs. For instance, uncleared land, oil,

natural gas, and minerals in the ground are not produced, at least not by people. On the other hand, some inputs are themselves produced. These are sometimes called intermediate goods. For instance, tools, machinery, trained workers, and buildings are all to various extends produced and then used to produce other goods. Finally, some of the goods in the production model are produced only for consumption. They are not used in the production of other goods. Examples of such pure outputs or final consumption goods are easy to name: ice cream, TV sets, some books, most records, clothing (except, perhaps uniforms and work clothes), trips to Disneyworld, and so on.

The existence of intermediate goods can make the analysis of production a little complicated. For instance, a drill bit is an output for the Hughes Tool Corporation, and an input for Exxon Corporation. Gasoline is an output for Exxon Corporation, but an input on a wheat farm in Nebraska. Wheat is an output for the wheat farm, but an input for Wonder Bread Bakeries. Bread is an output for the bakery company, and probably, most often, a genuine consumption good. (Unless, of course, some is eaten by a machinist in a cafeteria of the Hughes Tool Corporation.)

Fortunately, all this complexity is nicely sorted out in our production model. Firms are characterized in our model by sets of production vectors, or input-output vectors. We shall typically index firms with the letter k. We write

$$y_k = (y_{k1}, y_{k2}, \ldots, y_{km})$$

for a production vector for the kth firm. The quantity y_{k1} is an amount of good 1: y_{k2} is an amount of good 2, $\ldots$, y_{km} is an amount of good m. The following sign convention is crucial:

If firm k uses good j as an input, y_{kj} is negative.

If firm k produces goods j as an output, y_{kj} is positive.

This simple convention is the answer to the complications of inputs, outputs, pure inputs, intermediate goods, final goods, and so on. If an entry in a production vector is positive, that good is an output for that firm; if an entry is negative, that good is an input for that firm.

The set of production vectors that are technologically feasible for firm k is called firm k's *production set*, and is written Y_k.

Let's consider an example. Suppose firm 5 can produce 1000 cars from 500 tons of steel and 2000 man-days of labor, and that cars are good 1, steel is good 2, and labor is good 3. This combination of inputs (steel and labor) and output (cars) is represented by the production vector

$$y_5 = (y_{51}, y_{52}, y_{53}) = (1000, -500, -2000).$$

The auto (output) is positive, and the steel and labor (inputs) are negative. For firm 5, the production vector $y_5 = (1000, -500, -2000)$ is technologically feasible. That is, y_5 is in Y_5. But it might be impossible for that firm to produce 1001 cars from 500 tons of steel and 2000 man-days of labor. In that case, the production vector $(1001, -500, -2000)$ is not in Y_5.

As we observed above, the sign convention sorts out inputs from outputs. But it does more: it also allows us to sensibly add up different production vectors for different firms. Suppose again that good 1 is cars, good 2 is steel, and good 3 is labor. Suppose now that we have two firms with the following production vectors:

$$y_1 = (-1, 3, -6)$$
$$y_2 = (7, -3, -14)$$

That is, firm 1 uses one car and six man-days to produce 3 tons of steel. Firm 2 uses three tons of steel and fourteen man-days to produce seven cars. Now let's add the two vectors component by component. The result we get is

$$y_1 + y_2 = (-1 + 7, 3 - 3, -6 - 14) = (6, 0, -20).$$

This says that the net output of cars between the two firms is 6, the net output of steel is 0, and the net input of labor is -20 (or the total use of labor is 20 man-days). So the vector sum, the vector derived by adding the two vectors component by component, shows net quantities.

In the analysis that follows we make several assumptions about production sets. First, we assume they are fixed; there is no technological change or progress in this simple model. Second, we assume they are independent of each other. That is, firm 1's production set Y_1 has no bearing on firm 2's production set Y_2. The firms do not interfere with or help each other technologically; there are no externalities here. One firm's choice of a production vector does not affect the technological possibilities of another firm; it does not affect the set of production vectors feasible for the other firm.

We shall also assume in our model of production that there are K firms altogether. The list of firms is fixed. (This is not as rigid an assumption as it might seem on first sight, since firms can choose to produce nothing — they need not be actively in business.) Since there are K firms, the index k for firms runs from 1 to K. A list of production vectors, with y_k in Y_k for $k = 1, 2, \ldots K$, is called a *production plan* for the economy, and is written

$$y = (y_1, y_2, \ldots, y_K).$$

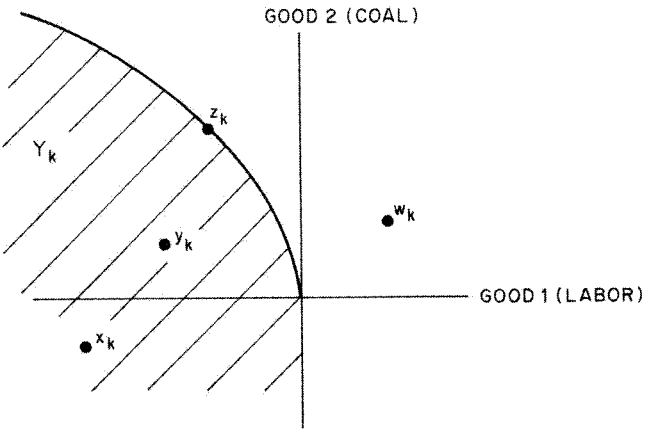


Figure 5.1.

A production plan for the economy shows what each and every firm is using as an input and producing as an output, and in what quantities. It details the state of every firm. And it is probably what in production theory most resembles an allocation in exchange theory.

With these terms defined, we are now ready to proceed to some analysis.

2. Optimal Production Plans

Figure 5.1 illustrates a production set for a firm that uses labor, good 1, to produce coal (good 2). (Obviously, most mining firms use more than labor, but for diagrams we are limited to two (or at most three goods.)

In the diagram, the cross-hatched area is firm k's production set, Y_k. That is, every production vector in that set is *technologically feasible* for the firm, and none outside that set is feasible. Evidently, this firm can take labor, good 1, and transform it into coal, good 2. A typical feasible production vector like y_k involves a negative quantity of labor and a positive quantity of coal. Labor is then the input, and coal is the output. But x_k is also feasible. At the point x_k, both the labor and coal coordinates are negative. This means the firm is using coal and labor and producing nothing. (It might be buying coal from a neighboring mine, and using labor to burn the coal it buys!) Clearly, x_k is a silly point; no firm would want to be there. But it is feasible. On the other hand, w_k is not feasible. At w_k, both the labor and the coal coordinates are positive. That is, the firm is using nothing to produce man-hours

and coal. The magical formula for doing this has yet to be discovered, so w_k must lie outside of Y_k.

Now the production vector y_k makes a lot more sense than the silly production vector x_k. However, y_k itself is undesirable in the sense that other feasible production plans, such as z_k, use less labor and produce more coal. Given our sign convention, both coordinates of preferable points like z_k are greater than or equal to the corresponding coordinates of y_k, and at least one is greater:

$$y_{kj} \leq z_{kj} \text{ for all } j$$

and

$$y_{kj} < z_{kj} \text{ for at least one good } j$$

Points like z_k which are not undesirable in the above sense, are called *technologically efficient* or *optimal production vectors* for the firm. They lie on the northeast frontier of the production set. These are the points with the property that the output of one good cannot be increased unless the output of another good is decreased, or the input of yet another good is increased.

The characterization of optimal production plans for the economy is more complicated, because what is an output for one firm might be an input for another.

Let $y = (y_1, y_2, \ldots, y_K)$ be a production plan for the economy. Since $y_{kj} > 0$ if firm k produces good j while $y_{kj} < 0$ if firm k uses good j,

$$\sum_{k=1}^{K} y_{kj}$$

is the net output in the economy of good j, if it is positive. If it is negative, it is the net input in the economy of good j. If it is zero, the total amount of good j produced by various firms equals the total used by other firms. If z is another production plan, and

$$\sum_{k=1}^{K} y_{kj} \leq \sum_{k=1}^{K} z_{kj},$$

then either:

1. Good j is a net output under both plans, and there is an equal or larger net output under z than under y.

2. Good j is a net input under both plans, and the net use of the good is equal or less under z than under y.

3. Good j is a net input under y, but a net output under z.

We shall assume that all goods, whether pure inputs, intermediate goods, or pure outputs are desirable. Because of this, it is always better to produce more (if a net output) or use less (if a net input). Therefore, if we consider only good j, we conclude that production plan z is at least as good as production plan y if

$$\sum_{k=1}^{K} y_{kj} \leq \sum_{k=1}^{K} z_{kj}.$$

If we think of all goods together, we must conclude that production plan z is at least as good as production plan y if

$$\sum_{k=1}^{K} y_{kj} \leq \sum_{k=1}^{K} z_{kj} \text{ for all } j.$$

For if the inequality holds for all goods, the net quantities of produced goods are as large under z as under y, and the net quantities of goods used are as small under z as under y.

Next, we say that a production plan z *technologically dominates* (and we'll call it *dominates* for short in this chapter) a production plan y if

$$\sum_{k=1}^{K} y_{kj} \leq \sum_{k=1}^{K} z_{kj} \text{ for all } j$$

and

$$\sum_{k=1}^{K} y_{kj} < \sum_{k=1}^{K} z_{kj} \text{ for at least one good } j.$$

And finally, we say that a *production plan* for the economy is *optimal* if there exists no other production plan for the economy that dominates it. If z is optimal, there is no way to get more of an output without using more of some input or producing less of some other output. Every change from z has a real cost attached to it. There is no "free lunch."

Let's consider our two-firm, three-goods example again, in which good 1 is cars, good 2 is steel, and good 3 is labor. Recall that the production plan y is given by

$$y_1 = (-3, 3, -6)$$
$$y_2 = (7, -3, -14).$$

Now let

$$z_1 = (-2, 4, -6)$$
$$z_2 = (8, -3\frac{1}{2}, -14).$$

Assume that firm 1's production set is $Y_1 = \{y_1, z_1\}$. That is, firm 1 has two and only two production vectors available to it. The first production vector is y_1. The second z_1 uses two cars and six man-days to generate 4 tons of steel. Since these are, by assumption, the only possibilities available to firm 1, they are clearly both optimal production vectors for firm 1. Assume that firm 2's production set is $Y_2 = \{y_2, z_2\}$. That is, firm 2 has two and only two production vectors available to it. We are familiar with y_2; z_2 uses three and a half tons of steel and fourteen man-days to produce eight cars. Again, both production vectors are clearly optimal for the firm.

Now consider the production plan $y = (y_1, y_2)$. Is it optimal? Although each firm's vector is optimal for it, the combination is not optimal for the economy. For the net quantities under y are

$$y_1 + y_2 = (6, 0, -20).$$

The net output of cars is 6, the net output of steel is 0, and the net input of labor is 20. However, the net amounts under z are

$$z_1 + z_2 = \left(-2 + 8, 4 - 3\frac{1}{2}, -6 - 14\right) = \left(6, \frac{1}{2}, -20\right).$$

Under plan z there is a net output of steel equal to $1/2$ ton, while the net output of cars and the net input of labor are the same as under y. So z dominates y and y is not an optimal production plan for the economy. The plan z is optimal, however.

3. Competitive Equilibrium Production Plans

In the section above there was no explanation of why firms might want to choose one production vector rather than another. Now we examine the behavioral assumption, the motive: We shall assume that owners or managers of firms attempt to maximize profits. We shall also assume that they act competitively, that is, they take the prices of goods as given by the market.

What is profit? The general definition is revenue less costs. Revenue for a firm is the aggregate value of the goods that the firm sells. Cost is the aggregate value of the goods that the firm buys. In terms of our model, revenue is the aggregate value of the outputs of a firm. Cost is

the aggregate value of the inputs. Now if good j is an output for firm k, then $y_{kj} > 0, p_j y_{kj}$ represents revenue, and the contribution to profit from the firm's sale of good j is

$$p_j y_{kj}.$$

On the other hand, if good j is an input for firm k, then $y_{kj} < 0, p_j y_{kj}$ represents cost, but the contribution to profit from the firm's use of good j is again

$$p_j y_{kj}.$$

Note that $p_j y_{kj}$ is a negative number, since y_{kj} is negative.

Now consider the sum of terms

$$p_1 y_{k1} + p_2 y_{k2} + \ldots + p_m y_{km} = \sum_{j=1}^{m} p_j y_{kj} = p \cdot y_k.$$

Some of the terms are positive and some of the terms are negative. The positive numbers are the contributions to revenue, and the negative numbers are the contributions to cost. Consequently,

$$p \cdot y_k$$

is equal to profit, or revenue less cost, for firm k, when it produces y_k.

We assume in what follows that firm k chooses y_k to maximize $p \cdot y_k$.

To illustrate profit maximization, we consider the firm represented in Figure 5.1. That firm uses labor, good 1, to produce coal, good 2. It maximizes

$$p \cdot y_k = p_1 y_{k1} + p_2 y_{k2}.$$

Now $p_1 y_{k1}$ is a negative number, representing expenditures on labor. On the other hand, $p_2 y_{k2}$ is a positive number, representing receipts from coal sales. If we set

$$p_1 y_{k1} + p_2 y_{k2} = c$$

we get the equation for the locus of production vectors for which profit equals the constant c. In this two-good case, that locus of points is a straight line with slope $-p_1/p_2$. Such straight lines, which are geometrically analogous to budget lines in exchange theory, are called *isoprofit lines*. Figure 5.2 shows the production set Y_k and isoprofit lines for firm k.

In the figure, ℓ_1, ℓ_2 and ℓ_3 are three different isoprofit lines. It should be clear that lines farther to the right represent higher profit levels (or higher c's in the equation $p_1 y_{k1} + p_2 y_{k2} = c$), at least when prices are positive. Consequently, ℓ_3 represents the highest profit level, of the three

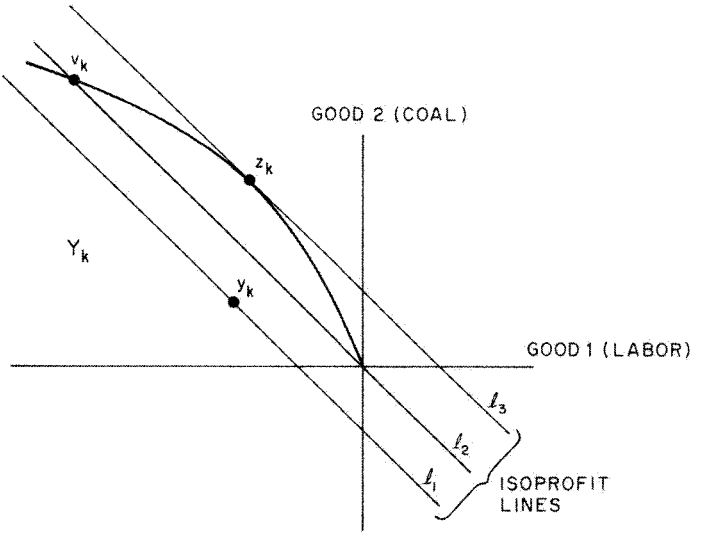


Figure 5.2.

isoprofit lines drawn. The firm's assumed behavior is to choose the production vector which puts it on the highest possible isoprofit line. In Figure 5.2, this means the firm will choose the production vector z_k.

Note that profits at z_k are higher than at the optimal production vector v_k, or at the nonoptimal production vector y_k. Also note that the profit-maximizing point z_k must be an optimal production vector for firm k. So profit maximization causes the firm to choose a vector optimal for it. But we've seen above that two production vectors optimal for each of the two firms might not be optimal for the economy. Optimality for a production plan for the economy requires more than just optimality for each production vector of each firm. It requires coordination among the firms.

In the model of an exchange economy, a competitive equilibrium is a state in which all individuals are maximizing their utilities subject to their budget constraints. In the production model, we define a competitive equilibrium as follows.

Let p be a given vector of prices of the m goods. Suppose that, for $k = 1, 2, \ldots, K$, the production vector y_k maximizes the k^{th} firm's profit. That is, $p \cdot y_k$ is at a maximum for production vectors in Y_k. Let y be the production plan $(y_1, y_2, \ldots, y_k)$. Then y and p constitute a *competitive equilibrium* for the production economy. The essence of a competitive

equilibrium is that there is a price vector p, and every firm is maximizing its profits given p.

Parallels with the exchange model are clear. In both the exchange model and the production model there is one price vector that everyone takes as given. In the exchange model, consumers maximize utility. In the production model, firms maximize profit. However, utility maximization is done subject to the budget constraint. Profit maximization is done subject to the constraint that y_k be technologically feasible, that is, in Y_k. In fact, there is no budget constraint for a firm. Also, in the exchange model, the total quantities of all the goods are fixed. In the production model, the total quantities are not fixed; only the Y_k's are fixed.

In a competitive equilibrium, each firm maximizes its profit subject to the given price vector p. We've already observed in Figure 5.2 that this implies that each firm will choose a production vector optimal for it. The question we must now ask is whether the one price vector, taken as given by all the firms, will serve to coordinate the firms' production decisions. Will y be an optimal production plan for the economy? Or, like the (y_1, y_2) in the two-firm, three-goods example above, will it be dominated by some other production plan?

4. The First Fundamental Theorem of Welfare Economics, Production Version

We are now ready to state and prove the production version of the First Fundamental Theorem. This extremely important result says that a competitive equilibrium production plan must be optimal. The competitive equilibrium price vector brings about perfect coordination of the firms' activities. In other words, in a competitive equilibrium there is no slack. It is impossible to increase the net output of one good without reducing the net output of another or increasing the net input of a third. Even though each firm is acting independently of the others, there is no possibility of unambiguous benefit through further coordination or through central planning.

The remarkable nature of this result must be emphasized. In the United States there are thousands of firms operating in fifty states. If they are profit-maximizing competitive firms and they see the same vector of prices, their decisions are optimal for the economy. Most will never communicate with each other. The firms in Rhode Island will, by and large, be unaware of the existence of the firms in Idaho. Yet, if a Central Authority had information about every firm on its computer, it could not unambiguously improve upon the competitive equilibrium. The best computer coupled with the largest staff of planners could not find slack

in the system. Even though the competitive mechanism is decentralized and apparently chaotic, it generates an optimal plan for the economy.

Here is the formal statement and proof of the theorem.

First Fundamental Theorem of Welfare Economics, Production Version. Suppose all prices are positive, and (y, p) is a competitive equilibrium. Then y is an optimal production plan for the economy.

Proof. Suppose instead that y is not an optimal production plan. We will show this leads to a contradiction.

If y is not a production optimum, there is a production plan for the economy $z = (z_1, z_2, \ldots, z_K)$ that dominates it. This means that

$$z_k \text{ is in } Y_k \text{ for } k = 1, 2, \ldots, K \tag{i}$$

$$\sum_{k=1}^{K} z_{kj} \geq \sum_{k=1}^{K} y_{kj}, \text{ for all goods } j \tag{ii}$$

$$\sum_{k=1}^{K} z_{kj} > \sum_{k=1}^{K} y_{kj} \text{ for at least one } j. \tag{iii}$$

If we multiply each of the inequalities in (ii) and (iii) by the corresponding positive price p_j, we get

$$p_j \sum_{k=1}^{K} z_{kj} \geq p_j \sum_{k=1}^{K} y_{kj} \text{ for all } j$$

with the strict inequality holding for at least one. Adding over all j then gives

$$\sum_{j=1}^{m} p_j \sum_{k=1}^{K} z_{kj} > \sum_{j=1}^{m} p_j \sum_{k=1}^{K} y_{kj}$$

or

$$\sum_{k=1}^{K} \sum_{j=1}^{m} p_j z_{kj} > \sum_{k=1}^{K} \sum_{j=1}^{m} p_j y_{kj}$$

or

$$p \cdot z_1 + p \cdot z_2 + \ldots + p \cdot z_K >$$

$$p \cdot y_1 + p \cdot y_2 + \ldots + p \cdot y_K.$$

Consequently, for at least one firm, say the k^{th},

$$p \cdot z_k > p \cdot y_k.$$

But since z_k is feasible by (i), this means firm k is not maximizing its profits at y_k, a contradiction. Q.E.D.

The First Fundamental Theorem obviously supports a *laissez faire* economic policy, since it implies that without interference, profit maximizing by competitive firms will bring about an optimal production plan. It seems to suggest that policy makers should leave competitive economic systems alone. They work. As the Sage said, if it works, don't fix it.

However, there are generally many optimal production plans. Some might involve large net outputs of housing, hospitals, food, and eduction. Some might involve large net outputs of automobiles, amusement parks, cigarettes, and liquor. To say that a production plan is optimal might not be enough. Most of us feel that there are good optimal production plans and there are not-so-good optimal production plans. Consequently, the First Fundamental Theorem leaves questions unanswered. Perhaps *laissez faire* isn't the best approach after all. It is a fine thing that the competitive mechanisms will bring the economy to a production optimum. But what if we want a different production optimum? These doubts about the first theorem bring us to the second.

5. The Second Fundamental Theorem of Welfare Economics, Production Version

In an economy it is wasteful for the productive sector to operate in a nonoptimal way. There is no point in accepting one production plan for the economy when there is an alternative plan for which outputs are greater and/or inputs are less. Whether all decisions are made by individual owners and managers of firms, operating in a climate of *laissez faire*, or whether crucial decisions are made by a Central Planning Board attempting to attain socially desirable production goals, optimality for production is necessary for sensible decisions. The most preferred production plan must be an optimal one.

At this point we need some more notation. Let y be a production plan for the economy. In order to represent the net inputs and outputs given y, we use the following definitions:

$$\begin{aligned} y_1^* &= y_{11} + y_{21} + \ldots + y_{K1} \\ y_2^* &= y_{12} + y_{22} + \ldots + y_{K2} \\ &\vdots \\ y_m^* &= y_{1m} + y_{2m} + \ldots + y_{Km} \end{aligned}$$

That is, y_1^* is the net amount of good 1 produced (if it is positive) or used (if it is negative) under the production plan y. It is the bottom line quantity for good 1. Similarly, y_2^* is the net amount of good 2 produced (if it is positive) or used (if it is negative) under the production plan y. It is the bottom line quantity for good 2. And so on, through good m. In mathematical language, we are defining $y^* = (y_1^*, y_2^*, \ldots, y_m^*)$ to be the *vector sum* of $y_1, y_2, \ldots, y_K$:

$$y^* = y_1 + y_2 + \ldots + y_K.$$

Now suppose that the competitive equilibrium production plan in a *laissez faire* economy, an economy with no political direction, is x. Then the corresponding list of net amounts is x^*. Suppose that x^* is heavy on luxury cars, liquor, and so on, and light on housing, food, etc. Let us assume that the nation's political leaders reject x^*, and the associated x. Suppose they decide that some other list of net amounts y^* is superior; let y be the associated production plan for the economy.

We are assuming here that y^* is possible, in the sense that there does exist an associated production plan y for the economy, with y_k in Y_k for every k. Sometimes some leaders decide on nonfeasible quantities of goods, in which case there is just no way that their plans can be realized. We'll stay away from the Rolls-Royce-in-every-garage, pie-in-the-sky production plans, often promised in electoral campaigns, in this analysis. We are also assuming that y is optimal; it is silly to accept a non-optimal goal.

One way to achieve y^* would be to have the economy's Central Planning Board send directives to each and every firm telling that firm exactly how much of each input it should use, and exactly how much of each output it should produce. The Board could, in other words, tell every firm what y_k to produce. Let's call this system *centralized socialism.* This system was the one used in the defunct Soviet Union to organize many of its productive activities.

Is centralized socialism practical? There are two reasons why it is not. First, it requires that the Central Planning Board send out enormous quantities of information. All the amounts of all goods used and produced by all firms are decided upon, and emanate from one Board. The information processing and transmitting requirements are prohibitive. Second, it destroys the incentives of the owners and managers of the firms. The people on the spot, directing day-to-day operations, are passive; they receive instructions and simply implement them. Consequently, they have no reasons to search for superior production processes, they have no incentives to increase outputs or reduce inputs. If the Central Planning Board mistakenly sends firm k a production vector that

is not optimal for that firm, the manager has no motive to find a production vector that dominates it. Often the on-the-spot manager has information about his firm that the Central Planner doesn't, but under centralized socialism he has no incentive to use superior information or to reveal it to the Central Planner.

These practical difficulties of centralized socialism bring us back to the question. Suppose the vector of net amounts x^* associated with *laissez faire* is objectionable, and another list of net amounts y^*, associated with an optimal production plan y, is desired by the planners in an economy. Is there some way to achieve y^*? The Second Fundamental Theorem of Welfare Economics says that there is a way. The alternative to centralized socialism is to have the Central Planning Board issue a list of prices p, and instruct the managers of firms (or state enterprises) to maximize revenues less costs, or profits, given the prices p. We shall call this system *decentralized socialism.* Decentralized socialism re-injects the profit motive into the planned economy. It harnesses competitive forces to attain a desired list of net quantities y^*. It avoids massive intervention in the minute decisions of every firm. Many of the reforms introduced in China starting in the 1990's are based on decentralized socialism.

Formally, the theorem says that given any optimal y, there exists some price vector p so that y and p constitute a competitive equilibrium.

In practice, the Planning Board has a y^* in mind. The theorem says that if the associated y is optimal (which makes sense), the Board can achieve it via a decentralized system, if it can find the right price vector p. The way p might actually be found probably involves trial and error; over a period of months or years the board might gradually adjust p until the desired y^* finally appears. Once the right p is hit upon, small adjustments can be made to bring about further small desired changes in y^*.

In order to provide a formal statement of the theorem, we need to define two technical terms. First, a production set Y_k is said to be *convex* if the following is true: For any two points x_k and y_k in Y_k, the straight line segment connecting those two points is entirely contained in Y_k. The reader can refer to Figure 5.1 to see a convex production set. (The reader can also refer to the definition of a convex indifference curve and Figure 3.6, which illustrates that definition. In Figure 3.6a, the set of bundles on or above the indifference curve is a convex set, while in Figure 3.6b the set of bundles on or above the indifference curve is not convex.) Second, a production set Y_k is said to be *closed* if it includes its boundaries. The Y_k in Figure 5.1 is closed since the northeast boundary

(including z_k) is part of Y_k, as is the east boundary, the lower half of the vertical axis.

Now we turn to the formal statement of the theorem. The proof is in the Appendix to this chapter.

Second Fundamental Theorem of Welfare Economics, Production Version. Suppose all production sets are convex and closed. Let y^* be any vector of net quantities whose associated production plan y is optimal.

Then there exists a price vector p, not equal to the zero vector, such that y^* could be achieved using p through decentralized socialism. That is, there exists a p such that (y, p) is a competitive equilibrium.

6. Extending the Production Model, and Combining Production and Exchange

The discussion above leaves certain important issues unclear. Let y be a production plan, and y^* the corresponding list of net quantities. Then $y^* = (y_1^*, y_2^*, \ldots, y_m^*)$ includes some positive numbers (for goods that are net outputs in the economy) and some negative numbers (for goods that are net inputs). Where then do the net inputs come from? Where do the net outputs go? The model is incomplete.

One way to make it more complete is to assume that there are certain fixed amounts of net inputs available to the economy, and that these amounts are owned by the Central Planning Board. This clearly makes for a fuller model in the socialist spirit, but it should be observed that this fuller model remains unfinished: the question of where the net outputs go remains unanswered. Let $s^* = (s_1^*, s_2^*, \ldots, s_m^*)$ be a vector of quantities of goods owned by the Board. All the s_j^*'s are non-negative; some might be zero. The positive quantities can be interpreted as amounts of available natural resources, like mineral ores, and farmland, or as capital equipment, and so on. One good might be interpreted as labor, if the Board owns labor in some meaningful sense. Now there is a new feasibility condition that any production plan must fulfill. In addition to the requirement that y_k be in Y_k for every k, it is necessary that the net inputs of any production plan be covered by the amounts available. That is, for any y we must have

$$s_j^* + y_j^* \geq 0 \text{ for all } j.$$

The immediate effect of this condition is to constrain the set of feasible net or aggregate production vectors; it makes the aggregate production set smaller. It also changes the set of optimal production plans, since to

be optimal a production plan must first be feasible. The second effect of the condition is to modify the notion of a competitive equilibrium. Up till now, a competitive equilibrium production plan has simply been one with the property that y_k maximizes firm k's profits, for all k. But the constraint that $s_j^* + y_j^* \geq 0$ for all j introduces a supply and demand condition for net input goods. If for some price vector p the aggregate net demand for good j exceeds the supply s_j^*; that is, if $y_j^* < 0$ and $s_j^* + y_j^* < 0$, then the equilibrium is physically impossible. Consequently, $s_j^* + y_j^* \geq 0$ for all j is a necessary condition for a competitive equilibrium. We are led, then, to the following revision of the definition of a competitive equilibrium; y and p constitute a competitive equilibrium if (1) for all k, y_k maximizes firm k's profits over the production set Y_k, and (2) for all goods j, $s_j^* + y_j^* \geq 0$.

With these modified definitions of optimal production plans and competitive equilibrium, the two fundamental theorems go through essentially unchanged.

We should observe at this point that this model might be desocialized in spirit by assuming that s^* is privately owned, rather than owned by a Central Planning Board.

Private ownership of resources, however, brings us to a yet more complex, and fuller, model. For if resources are owned by individuals, those individuals presumably take the payments they receive for what they sell (or rent) to firms, and spend those payments on the net outputs of firms. To reflect this complexity, the models of exchange and of production must be merged. We shall indicate in general terms what such a merged production/exchange private-ownership model looks like.

Let us now assume that goods used as net inputs, and the firms themselves, are owned by private individuals. Person i's initial bundle is, as before, ω_i. We assume, as before, that $i = 1, 2, \ldots, n$. Now the initial quantities of goods in the economy are given by the following equations:

$$\begin{aligned} s_1^* &= \omega_{11} + \omega_{21} + \ldots + \omega_{n1} \\ s_2^* &= \omega_{12} + \omega_{22} + \ldots + \omega_{n2} \\ &\vdots \\ s_m^* &= \omega_{1m} + \omega_{2m} + \ldots + \omega_{nm} \end{aligned}$$

In easy vector shorthand, we have

$$s^* = \omega_1 + \omega_2 + \ldots + \omega_n = \sum_{i=1}^{n} \omega_i.$$

But person i owns more than ω_i at the start, he also owns shares of firms. A share of a firm gives him a right to a fraction of the profit of

that firm. Let θ_{ik} be person i's fractional ownership of firm k. Since the sum of all fractional shares must be equal to 1, we have

$$\sum_{i=1}^{n} \theta_{ik} = 1, \text{ for } k = 1, 2, \ldots, K.$$

Now if firm k makes a profit of $p \cdot y_k$, person i receives $\theta_{ik}(p \cdot y_k)$. Since person i owns a fraction of every firm (some of these fractions obviously might be zero), his total income from profits of firms is

$$\sum_{k=1}^{K} \theta_{ik}(p \cdot y_k).$$

This profit income must appear on the income side of i's budget constraint. Therefore, in the production exchange private-ownership model, person i's budget constraint is

$$p \cdot x_i \leq p \cdot \omega_i + \sum_{k=1}^{K} \theta_{ik}(p \cdot y_k).$$

As before, x_i is a consumption bundle for i. It is assumed, of course, that i tries to maximize his utility subject to his budget constraint, so:

I. Each person i chooses x_i to maximize $u_i(x_i)$, subject to

$$p \cdot x_i \leq p \cdot \omega_i + \sum_{k=1}^{K} \theta_{ik}(p \cdot y_k).$$

Firms, of course, attempt to maximize profits. So we have:

II. Each firm k chooses y_k in Y_k to maximize $p \cdot y_k$.

Finally, goods must add up. Total demands by individuals for the m goods are given by:

$$\begin{aligned} &x_{11} + x_{21} + \cdots + x_{n1}, \text{ for good 1} \\ &x_{12} + x_{22} + \ldots + x_{n2}, \text{ for good 2} \\ &\vdots \\ &x_{1m} + x_{2m} + \ldots + x_{nm}. \text{ for good } m \end{aligned}$$

In vector notation, this list of total demands is

$$x_1 + x_2 + \ldots + x_n, \text{ or } \sum_{i=1}^{n} x_i.$$

Total initial quantities are given by

$$s^* = \sum_{i=1}^{n} \omega_i.$$

Net amounts produced (or used, if the quantities are negative) by the firms are given by the vector equation

$$y^* = y_1 + y_2 + \ldots + y_K = \sum_{k=1}^{K} y_k.$$

The supply equals demand condition can then be written in compact vector form as follows:

III. $$\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} \omega_i + \sum_{k=1}^{K} y_k$$

Note that the sign convention on the y_{kj}'s ensures that the righthand side of III makes sense: if the firms are, in aggregate, producers of good j, then the j^{th} component of

$$\sum_{k=1}^{K} y_k \text{ is positive;}$$

if the firms are, in aggregate, users of good j, then the j^{th} component of

$$\sum_{k=1}^{K} y_k \text{ is negative.}$$

In this production/exchange private-ownership model, a *competitive equilibrium* is (I) a price vector p, (II) a vector of consumption bundles $x = (x_1, x_2, \ldots, x_n)$, one for each person, and (III) a production plan for the economy $y = (y_1, y_2, \ldots, y_K)$, that satisfy I, II, and III above.

Let's now call a consumption-production plan given by a vector of consumption bundles $x = (x_1, x_2, \ldots, x_n)$ and a production plan for the economy $y = (y_1, y_2, \ldots, y_K)$ feasible if: $x_{ij} \geq 0$ for all i and j, y_k is in Y_K for all k, and III is satisfied. A feasible consumption-production plan x, y is *dominated* if there is another feasible consumption-production plan x', y' with the property that all individuals like x' as well as x, are some individuals like x' better than x. Note that dominance is once again defined in terms of individuals' utility levels, as it was in the model of exchange, rather than in terms of output quantities. Finally, we call feasible consumption-production plan (x, y) *Pareto optimal* if it is undominated by any feasible plan.

With this rather general model, and with these definitions, the two fundamental theorems can be rigorously formulated and proved. In essence, they are similar to the partial versions given above. The First Fundamental Theorem says that a competitive equilibrium allocation in the production/exchange economy is Pareto optimal (this statement can be strengthened to get inclusion in the core, for some appropriate definition of blocking in economies with production). And the Second Fundamental Theorem says that virtually any Pareto optimal consumption-production plan can be reached via a modified competitive mechanism. With the appropriate cash transfers, and/or stock ownership transfers, virtually any Pareto optimal arrangement can be achieved.

7. An Algebraic Example in a Simple Production Model

We return in this section to a simple production model and provide an algebraic example. Suppose there are two firms, and assume there are just two goods. We assume that firm 1 can use good 1 to produce good 2 according to a square root production function: its (maximum) output of good 2 is equal to the square root of the input of good 1. Formally the production function is

$$y_{12} = \sqrt{-y_{11}}.$$

We need the minus sign before the y_{11} because good 1 is used by firm 1 as an input, so y_{11} is a negative number. To define firm 1's production set Y_1, we allow for the fact that firm 1 can waste some of the input, so the output is at most equal to the square root of the input. Y_1 is then the set of points satisfying

$$y_{12} \leq \sqrt{-y_{11}}.$$

The set Y_1 is sketched in Figure 5.3a.

It turns out that the absolute value of the slope of the northeast boundary of Y_1 is given by the formula

$$\frac{1}{2}\frac{1}{\sqrt{-y_{11}}}.$$

When firm 1 maximizes its profits, it finds the point where its production set touches the highest isoprofit line ℓ_2. At this point, the boundary of Y_1 is tangent to the isoprofit line ℓ_2; that is, the slope of the boundary equals the slope of the isoprofit line, or

$$\frac{1}{2}\frac{1}{\sqrt{-y_{11}}} = \frac{p_1}{p_2}.$$

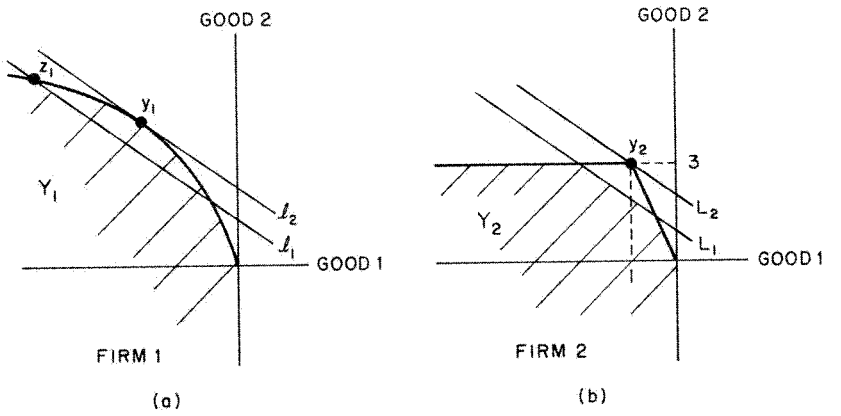


Figure 5.3.

For convenience, we take p_2 to be 1, and solve for $-y_{11}$:

$$-y_{11} = \frac{1}{4p_1^2}.$$

Let's assume that the production set Y_2 has the angular shape sketched in Figure 5.3b (which may not be drawn to the same scale as Figure 5.3a.) Formally, Y_2 is the set of points satisfying

$$y_{22} \leq \begin{cases} -3y_{11} \text{ for } -1 \leq y_{11} \leq 0 \\ \\ 3 \text{ for } y_{11} < -1 \end{cases}$$

Firm 2 maximizes its profits by finding the point where its production set touches the highest isoprofit line L_2. At this point, however, the slope of the isoprofit line is not equal to the slope of the boundary of Y_2, since there is a kink in that boundary at y_2 and its slope is undefined.

The isoprofit lines ℓ_1, ℓ_2 and L_1, L_2 have been drawn with the same slope since we assume both firms face the same price vector p. The pair $y = (y_1, y_2)$ and p constitute a competitive equilibrium in this economy. By the First Fundamental Theorem, y is an optimal production plan.

By the Second Fundamental Theorem, any optimal production plan can be achieved as a competitive equilibrium. That is, for any optimal production plan y there is a price vector p such that y and p are a competitive equilibrium. For instance, the production plan (z_1, y_2) happens to be optimal. The theorem says the two firms can be induced to operate at (z_1, y_2), given the right prices. What prices? At z_1, the absolute

value of the slope of the boundary of Y_1 is

$$\frac{1}{2}\frac{1}{\sqrt{-z_{11}}}.$$

If p_1/p_2 is chosen equal to this, one of firm 1's isoprofit lines will just touch Y_1 at z_1. So firm 1 will want to produce at z_1. Firm 2, with new flatter isoprofit lines, will still want to produce at y_2. Thus

$$(p_1, p_2) = \left(\frac{1}{2}\frac{1}{\sqrt{-z_{11}}}, 1\right)$$

and (z_1, y_2) constitute a competitive equilibrium, and the Central Planning Board can move the economy to (z_1, y_2) by announcing the new prices.

8. Exercises

1. There are three firms and three commodities. Good 1 is a composite consumption good, good 2 is capital, and good 3 is labor. Firm 1 can produce the consumption good from capital and labor according to the (Cobb-Douglas) rule:

$$y_{11} \leq (-y_{12})^{1/3}(-y_{13})^{2/3}$$

Firm 2 can produce the consumption good from capital and labor according to the rule:

$$y_{21} \leq -\frac{y_{22}}{3} - \frac{2y_{23}}{3}$$

Firm 3 can produce capital from labor according to the rule:

$$y_{32} \leq (-y_{33})^{1/2}$$

   a. Find a production vector for each firm (other than (0,0,0)), which is feasible for that firm, and which lies on the northeast boundary of the production set for that firm.

   b. Given the production vectors of part (a), what are the net inputs and outputs of the three goods, for the entire economy? Can you adjust your plan so the net output (or input) of capital is zero?

   c. Show that the production plan for the economy which you have chosen is not optimal. (Hint: This can be done by trial and error: you might start by having firm 1 use 1 unit less of capital and firm 2 use 1 unit more; if this doesn't work, continue with similar small switches. The odds are very small that the y_1, y_2, and y_3

you picked in part (a) constitute an optimal production plan for the economy.)

2. In the economy of problem 1, suppose the prices of the three goods are (3, 1, 2). Find profit-maximizing production vectors for the three firms.

   (Note: This problem requires some familiarity with the theory of the firm and profit maximization. Those familiar with these topics can solve the problem if they recall that for a Cobb-Douglas production function $q = K^{1/3}L^{2/3}$, the marginal rate of substitution of capital for labor is $2K/L$.)

9. Appendix

In order to prove the Second Fundamental Theorem, we need to use a mathematical result known as Minkowski's Theorem. Let $S_1, S_2, \ldots, S_K$ be sets of m dimensional vectors. The sum S of the sets S_1 through S_K is defined as follows: An m dimensional vector x is in the sum, or

$$x \text{ is in } S = S_1 + S_2 + \ldots + S_K$$

whenever

$$x = x_1 + x_2 + \ldots + x_K$$

and x_1 is in S_1, x_2 is in $S_2, \ldots, x_K$ is in S_K

The version of Minkowski's Theorem that we need goes as follows:

Minkowski's Theorem. Let $S_1, S_2, \ldots, S_K$ be closed convex sets of m dimensional vectors. Suppose $S = S_1 + S_2 + \ldots + S_K$, and assume that the vector s^* is on the boundary of S. Let $s^* = s_1 + s_2 + \ldots + s_K$, with s_1 in S_1, s_2 in S_2, and so on.

Then there exists an m dimensional vector p, not equal to the zero vector, such that, for $k = 1, 2, \ldots, K$,

$$p \cdot x_k \leq p \cdot s_k \text{ whenever } x_k \text{ is in } S_k.$$

Proof of the Second Fundamental Theorem, Production Version. Define the aggregate production set Y as follows:

$$Y = Y_1 + Y_2 + \ldots + Y_K$$

The Central Planning Board wants to achieve y^*, a vector in Y, through decentralized socialism. Let $y = (y_1, y_2, \ldots, y_K)$ be the associated production plan. Then

$$y^* = y_1 + y_2 + \ldots + y_K.$$

Also, y_1 is in Y_1, y_2 is in Y_2, and so on, and, by assumption, y is an optimal production plan.

To apply Minkowski's Theorem, we need to establish that y^* is on the boundary of Y. If to the contrary y^* is not on the boundary of Y, there is a point z^* in Y such that $z_j^* \geq y_j^*$ for $j = 1, 2, \ldots, m$ and $z_j^* > y_j^*$ for at least one j. But if z^* is in Y, there is a production plan $z = (z_1, z_2, \ldots, z_K)$, with z_1 in Y_1, z_2 in $Y_2, \ldots, z_K$ in Y_K, for which

$$z_j^* = \sum_{k=1}^{K} z_{kj} \text{ for all } j.$$

Now we have

$$\sum_{k=1}^{K} z_{kj} \geq \sum_{k=1}^{K} y_{kj} \text{ for all } j,$$

with the strict inequality holding for at least one j. This means the production plan z dominates the production plan y, a contradiction. Consequently, y^* must be on the boundary of Y.

By Minkowski's Theorem, there exists a nonzero price vector p such that, for all k,

$$p \cdot x_k \leq p \cdot y_k \text{ for all } x_k \text{ in } Y_k$$

This means y_k maximizes profits for the k^{th} firm. Therefore (y, p) is a competitive equilibrium. Q.E.D.

10. Selected References

(Items marked with an asterisk (*) are mathematically difficult.)

*1. G. Debreu, *Theory of Value*, John Wiley and Sons, Inc., New York, 1959, Chapters 3, 6.

Chapter 3 is an axiomatic treatment of production. It covers, among other things, assumptions on production sets, and profit maximization. The material is at a mathematically advanced level. Chapter 6 deals with competitive equilibrium and optimality — in other words, the fundamental theorems.

2. T.C. Koopmans, "Allocation of Resources and the Price System," *Three Essays on the State of Economic Science*, McGraw-Hill, New York, 1957.

Koopmans' essay is an excellent reference and it doesn't require great mathematical sophistication to understand. Chapter 1 is a good

treatment of sets, the sign convention, "separating hyperplane theorems," and the virtues of decentralization. The references in Part 1.8 are a useful guide to the history of the subject. Chapter 2 covers the fundamental theorems. In 2.1 there is a useful brief history of the idea of Pareto optimality, with good references.

*3. K. Lancaster, *Mathematical Economics*, Macmillan, New York, 1968, Chapter R4.

The student with a little mathematical sophistication who wants to understand Minkowski's Theorem can use this chapter of Lancaster's text, especially R4.3. Lancaster gives a simpler version of Minkowski's Theorem than the one we use, but the extension of his version to ours is not hard.

4. O. Lange and F.M. Taylor, *On the Economic Theory of Socialism*, The University of Minnesota Press, Minneapolis, 1939.

Oscar Lange and Abba Lerner are given much of the credit for developing the theory of what we call decentralized socialism, and for developing the Second Fundamental Theorem. Lange's essay in this book is reprinted from the *Review of Economic Studies*, V. II, October, 1936 and February, 1937. It starts out by giving credit to Ludwig Von Mises, who had previously criticized the socialists for ignoring the informational problems of running an economy from a centralized bureau. Mises's position was that economic accounting was impossible under socialism, since socialism abolished the prices that are attached to capital goods. Lange argues that the Central Planning Board can and should attach accounting prices to capital goods, and that, in fact, by attaching prices to goods the Board can direct the economy to an optimum through decentralized socialism. The Planning Board "has to fix prices and see to it that all managers of plants, industries, and resources do their accounting on the basis of the prices." "Thus, the accounting prices in a socialist economy, far from being arbitrary, have quite the same objective character as the market prices in a regime of competition."

Although Lange claims that the managers of production are "no longer guided by the aim of maximizing profit," he has the Central Planning Board impose the following rules on them: (1) cost must be minimized, and (2) price must equal marginal cost. These two rules are, of course, the ones that a profit maximizing competitive firm follows.

5. A.P. Lerner, *The Economics of Control*, The MacMillan Company, New York, 1944.

This book is a theory of what Lerner calls a "controlled" economy, which should be contrasted with a *laissez faire* economy. A controlled economy is what might now be called a "mixed" one — with public and private production sectors — although Lerner rejects that term. A controlled economy is not necessarily a collectivist, or centralized socialist, economy. It is an economy in which the "state uses its control to enable that method to prevail in each particular case which best serves the public interest."

The student should pay particular attention to Chapters 5, 6 and 7, which outline a simple theory of optimality and production in the collectivist, perfectly competitive, capitalist and controlled regimes. Chapter 5 provides a few nice remarks about the practical impossibility of centralized socialism (or collectivism), and gives a rule for successful operation of decentralized socialism (the controlled economy). The rule is, of course, logically equivalent to the rule for profit maximization for a competitive firm. Chapter 6 gives the marginal conditions necessary for an optimum in production, and indicates how perfect competition satisfies those conditions. Chapter 7 argues that the assumptions of perfect competition are in fact rarely met. Too many firms have monopoly power. Consequently, *laissez faire* will not bring about an optimum. The solution? A controlled economy in which government enterprises compete with private firms.

*6. E. Malinvaud, *Lectures on Microeconomic Theory*, American Elsevier Publishing Co., New York, 1972, Chapters 3, 4.

This is a succinct treatment of production and optimality theorems, on a graduate textbook level. It does require some mathematics, although not nearly as much as Debreu. In particular, it requires familiarity with Lagrange multipliers. Chapter 4 is especially close to the approach we have taken.

*7. A. Mas-Colell, M. Whinston and J. Green, *Microeconomic Theory*, Oxford University Press, 1995.

Chapter 4 of this excellent graduate textbook is a modern presentation of producer's theory.

Chapter 6

UNCERTAINTY IN EXCHANGE

1. Introduction

In this chapter we consider economies that operate in uncertain environments. Therefore, it will be necessary to employ the tools of decision theory under uncertainty. We will use the expected utility preferences, as defined in Chapter 1. Our concern now is how uncertainty may affect the economic activity. What are the implications of uncertainty for the different economic institutions that we have studied, such as pure barter, price-governed exchange, production decision making, etc.? Can one recover versions of the welfare theorems in these contexts? For our purposes, there is no loss of generality in excluding production, so we shall concentrate on exchange economies for simplicity.

As we shall see, what will be important for the extension of the welfare theorems is that, despite there being uncertainty, no person in the economy holds private information with respect to the others. Instances of such private information give something of a "monopoly power" to the person that holds it, and this causes difficulties for the performance of market institutions (think of insider trading, for instance). On the other hand, the conclusions are rather different If uncertainty exists but the information is always held symmetrically among individuals (e.g., today no one knows the true state of the world, while tomorrow it is observable by everyone). For this case, financial markets may work very well to deliver an efficient allocation of risk in the economy. We begin by describing the basic model of an economy with uncertainty, which uses the notion of states of the world.

2. States of the World

Suppose there are two dates in our model. At date 0 there is uncertainty concerning the true state of the world (or simply, state) that will be realized at date 1. At date 1 no uncertainty remains because the true state of the world is observable by all agents.

A state or *state of the world* is therefore a full description of everything relevant to the economy, i.e., the true preferences of all economic agents, their true initial endowments, and the true technology available to each firm (if we also included production in our model, which is not the case here). That is, in our exchange economy, there may be uncertainty at date 0 around the agents' true preferences and endowments. For example, suppose the weather at date 1 is uncertain at date 0: it might be rainy or sunny. We shall say then that there are two states of the world: the rainy state and the sunny state. In the rainy state (i.e., if it rains at date 1), agents' date 1 preferences over sun glasses and umbrellas may be quite different from those in the sunny state (if it turns out to be a beautiful sunny day). In another example, uncertainty could involve endowments. (For example, suppose there are four possible states, which vary according to the proportion of the population with a college degree. In the first state, the proportion is between 0 and 1/4; in the second, between 1/4 and 1/2; in the third, between 1/2 and 3/4; and in the fourth between 3/4 and 1. In each of these four states, there are different aggregate endowments in the population, in terms of skilled labor versus unskilled labor).

For simplicity, we shall assume that there is a finite set of possible states, $\{1, \ldots, k\}$. Each state s consists of an exchange economy with the same n traders and the same m goods. In state s, trader i's preferences over the m goods are represented by the utility function $u_{is}(x_{is})$, where $x_{is} = (x_{i1s}, \ldots, x_{ims})$ is the bundle of the m goods that trader i will consume at date 1 if state s is realized. In state s, $\omega_{is} = (\omega_{i1s}, \ldots, \omega_{ims})$ is person i's endowment of goods.

At date 0, when agents are uncertain about the state of the world to be realized at date 1, agents have expected utility preferences. That is, person i maximizes $\sum_{s=1}^{k} q_{is} u_{is}(x_{is})$, where q_{is} is the probability that person i assigns to state s.

At this point it will be useful to introduce two alternative assumptions on information that will be used in the chapter. First, we speak of *symmetric uncertainty* if the situation is such that at date 0 all agents think that each state might happen at date 1. They may differ in the probability assignments given to each state, as long as all of these probabilities are positive. (If all agents agree on the probability of a state being 0, this is also symmetric uncertainty, and without losing anything,

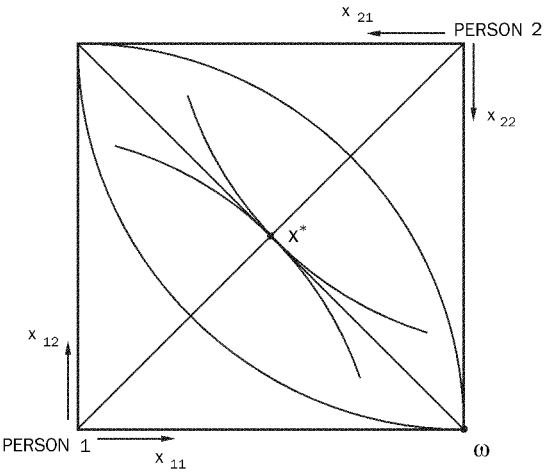


Figure 6.1.

one can remove such states from the analysis). In contrast, we speak of *asymmetric information* if for some individual, but not for all, the probability he assigns to at least one state s is 0. This means that different people have different information, in the sense that some of them know that certain things are just not true, will not happen, while other people still think that they might.

3. Symmetric Uncertainty

Let's begin with a simple example, in which we have $n = 2$ traders, $m = 1$ consumption good in each state, and $k = 2$ states. Suppose that the utility function for each person i and each state s is $u_{is}(x_{is}) = \sqrt{x_{is}}$. There is uncertainty concerning the endowment of the (only) consumption good. In state $s = 1$, agent 1 will hold one unit of the good, and agent 2 will initially have nothing. Exactly the opposite will happen in state $s = 2$. Suppose each person believes that the two states are equally likely at date 1.

First, it is interesting to note that we can still use the Edgeworth box diagram to represent this economy. See Figure 6.1. In it, the endowment is the lower right corner of the Edgeworth box, and the indifference curves depicted correspond to the expected utility function $(1/2)\sqrt{x_{i1}} + (1/2)\sqrt{x_{i2}}$ for each person i.

These individuals have an important problem to solve. If they do nothing about it, tomorrow one of them will hold all the consumption good and the other will have to starve. Furthermore, when they get

to date 1 and the uncertainty is resolved, one should see no trade if these individuals are self-interested: for each person i, his *ex-post utility function* is $\sqrt{x_{is}}$, so there is no incentive for the agent who holds the unit of the good to give any to the other.

From the perspective of date 0, this is not a good situation. Each person may well be scared at the prospect of having to starve with probability one half. Note in Figure 6.1 that the endowment point gives each of them an expected or *ex-ante utility* of $1/2$, while other points of the Edgeworth box, such as its center x^*, generate higher levels of ex-ante utility. Could they do something at date 0 to prevent this situation? Well, they could sign a contract to try to provide insurance to each other. Such a contract is a state contingent contract.

A *state contingent commodity* is a title, a piece of paper, that reads "whoever holds this piece of paper is entitled to having x units of good m if and when state s happens." The idea is to create markets where these titles, these assets, will be traded. Note how, for these state contingent commodities to be worth anything and make sense, the assumption of observability of the state at date 1 is important. Otherwise, the execution and enforcement of these contracts would be impossible. In the example of Figure 6.1, a state contingent contract could be the following: if state 1 occurs, person 2 has a claim to receive half a unit of the good to be delivered by person1, and vice versa if state 2 occurs.

We shall assume the existence of *complete markets*, which is to say that there is a market for each such state contingent commodity. There are km such markets because that's the number of state contingent commodities (the product of the number of goods times the number of states of the world). Traders will participate in these markets for financial assets at date 0. Let x_{isj} denote the number of units of contingent commodity sj purchased by agent i, i.e., if agent i purchases x_{isj} units of that state contingent commodity, he is effectively purchasing the right to consume those same units of good j if state s occurs at date 1. Let p_{sj} be the price of such a state contingent commodity.

The income that person i has to participate in these financial markets is simply the market value of his initial endowments, also interpreted as state contingent commodities, i.e., $\sum_{s=1}^{k} p_s \cdot \omega_{is}$, which can be written out as $\sum_{s=1}^{k} p_{s1}\omega_{is1} + \ldots + p_{sm}\omega_{ism}$.

A competitive equilibrium in this setting is known as an *Arrow-Debreu equilibrium*, named after economists Kenneth Arrow and Gerard Debreu. It is defined as follows: a vector of prices of state contingent commodities $p = (p_{11}, \ldots, p_{1m}, \ldots, p_{k1}, \ldots, p_{km})$ and an allocation of state contingent commodities $x = (x_1, \ldots, x_n)$, where each

$x_i = (x_{i11}, \ldots, x_{i1m}, \ldots, x_{ik1}, \ldots, x_{ikm})$ is an Arrow-Debreu equilibrium if two conditions are satisfied:

> Expected utility maximization: for each agent i, the bundle x_i of state contingent commodities maximizes his expected utility $\sum_{s=1}^{k} q_{is} u_{is}(z_{is})$ subject to the budget constraint $\sum_{s=1}^{k} p_s \cdot z_{is} \leq \sum_{s=1}^{k} p_s \cdot \omega_{is}$.
>
> Market clearing: in the market for each state contingent commodity sj, $\sum_{i=1}^{n} x_{isj} = \sum_{i=1}^{n} \omega_{isj}$.

Observe that, if each person i's ex-post utility function in state s is monotonic, then the budget constraint will be satisfied with equality in equilibrium, because at date 0 each agent will want to allocate all his income among the different state contingent commodities.

Let's try to understand the logic behind the notion of Arrow-Debreu equilibrium. According to this notion, agents try to eliminate the effects of uncertainty by insuring against it. The insurance that each person buys and sells is formulated in terms of his demand for the different state contingent commodities, which will give him the right to receive the real goods in each state. These demands take into account all relevant factors in his decision making, such as his endowment of the different commodities in each state, his beliefs about which states are more likely, and his preferences in each state. Thus, if many people believe that a particular state is very likely to happen, the prices of the state contingent commodities pertinent to that state will be high because many people will demand them, and vice versa if a state is generally deemed unlikely. Similarly, if almost no one likes to wear sun glasses when it rains, the price of that state contingent commodity (sun glasses if it rains) will be low; or if a particular good is expected to be abundant in some states, the corresponding state contingent commodities will also have a low price. In equilibrium, given all these different factors, the prices of all state contingent commodities will adjust to eliminate any excess demand or excess supply.

What can one say about the welfare properties of the Arrow-Debreu equilibrium? As we shall indicate below, the two fundamental theorems hold for this concept. But before we do that, we should spend some time talking about the definitions of efficiency. When one asks the efficiency question in these settings, the issue is the timing at which the Central Authority wishes to perform the improvement. Should it think of making the improvement at date 0, when the state is uncertain, or at date 1, when the state is known? In what follows, we refer to decisions made at date 0 as *ex-ante* or "before the fact," and decisions made at date 1

as *ex-post* or "after the fact." We have two different concepts of Pareto optimality or efficiency.

We say that an allocation of state contingent commodities x is not ex-ante efficient if there is another allocation of state contingent commodities y such that

$$\sum_{s=1}^{k} q_{is} u_{is}(y_{is}) \geq \sum_{s=1}^{k} q_{is} u_{is}(x_{is}) \text{ for all } i = 1, 2, \ldots, n$$
$$\sum_{s=1}^{k} q_{is} u_{is}(y_{is}) > \sum_{s=1}^{k} q_{is} u_{is}(x_{is}) \text{ for at least one } i.$$

If there is no such alternative allocation, x is *ex-ante efficient.*

That is, at an ex-ante efficient allocation, one cannot make an improvement on the entire economy, even when all the financial markets for state contingent commodities are open, by redistributing assets in some fashion. If one performs any such redistribution and makes some people better off in ex-ante or expected utility terms, it must be at the expense of making someone else's ex-ante utility lower. Since the Arrow-Debreu equilibrium is based on ex-ante trade in markets when uncertainty is present, one should evaluate its efficiency properties by using this notion.

Now we turn to *ex-post* or "after the fact."

We say that an allocation of state contingent commodities x is not ex-post efficient if there exists a state s and another allocation of state contingent commodities y such that

$$u_{is}(y_{is}) \geq u_{is}(x_{is}) \text{ for all } i = 1, 2, \ldots, n$$
$$u_{is}(y_{is}) > u_{is}(x_{is}) \text{ for at least one } i.$$

If there is no such state and alternative allocation, x is *ex-post efficient.*

The question that ex-post efficiency answers is whether one could perform an improvement on the entire economy after uncertainty has disappeared, by operating on the markets for goods. One important fact is the relationship between these two concepts of efficiency, sometimes referred to as the *no-trade theorem.* This says that, if the economy is at an ex-ante efficient allocation and information arrives so that uncertainty vanishes, no trade will be observed in the ex-post markets, because the allocation is also ex-post efficient. Thus, additional information does not result in new trade. We prove this result next:

No-Trade Theorem. If an allocation x is ex-ante efficient, it is also ex-post efficient.

Proof: We argue by contradiction. Suppose that x is ex-ante efficient, but not ex-post efficient. Since it is not ex-post efficient, there exists a state s' and an alternative allocation y that makes everyone in that state s' at least as good as does x, and at least one individual strictly better.

Now define the following allocation z:

$$z = \begin{cases} y_s \text{ if } s = s', \\ x_s \text{ otherwise} \end{cases}$$

Since z changes with respect to x only in state s', it follows that

$$\sum_{s=1}^{k} q_{is} u_{is}(z_{is}) \geq \sum_{s=1}^{k} q_{is} u_{is}(x_{is}) \text{ for all } i = 1, 2, \ldots, n$$

$$\sum_{s=1}^{k} q_{is} u_{is}(z_{is}) > \sum_{s=1}^{k} q_{is} u_{is}(x_{is}) \text{ for at least one } i.$$

To see the strict inequality, recall that for all i and for all s, $q_{is} > 0$. But this means that x is not ex-ante efficient either, which is a contradiction. Q.E.D.

Similarly, one can also provide definitions for the ex-ante core and the ex-post core of the economy, as a function of whether coalitional improvements are envisioned before or after the resolution of uncertainty. We define the ex-ante core next, which will be relevant for our result:

An allocation y of state contingent commodities is *ex-ante feasible* for coalition S if for every state s, $\sum_{i \in S} y_{is} = \sum_{i \in S} \omega_{is}$. That is, if the group S of individuals get together, they may think of trading state contingent commodities that add up to their endowments of goods in each state.

Then, if x is a proposed allocation of state contingent commodities and S is a coalition, we will say that S can *ex-ante block* x if there is a feasible allocation y for S, such that:

$$\sum_{s=1}^{k} q_{is} u_{is}(y_{is}) \geq \sum_{s=1}^{k} q_{is} u_{is}(x_{is}) \text{ for all } i \in S$$

$$\sum_{s=1}^{k} q_{is} u_{is}(y_{is}) > \sum_{s=1}^{k} q_{is} u_{is}(x_{is}) \text{ for at least one } i \in S$$

The *ex-ante core* is the set of allocations that cannot be ex-ante blocked by any coalition.

To define the notion of ex-post core, one should think that the coalition formation process may start before uncertainty, but coalitions need not act until it is resolved. Ex-post feasibility for a coalition in a state is equivalent to feasibility in an economy with no uncertainty. Thus, if x is an allocation of state contingent commodities and S is a coalition, we will say that S can *ex-post block* x if there is a state s and an ex-post feasible allocation y for S, such that:

$$u_{is}(y_{is}) \geq u_{is}(x_{is}) \text{ for all } i \in S$$
$$u_{is}(y_{is}) > u_{is}(x_{is}) \text{ for at least one } i \in S$$

The *ex-post core* is the set of allocations that cannot be ex-post blocked by any coalition.

Using similar steps as in the proof of the no-trade theorem, one can establish that the ex-ante core is a subset of the ex-post core: if coalition S finds that they can ex-post block x with y in state s', at the ex-ante stage they can commit to stay with the grand coalition at x if the realized state is other than s', and move to y otherwise. In addition, the ex-ante core is contained in the set of ex-ante efficient allocations, and the ex-post core is contained in the set of ex-post efficient allocations. This is so because efficiency concepts rely on the coalition of all agents finding an improvement, while core notions use every possible coalition.

By making the assumptions on preferences of Chapter 3, one can then establish the two fundamental theorems of welfare economics. We next provide their statements. We shall not prove them, although we note that the proofs are identical to the corresponding theorems seen in that chapter, once one takes account of the different notation (essentially, the distinction between goods versus state contingent commodities, and utility versus expected utility).

> *First Fundamental Theorem of Welfare Economics for Economies with Uncertainty.* Suppose uncertainty is symmetric and there are complete markets (i.e., there is a market for each state contingent commodity). Suppose also that all traders have monotonic self-interested utility functions in each state, and are expected utility maximizers. If (x, p) is an Arrow-Debreu equilibrium, then x is in the ex-ante core (and is, therefore, ex-ante efficient as well).

For the second theorem, we shall need the following property of utility functions: a utility function in state s u_{is} is *concave* whenever for every pair of bundles x_{is}, y_{is} and for every $\alpha \in [0, 1]$,

$$u_{is}([\alpha x_{is} + (1 - \alpha)y_{is}]) \geq \alpha u_{is}(x_{is}) + (1 - \alpha)u_{is}(y_{is}).$$

This just says that the utility derived from a bundle that is a convex combination of two bundles is no less than the convex combination of the utilities of those bundles. Concavity implies the condition of convex indifference curves, assumed earlier in the book. We are now ready to state the second theorem.

> *Second Fundamental Theorem of Welfare Economics for Economies with Uncertainty.* Suppose uncertainty is symmetric and there are complete markets (i.e., there is a market for each state contingent commodity). Suppose also that all traders have self-interested, monotonic and concave utility functions in each state, and are expected utility maximizers. Let y be any ex-ante efficient allocation. Then there exists a vector of bank balance transfers $(T_1, T_2 \ldots, T_n)$ and a price vector $p = (p_{11}, \ldots, p_{km})$ such that (y, p) is an Arrow-Debreu equilibrium given the transfers. That is, for all i, y_i maximizes person i's expected utility $\sum_{s=1}^{k} q_{is} u_{is}(z_{is})$ subject to $p \cdot z_i \leq p \cdot \omega_i + T_i$.

Note that most of the assumptions made here are the same as in Chapter 3. The only difference is the requirement of concave ex-post utility functions, instead of the weaker assumption of convex indifference curves. This is related to the fact that expected utility is preserved only under positive affine transformations of utility functions. Thus, we are led to assume concavity, a cardinal property of a function, instead of convex indifference curves, a property that would be preserved through any monotone transformation. The good news is that concave ex-post utility functions admit a nice economic interpretation, because they amount to risk aversion or risk neutrality (i.e., they rule out risk lovers from the analysis, those "nutty" people for whom facing risky situations is exciting; for most people, risk is something to be avoided).

We now know from the first welfare theorem that if the economy has enough markets - complete markets for all state contingent commodities - the market equilibrium, driven by the forces of demand and supply, will yield an ex-ante core allocation. That means that, even after the arrival of information, after the resolution of uncertainty, no coalition will be able to benefit from reallocating its endowments to improve upon the equilibrium allocation. (Recall that the ex-ante core is contained in the ex-post core). The markets for state contingent commodities perform an efficient transfer of income across states for each agent, as a function of how much consumption each of them wishes to have in each state. This is the role of the ex-ante budget constraint. From the second welfare theorem, we know that any ex-ante efficient allocation can be supported with Arrow-Debreu equilibrium prices, provided that

one makes the appropriate cash transfers (which amount to reallocations of state contingent commodities).

4. Examples

To understand the concept of Arrow-Debreu equilibrium and its welfare properties, it is useful to look at some examples. First, the reader can easily check that in the example of Figure 6.1 the Arrow-Debreu equilibrium allocation is x^*, which allocates the state contingent bundle $(1/2, 1/2)$ to each trader, associated with prices $p = (1, 1)$. That is, given the concavity of the ex-post utility functions (which is equivalent to risk aversion) and given that both traders expect each state to be as likely as the other, they provide complete insurance to one another by promising to split equally the endowment of the good in each state.

Consider next the example shown in Figure 6.2. In it, we still have two agents, two states and one good per state. The ex-post utility function of each trader continues to be $\sqrt{x_{is}}$. Endowments are also as before, so that trader i will only have one unit of the good in state i. However, suppose that now traders have different probability assessments about the states. Let's say that trader 1 believes $q_1 = (1/3, 2/3)$, while trader 2 believes $q_2 = (2/3, 1/3)$. Now the Arrow-Debreu equilibrium allocation, labeled as x^* in the figure, assigns the bundle $(1/5, 4/5)$ to trader 1 and $(4/5, 1/5)$ to trader 2. The equilibrium relative price is still 1 (exactly half way between each agent's relative probability assessment – the market takes those assessments into account to calculate the value of state contingent commodities). Note how now the allocation is strongly driven by the different beliefs: each trader plans a higher consumption in that state that he believes more likely.

To finish this section, consider Figure 6.3. Here, we have an economy with the same endowments as before. Agents hold the same beliefs as in the last example, so that $q_1 = (1/3, 2/3)$ and $q_2 = (2/3, 1/3)$. the ex-post utility function of agent 1 continues to be the same as before, but the one for agent 2 is now linear, so that he is risk neutral (he cares about the expected consumption, even if it varies greatly from state to state). That is, the expected utility functions for each of them are $(1/3)\sqrt{x_{11}} + (2/3)\sqrt{x_{12}}$ for agent 1, and $(2/3)x_{21} + (1/3)x_{22}$ for agent 2.

How does this change in agent 2's preferences affect the equilibrium? (To find the equilibrium, you should draw the *offer curves* of both traders and find their intersection. We leave this as an exercise). Note that at the equilibrium allocation of Figure 6.2, now the marginal rate of substitution of agent 2, which is 2, exceeds the old equilibrium relative price of 1. If the relative price were 1, agent 2 would now demand more units of the good in state 1 and less in state 2. This causes the price of

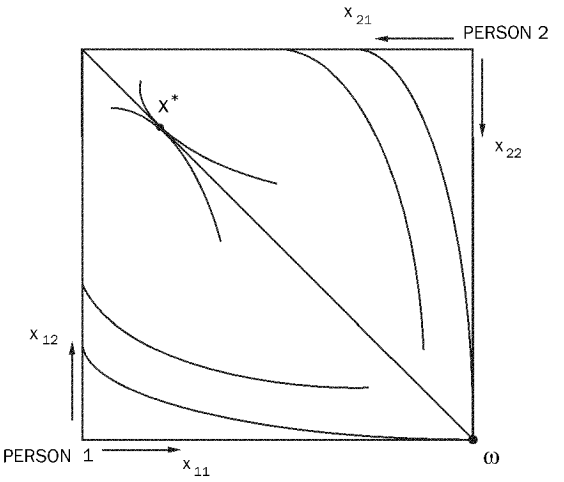


Figure 6.2.

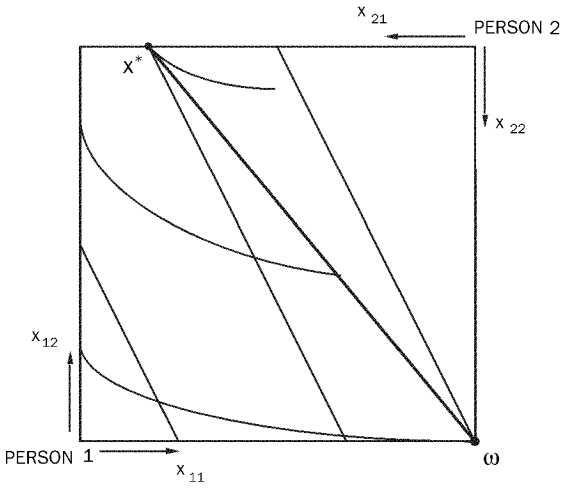


Figure 6.3.

good 1 in terms of good 2 to go up. The new equilibrium has a relative price of $p = 1/(\sqrt{8} - 2) > 1$, and the equilibrium allocation assigns the bundle $(3 - \sqrt{8}, 1)$ to agent 1 and $(\sqrt{8} - 2, 0)$ to agent 2. That is, agent 2 is happy to give up all his consumption in state 2 to fully insure agent 1: note that only for agent 1 is it true that his marginal rate of substitution in equilibrium equals the relative price of state contingent commodities (thus, he is fully insuring himself in the sense that he is planning his

consumption in both states so that the ratio of his marginal utilities exactly coincides with the market equilibrium relative price). On the other hand, agent 2, at the equilibrium price, would be content with demanding more consumption in state 1 and less in state 2, but he can't because he is already planning a zero consumption in state 2.

In all of these examples, the equilibrium allocation is in the ex-ante core. First, it is ex-ante efficient, thanks to the coordination in trade via the signals provided by prices. And second, no individual can block, since for each of them the equilibrium bundle maximizes expected utility over the budget set, which contains the endowment point.

5. Asymmetric Information

Now we move on to the study of economies in which some people have more information than others. For example, suppose the two states of the world concern tomorrow's weather (either sunny or rainy), one of the agents is a weather man who already knows today which of the two will occur, while the rest of agents are still uncertain. Or the government of a country knows whether a foreign dictator holds "nukular" weapons, while other countries' governments do not know.

At this point, we shall focus on a model with three dates. At date -1, no one has private information. Date -1 is the *ex-ante* stage (similar to date 0 in the previous sections). At date 0, some people have private information, and so this is the stage in which information is asymmetric. This is called the *interim* stage. At date 1 the true state of the world is resolved, it is verifiable by all agents and any information superiority disappears. This is the *ex-post* stage. We shall concentrate on models in which decisions are made at the interim stage, that is, in the presence of asymmetric information.

Formally, the way we model the arrival of information at date 0 is by assuming that each agent i receives a signal σ_i. Of course, in different states of the world, agent i can receive different signals. Recall that the set of possible states of the world is $\{1, \ldots, k\}$. The signal σ_i will allow agent i to rule out some states for sure, and hence his beliefs will be updated over the set of remaining states. We shall model the signal as simply a subset of the original set of states (the set of states that have not been ruled out): $\sigma_i \subseteq \{1, \ldots, k\}$. The interpretation is that agent i some how finds out that the true state is in σ_i, thereby being able to rule out states not in σ_i. We shall assume that there is no false information in the signals (a state s that has been zeroed out by the signal σ_i does not become ultimately the actual state at date 1).

An example will serve to clarify these ideas. Suppose there are five possible states: $\{1, 2, 3, 4, 5\}$, corresponding to the number of guests that

will show up to a party that you are hosting. (Suppose you care only about the number of people that show up, and not who those are). Let your ex-ante probability assessment about the five states be $q_i = (0.1, 0.2, 0.2, 0.1, 0.4)$. That is, at time -1 you believe that the most likely state is that all five guests show up, and this has probability 0.4. This is four times more likely that only one of them does not show up or that only one shows up. The intermediate cases (two or three guests) are believed to be equally likely, with probability 0.2 each.

To help you figure out how many will attend, they are supposed to RSVP to your invitation. Let's say that only three possible events might happen (and hence, there are only three possible signals that you might receive), which exhaust all possibilities: either you get exactly three affirmative RSVP's, or more than three, or less than three. Indeed, suppose you get a signal that informs you that three of them are for sure not coming. (E.g., only three phone calls, all of them declining your invitation). Let's call this signal σ_i. This signal allows to update your beliefs, which now are $q_i|\sigma_i = (1/3, 2/3, 0, 0, 0)$. This expression is read "belief q_i given signal σ_i," and updating is done using Bayes' rule. That is, given the signal σ_i, we can safely rule out the last three states; and for the rest (one or two guests are still possible), you continue to believe that a two-guest party is twice as likely as a party with only one; since the revised probabilities must add up to 1, these are $q_i(1|\sigma_i) = 1/3$ and $q_i(2|\sigma_i) = 2/3$. If the party ends up having three or more guests, you would not have received signal σ_i, since we are assuming that there are no false signals. As we said above, in that case you would have received either signal σ_i' or signal σ_i'': signal σ_i' is fully informative (you get exactly three affirmative and two negative RSVP's to the party), while signal σ_i'' still leaves you with some uncertainty (you receive four affirmative RSVP's and the fifth person forgets to call). Of course, the use of Bayes' rule to update your ex-ante beliefs gives $q_i|\sigma_i' = (0, 0, 1, 0, 0)$ and $q_i|\sigma_i'' = (0, 0, 0, 1/5, 4/5)$, respectively.

As a function of whether you receive signal σ_i, σ_i' or σ_i'', you may want to make different decisions regarding the number of potato chip bags, cans of soda, and pounds of other junk food to get.

We now return to our exchange economy with three dates, but where no trade takes place at date -1. At date 0, the interim stage, each trader may have received an informative signal (not receiving any signal is also a possibility, in which case we shall say that σ_i coincides with the entire set of states). Trade may take place and contracts may be written at this stage. The execution of these contracts and the consequent consumption of goods takes place at date 1, after uncertainty vanishes. (Alternatively,

as we will point out, one can also think of consumption taking place at the interim stage, at date 0).

The notion of market equilibrium most used in these contexts is the *rational expectations equilibrium* (REE), which we will define in the next paragraph. The idea of an REE is the following: we'll have an equilibrium price function p^*, which will assign a vector of prices of goods for each state s, i.e., $p^*(s) = (p_1^*(s), \ldots, p_m^*(s))$ for $s = 1, \ldots, k$. That is, in different states, given the different information held by traders, prices could incorporate parts of this information. We shall say that a price function is *fully revealing* if for every two distinct states s and s', $p^*(s) \neq p^*(s')$. In this case, all traders, who rationally expect the equilibrium price function, will be able to update their beliefs and learn the state completely just by looking at the market prices. Otherwise, such complete learning will not be possible and we will speak of a *nonrevealing* price function. (Note the slight change of notation with respect to earlier sections: to emphasize the concept of price function, we write the state as an argument, and not in the subscript as before; we shall do this also for allocations.) Given an REE price function, traders formulate their demands by maximizing their conditional expected utility (conditional on the information that they have received, via the private signal and the market prices). Furthermore, the prices must be such that in each state markets must clear. Many economists contributed to this notion, but among the first we should name Jerry Green, Sanford Grossman, Robert Lucas and Roy Radner. We give the definition next.

A price function p^* that consists of prices $(p_1^*(s), \ldots, p_m^*(s))$ in each state s, and an allocation $x = (x_1, \ldots, x_n)$, where each x_i describes for each state s the bundle $x_i(s) = (x_{i1}(s), \ldots, x_{im}(s))$ assigned to trader i is a *rational expectations equilibrium* (REE) if two conditions are satisfied:

> Conditional expected utility maximization: given signal σ_i and p^*, let T be the set of states that agent i still assigns positive probability. Then, for each agent i and each signal σ_i, the bundles x_i maximize his conditional expected utility $\sum_{s=1}^{k} q_i(s|\sigma_i, p^*) u_{is}(z_i(s))$ subject to the budget constraint $p^*(s) \cdot z_i(s) \leq p^*(s) \cdot \omega_i(s)$ for all $s \in T$, and $z_i(s) = z_i(s')$ for every $s, s' \in T$.
>
> Market clearing: in the market for each good j in each state s, $\sum_{i=1}^{n} x_{ij}(s) = \sum_{i=1}^{n} \omega_{ij}(s)$.

Note how the first requirement incorporates the updating of beliefs given the signal and the market prices. Also, we require that, if an agent cannot rule out either of two states, his consumption bundles in those two states must be the same. This is why the concept can also

be understood as one in which consumption may take place at date 0, before the resolution of uncertainty.

In terms of evaluating the welfare properties of a concept like REE, we have to introduce another definition of efficiency, in which improvements can be performed at the interim stage, given a fixed structure of signals σ_i's. The new notion is called *interim efficiency* and we define it now.

Given a fixed structure of signals, an allocation x is not interim efficient if there is another allocation y such that

$$\sum_{s=1}^{k} q_i(s|\sigma_i)u_{is}(y_i(s)) \geq \sum_{s=1}^{k} q_i(s|\sigma_i)u_{is}(x_i(s))$$

for all $i = 1, 2, \ldots, n$ and for all σ_i,

$$\sum_{s=1}^{k} q_i(s|\sigma_i)u_{is}(y_i(s)) > \sum_{s=1}^{k} q_i(s|\sigma_i)u_{is}(x_i(s))$$

for at least one pair i, σ_i.

If there is no such alternative allocation, x is an *interim efficient* allocation.

That is, at an interim efficient allocation, one cannot make an improvement on the entire economy at the time each agent has received some private information. If at an interim efficient allocation one wishes to perform any redistribution of goods, at least one of the agents, given his private information, will suffer from the redistribution and obtain a lower interim or conditional expected utility.

It is important to note that the set of allocations that is interim efficient falls between the ex-ante and the ex-post efficiency sets. That is, a no-trade theorem holds for the asymmetric information model. We state it here and leave the proof as an exercise.

> *No-Trade Theorem; Asymmetric Information*: If an allocation x is ex-ante efficient, it is also interim efficient. If an allocation x is interim efficient, it is also ex-post efficient.

Our next question is whether one can extend the welfare theorems to the asymmetric information model. The results here will be largely negative. We shall concentrate on the first welfare theorem, and show the following results:

> *First Fundamental Nontheorem of Welfare Economics; Asymmetric Information – Part I.* Under the usual assumptions on preferences, the conclusions of the first fundamental theorem of welfare economics may be violated under asymmetric information in terms of ex-ante

efficiency. Specifically, a rational expectations equilibrium allocation may fail to be ex-ante efficient.

Proof: We shall provide an example to make this point. Consider the following economy. There are two agents, two goods per state, and two equally likely states $s = 1, 2$. Agent 1 is fully informed, which means he can receive two possible signals at time 0, σ_1 (i.e., state $s = 1$ will hold) and σ_1' (i.e., $s = 2$ will hold): therefore, $q_1|\sigma_1 = (1, 0)$ and $q_1|\sigma_1' = (0, 1)$. Agent 2 is uninformed, his signal does not provide any new information so that $q_2|\sigma_2 = (1/2, 1/2)$. Preferences in state $s = 1, 2$ are represented by the utility functions

$$u_{1s}(x_1) = x_{11} + s \cdot \ln(x_{12}), \quad u_{2s}(x_2) = \ln(x_{21}) + s \cdot \ln(x_{22}).$$

As for endowments, they are state independent:

$$\omega_1 = (1, 1), \quad \omega_2 = (2, 1).$$

Figure 6.4 illustrates this economy in two Edgeworth box diagrams. Note that Figure 6.4a is for state 1, and Figure 6.4b is for state 2.

It turns out that there is only one REE in this economy (this uniqueness is not essential for our current argument, and we leave it as an exercise). Moreover, it is fully revealing. Thus, it corresponds to the unique ex-post competitive equilibrium in each of the states. We call the unique REE allocation x. Normalize the price of good 2 to 1 in each state. Then, the equilibrium price function is $p^*(1) = (3/4, 1)$ and $p^*(2) = (2/5, 1)$. We provide the description of the economy, the REE allocation x and the utility levels it yields in each state in the table below.

State 1	State 2
$u_{11}(\cdot) = x_{11}(1) + \ln(x_{12}(1))$	$u_{12}(\cdot) = x_{11}(2) + 2\ln(x_{12}(2))$
$u_{21}(\cdot) = \ln(x_{21}(1)) + \ln(x_{22}(1))$	$u_{22}(\cdot) = \ln(x_{21}(2)) + 2\ln(x_{22}(2))$
$\omega_1(1) = (1, 1), \omega_2(1) = (2, 1)$	$\omega_1(2) = (1, 1), \omega_2(2) = (2, 1)$
$x_1(1) = (x_{11}(1), x_{12}(1)) = (4/3, 3/4)$	$x_1(2) = (x_{11}(2), x_{12}(2)) = (3/2, 4/5)$
$x_2(1) = (x_{21}(1), x_{22}(1)) = (5/3, 5/4)$	$x_2(2) = (x_{21}(2), x_{22}(2)) = (3/2, 6/5)$
$u_{11} = (4/3) + \ln(3/4) = 1.0457$	$u_{12} = (3/2) + 2\ln(4/5) = 1.0537$
$u_{21} = \ln(5/3) + \ln(5/4) = 0.7340$	$u_{22} = \ln(3/2) + 2\ln(6/5) = 0.7701$

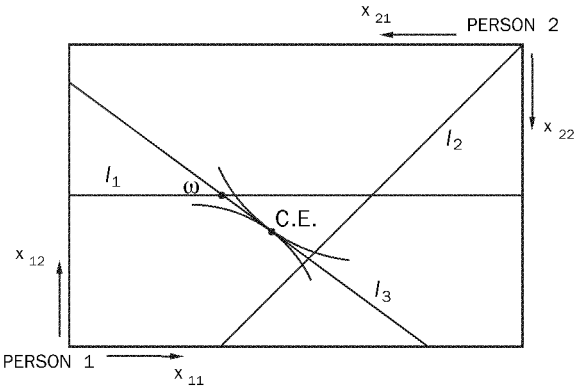


Figure 6.4a.

State 1. Along line l_1, $MRS_1 = 1$; along line l_2, $MRS_2 = 1$; along line l_3, absolute slope = the price ratio = 3/4.

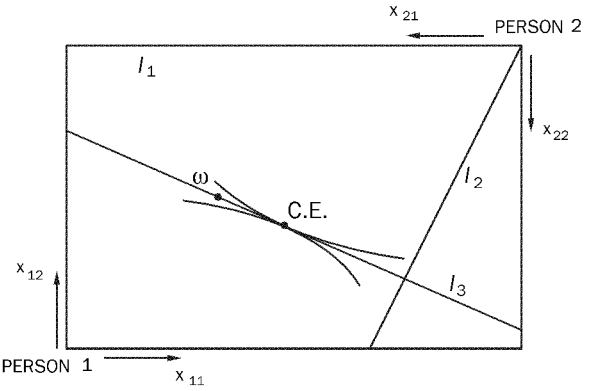


Figure 6.4b.

State 2. Along line l_1, $MRS_1 = 1$; along line l_2, $MRS_2 = 1$; along line l_3, absolute slope = the price ratio = 2/5.

The ex-ante expected utility for person 1 from the REE allocation x is

$$(1/2) \cdot (1.0457) + (1/2) \cdot (1.0537) = 1.0497.$$

And the ex-ante expected utility for person 2 from the REE allo-

cation x is

$$(1/2)\cdot(0.7340)+(1/2)\cdot(0.7701)=0.7525.$$

We need to show that this REE allocation is not ex-ante efficient. We propose the improvement x'. This new allocation and the utility levels it yields in each state is also shown in table form:

State 1	State 2
$x'_1(1)=(x'_{11}(1),x'_{12}(1))=(17/12,3/4)$	$x'_1(2)=(x'_{11}(2),x'_{12}(2))=(17/12,4/5)$
$x'_2(1)=(x'_{21}(1),x'_{22}(1))=(19/12,5/4)$	$x'_2(2)=(x'_{21}(2),x'_{22}(2))=(19/12,6/5)$
$u'_{11}=(17/12)+\ln(3/4)=1.1290$	$u'_{12}=(17/12)+2\ln(4/5)=0.9704$
$u'_{21}=\ln(19/12)+\ln(5/4)=0.6827$	$u'_{22}=\ln(19/12)+2\ln(6/5)=0.8242$

Now the ex-ante expected utility for person 1 from the allocation x' is

$$(1/2)\cdot(1.1290)+(1/2)\cdot(0.9704)=1.0497.$$

And the ex-ante expected utility for person 2 from x' is

$$(1/2)\cdot(0.6827)+(1/2)\cdot(0.8242)=0.7534.$$

Since the improvement does not change the allocation of good 2, only changes in utility through good 1 must be evaluated. Since that part of 1's utility function is linear, note how he is exactly indifferent in ex-ante terms (i.e., before he receives his signal, before he knows the state) because 17/12 is the average of the amounts of good 1 that he was receiving in each state at the REE allocation (4/3 and 3/2). However, since the utility of agent 2 for good 1 is a logarithmic function, which is strictly concave, agent 2 is strictly better off in ex-ante terms with this improvement.

Therefore, the allocation x' is an ex-ante improvement over the REE allocation x. Q.E.D.

First Fundamental Nontheorem of Welfare Economics; Asymmetric Information – Part II. Under the usual assumptions on preferences,

the conclusions of the first fundamental theorem of welfare economics may be violated under asymmetric information in terms of interim efficiency. Specifically, a rational expectations equilibrium allocation may fail to be interim efficient.

Proof: We shall modify the example in the earlier proof slightly to make this point.

Allocation x' is not yet an interim improvement over the REE allocation x. This is because agent 1 in state $s = 2$ is made actually worse off in interim terms: $17/12 < 3/2$. Note, however, how agent 1 in $s = 1$ is made better off in interim terms at x'. Agent 2 is also made better off in interim terms (for him, ex-ante and interim terms are the same).

Here's how we modify the example: suppose we look at the economy with $k + 1$ copies of agent 1 and $k + 1$ copies of agent 2 in terms of preferences and endowments in the two states. However, only the original agent 1 is informed about the state. All other copies of agent 1 and all copies of agent 2 (including the original one) are uninformed.

For this economy, the original REE continues to be an REE, and it is fully revealing (that is, the ex-post competitive equilibrium allocation of this replica economy is the replica of the original equilibrium allocation). To see this, notice that the fully revealing nature of the equilibrium makes the apparent asymmetry in the different copies of agent 1 disappear.

But now we can take away an amount ϵ of good 1 from the improvement made on each copy of agent 2 to make him receive $(19/12) - \epsilon$ in each state. We choose ϵ small enough so that this still constitutes an interim improvement for each of the copies of agent 2 (and note that this ϵ does not depend on k). Now give this extra amount of good 1 to the original agent 1. The one we were concerned with, that is, the original agent 1 in state $s = 2$ is now receiving an amount $(17/12) + (k+1)\epsilon$. Thus, it suffices to choose k large enough to guarantee that $(17/12) + (k + 1)\epsilon \geq 3/2$. Now the resulting allocation is indeed in interim improvement, so that the REE allocation of this replica economy is not interim efficient. Q.E.D.

First Fundamental Theorem of Welfare Economics; Asymmetric Information – Part III. Under the usual assumptions on preferences, the conclusions of the first fundamental theorem of welfare economics almost always hold under asymmetric information in terms of ex-post efficiency. Specifically, a rational expectations equilibrium allocation in almost every economy is ex-post efficient.

Proof: To prove the "almost always" bit rigorously is beyond the scope of this book. However, here's an intuition. Constructing a non-revealing REE is difficult. In it, one must require that an uninformed agent who cannot distinguish among several states (recall that he faces the same prices over those states) maximizes expected utility subject to the budget constraint in each state. The solution to this problem will typically yield different bundles in each such state, violating the REE requirement that consumption must be the same.

Thus, we learn that almost every REE must be fully revealing. But then, in each state, the projection of the REE to that state is an ex-post Walrasian equilibrium. By the standard version of the first welfare theorem, the corresponding ex-post allocation is then ex-post efficient. Q.E.D.

A few remarks are in order:

First, what we have shown is a strong violation of the first welfare theorem in these settings. Given that REE is a concept that concerns interim trade, the most relevant result of the ones we have just shown is the second nontheorem. Since REE allocations are not necessarily interim efficient, they will also fail to be elements of any interim core. The interim core will be contained in the set of interim efficient allocations. Defining interim cores, though, is not a straightforward task, as one needs to take account of the amount of information transmission that may take place within coalitions that consist of asymmetrically informed individuals.

Second, the first part of our nontheorem similarly implies that REE allocations may fail to be in the ex-ante core. It is important to point out that the reason for the lack of interim and ex-ante efficiency of REE's is that "too much" information is revealed by the equilibrium price function, which kills insurance possibilities.

Third, as argued above, an REE is "almost always" fully revealing. That is, in almost every economy all REE's yield a fully revealing price function. Whenever this is the case, the corresponding REE allocations will be in the ex-post core, simply because they correspond to ex-post competitive equilibria. The fact that an REE typically gives away all the information and allows people to learn the state perfectly is a very intriguing result. To try to understand how all the private information ends up being incorporated in the price function, different authors have investigated alternative trading procedures, in which traders choose how much information to transmit to the market (see for example Wolinsky (1990) and Gottardi and Serrano (2005)).

6. Exercises

1 Show that the set of ex-ante efficient allocations is contained in the set of interim efficient allocations. Show that the set of interim efficient allocations is contained in the set of ex-post efficient allocations.

2 Consider the economy depicted in Figure 6.1. Find the entire set of ex-ante efficient allocations. Find the entire set of ex-post efficient allocations.

3 Consider again the economy of Figure 6.1, but now allow any beliefs. That is, let trader i's belief be $q_i = (q_{i1}, q_{i2})$. Calculate the Arrow-Debreu equilibrium in each of these economies.

4 In the economy with asymmetric information depicted in Figure 6.4, show that there is a unique REE (which is the one given in the text).

7. Selected References

(Items marked with an asterisk (*) are mathematically difficult.)

*1. K. Arrow, "The Role of Securities in the Optimal Allocation of Risk Bearing," *Review of Economic Studies* V. 31, 91-96, 1964.

In this article, Arrow introduces the so-called Arrow securities, a simpler way to get the results obtained under complete markets. Arrow argues that, instead of having km state contingent commodities, it suffices to have as many assets as states (i.e., only k assets) to be able to transfer wealth across states.

*2. G. Debreu, *Theory of Value*, Wiley, New York, 1959.

Debreu's book also covers the case of symmetric uncertainty as part of the model of a general equilibrium economy.

*3. F. Forges, E. Minelli and R. Vohra, "Incentives and the Core of an Exchange Economy: A Survey." *Journal of Mathematical Economics*, V. 38, pp. 1-41, 2002.

This article is an excellent survey of the recent literature on the core with asymmetric information. Ex-ante and interim cores are described, and the emphasis is on economies in which states are not verifiable, even ex-post. For these cases, incentive constraints must be imposed, as we shall do in the last chapter of this book.

*4. P. Gottardi and R. Serrano "Market Power and Information Revelation in Dynamic Trading," *Journal of the European Economic Association* V. 3, December 2005.

This article investigates different models of information revelation in a dynamic market game setup. The key question is to identify conditions under which the equilibria of the market game converge to REE's.

*5. J. Green, "Information, Efficiency, and Equilibrium, Discussion paper no.284. Harvard Institute for Economic Research, 1973.

This is one of the first papers, together with others by Grossman and Lucas, in which the definition of REE is proposed.

*6. S. Grossman, *The Informational Role of Prices.* The MIT Press, Cambridge, MA, 1989.

This book is a presentation of many issues related to information and equilibrium prices, especially REE.

*7. B. Holmstrom and R. Myerson, "Efficient and Durable Decision Rules with Incomplete Information." *Econometrica* V. 51, pp. 1799-1819, 1983.

This classic article introduces the different notions of efficiency under asymmetric information. These consist of the ex-ante, interim and ex-post sets of efficient allocations. Definitions are provided both for the case in which states are or are not ex-post verifiable.

*8. J.-J. Laffont, "On the Welfare Analysis of Rational Expectations Equilibria with Asymmetric Information," *Econometrica* V. 53, pp. 1-29, 1985.

In this article, Laffont analyzes the welfare properties of REE. The example we use to prove that REE may not be interim efficient is a simpler version of one of his examples.

*9. R. Lucas, "Expectations and the Neutrality of Money," *Journal of Economic Theory* V. 4, pp. 103-124, 1972.

This is, together with the contributions of Green and Grossman, one of the first articles introducing REE.

*10. A. Mas-Colell, M. Whinston and J. Green, *Microeconomic Theory*, Oxford University Press, 1995.

This is an excellent up-to-date textbook on microeconomic theory. In particular, Chapter 19 has a very clear treatment of economies under uncertainty.

*11. M. McGill and W. Shafer, "Incomplete Markets," Chapter 30 in *Handbook of Mathematical Economics, Vol. IV*, edited by W. Hildebrand and H. Sonnenschein, North Holland, 1991.

This article surveys the literature on incomplete markets. That is, what happens if the economy does not have enough assets to perform transfers of wealth across states. In general, no extension of the welfare theorems can be achieved: a very strong suboptimality result can be established.

*12. P. Milgrom and N. Stokey, "Information, Trade and Common Knowledge," *Journal of Economic Theory* V. 26, pp. 17-27, 1982.

This article deals with several versions of the no-trade theorem.

*13. R. Radner, "Equilibrium under Uncertainty," Chapter 20 in *Handbook of Mathematical Economics*, V. II, edited by K.J. Arrow and M.D. Intriligator, North-Holland, 1982.

This is a survey of different market equilibrium concepts under uncertainty. The cases of symmetric uncertainty and asymmetric information are covered.

*14. R. Serrano, R. Vohra and O. Volij, "On the Failure of Core Convergence in Economies with Asymmetric Information," *Econometrica* V. 69, pp. 1685-1696, 2001.

This article shows that in economies at the interim stage different versions of interim cores fail to converge to price equilibrium notions, including REE. Thus, asymmetric information remains as a source of bargaining power that prevents competitive outcomes even in large economies.

*15. R. Vohra, "Incomplete Information, Incentive Compatibility and the Core," *Journal of Economic Theory* V. 86, pp. 123-147, 1999.

This paper extends the first concept of interim core, the so-called coarse core, to economies in which states are not ex-post verifiable. The main result is negative: Vohra shows robust examples in which the coarse core is empty.

*16. R. Wilson, "Information, Efficiency and the Core of an Economy," *Econometrica* V. 46, pp. 807-816, 1978.

This is the first paper that provides notions of the core under asymmetric information. It studies economies in which states will be verifiable at the ex-post stage. Two extreme situations are analyzed at the interim stage. In the first, no information is shared among members of a blocking coalition. This idea leads to the coarse core, and Wilson establishes a general result of nonemptiness. In the second, members of a blocking coalition can share any part of the private

information that each of them holds. This behavior serves to define the fine core, which is shown to be empty in many economies.

*17. A. Wolinsky "Information Revelation in a Market with Pairwise Meetings," *Econometrica* V. 58, 1990, pp. 1-23.

In this article, Wolinsky studies a trading procedure in which transactions take place in pairwise meetings of agents. Information is heavily restricted: traders have no access to observing any signal of market performance, such as a price function. Each trader tries to learn the state from his past history of observed prices when he visits different stores. The result is that, even when frictions - search costs - are removed from the model, its equilibria fail to converge to an REE.

Chapter 7

EXTERNALITIES

1. Introduction

The fundamental results of the previous chapters, the results that link optimality and competition, depend on assumptions that (1) people have self-interested utility functions, and (2) firms' production sets are unaffected by other firms' production decisions. In many actual cases these assumptions break down. When person A's utility depends on what person B consumes, or when A's technological production possibilities depend on what firm B does, there is an external effect. The decisions of one person or firm have a tangible, nonmarket impact on a different person or firm.

For example, suppose person A knows person B, and feels that person B eats too little or too much. Then person B's eating has a direct effect on A's utility level. This we call an external effect. On the other hand, if A does not know B and does not care how much B eats, B's consumption might still have some impact on A. For whenever B buys food he affects the equilibrium price of food (perhaps by a minute amount), and therefore, B's appetite indirectly influences A's utility level. However, we do not call this an external effect. In the case of consumers, an external effect is a direct effect of one person's consumption on another person's utility level, not an indirect one that operates via the price mechanism.

What is the difference? When one person's consumption affects another's welfare through the price mechanism, when B bids up the price of food for A, the system is working in the way assumed by the two fundamental theorems. If B gets more food at the expense of A because he is willing to pay more, then the price mechanism is directing the food to the person who wants it most. And the distribution of goods that re-

sults is efficient. However, if B's consumption of goods affects A's utility directly, irrespective of prices, then the price mechanism gives inappropriate signals. When B consumes food, he thinks only about his utility; he looks at the prices, and then makes a decision. But this decision has a direct impact on A's utility, and the price that B pays for food does not reflect this impact on A. Consequently, the price mechanism does not tell B of the total social benefits and costs of his actions, and the resulting distribution of goods is not, in fact, efficient.

Once we know what to look for, we can discover external effects all around us. Many of us are directly affected when we learn that other people don't have enough food to eat. When we hear that a child a thousand miles away is starving, we are worse off. When we learn that people at the other end of the country are living in tar paper shacks, we are worse off. There are externalities, then, in the consumption of food, of housing, perhaps even of medical care. Many of us are worse off when we discover that others are in severe distress because of inadequate consumption of some vitally important good.

There are also myriad mundane consumption externalities. Nonsmokers are bothered by smokers. When A smokes, B's utility level drops. Nondrinkers are occasionally bothered by drinkers. When A drinks to excess, B, a member of the Women's Christian Temperance Union, feels worse off. Those who prefer classical music are bothered by those who play rock music. Many of us are bothered by loud exhaust noises of cars, trucks, and motorcycles. When A drives his car with modified (amplified) exhaust pipes through town, hundreds or thousands of B's might be made briefly worse off.

Externalities among firms are common. One standard story has two firms located on the same river. Firm A, the upstream firm, dumps its wastes in the river, while firm B, the downstream firm, uses river water for washing and otherwise processing its outputs. If firm A increases its output (and its wastes), firm B's production suffers. To produce the same output with dirtier water, firm B must use more chemical agents, more labor, and more electricity. That is, firm B's production set shrinks. If firm A pays nothing for dumping wastes in the river, it receives no information from the price system about the external costs it is imposing on firm B. The consequence is that the price mechanism no longer ensures efficiency. (Note that if firm A were downstream, this externality problem might not arise.)

A second standard story has two firms sharing the same air. Firm A is an old-fashioned electric generating facility that burns coal and uses no scrubbing or antipollution devices. Consequently, every kilowatt hour produced results in a belch of black smoke. Firm B is a laundry located

nearby. When firm A produces more electricity (and more smoke), firm B has to cope with more dirt and grime settling down on its plant, in its machines, on its tables and presses, and on the clothing being cleaned. So firm B must use more soap, more labor, more wrapping paper, and so on, to produce the same output of clean garments. In other words, when A's output rise, firm B's production set shrinks. But A does not take these costs for B into account in its decisions, so the price mechanism provides it with misleading information. It acts as if the air is free.

A very important type of externality occurs when a firm's production decisions have direct nonmarket effects on a person's utility level. For example, a firm that stripmines for coal without reclaiming the land affects the utility levels of people who see the results. A firm that produces smoke affects the utility levels of people who breathe the smoke. The residents of Chicago, Illinois are directly affected by the output decisions of steel mills in Gary, Indiana, whenever the wind is from the east. People who live near the Three Mile Island Nuclear Reactor in Pennsylvania might be affected by the production decisions of the firm that operates that plant. People who live near coal-burning electric generators are often affected by the output decisions of those firms.

Not all externalities of this firm-person type are harmful or negative. People who live downwind from a bakery might be happier when bread production is high. Firms that build attractive plants or office buildings make people who look at those buildings better off. Many of the impressive and exciting sights of a large city are the skyscrapers built by private firms, such as the Empire State Building and the Chrysler Building in New York, the Sears and John Hancock towers in Chicago. Much of New England is dotted with handsome nineteenth century mill buildings, which still provide viewing pleasure long after the firms that built them went bankrupt or moved away.

In all of these cases, whether the externalities are in consumption or in production, whether they are positive (beneficial) or negative (hurtful), the price mechanism does not provide complete enough information to the decision maker. In the case of negative externalities, the price mechanism does not tell the decision maker how much his decision really costs. In the case of positive externalities, the price mechanism does not tell the decision maker how much his decision really helps. And it follows that the link between competition and optimality is broken.

In this chapter we shall carefully analyze two examples of external effects, one in an exchange model and one in a production model. The examples will illustrate how the external effects destroy the optimality of a competitive equilibrium.

However, the existence of externalities does not mean that markets must be disbanded. Abolishing the price mechanism because air pollution is bothersome and because flower gardens are pleasurable would be throwing the baby out with the bath water.

Since the early twentieth century economists have advocated taxes and subsidies to correct important externality-induced inefficiencies. The idea is that those who harm others through their production or consumption decisions should pay a tax to reflect that harm. The size of the tax should depend on the extent of the harm: these are not lump sum taxes like the bank balance transfers of the Second Fundamental Theorem. With the tax in their figuring, the decision makers would be led, via the tax-modified price system, to the right decisions. They would take into account the real social costs of their decisions. Similarly, consumers and firms that create external benefits should be subsidized to reflect those benefits. Again, the extent of the subsidies should depend on the extent of the benefits. With the subsidies in their figuring, the decision makers would be led, again, to the right decisions.

We will show how the appropriate taxes or subsidies ought to be figured in each of our examples, and we will show how the tax- or subsidy-modified price mechanism once again produces an optimal distribution of goods, or an optimal production plan.

2. Externalities in an Exchange Economy: An Example

We now look at what happens when the self-interestedness assumption is relaxed in an exchange economy. To illustrate the problem, we construct a simple two-person two-goods example.

Let

$$\begin{aligned} u_1(x) &= x_{11}x_{12} + x_{21} \qquad \omega_1 = (10, 0) \\ u_2(x) &= u_2(x_2) = x_{21}x_{22} \; \omega_2 = (0, 10). \end{aligned}$$

Person 1 is altruistic; he gets some pleasure out of 2's consumption of good 1. Person 2, on the other hand, is self-interested.

To start the analysis, we solve for the set of Pareto optimal allocations. First, we rewrite u_1 as

$$u_1(x) = x_{11}x_{12} + x_{21} = x_{11}x_{12} + (10 - x_{11}).$$

Next we calculate the marginal rate of substitution of good 2 for good 1 for person 1:

$$\text{MRS person 1} = \frac{\text{MU of good 1}}{\text{MU of good 2}} = \frac{x_{12} - 1}{x_{11}}$$

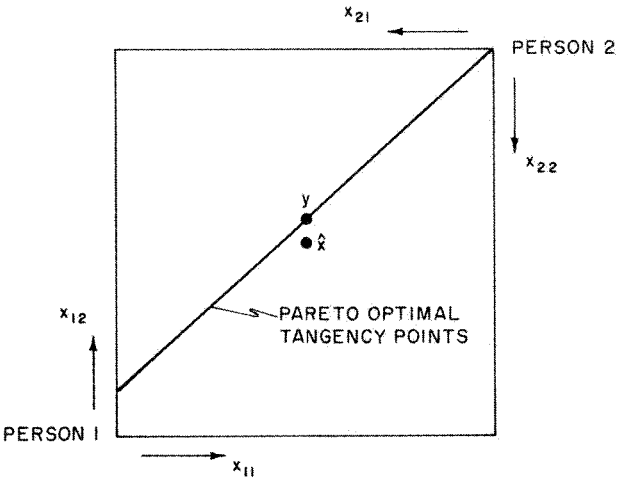


Figure 7.1.

and for person 2:

$$\text{MRS person 2} = \frac{\text{MU of good 1}}{\text{MU of good 2}} = \frac{x_{22}}{x_{21}}.$$

Then, to find the locus of tangency points of the two individuals' indifference curves, we set MRS for person 1 equal to MRS for person 2:

$$\frac{x_{12}-1}{x_{11}} = \frac{x_{22}}{x_{21}} = \frac{10-x_{12}}{10-x_{11}}.$$

Solving this equation for x_{12} in terms of x_{11} gives

$$x_{12} = 1 + \frac{9}{10}x_{11}.$$

Figure 7.1 shows this locus of Pareto optimal tangency points and two other allocations to which we shall soon refer.

The next step is to calculate a competitive equilibrium. We assume that 1 and 2 act as price takers, and at this point we also suppose that 1 does not know that $x_{11} + x_{21} = 10$, and does not know that

$$u_1(x) = x_{11}x_{12} + (10 - x_{11}).$$

We must make this somewhat artificial assumption in order to discuss external effects in the simple two-person case; if we do not make it, all apparently altruistic (or malevolent) utility functions could be rewritten

and solved as self-interested ones. If there were three or more people this artificiality would disappear.

Person 1 wants to maximize $x_{11}x_{12} + x_{21}$ subject to his budget constraint. Since he cannot choose x_{21} himself (2 has something to say about it too), we suppose he contemplates buying a quantity g of the second good, and giving it to 2. Person 2's consumption of the first good will be the quantity he purchases, which 1 cannot control, plus the gift g. Therefore, 1 wants to maximize $x_{11}x_{12}$ plus g, subject to the budget constraint

$$p_1x_{11} + p_2x_{12} + p_1g \leq p_1 \cdot 10 + p_2 \cdot 0 = 10p_1.$$

He will clearly want to spend all his income 10 p_1, and we can therefore ignore the inequality.

Person 2, on the other hand, simply wants to maximize $u_2(x_2) = x_{21}x_{22}$, subject to $p_1x_{21} + p_2x_{22} = p_1 \cdot 0 + p_2 \cdot 10 = 10p_2$.

A competitive equilibrium in this economy is a price vector $p = (p_1, p_2)$ and consumption (+ gift) vectors $(\hat{x}_{11}, \hat{x}_{12}, g)$, $(\hat{x}_{21}, \hat{x}_{22})$, such that $(\hat{x}_{11}, \hat{x}_{12}, g)$ maximizes u_1 subject to 1's budget equation, and $(\hat{x}_{21}, \hat{x}_{22})$ maximizes u_2 subject to 2's budget equation.

To solve for the equilibrium, note that if x_{12} is grater than 1, person 1 will choose $g = 0$, that is, he will give nothing. This is so because the marginal utility of 1's private consumption of good 1 is x_{12}, whereas the marginal utility of 1's charity is 1. But it is obvious from inspection that x_{12} will be chosen greater than 1. Therefore, 1 will attempt to maximize the private part of his utility function $x_{11}x_{12}$ subject to the constraint

$$p_1x_{11} + p_2x_{12} = 10p_1.$$

The perfect symmetry of the two individuals' maximization problems, and the symmetry of ω, lead to the conclusion that the competitive equilibrium is

$$\begin{aligned} p &= (1, 1) \\ \hat{x}_{11} = \hat{x}_{12} &= \hat{x}_{21} = \hat{x}_{22} = 5 \\ g &= 0. \end{aligned}$$

The allocation $\hat{x}$ is shown in Figure 7.1. Note that $\hat{x}$ is not on the locus of tangencies in that figure; it does not satisfy the tangency condition

$$x_{12} = 1 + \frac{9}{10}x_{11}.$$

Consequently, it is not Pareto optimal. When there are externalities present, a competitive equilibrium allocation need not be Pareto optimal, and the First Fundamental Theorem breaks down.

Recall that the exchange version of the Second Fundamental Theorem of Welfare Economics says that any Pareto optimal point can be viewed as a competitive equilibrium allocation given appropriate cash transfers. Now consider the allocation y defined by

$$y_1 = (5, 5.5), y_2 = (5, 4.5).$$

Note that

$$5.5 = y_{12} = 1 + \frac{9}{10} y_{11} = 1 + \frac{9}{10} \cdot 5.$$

That is, y satisfies the tangency condition; it is on the locus of tangencies in Figure 7.1. Therefore, y is Pareto optimal. Do there exist cash transfers T_1 and T_2, such that individual 1 will finish at y_1, and individual 2 will finish at y_2, when they maximize their utilities subject to their (adjusted) budget constraints

$$p_1 x_{11} + p_2 x_{12} + p_1 g \leq 10 p_1 + T_1$$

and

$$p_1 x_{21} + p_2 x_{22} \leq 10 p_2 + T_2?$$

It should again be clear that the gift 1 chooses will be zero, so g may be ignored. To see whether we can induce 1 and 2 to go to y via the competitive mechanism with cash transfers, we first see what is required to induce person 2 to go to y_2:

> Person 2 wants to maximize $u_2(x_2) = x_{21} x_{22}$
> subject to $p_1 x_{21} + p_2 x_{22} \leq 10 p_2 + T_2$.

He sets his marginal rate of substitution x_{22}/x_{21} equal to the price ratio. But his MRS at y_2 is $4.5/5$, so we have

$$\frac{4.5}{5} = \frac{9}{10} = \frac{p_1}{p_2}.$$

In order to get person 2 to the point y_2, the price ratio must be $9/10$. Let's normalize prices by setting $p_2 = 10$. Then p must equal $(9, 10)$. Substituting these prices and quantities in 2's budget equation gives

$$9 \cdot 5 + 10 \cdot 4.5 = 10 \cdot 10 + T_2$$

so $T_2 = -10$. In short, person 2 will choose y_2 if $p = (9, 10)$ and $T_2 = -10$.

However, if the prices are $(9, 10)$ person 1 will *not* choose the point $y_1 = (5, 5.5)$. Since he will not give a gift, he will attempt to maximize the private part of his utility function, that is, $x_{11} x_{12}$. His private

marginal rate of substitution at y is $5.5/5 = 11/10$, which differs from the price ratio $9/10$. He will never choose y_1 when $p = (9, 10)$. Since person 2 will choose y_2 only if $p = (9, 10)$ and person 1 will never choose y_1 when $p = (9, 10)$, the two people cannot be induced to move to y through a cash transfer arrangement. When there are externalities present, a Pareto optimum need not be achievable via the cash transfer modified competitive mechanism, and the Second Fundamental Theorem breaks down.

3. Pigouvian Taxes and Subsidies: The Exchange Example Continued

Externalities weaken the link between competition and optimality. But all is not lost. There is remedy that is consistent with a decentralized price mechanism. It is not necessary to have a Central Bureau (a super Environmental Protection Agency) to direct the consumption decisions of all individuals. The decentralized remedy is the introduction of per-unit (or marginal) *taxes* or *subsidies* on the consumption of the goods that induce the external effects. These taxes or subsidies cannot be of the lump sum cash transfer variety; we saw above that simple cash transfers won't get us to an optimal allocation. They must depend on the quantities actually consumed, for they must affect the relative prices paid by different individuals. They must be designed to encourage a person's consumption of a good if that consumption has positive external effects, and to discourage a person's consumption of a good if that consumption has negative external effects.

With this general motivation in mind, let's return to the example of the previous section. In that example, person 2's consumption of good 1 confers an external benefit on person 1. But when person 2 makes his consumption decisions in the standard competitive equilibrium model, he does not take the external benefit into account. This suggests that this consumption of good 1 ought to be subsidized.

Let s be a subsidy paid to person 2 for each unit of good 1 that he consumes. The link between competition and optimality will be rebuilt if there exist an s, a price vector p, and cash transfers T_1 and T_2, so that, when 1's budget constraint is

$$p_1 x_{11} + p_2 x_{12} \leq 10p_1 + T_1$$

and 2's budget constraint is

$$(p_1 - s)x_{21} + p_2 x_{22} \leq 10p_2 + T_2$$

the two will move to a Pareto optimal allocation through the competitive mechanism. In fact, there do exist such s, p, T_1, T_2, and, with the

appropriate choice of T_1 and T_2, one can move to whatever interior Pareto optimal allocation one desires.

For example, suppose

$$y_1 = (5, 5.5) \text{ and } y_2 = (5, 4.5)$$

is again the goal. It can be achieved this way. First consider person 1. He will again maximize the self-interested part of his utility function. The marginal rate of substitution condition is

$$\frac{x_{12}}{x_{11}} = \frac{p_1}{p_2}$$

and his budget equation is

$$p_1 x_{11} + p_2 x_{12} = 10p_1 + T_1.$$

If we let $T_1 = 0$, the two equations imply $x_{11} = 5$, and $x_{12} = 5p_1/p_2$. We want him to choose $y_{11} = 5$, $y_{12} = 5.5$. Therefore, if we normalize prices by setting $p_2 = 10$, we must have $p_1 = 11$.

Now turn to person 2. The marginal rate of substitution condition and the budget constraint for 2 are

$$\frac{x_{22}}{x_{21}} = \frac{p_1 - s}{p_2}$$

and

$$(p_1 - s)x_{21} + p_2 x_{22} = 10p_2 + T_2.$$

But $p_1 = 11$ and $p_2 = 10$. Moreover, we want person 2 to choose the point $y_{21} = 5$, $y_{22} = 4.5$. Substituting these values in the above equations gives

$$\frac{4.5}{5} = \frac{11 - s}{10}$$

and $(11 - s) \cdot 5 + 10 \cdot 4.5 = 10 \cdot 10 + T_2$. Consequently, we can let $s = 2$ and $T_2 = -10$.

In short, if $T_1 = 0$, $T_2 = -10$, and if $s = 2$, the competitive mechanism, modified by T_1, T_2 and s, will take the economy to the Pareto optimal allocation y. Thus the introduction of the subsidy s re-establishes the link between competition and optimality.

These particular calculations are not especially intuitive; for policy applications we ought to have some simpler concepts to guide the choice of taxes and subsidies. To derive those concepts, we shall carry the example through a few steps further.

Recall that person 1's utility function is $u_1(x) = x_{11}x_{12} + x_{21}$. The marginal utility to person 1 of person 2's consumption of good 1 is therefore 1. The marginal utility to person 1 of his own consumption of good

1 is, in general, equal to x_{12}. At the allocation $y = ((5, 5.5), (5, 4.5))$, the marginal utility to person 1 of his own consumption of good 1 is then 5.5. Now if person 2 reduced his consumption of good 1 by a unit, person 1 would have to increase his consumption of good 1 by 1/5.5 units in order to remain as well off as before. That is, there is a well-defined marginal rate of substitution of 1's own consumption of good 1 for 2's consumption of good 1, equal to person 1's marginal utility from his own consumption of good 1 divided by person 1's marginal utility from person 2's consumption of good 1. We call this particular marginal rate of substitution the *marginal external benefit* or MEB of person 2's consumption of good 1:

$$\text{MEB} = \frac{\text{MU to person 1 of person 2's consumption of good 1}}{\text{MU to person 1 of his own consumption of good 1}}.$$

This gives us a measure of the benefit provided to 1 by 2's consumption, in terms of units of good 1. At y, the MEB = 1/5.5.

Now suppose we ask this question: How many dollars (or other units of currency) would person 1 have to be given to just compensate him for person 2's reducing his consumption of good 1 by 1 unit? The answer, of course, is $p_1 \cdot$ MEB, which we shall call MEB in dollars. At the point y, if $p_1 = 11$, then

$$\text{MEB in dollars } = p_1 \cdot \text{ MEB } = 11 \cdot \frac{1}{5.5} = 2.$$

But we found above that the required subsidy s was 2 (that is, \$2/unit). In fact, the intuitive rule for finding the right subsidy is given by the formula

$$\text{MEB in dollars } = s.$$

The subsidy should just equal the value (in dollars) of the (marginal) external benefit. This makes sense: if person 2 is causing \$2 worth of external good for every extra unit of good 1 he consumes, then the appropriate way to achieve optimality is through a subsidy of \$2 per unit on each extra unit he consumes. The externality is in this way internalized; it is plugged back into the calculation of the decision maker.

What would happen if the externality were negative? Suppose, for instance, that person 1's utility function were

$$u_1(x) = x_{11}x_{12} - x_{21}.$$

Then every extra unit of good 1 that person 2 consumes would make person 1 worse off, and person 1 would need to be given more of good 1 to compensate him for an increase in good 1 consumption by person

2. We would then have a *marginal external cost* associated with 2's consumption, defined as follows:

$$\text{MEC} = \frac{\begin{array}{c}\text{Marginal disutility to person 1 of} \\ \text{person 2's consumption of good 1}\end{array}}{\text{MU to person 1 of his own consumption of good 1}}.$$

(Or, equivalently, we could have a negative MU in the numerator and therefore a negative MEB.) We would find MEC in dollars the same way as before:

$$\text{MEC in dollars} = p_1 \cdot \text{MEC}.$$

And the tax required to correct the externality problem would be

$$\text{MEC in dollars} = t.$$

For each extra unit of good 1 that he consumes, person 2 would be required to pay a tax of t, and this tax, like the subsidy above, would internalize the externality. That is, it would be plugged into the calculations of the decision maker who is responsible for the externality.

To summarize these results: In order to re-establish the connection between competition and optimality when there are externalities present, per-unit taxes or subsidies can be imposed on consumption, and they should be chosen so that

$$\text{MEC in dollars} = t$$

or

$$\text{MEB in dollars} = s.$$

This idea of introducing taxes to take care of the market failure caused by the presence of externalities was developed by Arthur Pigou in the 1920's.

4. Pigouvian Taxes and Subsidies: A Production Example

Now let's work through a simple production externality example. We assume there are two firms and two goods. Both firms use good 1 to produce good 2. Firm 1 can be viewed as the "upstream" firm. Its production set Y_1 is determined by

$$y_{12} \leq \sqrt{-y_{11}}.$$

For instance, with nine units of the input good 1 ($y_{11} = -9$), it can produce up to three units of the output good 2. Firm 2 can be viewed

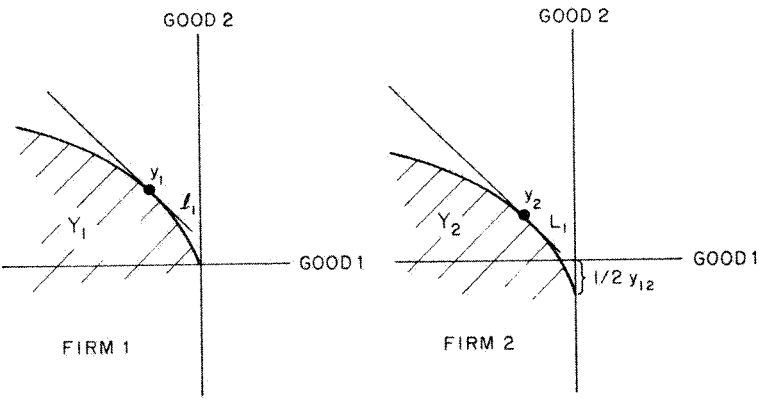


Figure 7.2.

as the "downstream" firm. Its production set Y_2 is determined by

$$y_{22} \leq \sqrt{-y_{21}} - \frac{1}{2}y_{12}.$$

That is, its production function is basically like firm 1's, except its output is shifted down by an amount depending on firm 1's output. The more firm 1 produces, the larger the shift $1/2\, y_{12}$ and the more trouble firm 2 is in. Figure 7.2 illustrates the production sets of the two firms, as well as an (unmodified) competitive equilibrium.

In the figure the production sets Y_1 and Y_2 have the same shape, but set Y_2 has been translated down by an amount $1/2\, y_{12}$.

Now suppose the prices of goods 1 and 2 are given by $p = (1, 1)$. If the isoprofit lines ℓ_1 and L_1 in the figure have slope equal to 1, in absolute value, then the firms choose the tangency points y_1 and y_2 shown. Formally, firm 1 wants to maximize $p \cdot y_1 = y_{12} + y_{11}$ subject to the constraint $y_{12} \leq \sqrt{-y_{11}}$. The solution to this maximization problem is

$$y_{11} = -\frac{1}{4}$$
$$y_{12} = \frac{1}{2}.$$

These are the coordinates of y_1 in Figure 7.2. With $y_1 = (-1/4, 1/2)$, firm 1's profit is $p \cdot y_1 = 1/4$. Firm 2 wants to maximize $p \cdot y_2 = y_{22} + y_{21}$ subject to the constraint

$$y_{22} \leq \sqrt{-y_{21}} - \frac{1}{2}y_{12} = \sqrt{-y_{21}} - \frac{1}{4}.$$

The solution to this maximization problem is

$$y_{21} = -\frac{1}{4}$$
$$y_{22} = \frac{1}{4}.$$

These are the coordinates of y_2 in Figure 7.2. With this production vector, firm 2's profit is $p \cdot y_2 = 0$.

In the unmodified competitive production economy, the net output of good 2 is $1/2 + 1/4 = 3/4$, while the net input of good 1 is, in absolute value, $1/4 + 1/4 = 1/2$. Is y an optimal production plan? Or is it possible to get more good 2 produced with the same total input of good 1? Let's consider an alternative production plan in which firm 1's input of good 1 is $z_{11} = -1/8$, and firm 2's input of good 1 is $z_{21} = -3/8$. Then firm 1's output of good 2 is given by

$$z_{12} = \sqrt{-z_{11}} = \sqrt{1/8}.$$

Firm 2's output of good 2 is given by

$$z_{22} = \sqrt{-z_{21}} - \frac{1}{2}\sqrt{z_{12}} = \sqrt{3/8} - \frac{1}{2}\sqrt{1/8}.$$

The net quantity of good 2 being produced by the two firms is then

$$z_{12} + z_{22} = \sqrt{1/8} + \sqrt{3/8} - \frac{1}{2}\sqrt{1/8} = .79.$$

In short, the production plan z yields a higher net output of good 2 than does the production plan y, and it uses the same net input of good 1. Therefore y is not technologically optimal. Because of the externality, the competitive equilibrium production plan is not optimal, and the production version of the First Fundamental Theorem breaks down.

It is not difficult to establish that the production version of the Second Fundamental Theorem also fails when the externality is present. The existence of externalities, then, breaks the link between competition and optimality.

What can be done? We could turn to a centralized socialist system, in which the Central Planning Board — aware of external effects just as it is aware of everything else — makes all production decisions for all firms. But this is the truly impractical solution. There is a decentralized solution exactly analogous to the tax-subsidy solution for consumption externalities. We turn to that tax system now.

If firm 1 increases its output of good 1 by 1 unit the northern frontier of firm 2's production frontier is shifted down by 1/2 unit. The slope of

that frontier won't change anywhere, and, providing relative prices are constant and firm 2 doesn't decide to go out of business altogether, firm 2 will still maximize its profits with the same input level. But its output will change. That is, in Figure 7.2 the point y_2 will shift down, but not sideways. Consequently, the effect on firm 2 of a unit increase in y_{12} is simply to reduce firm 2's output of good 2 by 1/2 unit, and therefore, to reduce firm 2's profit by $p_2 \cdot 1/2$ dollars. In our example, $p_2 = 1$, so the effect on firm 2 is to reduce its profit by \$1/2.

In the discussion of consumption externalities, we saw that a positive externality should be subsidized an amount equal to the marginal external benefit in dollars; and that a negative externality should be taxed an amount equal to the marginal external cost in dollars. In our production externality example, the marginal external cost to firm 2 of firm 1's production of good 2 is equal to 1/2 dollars. The way to correct the externality is once again to internalize it, that is, to force firm 1 to pay this marginal cost. Let us therefore impose a tax on firm 1 of

$$t = \frac{1}{2}.$$

That is, for every unit of good 2 that firm 1 produces, it must pay 1/2 dollar to a central authority. The central authority might redistribute the tax proceeds through lump sum grants, or use it in some other way. What happens when the tax is imposed?

Now firm 1 wants to maximize $p \cdot y_1 - ty_{12}$, subject to the constraint that y_1 be in Y_1. That is, it wants to maximize $y_{11} + y_{12} - 1/2\, y_{12}$ or $y_{11} + 1/2\, y_{12}$, subject to $y_{12} \leq \sqrt{-y_{11}}$. The solution to this maximization problem is

$$y_{11} = -\frac{1}{16}$$
$$y_{12} = \frac{1}{4}.$$

Firm 2 wants to maximize $p \cdot y_2$ subject to the constraint that y_2 be in Y_2. That is, it wants to maximize $y_{21} + y_{22}$ subject to

$$y_{22} \leq \sqrt{-y_{21}} - \frac{1}{2} y_{12} = \sqrt{-y_{21}} - \frac{1}{8}.$$

The solution to this maximization problem is

$$y_{21} = -\frac{1}{4}$$
$$y_{22} = \frac{3}{8}.$$

Under this new y, the total output of good 2 is $1/4 + 3/8 = 5/8$, while the total input of good 1 is, in absolute value, $1/16 + 1/4 = 5/16$.

Is this production plan optimal? It requires a few lines of elementary calculus to establish it, but the answer is Yes. There is no way to increase total output of good 2 without increasing the total input of good 1.

The introduction of the tax t saves the First Fundamental Theorem. The modified competitive equilibrium production plan is optimal. For the Second Fundamental Theorem we have a similar result, although now our example would have to be slightly generalized. In the example above, $p = (1, 1)$. For the Second Fundamental Theorem we would have to take $p = (p_1, p_2)$. The tax t would have to be set at $p_1 \cdot 1/2$, so that $t =$ MEC in dollars is preserved. With these modifications, we would have the result: For any optimal production plan y, there is a price vector $p = (p_1, p_2)$ and a tax rate t so that the modified market mechanism brings the production economy to y.

5. Exercises

1. An economy is made up of two people. The utility functions are:

$$u_1(x_1) = x_{11}x_{12}$$

and

$$u_2(x) = 2x_{21} + 2x_{22} - x_{11}.$$

The initial bundles are:

$$\omega_1 = (1, 0)$$
$$\omega_2 = (0, 1)$$

Although 2 suffers from 1's consumption of good 1, he cannot control it, nor does he realize that the total quantity of good 1 available is 1.

   a. Calculate a competitive equilibrium from ω. Draw an Edgeworth box diagram to illustrate your answer.

   b. Find the locus of interior Pareto optimal points.

   c. Calculate prices p_1 and p_2, a per-unit subsidy s or tax t, and lump sum cash transfers T_1 and T_2 to bring the economy to the allocation $y_1 = (1/3, 1/2), y_2 = (2/3, 1/2)$.

2. Consider an economy with the following characteristics:

$$u_1(x) = x_{11}x_{12} - x_{21}x_{22} \qquad \omega_1 = (1, 0)$$

$$u_2(x_2) = x_{21}^{1/3}x_{22}^{2/3} \qquad \omega_2 = (0, 1)$$

Person 1 hates person 2 and projects his own tastes on him. Assume that 1 does not know the total quantities of the goods available are (1,1).

Discuss the optimality (or non-optimality) of a competitive equilibrium allocation. Can you calculate taxes (or subsidies) and lump sum transfers which will bring about an efficient allocation?

Hints: 1 can't steal from 2. Also, for 2, we have the following marginal utilities:

$$\text{MU of good 1} = \frac{1}{3}x_{21}^{-2/3}x_{22}^{2/3}$$
$$\text{MU of good 2} = \frac{2}{3}x_{21}^{1/3}x_{22}^{-1/3}$$

3. (This problem requires some knowledge of calculus.) Suppose there are two firms in a production economy, each using good 1 as an input to produce good 2. They are both located in the same town, and both are bad polluters, so each one's operation has a deleterious effect on the other. Firm 1's production set Y_1 is given by

$$y_{12} \leq \sqrt{-y_{11}} - \frac{3}{4}y_{22}.$$

Firm 2's production set Y_2 is given by

$$y_{22} \leq \sqrt{-y_{21}} - \frac{3}{4}y_{12}.$$

Suppose the market prices of goods 1 and 2 are given by $p = (1, 1)$. Assume that each firm takes the other's output as given and fixed.

   a. Calculate the competitive equilibrium production plan for the economy.

   b. Show that the competitive equilibrium is not an optimum.

6. Selected References

1. F.M. Bator, "The Simple Analytics of Welfare Maximization," *American Economic Review*, V. 47, 1957, pp. 2350.

   This is listed here because it's a good companion piece for the Bator article listed below. On its own, it would belong with Chapters 3 and 4. It's basically an exposition of the two fundamental theorems, done with graphs, and without rigorous proofs

2. F.M. Bator, "The Anatomy of Market Failure," *Quarterly Journal of Economics*," 1958, pp. 351-379.

   In this article Bator discusses various different types of externalities, as well as other sources of market failures, or what we would call failures of the two fundamental theorems.

3. W.J. Baumol and W.E. Oates, *The Theory of Environmental Policy*, Prentice-Hall, Inc. Englewood Cliffs, New Jersey, 1975.

   This book provides an excellent interpretative summary of externalities, Pigouvian taxes and subsidies, environmental policy, and related issues.

4. J.M. Buchanan and W.C. Stubblebine, "Externality," *Economica*, V. 29, 1962, pp. 371-384.

   This paper attempts to clarify the definition of externality. The theoretical parts requires some calculus, but there are some nice examples that don't.

5. R.H. Coase, "The Problem of Social Cost," *Journal of Law and Economics*, V. 3, 1950, p. 1-44.

   This article solves the externality problem in a fundamentally different way. Coase suggests that many externality problems can be remedied by an appropriate assignment of rights to amenities like clean air, rather than by tax-subsidy schemes. That is, if property rights are well defined and markets in those rights function smoothly, Pareto optimality will be re-established.

6. J.E. Meade, "External Economies and Diseconomies in a Competitive Situation," *The Economic Journal*, V. 62, 1952, pp. 54-67.

   Meade shows how externality taxes and subsidies might be calculated in a production model.

7. A.C. Pigou, *The Economics of Welfare*, MacMillan and Co., Ltd. London, 1920, Part II, Chapter VI.

   Part II of Pigou's classic treatise is on "The Magnitude of the National Dividend and the Distribution of Resources Among Different Uses." Chapter VI of Part II is on divergences between private benefit and social benefit, or what we call externalities. To solve the maldistributions created by those divergences, Pigou suggests bounties or taxes, or what we now call Pigouvian subsidies and taxes.

Chapter 8

PUBLIC GOODS

1. Introduction

In the last chapter we analyzed some examples of external effects, and discussed the calculation of appropriate taxes and subsidies to correct externality problems. In this chapter we carry the externality phenomenon to its logical extreme. We shall examine the theory of the production and consumption of goods whose character is essentially public, rather than private.

What do we mean by a public good? Some goods have the property that when one person uses them, all people use them. That is, their use is nonexclusive; if the goods are available to one, they are available to all. There is no practical way for one person to use them alone.

Goods that aren't like this, goods that are really private, or exclusively used by one person, are easy to think of: a glass of beer, a set of false teeth, a pair of socks, a hamburger. When A is using or consuming one of these things, then, necessarily, B isn't. Goods whose use is necessarily nonexclusive are less common, but there are many important ones: Radio and television broadcasts (unless scrambled) are nonexclusive. If A can get the TV signal and B lives nearby, then B can get the TV signal also. (Note that we are talking about the signal, not the TV set, which is a private good. Also note that we are not talking about cable TV, the access to which is again a private good.) There is no practical way to deliver radio waves to A without simultaneously delivering them to B. The signals of a lighthouse are nonexclusive. If a lighthouse is warning ship A to stay away from the shoals, then it is also necessarily warning any nearby ship B to stay clear. The outside of the Washington Monument is nonexclusive. If person A can view it and enjoy it, person

B can do so as well. It wouldn't be practical to screen it and charge to let people inside the screen. On the other hand, it is practically possible to exclude people from the inside of the Washington Monument. The Statue of Liberty, viewed from outside, is a nonexclusive good. If it's available for one to see, it's available to all.

National defense is an important nonexclusive good. If the person and property of U.S. resident A are being protected from foreign armies, then the person and property of U.S. resident B are also necessarily being protected from those armies.

Another important example is scientific and technological knowledge. Some technological knowledge is patentable and its use can be restricted, but a larger part is not. The technology of the internal combustion engine is nonexclusive. If it is available in person A's library, it is probably available in B's. Medical knowledge is partly nonexclusive. If a cure for infection, like penicillin, is known to A's doctor, then it is known to B's. When someone discovers a cure for some form of cancer, that cure will ultimately be public knowledge, and nonexclusive in its use. We say ultimately because it will likely be patented and restricted in use initially.

In each of these cases, when the good is there for one, it is necessarily there for all. Goods with this property are called public goods.

What are the efficiency implications of public goods? Like externalities, public goods undermine optimality in a standard competitive equilibrium. What then should be done? What are the optimal quantities of public goods? How should public goods be financed? We shall explore these and similar questions in this chapter.

2. The Public Goods Model

In this section we shall develop a rather special, and rather different model to analyze the problems of public goods. A good is public if it is by nature available to all: if one man uses it, everyone can use it. Public goods can be viewed as goods with extreme external effects: if person i's consumption of the good is X, then X appears in each and every person's utility function. However, we won't continue the externality and Pigouvian tax/subsidy analysis of the last chapter: it is more convenient to start anew.

The model we use here has both production and consumption, because one principal question we want to answer is this: How much of the public good should be produced? And the answer to the question depends both on people's demand for it, and on the nature of the productive sector of the economy. But in order to avoid notational and analytical complexity, our model will be exceedingly simple. We assume that there are only two

goods, one private, and one public. Also we assume that the productive sector of the economy can transform units of the private good into units of the public good, in the ratio of one to one. And, therefore, we assume the equilibrium prices of the two goods are 1 and 1.

Our model will also make a crucial simplifying assumption about the nature of utility functions. We assume that person i's utility is the sum of the quantity of the private good he consumes, plus a well-behaved function of the quantity of the public good produced and available to all, including i. Such a utility function is said to be *separable* between private and public consumption. Some of the analysis below hinges on this special assumption; some does not. We make the assumption for two reasons: (1) it greatly simplifies all the mathematics, and (2) the discussion of demand revealing taxes breaks down without it.

Now let's develop some of the notation. It should be observed that this notation differs slightly from what is used in the exchange and production models treated above. First, we let

$$x = \text{ the quantity of the public good.}$$

Note that x is a scalar, not a vector. Also, note that x can be viewed as the quantity (or size) of the public good in physical units, or in dollars, since we assume that the prices of both the public and the private good are one. Second, we let

$$y_i = \text{ person } i\text{'s quantity of the private good.}$$

Note that y_i is a scalar. We assume that person i's utility function u_i can be written

$$u_i = v_i(x) + y_i.$$

That is, i's utility is the sum of the function v_i, which depends only on x, plus i's quantity of the private good. We also assume that v_i is continuous, smooth, monotonic, and concave; that is, it looks like the one in Figure 8.1. (Actually, monotonicity can largely be relaxed. We could assume instead that only *some* individuals' utility functions are monotonic.)

In the figure, the function v_i is smooth and concave, that is, it bends downward. The intercept a might be positive or negative, or v_i might even be asymptotic to the (negative half of the) vertical axis. At the point P, the ratio $\Delta v_i/\Delta x$ is person i's *marginal utility* from the public good, or approximately the amount by which his utility rises if the quantity of the public good is increased by 1, while his private consumption remains fixed. Note that for sufficiently small Δ's, $\Delta v_i/\Delta x$ and the slope of the v_i function at P are equal. Instead of writing MU of the

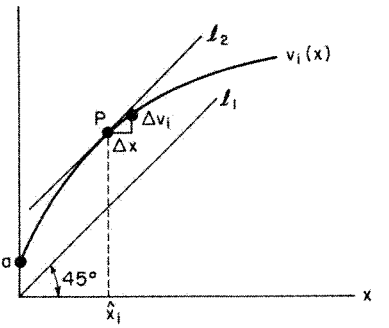


Figure 8.1.

public good for i, we now use this simpler notation:

$$v_i'(x) = \text{ MU of the public good for } i, \text{ when the quantity of the public good is } x.$$

We also assume that each person starts with an initial endowment of the private good. We let

$$\omega_i = i\text{'s initial quantity of the private good.}$$

Note that ω_i is also a scalar.

In order to be *feasible*, a vector of public and private good consumption levels $(x, y_1, y_2, \ldots, y_n)$ must satisfy this condition:

$$x + \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \omega_i.$$

(Later in this chapter, we will relax the feasibility condition to $x + \sum_{i=1}^{n} y_i \leq \sum_{i=1}^{n} \omega_i$.)

3. The Samuelson Public Good Optimality Condition

Let's now see how a Pareto optimal output for the public good can be found. In order to do this, we start by considering the inequality

$$v_1'(x) + v_2'(x) + \ldots + v_n'(x) > 1.$$

That is, we consider an output x of the public good, such that the sum of the marginal utilities of all individuals at x exceeds 1. Let's assume for mathematical simplicity that each person's private consumption y_i

exceeds his marginal utility from public consumption $v_i'(x)$. Now suppose we reduce each person's consumption of the private good by an amount $v_i'(x)$; that is, we define a new amount of private good consumption for each i:

$$\bar{y}_i = y_i - v_i'(x).$$

If we stop here, then each person's utility must drop by the amount of the private good he loses, $v_i'(x)$. But we have extracted

$$v_1'(x) + v_2'(x) + \ldots + v_n'(x) = 1 + \Delta$$

units of private good from the economy, where Δ is some number greater than zero. Now suppose we take 1 unit of private good from this total (leaving Δ) and send it to the productive sector of the economy (or the firm) to be transformed into 1 unit of public good. Then we get a new public good output of

$$\bar{x} = x + 1.$$

But increasing the public good available by a unit increases each person's utility by an amount approximately equal to the marginal utility of the public good, or, for $i, v_i'(x)$. Therefore, at $(\bar{x}, \bar{y}_1, \bar{y}_2, \ldots, \bar{y}_n)$ each person is as well off as he was at $(x, y_1, y_2, \ldots, y_n)$. But Δ units of the private good are left over. This amount can be redistributed among the individuals to make some (or all) better off than they used to be at $(x, y_1, y_2, \ldots, y_n)$. Consequently, if

$$v_1'(x) + v_2'(x) + \ldots + v_n'(x) > 1,$$

it is possible to make some people better off and no one worse off through an appropriately financed increase in the public good x. Therefore, x is not a Pareto optimal output for the public good.

This result shows us when the output of the public good ought to increase. But as x increases, $v_i'(x)$ decreases for every i, because of the shapes of the v_i functions assumed in Figure 8.1. Unless all the private good is exhausted first, if x is continually increased, we eventually reach a point where

$$v_1'(x) + v_2'(x) + \ldots + v_n'(x) < 1.$$

An argument similar to the one made above establishes that, when this inequality holds, it is possible to make some people better off and no one worse off through a decrease in public good x, with the savings appropriately distributed among the individuals. Therefore, x is not the Pareto optimal level of output for the public good.

We have seen that when

$$v_1'(x) + v_2'(x) + \ldots + v_n'(x) > 1$$

x is too small, and is not Pareto optimal, and that when

$$v_1'(x) + v_2'(x) + \ldots + v_n'(x) < 1$$

x is too large, and is not Pareto optimal either. A necessary condition for Pareto optimality, therefore, is

$$v_1'(x) + v_2'(x) + \ldots + v_n'(x) = \sum_{i=1}^{n} v_i'(x) = 1.$$

That is, the sum of the marginal utilities for the public good should equal the marginal cost of producing the public good, in terms of foregone private good. Named after Paul Samuelson, this is called the *Samuelson optimality condition for public goods.*

There is another simple way to view the Samuelson optimality condition. Consider the expression

$$v_1(x) + v_2(x) + \ldots + v_n(x) - x = \sum_{i=1}^{n} v_i(x) - x.$$

This can be interpreted as the *aggregate net benefit* of the public good output level x. Now it's rather clear that if x does not maximize aggregate net benefit it cannot be Pareto optimal: If there is an alternative level of output $\bar{x}$ that gives a higher aggregate net benefit than does x, there must be some way to shift from x to $\bar{x}$ and make everyone better off in the shift. In fact, a simple application of calculus indicates that maximizing

$$v_1(x) + v_2(x) + \ldots + v_n(x) - x$$

leads directly to the Samuelson condition.

In short, maximizing aggregate net benefit

$$\sum_{i=1}^{n} v_i(x) - x$$

is necessary for achieving the Pareto optimal output of the public good, just as the Samuelson condition

$$\sum_{i=1}^{n} v_i'(x) = 1$$

is necessary for achieving the Pareto optimal x. Moreover, the two conditions are also sufficient, under our assumptions of concave utility for the public good and the equality in the definition of feasible allocations. Consequently, the Pareto optimal output of the public good is determined by the Samuelson condition, or, equivalently, the condition that aggregate net benefit be maximized.

4. Private Financing of the Public Good and the Free Rider Problem:

Now let's consider how the public good is financed. In this section we assume that it is privately purchased. This means individual i might pay for a certain amount of the public good, which would then be available to all. (Some examples of privately purchased public goods are privately owned parks open to the public, privately owned pieces of art on display in a public museum, and private contributions to charitable organizations.) We are assuming, then, that the unmodified (private) market mechanism is being used to supply the public good. In subsequent sections we'll analyze public (that is, government) financing of the public good.

To start the analysis, we suppose that person i takes the lead; he makes the first purchase of the public good. How much does he buy? He wants to choose an x to maximize his utility

$$u_i = v_i(x) + y_i$$

subject to his budget constraint

$$1 \cdot x + 1 \cdot y_i = 1 \cdot \omega_i.$$

Substituting for y_i, person i wants to maximize $u_i = v_i(x) - x + \omega_i$. The graphical solution to the problem can be seen in Figure 8.1. In that figure, the lines ℓ_1 and ℓ_2 have slope 1. Maximizing $v_i(x) - x + \omega_i$ is equivalent to maximizing $v_i(x) - x$, the vertical distance between the v_i function and the line ℓ_1. This vertical distance is greatest at the point P, where v_i is tangent to the line ℓ_2. At the tangency point, the slope of the v_i function equals the slope of the line ℓ_2, or

$$v_i'(x) = 1.$$

We shall let $\hat{x}_i$ be the quantity of the public good that i would choose to purchase privately. Note that the subscript i serves here only to remind us that i is making a private purchasing decision; the good is still public; and i cannot exclude others from enjoying the benefits of his purchase.

Now if person i has purchased $\hat{x}_i$ units of the public good what do the others do? Each of the others is enjoying the benefits from i's purchase without paying for those benefits. To be more precise, let's suppose first that $v_j'(\hat{x}_i) < 1$: That is, the marginal utility to person j from the public good is less than the marginal cost of the public good, given person i's (generous) decision to purchase $\hat{x}_i$ (and provide it to all). In this case, if j were to curtail his own private consumption by 1 unit, and purchase an additional unit of the public good for his (and everyone

else's) extra enjoyment, he would be worse off. So he won't do it. That is, if $v_j'(\hat{x}_i) < 1$, person j will be content to let i buy the public good, he will not buy any himself. He will take a free ride on i's consumption of the public good. On the other hand, if $v_j'(\hat{x}_i) > 1$, then j would benefit if he curtailed his private consumption and purchased more units of the public good for his (and everyone else's) extra enjoyment. How much would he purchase? He would want the quantity of the public good increased until $v_j'(x) = 1$, since whenever $v_j'(x) > 1$, person j makes himself better off by reducing his consumption of the private good and increasing his (and everyone else's) consumption of the public good.

The final equilibrium in the unmodified private market for public and private goods will have these properties: For (at least) one person i, we will have $v_i'(x) = 1$. For all i, we will have $v_i'(x) \leq 1$. And the financing — the details of who pays how much — will largely be determined by who made the first purchase, who the second, and so on. But if $v_i'(x) = 1$ for some i, and if (as we are obviously assuming) $v_i'(x) > 0$ for the other i's, then

$$v_1'(x) + v_2'(x) + \ldots + v_n'(x) > 1.$$

That is, the sum of the marginal utilities will exceed the marginal cost of the public good and x will not be Pareto optimal. It will generally be possible to make some people better off and no one worse off through an appropriately financed increase in the public good. There will not be enough of the public good produced.

To get a clear intuitive idea of what's causing this insufficiency of public good production, think of the case where v_i is the same function for all i. In other words, everyone's tastes are the same. Now if a particular person, say person 1, takes the first step, and purchases $\hat{x}_1$ of the public good, what do the others do? They all take a free ride, since $\hat{x}_1 = \hat{x}_2 = \ldots = \hat{x}_n$. Persons 2 through n are free riders on the purchase of the public good by person 1: they enjoy the benefits and pay none of the costs. It's no surprise that 1 doesn't choose the optimal quantity of the public good; he figures his own benefit against the total cost, and pays no attention to the benefits enjoyed by the free riders. And in this particular instance, there is a very large discrepancy between

$$\sum_{i=1}^{n} v_i'(x) \text{ and the marginal cost of the public good, since}$$

$$\sum_{i=1}^{n} v_i'(x) = 1 + 1 + \ldots + 1 = n \gg 1.$$

Consequently, there is probably a large difference between $\hat{x}_1$ and the optimal quantity of the public good.

To sum up this section, the problem with the private provision of public goods is the problem of free riders. Those who enjoy the public good without paying for it never signal their desire for it. Consequently, not enough of the public good is provided. This is why public goods ought to be, in some way, publicly financed.

5. The Wicksell-Lindahl Tax Scheme

We now focus on a system in which the public good is publicly financed. A central government authority called the Public Good Board decides on the output of the public good x, and collects taxes to pay for x. What would we like this system to accomplish? First, it should somehow provide for a Pareto optimal output of the public good. Any x that is not Pareto optimal is unambiguously unsatisfactory. And, second, it ought to link a person's taxes to the benefits he receives. This is partly a matter of common sense and partly a matter of justice. A person's benefits from his private consumption are clearly linked to his payments for private consumption. And that linkage works well: it produces a Pareto optimal outcome in a private-good economy. So common sense suggests the linkage might be useful in an economy with a public good. Moreover, ever since Knut Wicksell wrote about "just" taxation in 1896, economists have occasionally suggested linking taxes and benefits because it's "just" to do so. Why should a person who gets little or no utility from the public good pay the same taxes as a person who gets lots of utility from it? As Wicksell wrote, "it would seem to be a blatant injustice if someone should be forced to contribute toward the costs of some activity which does not further his interests or may even be diametrically opposed to them."

Optimality and linkage are, then, the guiding principles in the taxation system named after Wicksell and Erik Lindahl. To explain the scheme we need a little more notation. Again, this notation is specific to our discussion of public goods, and should not be confused with similar notation we have used before.

We shall let T_i stand for person i's total tax payment to the Public Good Board. With the total tax T_i, i's budget becomes

$$y_i + T_i = \omega_i.$$

We are continuing to assume that the prices of the public and private goods are 1 and 1, respectively, so T_i can be viewed as a payment in dollars (or currency), or as a payment in units of the private good. We shall assume in this section that each person i has to bear a fractional share of the expenditure on the public good, and we shall let t_i stand for person i's share. Since the sum of the fractional shares of the individuals

must be 1, we have

$$\sum_{i=1}^{n} t_i = 1.$$

Also, if i's share is t_i and the total quantity of (or expenditure on) the public good is x, then necessarily

$$T_i = t_i x.$$

Now suppose we confront individual i with this question: "If your share $t_i = 1/4$, what quantity of the public good do you want produced?" Individual i thinks to himself: "I shall pay $T_i = 1/4 \cdot x$, so the x I want is the one that maximizes $u_i = v_i(x) + y_i$ subject to $y_i + T_i = \omega_i$, or $y_i + \frac{1}{4}x = \omega_i$. In short, I want the one that maximizes $v_i(x) - \frac{1}{4}x$." The graphical solution to this problem can be found by redrawing Figure 8.1 in such a way that the lines ℓ_1 and ℓ_2 have slope $= 1/4$, rather than slope $= 1$. If we confront individual i with this question: "If your share $t_i = 1/2$, what quantity of the public good do you want produced?" He will go through the same calculations, except with $1/2$ where $1/4$ used to be. And he will obviously come up with a different answer. In Figure 8.2 we have drawn a v_i function and lines ℓ_1 and ℓ_2 for the general case: ℓ_1 and ℓ_2 have slope $= t_i$. The symbol $\hat{x}_i(t_i)$ represents the quantity of the public good i wants produced, given that his share of the cost is t_i.

For the general case, i wants to maximize $u_i = v_i(x) + y_i$ subject to $y_i + T_i = \omega_i$ or $y_i + t_i x = \omega_i$. In short, he wants to maximize $v_i(x) - t_i x + \omega_i$, or, equivalently, $v_i(x) - t_i x$. This quantity is greatest when the vertical distance between the v_i function and the line ℓ_1 is greatest, which occurs at the point P where ℓ_2, a line parallel to ℓ_1, is tangent to v_i. At that tangency point, the slope of ℓ_2 equals the slope of v_i. But the slope of ℓ_2 is t_i, and the slope of v_i is i's marginal utility from the public good, or $v_i'(x)$. Therefore, the point P, and i's desired quantity of the public good $\hat{x}_i(t_i)$, are determined by the equation

$$v_i'(x) = t_i.$$

A careful examination of Figure 8.2 should convince the reader of this general result: The higher is person i's share t_i, the lower is the quantity of the public good $\hat{x}_i(t_i)$ he wants produced. This makes intuitive sense; it's analogous to the Law of Demand for private goods: the higher the price of a private good, the less the individual wants to purchase, all else equal.

It is crucial to note that if the actual output of the public good happens to coincide with i's desired output of the public good, or if

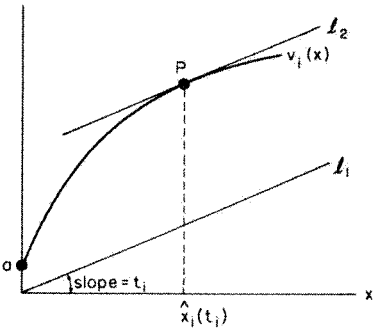


Figure 8.2.

$x = \hat{x}_i(t_i)$, then i's marginal utility from the public good $v_i'(x)$ equals his tax share t_i. That is, his tax share is linked to his marginal benefit.

But how can the actual output be made to agree with i's desired output? After all, each of the n people has his own $\hat{x}_i(t_i)$ function, his own schedule of desired outputs contingent on t_i, and given any list of fractional shares $(t_i, t_2, \ldots, t_n)$, each person will have his own particular desired output of the public good. How can the Public Good Board find a way to insure that each and every person's tax share is linked to his marginal benefit? The Wicksell-Lindahl tax scheme answers this question.

The trick of this tax scheme is to adjust the tax shares until every person agrees on the desired output of the public good. For instance, suppose there are just two people, and, when $t_1 = t_2 = 1/2$, person 1's desired public good output is $\hat{x}_1(1/2) = 10$, while person 2's desired public good output is $\hat{x}_2(1/2) = 20$. Given the shares $(1/2, 1/2)$, they disagree about the best level of output. No matter what level of output is actually chosen, for at least one of the two, his tax share will diverge from his marginal benefit. The solution? Gradually decrease t_1, and increase t_2. As t_1 decreases, 1 wants more and more of the public good produced. As t_2 increases, 2 wants less and less of the public good produced. Eventually, a point is reached where each thinks the same quantity of the public good should be produced. Say that point is reached when $t_1 = 1/3$ and $t_2 = 2/3$ and say $\hat{x}_1(1/3) = 14 = \hat{x}_2(2/3)$. Then, the Public Good Board assigns person 1 a tax share $t_1 = 1/3$, and person 2 a tax share $t_2 = 2/3$, and it has 14 units of the public good produced. Under these circumstances, each person's tax share is linked to the marginal utility he gets from the public good.

Formally, a *Lindahl equilibrium* is defined to be a vector of tax shares $(t_1, t_2, \ldots, t_n)$ and a level of output $\hat{x}$ for the public good, such that, for all i, when i's tax share is t_i his desired level of public good output equals $\hat{x}$. That is, for all i, $\hat{x}$ maximizes $u_i = v_i(x) + y_i$ subject to the budget constraint $y_i + t_i\, x = \omega_i$.

It is clear from its definition that a Lindahl equilibrium, or the Wicksell-Lindahl tax scheme, equates tax shares and marginal benefits for every person. That is, what people pay is connected to what they get. It should be emphasized that the connection is with marginal utility rather than total utility, but at least the linkage is there. What about optimality? Does a Lindahl equilibrium level of public good output $\hat{x}$ have to be optimal? The answer is Yes.

Recall that i's desired quantity of the public good $\hat{x}_i(t_i)$ is determined by the equation

$$v_i'(x) = t_i.$$

Therefore, since a Lindahl equilibrium $\hat{x}$ equals $\hat{x}_i(t_i)$ for every i, we must have

$$v_i'(\hat{x}) = t_i \text{ for all } i.$$

Summing over all the i's gives

$$v_1'(\hat{x}) + v_2'(\hat{x}) + \ldots + v_n'(\hat{x}) = t_1 + t_2 + \ldots + t_n = 1.$$

Consequently, the Samuelson condition is satisfied by $\hat{x}$, and $\hat{x}$ is the optimal quantity of the public good.

The Wicksell-Lindahl scheme then accomplishes the two things we set out for it: optimality and linkage. But it does have several drawbacks.

First, a Lindahl equilibrium might not exist. The adjustment process described above may not lead to a successful end. For example, if the utility functions of the two individuals are $v_1(x) = (3/4)x + \ln x$ and $v_2(x) = (1/4)x + \ln x$, there is no equilibrium. This is because the sum of the marginal utilities from the public good of the two individuals never equals 1. One can add an extra condition on preferences to guarantee existence: for example, that the sum of the marginal utilities of the individuals when x becomes large enough be less than 1. (This amounts to saying that as the public good x becomes very abundant, the social valuation for it falls below its marginal cost, something that sounds plausible).

Second, if Lindahl tax shares are viewed as "personalized prices" attached to "personalized commodities" (person i would be purchasing his amount of the public good), the equilibrium concept only makes good sense if the consumption of the public good is excludable, something

that goes against the nature of many public goods. See the discussion in Mas-Colell, Whinston and Green (1995, Chapter 11) on this point.

Third, the Lindahl equilibrium concept is one in which individuals are "price-takers," i.e., they take their tax shares, or personalized prices, as given. As we will see below, price-taking behavior is a heroic assumption in this setting: in a sense, there is a problem of bilateral monopoly between the Public Good Board and each individual, to determine $\hat{x}_i$.

Fourth and finally, there is a problem of incentives. When the Public Good Board calculates the Lindahl equilibrium, when it calculates the appropriate vector of tax shares $(t_1, t_2, \ldots, t_n)$ and the output for the public good $\hat{x}$, it relies on information it receives from the individuals. It needs each individual's $\hat{x}_i(t_i)$ schedule, which we might call i's demand schedule for the public good. Or, equivalently, it needs each individual's marginal utility function v_i', or his total utility function v_i. The three functions are really interchangeable, but the Public Good Board must ask for one of them. Now put yourself in the place of a bright individual who has just been asked to provide one of these schedules, one of these functions. Suppose you know how the Public Good Board operates; you know that you will end up with a tax share t_i equal to your marginal utility $v_i'(\hat{x})$. What will you report to the Board?

If you are at all devious, you will lie about your utility, or marginal utility, or demand function. You'll say the public good is useless to you; or you'll say that your marginal utility from the public good is zero for all relevant levels of output. You won't reveal your true demand. You'll misrepresent your preferences and take a free ride. The incentives here are not compatible with truthful answers.

And, of course, if everyone is lying like mad about his demand or marginal utility function, the Public Good Board is not likely to reach an equilibrium that links tax shares to real marginal utilities, or that is Pareto optimal given the real utility functions.

Let us note that there is also an incentive compatibility or a demand-revealing problem in a model with only private goods. But it's much less severe. It might well be to a person's advantage to understate his demand for a private good in order to put downward pressure on the price of that good. But if he does so he'll probably end up with significantly less of the good. And if n is large, if there are many people, his misrepresentation will have very little effect on the price. So if he says "I don't want the good, it's of no use to me" he will in fact end up with a lot less of it than he would have otherwise, and he'll pay almost as much per unit as he would have otherwise. Contrast this situation to the public good model we've just analyzed. Suppose there are lots of people, and a particular person says to the Public Good Board "I

don't want the good, it's of no use to me." His falsehood will induce the Board to produce a slightly smaller $\hat{x}$; but only slightly smaller, when n is large. And under the Wicksell-Lindahl tax scheme, his tax share will drop dramatically. So the person who is willing to conceal his demand for the public good will end up with almost as much of it as he would have otherwise, and he'll pay a lot less per unit than the would have otherwise!

The critical problem with the Wicksell-Lindahl tax scheme is this problem of demand-revelation. People will not want to reveal their true feelings to the Public Good Board. This is why we must look elsewhere for an ideal public finance scheme. We do so in the rest of this chapter, and more generally, in the last chapters of the book, on the theory of implementation.

6. Fixed Tax Shares and Majority Voting

Before continuing the search for a theoretically ideal tax scheme, let's digress slightly, and carefully examine a realistic method for financing the public good and determining the quantity that ought to be produced. In this scheme the tax shares $t_1, t_2, \ldots, t_n$ are fixed. And the amount to produce, the quantity x, is determined by majority voting.

As an example of a fixed tax shares scheme, we might have $t_i = 1/n$, for all i. That is, each person might pay an equal share of the cost of producing x. This is obviously a commonly used scheme. As another example, the t_i's might be proportional to the ω_i's. That is, we might have

$$t_i = \frac{\omega_i}{\sum_{i=1}^{n} \omega_i} \text{ for all } i.$$

The person who starts out the richest pays the highest tax; the poorest pays the lowest. Note that in both of these examples the t_i's sum to unity.

Now let's analyze what individual i wants under this system of finance. When informed that his share is t_i, i thinks of how to maximize $u_i = v_i(x) + y_i$ subject to $y_i + t_i x = \omega_i$. That is, he wants to maximize $v_i(x) - t_i x$. We've already seen in the analysis of the Wicksell-Lindahl tax scheme that person i would most prefer the level of public expenditure $\hat{x}_i(t_i)$ shown in Figure 8.2. However, if given a choice between any two public good levels x_1 and x_2, and asked to vote between those two, he would very probably vote for the one for which $v_i(x) - t_i x$ is greater.

How is an equilibrium found? Each person i has his favorite level of output $\hat{x}_i(t_i)$. For some i, this is small, for some i, it is large. For instance, if persons 1 and 2 have the same v_i function, but t_1 is big,

while t_2 is small, then 1 will have a small $\hat{x}_1(t_1)$, and 2 will have a large $\hat{x}_2(t_2)$. The one who pays a bigger share will want a smaller project. Now for the sake of mathematical simplicity, let's assume here that all the $\hat{x}_i(t_i)$'s are distinct, that the people are numbered in such a way that

$$\hat{x}_1(t_1) < \hat{x}_2(t_2) < \ldots < \hat{x}_n(t_n)$$

and that the number of people n is odd. Let person M be the *median person*, the person whose $\hat{x}_M(t_M)$ is in the middle. That is, there are as many $\hat{x}_i(t_i)$'s less than $\hat{x}_M(t_M)$ as there are $\hat{x}_i(t_i)$'s greater than $x_M(t_M)$.

In the fixed tax shares, majority voting system of public finance, the Public Good Board conducts a sequence of elections, elections in which the candidates are levels of output for the public good. For simplicity, we shall assume that the list of candidates is just the set $\{\hat{x}_1(t_1), \ldots, \hat{x}_n(t_n)\}$. The Public Good Board conducts these elections until it finds a level of output $\hat{x}_i(t_i)$ which wins a majority over any other level of output $\hat{x}_j(t_j)$. The winning $\hat{x}_i(t_i)$ is the *fixed tax shares, majority voting equilibrium.*

And it turns out that the equilibrium must be $\hat{x}_M(t_M)$, that is, the desired level of output of the median voter. Let's just briefly indicate why this ought to be the case. (For a fuller treatment see Chapter 12). In what follows we will write $\hat{x}$ instead of $\hat{x}_M(t_m)$. Now consider a vote between $\hat{x}$ and some $\hat{x}_i(t_i) < \hat{x}$. Some reflection should convince you that any person $j \geq M$ will want to vote for $\hat{x}$. But since M is the median, the people numbered M or above make up a majority. So $\hat{x}$ wins a majority vote over $\hat{x}_i(t_i)$. Similarly, in a vote between $\hat{x}$ and any $\hat{x}_i(t_i) > \hat{x}$, $\hat{x}$ wins another majority, since it gets the votes of everyone numbered M or below.

What are the advantages of this system of public finance? First of all, it is relatively simple and comprehensible. It can be easily understood by the people reporting their desired quantities of the public good, and voting on those quantities. Second, unlike the Wicksell-Lindahl scheme, the incentives for misrepresentation and duplicity don't stand out like a sore thumb. But if subtle, these incentives might still be there. We have deliberately been vague about the exact nature of the voting process, or what agenda the Board uses. It is possible that, given certain agendas, people might vote against a preferred expenditure level at one stage, in order to end up with a better outcome at a later stage. This possibility becomes a virtual certainty if there are two or more public good expenditure levels being chosen simultaneously. Nonetheless, the incentives to lie are not as glaring in the fixed tax share majority voting scheme as they are in the Wicksell-Lindahl scheme.

What are the disadvantages? First of all, there is no linkage between a person's tax share t_i and his utility, or marginal utility from the public good, except for the median person M. For M we know that t_M must equal the marginal utility from the public good at the equilibrium level of output $\hat{x}$. But for every $i \neq M$,

$$v_i'(\hat{x}) \neq t_i.$$

Half of the people have tax shares less than their marginal utilities, half have tax shares greater.

The second and crucial disadvantage is that $\hat{x}$ is generally not Pareto optimal. Let's consider the Samuelson test. In the sum

$$v_1'(\hat{x}) + v_2'(\hat{x}) + \ldots + v_{M-1}'(\hat{x}) + v_M'(\hat{x}) + v_{M+1}'(\hat{x}) + \ldots + v_n'(\hat{x}),$$

we know that the first $M-1$ numbers are all less than the corresponding t_i's; we know that $v_M'(\hat{x}) = t_M$; and we know that the last $M-1$ numbers are all greater than the corresponding t_i's. But we do not know, and in fact it is generally not true, that the sum of the n numbers is equal to

$$\sum_{i=1}^{n} t_i = 1.$$

That is, in general

$$v_1'(\hat{x}) + v_2'(\hat{x}) + \ldots + v_n'(\hat{x}) \neq 1.$$

Therefore, in general $\hat{x}$ fails the Samuelson test; it is not Pareto optimal. Whether $\hat{x}$ is too large, that is,

$$\sum_{i=1}^{n} v_i'(\hat{x}) < 1$$

or too small, that is,

$$\sum_{i=1}^{n} v_i'(\hat{x}) > 1$$

will depend on the circumstances. But the chances are slim that $\hat{x}$ will be just right.

7. The Demand-Revealing Tax Scheme

At this point we return to our search for a theoretically ideal tax scheme. Let's examine another approach, one whose basic virtues are

(1) that it leads to the optimal output for the public good, and (2) that it provides no incentives for individuals to misrepresent their demands for the public good. Because of the latter virtue (which is so dramatically missing in the Wicksell-Lindahl scheme), this public finance rule is said to solve the incentive problem or to be *incentive compatible*, or to be *demand-revealing*. The demand revealing tax scheme was developed independently by Edward Clarke and Theodore Groves in the 1970's. A similar scheme to discourage speculation in private goods markets had been developed by William Vickrey in 1961. The name "demand-revealing" was coined by T. Nicolaus Tideman and Gordon Tullock.

We depart in this and the next section from one important previous assumption. That is, we relax the feasibility equality $x + \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \omega_i$ to an inequality:

$$x + \sum_{i=1}^{n} y_i \leq \sum_{i=1}^{n} \omega_i.$$

This effectively means that the Public Good Board's budget is now not required to be exactly balanced. It may have a surplus. We'll come back to this issue in the sequel.

In the demand-revealing tax system, each individual sends a "message" to the Public Good Board, a message about his feelings for the public good. We shall assume that the message is a utility function for the public good, that is, a v_i function. This might be i's real utility-from-the-public-good function, or it might not. We shall show that i will not be able to gain by lying, so he will report his true function. But for now, when i reports a v_i function, it must be viewed with suspicion.

What does the Public Good Board do with the (possibly fraudulent) v_i functions? First, it derives from each v_i function a (possibly fraudulent) marginal utility function v_i', and, with these marginal utility functions it uses the Samuelson condition to solve for a level of expenditure on the public good. That is, it solves for an $\hat{x}$ that satisfies the equation

$$\sum_{i=1}^{n} v_i'(x) = 1.$$

We saw above that this exercise is equivalent to the maximization of aggregate net benefit from the public good. Therefore, we can also say that the Board finds an $\hat{x}$ that maximizes the expression

$$\sum_{i=1}^{n} v_i(x) - x.$$

Now we know that if the v_i functions are true, if individuals aren't misrepresenting their preferences, then $\hat{x}$ is a Pareto optimal level of output for the public good.

Second, the Public Good Board sets taxes. This, of course, is the delicate part, since we have seen in the case of the Wicksell-Lindahl scheme that the tax rule can tempt people to lie. Allowing the Public Good Board to run an (actual) surplus or a (provisional) deficit will help solve this problem. That is, in the demand-revealing scheme, the Board does not set tax shares or t_i's, it simply sets total tax payments, or T_i's. We shall assume for now that it does so according to the following rule:

$$T_i = \hat{x} - \sum_{j \neq i} v_j(\hat{x}).$$

The symbol "$\sum_{j \neq i} v_j(\hat{x})$" means "sum the v_j's over all the people except i."

That is, person i's tax equals the entire cost of the public good (or its level of output) less the aggregate utility accruing to other people from the public good. For instance, suppose that $\hat{x}$ is a \$1,000 bridge, and suppose there are five users who assign the following values to it:

$$v_1(\hat{x}) = 0, v_2(\hat{x}) = 500, v_3(\hat{x}) = 100, v_4(\hat{x}) = 200 \text{ and } v_5(\hat{x}) = 300.$$

Then, $T_1 = -\$100$, $T_2 = \$400$, $T_3 = 0$, $T_4 = \$100$, and $T_5 = \$200$. Notice that T_1 is negative, which means person 1 is getting a grant rather than paying a tax. Also notice that the sum of the T_i's equals only \$600, which is not enough to pay for the bridge! We'll fix up this problem of insufficient funds below.

Let's make an important observation about the rule for finding i's tax T_i: Person i's message, his reported v_i function, does not appear directly in the definition of T_i. His expressed demand for the public good does have an indirect effect on T_i, since it enters into the Board's determination of $\hat{x}$. But if i were to understate his demand, say by declaring that $v_i(x)$ were zero for all levels of x, he would still pay a tax, since T_i depends on the chosen quantity of the public good and everyone else's expressed valuation of that chosen quantity, but not i's.

Next let's see whether or not the actions of the Public Good Board provide each person with the proper incentives to honestly reveal his v_i function.

We consider a particular fixed i. Suppose persons $j = 1, 2, \ldots, i-1, i+1, \ldots, n$ have reported their v_j functions, which might be real or whimsical; i doesn't know. Person i of course wants to maximize his real utility

$$u_i = v_i(\hat{x}) + y_i$$

subject to $y_i + T_i = \omega_i$. That is, he wants to maximize $v_i(\hat{x}) - T_i + \omega_i$, or, equivalently, $v_i(\hat{x}) - T_i$. Of course, it is the Board rather than i that chooses $\hat{x}$ and T_i, which might seem to tie i's hands, but i does have the option of lying about v_i and thereby directly affecting $\hat{x}$ and indirectly affecting T_i.

Now person i, along with everyone else, knows the tax rule; he knows that the Board will always set

$$T_i = \hat{x} - \sum_{j \neq i} v_j(\hat{x}).$$

Consequently, i wants to maximize the following expression for his after-tax utility:

$$v_i(\hat{x}) - \left[\hat{x} - \sum_{j \neq i} v_j(\hat{x}) \right].$$

At this point let's use a $\sim$ to indicate a false function, or a quantity that results when i lies, and no $\sim$ to indicate a true function, or a quantity that results when i tells the truth. If i lies, or reports $\tilde{v}_i$, the Board comes up with $\tilde{x}$, and if i tells the truth, or reports v_i, the Board comes up with $\hat{x}$. The question is, can i lie in such a way that he is (truly!) better off?

Well, person i knows that the Public Good Board always chooses $\hat{x}$ to maximize aggregate net (reported) benefit. In particular, if i tells the truth, if he reports v_i, while the others are reporting (true or false) v_j's, the Board will choose $\hat{x}$ to maximize

$$\sum_{i=1}^{n} v_i(x) - x = v_i(x) + \sum_{j \neq i} v_j(x) - x = v_i(x) - \left[x - \sum_{j \neq i} v_j(x) \right].$$

That is, if i reports his true v_i, the Board will attempt to maximize precisely what i himself wants maximized: the Board's goal will coincide with i's, and i consequently has nothing to gain by lying.

More formally,

$$v_i(\hat{x}) - \left[\hat{x} - \sum_{j \neq i} v_j(\hat{x}) \right] \geq v_i(x) - \left[x - \sum_{j \neq i} v_j(x) \right]$$

for all x, since the Board chooses $\hat{x}$ to maximize aggregate net benefit. In particular,

$$v_i(\hat{x}) - \left[\hat{x} - \sum_{j \neq i} v_j(\hat{x}) \right] \geq v_i(\tilde{x}) - \left[\tilde{x} - \sum_{j \neq i} v_j(\tilde{x}) \right]$$

for any $\tilde{x}$ that results from a false $\tilde{v}_i$. Therefore, it is never advantageous for i to misrepresent his utility from the public good. And this holds true no matter what other people ($j \neq i$) might do; whether the v_j's are true or false is irrelevant to the argument above. Consequently, telling the truth is called a *dominant strategy* for i; that is, i will maximize his utility by reporting v_i truthfully, no matter what anyone else is doing.

Since truth is a dominant strategy under the demand-revealing tax scheme, we shall assume, without doing violence to common sense, that people in fact report their true v_i's.

So far, we've shown that the demand-revealing tax scheme induces people to be honest. It solves the misrepresentation problem. Also, since the Board is choosing $\hat{x}$ to maximize

$$\sum_{i=1}^{n} v_i(x) - x,$$

the demand-revealing tax scheme brings about the proper Pareto optimal level of output $\hat{x}$ for the public good. These are its two important advantages.

What are its disadvantages? We must refer back at this point to our \$1,000 bridge example, in which the sum of the tax payments

$$\sum_{i=1}^{5} T_i$$

was \$600. Funds were insufficient to build the bridge. As we have described it so far, then, the system simply might not work: tax collections might not cover the proposed expenditure $\hat{x}$.

The possibility of insufficient funds leads to a question. Is there a way to fix up the taxes so that we can always be sure that total tax receipts will cover expenditure, or

$$\sum_{i=1}^{n} T_i \geq \hat{x}$$

and so that incentives to tell the truth are preserved? The answer is Yes.

It is clear that i's incentive to tell the truth will remain if we add a term to i's tax that does not depend on v_i or $\hat{x}$. For such an addition to i's tax would remain fixed no matter what v_i he reported, and no matter what decision the Public Good Board took; i's report would have no effect on it. We shall proceed to add such a term to T_i.

Let's define

$$S_i = \max_x \sum_{j \neq i} \left[v_j(x) - \frac{x}{n} \right].$$

To understand the intuition here, imagine everyone is first assigned an equal $1/n^{\text{th}}$ share of the cost of x. Then person j's net benefit from x is $v_j(x) - x/n$. The sum of the net benefits of all persons except i is

$$\sum_{j \neq i} \left[v_j(x) - \frac{x}{n} \right],$$

a function of x, and S_i is the maximum value achieved by this function of x. If i were an entirely passive person, a pushover, and if he said to the Public Good Board, "Go ahead and maximize their aggregate net benefit, pay no attention to me, but do let me pay $1/n^{\text{th}}$ share of the cost," then the Board would maximize

$$\sum_{j \neq i} \left[v_j(x) - \frac{x}{n} \right].$$

Therefore, we can interpret S_i as the maximum aggregate net benefit for all the others, when i is passive.

Now we define a new tax for i by adding S_i to i's old tax. For the rest of this section we let

$$T_i = \hat{x} - \sum_{j \neq i} v_j(\hat{x}) + S_i.$$

This can be rewritten as

$$T_i = \frac{\hat{x}}{n} + S_i - \sum_{j \neq i} \left[v_j(\hat{x}) - \frac{\hat{x}}{n} \right].$$

Since S_i does not depend in any way on i's reported v_i, or on $\hat{x}$, the addition of S_i to i's tax does not affect the demand-revealing character of the tax. Telling the truth is still i's dominant strategy. Moreover, since

$$S_i = \max_x \sum_{j \neq i} \left[v_j(x) - \frac{x}{n} \right] \geq \sum_{j \neq i} \left[v_j(\hat{x}) - \frac{\hat{x}}{n} \right],$$

we have

$$T_i \geq \frac{\hat{x}}{n}, \text{ and so } \sum_{i=1}^{n} T_i \geq \hat{x}.$$

Therefore, we have cured the problem of possibly insufficient tax revenues.

Also, there is now a nice intuitive interpretation of T_i. Again, we imagine that everyone is first assigned an equal $1/n^{\text{th}}$ share of the cost of

x. We've noted above that S_i can be viewed as the maximum aggregate net benefit for all save i, when i is passive. What about the term

$$\sum_{j \neq i} \left[v_j(\hat{x}) - \frac{\hat{x}}{n} \right] ?$$

Remember that the Board chooses $\hat{x}$ using the information about demand it receives from everybody, including i. Consequently, this term can be interpreted as the aggregate net benefit for all save i, when i is speaking up for himself. Therefore, we can interpret the new T_i as the sum of (1) i's equal share of the cost $(1/n)\,\hat{x}$ plus (2) the difference between the aggregate net benefit for all others when i is passive, and the aggregate net benefit for all the others when i expresses his demand. This difference itself can be viewed as the loss i imposes on the others by expressing his demand, by not being passive. So, finally, T_i is the sum of (1) i's cost share $(1/n)\,\hat{x}$ plus (2) the loss that i imposes on the others through his demand for the public good.

Let's note at this point that the initial assignment of equal cost shares to all, in the definition of the S_i's, could easily be relaxed. We could start out by assigning positive shares to all, shares which sum to 1. Formally, everything would go through almost as it did before.

At this point we have a tax scheme with the following properties: (1) Honesty is a dominant strategy, so everyone always tells the truth. (2) The optimal quantity $\hat{x}$ of the public good is produced. (3) The tax revenues are always sufficient to pay for $\hat{x}$.

But the analysis is not yet done. As we have defined the T_i's, we know

$$\sum_{i=1}^{n} T_i \geq \hat{x}$$

must hold. But this allows the possibility of a budget surplus for the Public Good Board. In other words, we might have

$$\sum_{i=1}^{n} T_i - \hat{x} > 0.$$

What would become of this surplus?

In fact, the possibility of a surplus introduces a bizarre last twist in the analysis of the demand-revealing scheme. For suppose the Public Good Board were to return the surplus. We have been assuming throughout that individuals know exactly how the Public Good Board operates. So we must suppose that they also know the rule for returning the surplus. But if i knows that he will get back a rebate of, say, $F(i, \hat{x}, v_1, v_2, \ldots, v_n)$,

then we have been incorrectly calculating his tax; his tax ought to have been $T_i - F(i, \hat{x}, v_1, v_2, \ldots, v_n)$, rather than T_i. And if F actually depends on $\hat{x}$ or v_i, then the incentive problem rears its head again. In short, returning the surplus (1) might make the analysis we've done so far wrong, and (2) might destroy the demand-revealing character of the properly calculated tax.

What if the surplus is not returned? What if it is, say, donated to the citizens of another country, or simply destroyed? If the surplus is not returned, we have a paradox. For now we have

$$\hat{x} + \sum_{i=1}^{n} y_i < \sum_{i=1}^{n} \omega_i.$$

That is, some real goods are disappearing from the system. Some real wealth is wasted. It follows that $(\hat{x}, y_1, y_2, \ldots y_n)$ cannot be Pareto optimal, even if $\hat{x}$ satisfies the Samuelson condition, i.e., even if $\hat{x}$ is the Pareto optimal output for the public good. (Some practical authors, particularly Gordon Tullock, argue that it might be quite sensible to donate the surplus to another country, because the surplus will in fact be rather small. So the absence of Pareto optimality is not a big problem. We would be just pennies away from a Pareto optimal result here in the United States, even if the surplus were sent to New Zealand.)

The surplus problem raises a final question. Can a different, more clever definition of S_i be found, one that produces no surplus? That is, we can ask whether it is possible to find S_i functions so that, when we set

$$T_i = \hat{x} - \sum_{j \neq i} v_j(\hat{x}) + S_i,$$

we have the following three virtues:

1. Truth is a dominant strategy for all i,

2. $\hat{x}$ satisfies the Samuelson condition, and

3. $\sum_{i=1}^{n} T_i = \hat{x}$.

Unfortunately, if we want to allow for any reasonable v_i functions, the answer to the question is No. That is, there exists no tax scheme that always satisfies conditions (1), (2), and (3). For a proof of this distressing result, the reader is referred to the Leonid Hurwicz article listed in the References section.

8. The Groves-Ledyard Tax Scheme

The last public finance scheme we shall consider was introduced in the late 1970's by Theodore Groves and John Ledyard. It is similar in several ways to the demand-revealing system. In the Groves-Ledyard scheme, as in the demand-revealing scheme, people send messages to the Public Good Board; the Board decides on a level of output $\hat{x}$ on the basis of those messages; and the Board sets tax payments, or T_i's, to finance $\hat{x}$. In the Groves-Ledyard scheme, as in the demand-revealing scheme, the $\hat{x}$ chosen satisfies the Samuelson condition: it is the optimal level of output for the public good. However, in contrast to the demand-revealing scheme, the Groves-Ledyard scheme produces no problematic budget surplus. That is, it brings about a Pareto optimal allocation. This is its advantage over the demand-revealing scheme. Its limitations will unfold below.

Recall that in the demand-revealing scheme (and the Wicksell-Lindahl scheme as well) person i sends the Public Good Board his v_i function. In the Groves-Ledyard scheme i's message to the Board is a desired increment in output of the public good.

These messages are sent simultaneously by each person. That is, at the time each sends his message, he does not know what messages the others will be sending. Nevertheless, each person can try to forecast what those will be, and best-respond to those forecasts. In general, person i's best response, and hence the message that he wants to send, depends on those forecasts about the other people's messages. And eventually, for a profile of messages to make sense in person i's introspective process of analysis, a necessary requirement is that each message is a best-response to the others. Formally, this means that we replace the game theoretic concept of *dominant strategy equilibrium* with that of *Nash equilibrium*, sometimes referred to as *Cournot-Nash equilibrium*. This equilibrium notion has a long history in oligopoly theory, dating back to Augustin Cournot in 1838, and a necessarily shorter history in game theory, dating back to John F. Nash in 1950. Nash formulated the concept for its general use in any n-person game and proved that equilibria for such games exist.

In the Groves-Ledyard scheme, person i is sending a best-response message contingent on what he believes the others' messages will be. For example, if i believes the others have expressed an aggregate desire for, say 100 jet fighters in their nation's air force (a public good), i might choose to say: "I want ten more," or "I want three less."

The Nash equilibrium concept does not specify how players come to their beliefs about what the other players will do. This process might be purely introspective or might involve a stage of pre-play communications.

It is clear that for this scheme to work, either everyone is going to have to initially express mutually consistent desired increments, or else there is going to have to be some sort of iterative process that leads from an initial list of mutually-inconsistent desired increments to a final mutually-consistent list of desired increments. For instance, person 1 might start things going. He might propose a certain increment of the public good, contingent on everyone else's proposed increment being zero. Then person 2 might propose an increment, contingent on 1's proposal, and on all others being zero. Then 3 might propose an increment, contingent on 1's and 2's proposals. After a while, person n might propose an increment, contingent on 1's, 2's, $\ldots, n-1$'s. Next, person 1 would revise his increment, since he now knows of the proposals of 2 through n. Then person 2 would revise his increment. So would persons $3, \ldots, n$. Then person 1 would have to revise his increment again, as would $2, 3, \ldots, n$, and so on and so forth, possibly ad infinitum. To the best of our knowledge, it is presently an open question whether or not such an iterative process would ever converge. But if it does converge, it will converge to a Nash equilibrium.

Having noted this limitation, we move on. We now ignore these dynamic adjustment processes and simply assume that individuals play according to a Nash equilibrium.

We now let Δ_i represent i's desired increment in the output of the public good. Δ_i, a number, is the message that i sends to the Board. Person i's best message obviously depends on all the others' messages, and it depends on the rules governing the Board's behavior.

Let's turn to those rules. We assume that the Board sets the level of output of the public good according to this equation

$$\hat{x} = \sum_{i=1}^{n} \Delta_i.$$

Note that this is not what the Board does in the demand revealing scheme, but there the individuals' messages are v_i functions, and with v_i functions in hand, it makes sense to maximize aggregate net benefit. Here the messages are only increments, and the obvious thing to do with a collection of increments is to add them together, to get a public good total. We'll see below that this obvious thing to do is also the right thing to do.

We also assume that the Board sets taxes according to the following rule:

$$T_i = \frac{\hat{x}}{n} + \frac{\gamma}{2}\left\{\frac{n-1}{n}(\Delta_i - A_i)^2 - \sum_{j \neq i} \frac{1}{n-2}(\Delta_j - A_i)^2\right\}$$

where

$$A_i = \frac{1}{n-1}\sum_{j\neq i}\Delta_j$$

and γ is some positive number. Let's make a few remarks about T_i. First, the $\hat{x}/n$, which represents an equal share of the cost of $\hat{x}$, could be generalized to an arbitrary fixed tax share. That is, $1/n$ could be replaced with a fixed t_i, as long as

$$\sum_{i=1}^{n} t_i = 1$$

and i's total tax would then be viewed as a fixed share of the expenditure on the public good, $t_i\hat{x}$, plus an adjustment factor. Second, the term A_i is the mean of the $n-1$ other people's desired increments, and it therefore does not depend on Δ_i. Similarly, the term

$$\sum_{j\neq i}\frac{1}{n-2}(\Delta_j - A_i)^2$$

does not depend on Δ_i either.

With the specification of the rules for finding $\hat{x}$ and T_i, we have completely described the Board's behavior. Now let's turn to i's behavior.

We assume, as usual, that i wants to maximize his after tax utility, or, equivalently, $v_i(\hat{x}) - T_i$. He chooses a Δ_i to do this. And we assume that, in finding the solution to his maximization problem, i takes Δ_j as given and fixed for all $j \neq i$. This is the Nash equilibrium assumption.

Person i chooses Δ_i to maximize

$$v_i\left(\Delta_i + \sum_{j\neq i}\Delta_j\right) - T_i$$

or

$$v_i\left(\Delta_i + \sum_{j\neq i}\Delta_j\right) - \frac{1}{n}\left(\Delta_i + \sum_{j\neq i}\Delta_j\right) -$$

$$-\frac{\gamma}{2}\left\{\frac{n-1}{n}(\Delta_i - A_i)^2 - \sum_{j\neq i}\frac{1}{n-2}(\Delta_j - A_i)^2\right\}.$$

Under the Nash assumption, as far as i is concerned, the only variable here is Δ_i. Simple calculus yields the following maximization condition:

$$v_i'\left(\Delta_i + \sum_{j\neq i}\Delta_j\right) - \frac{1}{n} - \frac{\gamma(n-1)}{n}(\Delta_i - A_i) = 0$$

or

$$v_i'\left(\Delta_i + \sum_{j\neq i}\Delta_j\right) = \frac{1}{n} + \frac{\gamma(n-1)}{n}\left(\Delta_i - A_i\right).$$

So i chooses a Δ_i to satisfy the above equation.

Now we turn to the definition of an equilibrium in the Groves-Ledyard scheme. As we indicated above, the whole business can work only if all the Δ_i messages are mutually consistent. That is, it must be the case that Δ_1 maximizes $v_1(\hat{x}) - T_1$, or satisfies 1's maximization equation, given $\Delta_2, \ldots, \Delta_n$. And it must be the case that Δ_2 maximizes $v_2(\hat{x}) - T_2$, or satisfies 2's maximization equation, given $\Delta_1, \Delta_3, \ldots, \Delta_n$. And so on, down through person n. That is, a Nash equilibrium in the *Groves-Ledyard tax scheme*, or a *Groves-Ledyard equilibrium* for short, is a list of increments $(\Delta_1, \Delta_2, \ldots, \Delta_n)$ and a level of output

$$\hat{x} = \sum_{i=1}^{n}\Delta_i$$

such that, for all i, Δ_i maximizes person i's after tax utility, given Δ_1, $\Delta_2, \ldots, \Delta_{i-1}, \Delta_{i+1}, \ldots, \Delta_n$.

What are the properties of such an equilibrium? First, $\hat{x}$ satisfies the Samuelson condition. To see that this is the case, we sum the n maximization equations of individuals 1 through n.

$$\sum_{i=1}^{n} v_i'\left(\Delta_i + \sum_{j\neq i}\Delta_j\right) = \sum_{i=1}^{n}\left[\frac{1}{n} + \frac{\gamma(n-1)}{n}\left(\Delta_i - A_i\right)\right].$$

This gives

$$\sum_{i=1}^{n} v_i'(\hat{x}) = 1 + \frac{\gamma(n-1)}{n}\sum_{i=1}^{n}\left(\Delta_i - A_i\right) = 1 + 0 = 1.$$

Consequently, $\hat{x}$ is the optimal output for the public good.

Second, the Groves-Ledyard tax scheme produces no embarrassing and problematic budget surplus. If $(\hat{x} - \Delta_i)/(n-1)$ is substituted for A_i in the definition of T_i, one or two pages of unpleasant and unsubtle algebra will produce the following result:

$$T_i = \frac{\hat{x}}{n} + \frac{\gamma}{2n(n-2)}[n^2\Delta_i^2 - n\sum_{j=1}^{n}\Delta_j^2 + 2\hat{x}^2 - 2n\hat{x}\Delta_i].$$

If we then sum all the T_i's we get:

$$\sum_{i=1}^{n} T_i = \sum_{i=1}^{n}\frac{\hat{x}}{n} + \frac{\gamma}{2n(n-2)}[0] = \hat{x}.$$

The Groves-Ledyard scheme collects just enough taxes to finance the public expenditure, no more and no less. This is its principle advantage over the demand revealing scheme.

Because the Groves-Ledyard scheme is based on the simultaneous reporting of desired increments, its performance is vulnerable to sequentiality. That is, if moves are sequential instead of simultaneous, individuals will typically have incentives to change their messages, because they will understand that their moves are observable by the people that move after them, violating the Nash assumption. Moreover, as is well known, Nash equilibria are not necessarily coalitionally stable, and so coalitions of individuals could also upset the outcome of the tax scheme.

To sum up, the Groves-Ledyard scheme will produce an optimal $\hat{x}$ and no surplus, providing that individuals best-respond to each other, play according to a Nash equilibrium. But, like the Wicksell-Lindahl scheme, the demand-revealing scheme, and the fixed tax shares majority voting scheme, this scheme has its limitations.

9. Exercises

1 In a simple public goods model we let $x=$ the quantity of the public good, $y_i = i$'s consumption of the private good, $u_i(x, y_i) = v_i(x) + y_i$ is person i's utility function, and we assume one unit of public good can be produced with one unit of private good.

(a) Define aggregate net benefit from the public good.

(b) Derive the Samuelson optimality condition, by maximizing aggregate net benefit.

(c) Show that if only individual 1 buys the public good, and pays for it by himself, the output of the public good will generally not be optimal.

2 There are three people and two goods, one public and one private. Let $x=$ the quantity of public good, and y_i denote i's quantity of private good. Assume , as usual, that the private good can be transformed into public good unit for unit. Suppose
$u_1 = a \ln x + y_1$
$u_2 = b \ln x + y_2$
$u_3 = c \ln x + y_3$
(Note: The derivative of $\ln x$ is $\frac{1}{x}$.)

(a) Using the Samuelson optimality condition, solve for the optimal output of the public good.

(b) Find Lindahl-Wicksell tax shares.

(c) Explain the virtues and shortcomings of the Lindahl-Wicksell scheme.

3 Consider a public goods model where $u_1(x, y_1) = a \ln x + y_1$
$u_2(x, y_2) = b \ln x + y_2$.
Assume $a, b > 0$. Assume also, as usual, that one unit of private good is converted into one unit of public good.

(a) Solve for the optimal level of public good $\hat{x}$.

(b) Find Lindahl-Wicksell tax shares t_1 and t_2.

(c) Using your tax share t_1 as calculated above, solve for 1's after tax utility from the public good, $v_1(\hat{x}) - t_1\hat{x}$.

Depending on the exact values of a and b , person 1 might do better by claiming his valuation function is zero, i.e, $v_1(x) = 0$, rather than $v_1(x) = a \ln x$. What condition on a and b would ensure that 1 would do better by claiming $v_1(x) = 0$?

(d) Solve for demand revealing taxes T_1 and T_2. (You need not include the S_i terms.)

4 Three individuals have the following utility functions for the public good (x) and private consumption (y_i):
$u_1(x, y_1) = ax^{\frac{1}{2}} + y_1$
$u_2(x, y_2) = bx^{\frac{1}{2}} + y_2$
$u_3(x, y_3) = cx^{\frac{1}{2}} + y_3$.

(a) Suppose their fixed tax shares are $t_1 = t_2 = t_3 = \frac{1}{3}$. What is the majority voting equilibrium output for the public good ? (Assume $0 < a < b < c$.)

(b) What is the optimal output of the public good ?

(c) Under what condition will the answers to (a) and (b) be the same?

10. Selected References

(Items marked with an asterisk (*) are mathematically difficult.)

1. E.H. Clarke, "Multipart Pricing of Public Goods," *Public Choice*, V. 11, 1971, pp. 17-33.

This paper is one of the seminal papers on the demand-revealing tax scheme. Clarke develops a "two-part tariff" to finance the provision of a public good; it is made up of a "fixed charge" (or an assigned cost share times the cost of the public good) plus a "variable charge" that

depends on the difference between the individual's desired output and the actual output of the good. The scheme has three virtues: each individual is induced to reveal his correct demand schedule, an optimal output of the public good is chosen, and tax revenues equal or exceed the cost of the public good.

The analysis is mostly graphical and verbal, with little mathematics.

*2. J. Green and J-J. Laffont, "Characterization of Satisfactory Mechanics for the Revelation of Preferences for Public Goods," *Econometrica*, V. 45, 1977, pp. 427-438.

Green and Laffont extend the analysis of Groves and Loeb, that is, the analysis of demand-revealing mechanisms. Those mechanisms have the virtues that (1) truth is a dominant strategy for every individual and (2) that Pareto optimal levels of output of the public goods are always chosen. Green and Laffont show that any mechanism with these two properties must be a generalized Groves and Loeb mechanism.

3. T. Groves, "Information, Incentives, and the Internalization of Production Externalities," reprinted in S.A.Y. Lin, ed., *Theory and Measurement of Economic Externalities*, Academic Press, New York, 1976, pp. 65-86.

This paper has a model of production externalities which are remedied by a demand-revealing tax scheme. Each of n firms creates external effects for the others, and each firm sends a message to a coordinating center that indicates how its profits depend on all the externalities. The center wants to coordinate the activities of the firms to maximize total profits (the analog of the Samuelson condition for a public good), and it taxes and transfers money among the firms to elicit truthful messages.

*4. T. Groves and J. Ledyard, "Optimal Allocation of Public Goods: A Solution to the 'Free Rider' Problem," *Econometrica*, V. 45, 1977, pp. 783-809.

This is a rigorous and difficult paper that uses a general equilibrium model, with many private goods, many public goods, private firms, private markets, and a government. Section 3 of the paper analyzes the demand-revealing process in the formal general equilibrium model, and a version of the First Fundamental Theorem of Welfare Economics. Section 4 of the paper develops the model of what Groves and Ledyard call an "Optimal, Unbiased Government," or what is now called the Groves-Ledyard system. In it each individual's message to the government is interpreted as the increment of each public

good the individual would like added to (or subtracted from) the total. Groves and Ledyard rigorously prove the two fundamental theorems for the optimal, unbiased government. The first establishes that an equilibrium in the Groves-Ledyard scheme is Pareto optimal (and, in particular, optimal outputs of public goods are achieved and there is no government budget surplus).

In Section 5, the paper provides a survey of some of the important theoretical literature on the free rider problem.

5. T. Groves and J. Ledyard, "Some Limitations of Demand Revealing Processes," *Public Choice*, V. XXIX-2, 1977, pp. 107-124.

   This clear and well-written article lists five drawbacks of demand-revealing tax schemes: (1) the problem of the budget surplus, (2) the possibility that an individual's tax might exceed his initial endowment and therefore bankrupt him, (3) the problem of dynamic instability of the scheme if utility functions are not separable, (4) the problem of strategic, nontruthful, behavior if utility functions are not separable, and (5) the problem of strategic behavior by coalitions. At the very end of the paper Groves and Ledyard sketch their own mechanism.

6. T. Groves and M. Loeb, "Incentives and Public Inputs," *Journal of Public Economics*, V. 4, 1975, pp. 211-226.

   This clearly written article analyzes a model of a group of firms that use a public good input. Groves and Loeb provide a demand-revealing scheme in which a Center coordinates the activities of the firms. The scheme (1) always induces firms to reveal their true demands for the input, and (2) leads to a joint profit maximum for the firms.

*7. L. Hurwicz, "On the Existence of Allocation Systems Whose Manipulative Nash Equilibria are Pareto-Optimal," unpublished paper, 1975.

   This paper provides the theorem (Part A, Theorem 4) that says there is no mechanism for which (1) truth is always a dominant strategy for everyone and (2) the outcome is always Pareto optimal (and in particular there is no "budget surplus"). (See Reference 4 above, p. 795.)

8. E. Lindahl, "Just Taxation – A Positive Solution," ("Die Gerechtigkeit der Besteuerung," Lund, 1919), translated and reprinted in R.A. Musgrave and A.J. Peacock, eds., *Classics in the Theory of Public Finance*, MacMillan and Co., New York, 1958.

This extremely important article is about the distribution of the total cost of the collective goods between two categories of taxpayers, and about the appropriate extent of collective activity. The tax scheme Lindahl develops is one in which each type of taxpayer is assigned a certain fraction of the total cost, and given their assigned fractions, the two types of taxpayers agree on the appropriate extent of collective activity. It follows that the tax shares are linked to the taxpayers' valuations of the public services.

9. A. Mas-Colell, M. Whinston and J. Green *Microeconomic Theory*, Oxford University Press, 1995.

   Several sections of Chapter 11 of this graduate textbook deal with public goods. In particular, the interpretation of Lindahl equilibrium in terms of "personalized prices" is discussed.

10. R.A. Musgrave and A.T. Peacock, *Classics in the Theory of Public Finance*, MacMillan and Co., New York, 1958.

    In addition to providing translations of Lindahl and Wicksell, this volume has interesting classic essays in taxation and public goods by E. Barone, F.Y. Edgeworth, E. Sax, A. Wagner and other European authors. The introductory chapter by Musgrave and Peacock connects the principal contributions to the theory of public finance, starting with those of Adam Smith and ending with the ones made in the 1920's.

11. P.A. Samuelson, "The Pure Theory of Public Expenditure," *Review of Economics and Statistics*, V. 36, 1954, pp. 387-389.

    This seminal paper provides a formal optimality condition for the provision of public goods — now known as the Samuelson condition. Samuelson also raises the free rider problem, which he views as insurmountable in a decentralized economic system: "*No decentralized pricing system can serve to determine optimally these levels of collective consumption.* Other kinds of "voting" or "signalling" would have to be tried. But, and this is the point sensed by Wicksell but perhaps not fully appreciated by Lindahl, now it is in the selfish interest of each person to give *false* signals, to pretend to have less interest in a given collective consumption activity than he really has, etc." (His italics.)

12. T. Nicolaus Tideman, ed., *Public Choice*, V. XXIX-2, Special Supplement to Spring, 1977.

    This entire special issue of *Public Choice* is devoted to the demand-revealing and Groves-Ledyard mechanisms. Tideman's introduction

is good and provides a useful bibliography. The issue includes articles by Clark, Green and Laffont, Groves and Ledyard, Loeb, Tideman, Tullock, and others.

13. W. Vickrey, "Counterspeculation, Auctions, and Competitive Sealed Tenders," *Journal of Finance*, V. 16, 1961, pp. 8-37.

    Vickrey's early paper develops a mechanism to be used by a government agency that deals with monopolistic sellers and buyers. The mechanism is designed to elicit truthful information from the monopolists, and to bring about an optimal consumption and production equilibrium. This article gave rise to an important literature in auctions in the 1990's.

14. K. Wicksell, "A New Principle of Just Taxation," (Ein Neues Prinzip der Gerechten Besteuerung," Jena 1896), translated and reprinted in R.A. Musgrave and A.J. Peacock, eds., *Classics in the Theory of Public Finance*, MacMillan and Co., New York, 1958.

    Wicksell's seminal essay argues for the principle of taxation according to benefit, to replace the principle of taxation according to ability to pay. There are two reasons to prefer taxation according to benefit: First, it allows the taxing authorities to find an appropriate extent of public expenditures. If people are taxed according to their ability to pay, the tax authorities get no information about the usefulness of public expenditures from those who are presumably "consuming" the public expenditures, namely, the taxpayers. If people are taxed according to benefit, then the tax authorities must gather some information about the usefulness of public expenditures, they must connect benefits and taxes, and this connection should lead to an appropriate level of expenditures.

    Second, it is just to tax according to benefit. In fact, the greatest injustice of taxation is forcing someone to pay a tax to finance an activity which provides him no benefit, which he might actually disapprove.

    Wicksell is led by this principle to the espousal of a system in which public activities and the particular distribution of taxes used to finance them, must be approved by unanimous consent.

Chapter 9

COMPENSATION CRITERIA

1. Introduction

We now put aside the questions of the last two chapters, about reconciling external effects and public goods with the market system, and return to the basics, and the two fundamental theorems of welfare economics. The First Fundamental Theorem says, roughly, that the market mechanism leads to a Pareto optimum, barring externalities or public goods complications. The Second Fundamental Theorem says, roughly, that any Pareto optimum can be reached via an appropriately modified market mechanism, e.g., in the case of an exchange economy, the right lump sum cash transfers will allow us to get to any optimum we want.

But this second theorem raises a profound question: How is the particular Pareto optimum that the theorem says we can reach to be chosen? That is, how should society choose between two Pareto optimal arrangements? Or, more generally, how should society choose between any two economic arrangements, or, in the case of an exchange economy, any two allocations?

This raises the issue touched on in Chapter 1: Are there social preferences distinct from individual preferences? We are quite comfortable with the idea that any single person has tastes, or a utility function, or a preference relation R_i, any of which can indicate when x is better than y, for that person. We are comfortable with the idea that person i's preference relation R_i is complete, which means that he can always decide when x is better than y or y is better than x or they are equally good. We are comfortable with the idea that R_i is transitive. If person i likes x at least as well as y and y at least as well as z, then he likes x at least as well as z. Can the idea of a preference relation, or better, a

complete preference relation, or best, a complete and transitive preference relation, be transplanted from the analysis of individual choice to the analysis of social choice? Does a group of people have a reasonable preference relation? Or is it nonsensical to attribute the characteristics of an individual to a collection of individuals?

Let's return to the Second Fundamental Theorem. It says that if the Pareto optimal allocation y is desired by society, then there are cash transfers that will, with the market mechanism, get society to y. But why y rather than x? Does society have a preference for y over x? What does this mean? The Second Fundamental Theorem of Welfare Economics raises these questions but provides no answers. That is, the market mechanism provides no guide as to whether one Pareto optimal arrangement x is better or worse than another Pareto optimal arrangement y. In fact, it often provides no guide for the choice between x and y even if one or both is not Pareto optimal.

There are then economic questions, questions about the distribution of goods and services, that the competitive market mechanism does not answer. This chapter and the ones that follow are about attempts to answer these questions. In particular, this chapter is about economists' standard answers to the question of when one arrangement (or allocation) is better for society than another.

The question is obviously one with lots of practical ramifications. Government officials are often faced with the choice between x and y. Occasionally they consult with economists about the choice. For example, the choice might be to reduce tariffs on imported goods, or to leave them the way they are. Or the choice might be to impose a special excise tax on crude oil or not to impose it, or to subsidize production of certain agricultural commodities or not to subsidize them. Or the choice might be to allow more timber harvesting on government land, or not to allow it, or to build a highway, or not to build it, and so on, almost ad infinitum. Can economists offer some suggestions? Naturally lots of these questions have a public goods character, and partial answers can be found in the previous chapter, in particular answers about the optimal extent of public expenditure and appropriate tax mechanisms. But often the questions are of the type "Is x better for society than y?" It is this particular type of question on which we focus.

2. Notational Preliminaries

In this chapter we shall discuss social alternatives, labeled x, y, z, etc., which might be allocations in an exchange economy, or which might be more general arrangements, like combined allocation-production plans.

Or they might be simple mutually exclusive possibilities, like "build a road" or "don't build the road."

We shall lean heavily on the symbols of Chapter 1. R_i is person i's preference relation, so xR_iy means i likes x as well as i. P_i is the associated strict preference relation, and I_i is the associated indifference relation. Also, u_i is person i's utility function. We allow for external effects; so u_i is a function of x, rather than, say, x_i.

3. The Pareto Criterion

Our definition of Pareto optimality for an exchange economy, and our definition of optimality in the provision of public goods, both reduce to this: if there is an alternative y that is feasible, that everyone likes as well as x and someone likes better, then x isn't optimal.

This suggests one obvious answer to the question, "When is x better for society than y?" Here is the formal definition of the Pareto criterion:

> Let x and y be two alternatives. Then x is *Pareto superior* to y, which we write $x\mathcal{P}y$ if
>
> xR_iy for all individuals, and
> xP_iy for at least one individual i.

So x is Pareto superior to y if all like x as well as y, and some people actually prefer x to y. We can easily illustrate the Pareto criterion in an Edgeworth box diagram.

In Figure 9.1, both people prefer allocation x to allocation y. Consequently, x is Pareto superior to y, or $x\mathcal{P}y$. There is no ambiguity about x's being better for society than y. But it is not the case that $x\mathcal{P}z$, and it is not the case that $z\mathcal{P}x$. Should we then say that x and z are indifferent according to the Pareto criterion? Well, we don't. We just say that they are noncomparable.

What about y and z? Are they Pareto comparable? Is it the case that z is Pareto superior to y, or that y is Pareto superior to z? The answer to both questions is No. Person 1 prefers y to z, but person 2 prefers z to y. Whenever the interests of two people are opposed like this, neither $y\mathcal{P}z$ nor $z\mathcal{P}y$ can be true. (This was also true of x and z.)

At this point, it is useful to be a little bit more precise about the parallel between the Pareto criterion $\mathcal{P}$ and an individual's preference relation R_i. The reader should first observe that $\mathcal{P}$ is analogous to a strict preference relation, a P_i for an individual. When we say $x\mathcal{P}y$ we mean x is Pareto better than y, and we don't mean that x and y might be equally good. When we need a Pareto relation analogous to a weak preference relation R_i, we define it this way:

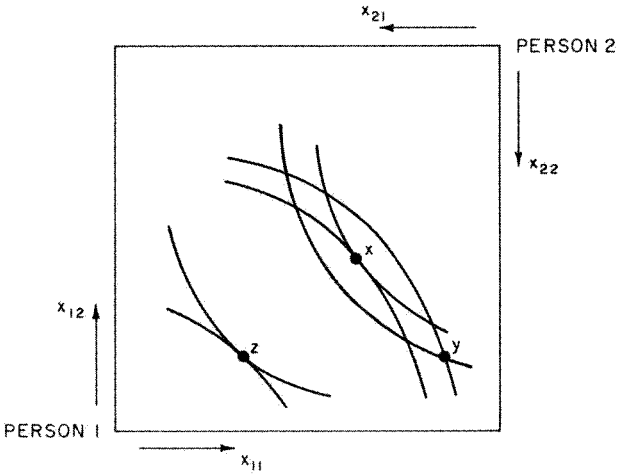


Figure 9.1.

Let x and y be two alternatives. Then x is *Pareto as-good-as* y if xR_iy for all i. From the Pareto as-good-as relation we can retrieve the Pareto superior relation $\mathcal{P}$: x is Pareto superior to y, or $x\mathcal{P}y$, if x is Pareto as-good-as y and y is not Pareto as-good-as x. For if x is Pareto as-good-as y, then xR_iy for all i. And if y is not Pareto as-good-as x, then for at least one i, not yR_ix, and since we assume all individuals' preference relations are complete, not yR_ix for at least one i implies xP_iy for at least one i.

The Pareto as-good-as relation is analogous to an individual's R_i, while the Pareto superior relation is analogous to an individual's P_i. We saw in Figure 9.1 that x and z were Pareto noncomparable in the sense that neither $x\mathcal{P}z$ nor $z\mathcal{P}x$ was true. This noncomparability is still with us when we use the weaker Pareto as-good-as relation. It is not the case that x is Pareto as-good-as z, nor is the case that z is Pareto as-good-as x. In short, the possibility of Pareto noncomparability means the Pareto as-good-as relation is not complete. And this makes it distinctly different from an individual's R_i relation.

We saw in Chapter 1 that completeness for a preference relation is crucial for the process of choice. Completeness is needed for the existence of best alternatives, for the nonemptiness of choice sets. This is true for social preference relations as well as for individual preference relations. And the lack of completeness, the pervasive problem of Pareto noncomparability, is the fundamental drawback of the Pareto criterion.

In Figure 9.1, most of the interesting allocations are Pareto noncomparable. In practical situations, it is rarely the case that the social choice is between an alternative x and an alternative y, such that everyone likes x as well as y. The Pareto criterion can rarely be used. So economists have developed other criteria to decide when one alternative is socially preferable to another.

4. The Kaldor Criterion

To illustrate the next criterion we first construct a model that is a little more abstract than our exchange, production, and public good models. At the end of this chapter we will return to the less abstract exchange economy model.

The usual models of an economy allow for the possibility of movement from any social state (e.g. allocation) to any other social state (e.g. allocation). Often, however, the policy maker is faced with a choice between two alternatives which are mutually exclusive—if he chooses the left fork of the road he will never be permitted the possibilities down the right fork. Some states are simply not accessible from other states.

For example, suppose the choice is to build a major road from point A to B, or not to build it. If it is built, and financed in a particular way, we have one alternative, say x. But there are many ways to finance the road, and there are many ways to transfer goods among individuals, given that the road is built. That is, given the road, or given x, there are lots of alternatives that are accessible from x, lots of distribution of goods among people contingent on the road's being built. We let $S(x)$ be the set of alternatives accessible from x. On the other hand, if the road is not built, and a particular distribution of goods holds, we have another alternative, say y. But there are many ways that the goods under y might be redistributed among individuals, or otherwise moved around; that is, given no road, there are lots of alternatives accessible from y. We let $S(y)$ be the set of alternatives accessible from y.

(As a matter of common sense it is surely the case that building the road, or x itself, is accessible from y, although y, the no-road situation, might not be accessible from x since it costs real resources to tear up a previously built road. We certainly allow that x might be in $S(y)$. If the choice is to have a tariff or not to have a tariff, then it might well be the case that x is in $S(y)$ and y is in $S(x)$.)

Now we turn to the social improvement criterion devised by Nicholas Kaldor and John Hicks. Formally , we say that x is *Kaldor superior* to y, or $x\mathcal{K}y$, if there exists a z in $S(x)$ such that $z\mathcal{P}y$. That is, there must be a state accessible from x, which is Pareto superior to y. The existence

of such a state allows for the theoretical possibility of everyone's being made better off after x is chosen, although there is no assurance that this would in fact happen.

Suppose in our road/no-road example that $x\mathcal{K}y$. This means that once the road is built (path x is chosen) there exists some tax/subsidy or compensation scheme, based on x, which would produce a z that is Pareto superior to the no-road situation y. Once the road is built, in other words, the people who gained from the construction of the road could potentially transfer goods to the people who lost from the construction of the road, so that nobody ends up worse off than before the road was built.

To illustrate the Kaldor criterion graphically, it is convenient to construct what are called utility frontiers. Given an alternative z in $S(x)$, there is a utility level $u_i(z)$ for each person i, and therefore, a utility vector that depends on z, $u(z) = (u_1(z), u_2(z), \ldots, u_n(z))$. So the set of alternatives $S(x)$ generates a set of utility vectors which we shall call $U(x)$. If there are just two people, the set of utility vectors $U(x)$ can be represented by an area in a graph that shows person 1's utility u_1 on the horizontal axis and person 2's utility u_2 on the vertical axis. The northeast frontier of that area is called the *utility frontier for* x.

In Figure 9.2, $U(x)$ is the shaded area, and the utility frontier for x is labelled. The particular vector of utilities $u(x) = (u_1(x), u_2(x))$ that corresponds to alternative x (build the road) lies on the frontier, as does another particular vector of utilities $u(w) = (u_1(w), u_2(w))$. The alternative w is in $S(x)$, that is, it's an alternative that comes about when the road is built and a certain tax/subsidy scheme, or compensation scheme, is instituted. Clearly, person 2 likes alternative w better than alternative x, so the move from x to w involves person 1 compensating person 2 in some fashion. The utility frontier for y (don't build the road) is also labelled in the figure, and the vector of utilities $u(y) = (u_1(y), u_2(y))$ lies on the frontier for y, as does the vector of utilities for another alternative v, accessible from y.

In terms of Figure 9.2, what is Kaldor superior to what? Given alternative x (build the road), it is possible to get to alternative w. But both people prefer alternative w to alternative y; that is, $u_1(w) > u_1(y)$ and $u_2(w) > u_2(y)$. So w is Pareto superior to y. Consequently, x is Kaldor superior to y.

On the other hand, given alternative y (no road), it is possible to get to alternative v. But both people prefer alternative v to alternative x; that is, $u_1(v) > u_1(x)$ and $u_2(v) > u_2(x)$. So v is Pareto superior to x. Consequently, y is Kaldor superior to x.

In short, $x\mathcal{K}y$ but also $y\mathcal{K}x$. That is, the criterion is inconsistent.

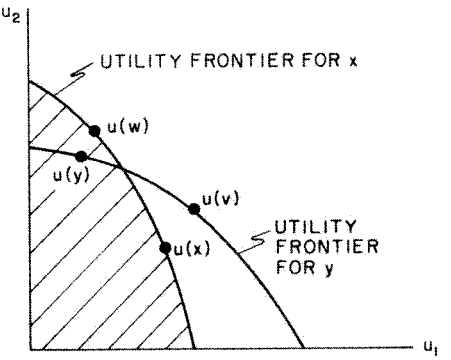


Figure 9.2.

As it stands, Kaldor's method is an imperfect guide to social policy. In this case, at least, it says: "Build the road — but also, don't build it." This possibility of Kaldor inconsistency leads us to the next compensation criterion.

5. The Scitovsky Criterion

Tibor Scitovsky devised a social improvement criterion which necessarily avoids Kaldor's inconsistency. It is formally defined as follows: x is *Scitovsky superior* to y, written $x\mathcal{S}y$, if $x\mathcal{K}y$ but not $y\mathcal{K}x$.

Since it must avoid nonsensical results like "build it and don't build it," Scitovsky's criterion improves upon Kaldor's. But it also has a shortcoming, to which we now turn.

In order to highlight this shortcoming it is useful to first provide a new and slightly different definition of Pareto optimality, a definition that conforms with the model of this chapter. We have been attaching to each alternative x a set of states $S(x)$ accessible from x. Once we are at x, the only feasible options are the alternatives in $S(x)$. Surely this possibly restrictive feasibility condition should be recognized in the definition of optimality. Consequently, we make the following definition:

> A social state x is *Pareto optimal* if there is no y in $S(x)$, that is no y accessible from x, such that $y\mathcal{P}x$.

Now let's consider what happens to the Scitovsky relation when, as in an exchange economy model, all states are accessible from each other. (When we say that all states are accessible from each other, or mutually accessible, we mean that if z is in $S(x)$ then z is also in $S(y)$, for all x, y and z.)

Proposition 1. Suppose all states are mutually accessible. Then $x\mathcal{S}y$ if and only if x is Pareto optimal and y is not.

Proof. Let $x\mathcal{S}y$. Then $x\mathcal{K}y$ and not $y\mathcal{K}x$. Since $x\mathcal{K}y$, there exists a z accessible from x and therefore from y such that $z\mathcal{P}y$. Therefore, y is not Pareto optimal. If x were not Pareto optimal, there would be a w, accessible from x and hence from y, such that $w\mathcal{P}x$, which would contradict not $y\mathcal{K}x$. Therefore, x is Pareto optimal.

Now suppose x is Pareto optimal and y is not. Since x is Pareto optimal, there is no z accessible from x or y with $z\mathcal{P}x$. Therefore, not $y\mathcal{K}x$. Since y is not Pareto optimal, there is a w, accessible from y and hence from x, such that $w\mathcal{P}y$. Therefore, $x\mathcal{K}y$. Q.E.D.

The following result is also obvious.

Proposition 2. Suppose all states are mutually accessible. Then $x\mathcal{K}y$ if and only if y is not Pareto optimal.

Evidently, at least in the case where all states can be reached from each other, the Scitovsky criterion does not yield any information beyond what is inherent in the notion of Pareto optimality. This is its shortcoming.

6. The Samuelson Criterion

We now turn to a compensation criterion developed by Paul Samuelson. The intuition of Samuelson's criterion is very clear, even if the formal definition isn't. We'll write down the formal definition first and then turn quickly to a diagram to see the intuition.

Formally, x is *Samuelson superior* to y, which we shall write $x\mathcal{N}y$, if the following is true: For any z in $S(y)$, $x\mathcal{K}z$.

To put it another way, x is Samuelson superior to y if, for any z accessible from y, there is a w accessible from x that is Pareto superior to z.

In terms of a utility frontier diagram for two individuals, for x to be Samuelson superior to y the utility frontier for x must lie to the northeast of the utility frontier for y. Figure 9.3 illustrates this criterion.

In the figure, the utility frontier for x lies outside the utility frontier for y. Consequently, for any alternative z accessible from y, there is a w accessible from x that provides at least one person with a higher utility level than z, and neither with a lower level of utility. That is, the utility vector $u(w)$ lies to the northeast of the utility vector $u(z)$. No matter what you do, starting from y, you could do better, if you started from x.

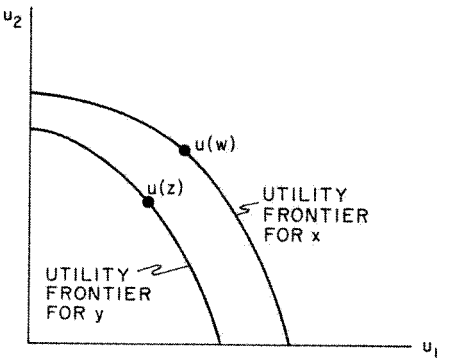


Figure 9.3.

The diagram should make it obvious that the Samuelson criterion cannot be inconsistent: it is impossible to have $x\mathcal{N}y$ and $y\mathcal{N}x$.

However, like the Kaldor and Scitovsky criteria, the Samuelson criterion has a shortcoming when all states are mutually accessible. If all alternatives are mutually accessible, there is really only one set of feasible alternatives — the set of all alternatives — and there is only one set of possible utility vectors. Consequently, the utility frontier is unique. (It is possible to construct odd mathematical examples where there are no utility frontiers, where $U(x)$ does not have a boundary or include its boundary; however, we won't worry here about those peculiar cases.) Now if there is only one utility frontier, a brief examination of Figure 9.3 should convince the reader that $x\mathcal{N}y$ can never occur.

More formally, suppose the set of possible utility vectors is unique and includes its frontier. Let y be any alternative. Find a z (necessarily accessible from y) on the utility frontier — a z such that there are no feasible utility vectors to the northeast of $u(z)$. Then $x\mathcal{N}y$ is clearly impossible, for any x, since $x\mathcal{K}y$ is impossible.

In short, if all alternatives are mutually accessible, the Samuelson criterion is entirely devoid of content. No state is Samuelson superior to any other.

Before ending this discussion we ought to briefly comment on the idea that all states are mutually accessible.

In one sense, this idea is obviously wrong, Once a road has been built, real resources have gone into it, and there is no way to recapture what might have been had the road not been built. So in this sense, our criticisms of the compensation criteria of Scitovsky and Samuelson are unfair.

However, in another sense, the idea is quite plausible. For these criteria are meant to be used by decision makers before decisions are made; before the roads, for example, are built. Before the crucial decisions are made, paths are not closed, and everything is possible. Even if the policymaker is contemplating a decision with irrevocable effects, everything remains accessible as long as he is just contemplating. So in this sense, our criticisms of the Scitovsky and Samuelson compensation criteria should be taken seriously.

7. Compensation Criteria in Exchange Economies

Let's revisit our compensation criteria in a less abstract way than what we have done above. In our definition of the Kaldor criterion, for example, we defined x to be Kaldor superior to y if there exists a z in $S(x)$, the set of alternatives accessible from x, such that z is Pareto superior to y.

The $S(x)$ notion is slightly vague, and to be more concrete we will now focus on exchange economies. Recall that in an exchange economy there is an initial allocation of goods $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$, where ω_i represents person i's initial bundle of goods. An allocation in an exchange economy is an $x = (x_1, x_2, \ldots, x_n)$, where x_i represents person i's bundle of goods, such that $x_{ij} \geq 0$ for all i and j, and such that the goods add up:

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} \omega_i.$$

We will now assume that the alternatives being considered are simply alternative allocations in one exchange economy and that $S(x)$ is, for any x, simply the set of all the allocations. That is, in the terminology of the previous sections, all states are mutually accessible.

We know from Propositions 1 and 2 above that, when all states are mutually accessible, $x\mathcal{K}y$ if and only if y is not Pareto optimal; that $x\mathcal{S}y$ if and only if x is Pareto optimal and y is not; and that the Samuelson criterion is vacuous.

We will now also assume that all individuals have utility functions that are self-interested and monotonic. That is, we make the standard assumptions we made for the exchange economy version of the first fundamental theorem of welfare economics.

At this point we can see Kaldor inconsistency in a simple Edgeworth box diagram. In Figure 9.4, consider a move from y to x. We want to know whether it is a Kaldor move, that is, is x Kaldor superior to y? In the abstract, the answer is obviously yes. This is because there exist allocations based on x (all allocations in the box are allocations based

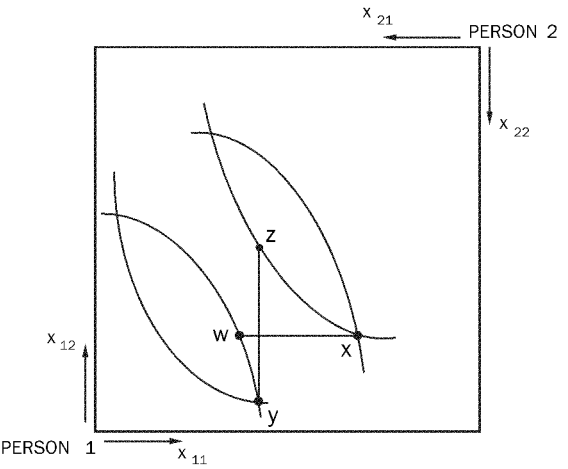


Figure 9.4.

on x) that are Pareto superior to y. Next consider a move from x to y. We want to know whether it is a Kaldor move, that is, is y Kaldor superior to x? By an identical argument, the answer is yes. Ergo we have Kaldor inconsistency in an exchange economy.

But this abstract notion, that all points in the Edgeworth box are mutually accessible, differs somewhat in spirit from Kaldor's intent. Kaldor was analyzing an important policy change in England (the 1846 repeal of the corn laws), which hurt a certain group of people (farmers and landlords) but helped another group (consumers of bread).

In Kaldor's view, when society moves from y to x, there are people who gain from the change, and people who suffer from the change. (We will call the former "winners" and the latter "losers.") In Kaldor's opinion, the move from y to x is beneficial to society if the winners could, in theory, compensate the losers for their losses, and remain better off than they were originally.

The argument in Figure 9.4 that is closer in spirit to Kaldor's intent is the following: In the move from y to x, person 1 is the winner and person 2 is the loser. In theory, the winner, person 1, could give up some of (say) good 1 to person 2, putting the economy at w. Now person 2 has been fully compensated for his loss, and person 1 remains better off than he was at y. Therefore y to x is a Kaldor move. But by the same token, consider a move from x to y. Now person 2 is the winner and person 1 is the looser. The winner, person 2, could in theory give some of (say) good 2 to person 1, putting the economy at z. Now person 1 has been fully compensated for his loss, and person 2 remains better off

than he was at x. This shows Kaldor inconsistency in a way that's closer to Kaldor's original story.

But this description is still not *exactly* like Kaldor's story, because Kaldor, like other economists, thought of the hypothetical compensation paid by winners to losers in terms of *money*, rather than in terms of commodity transfers. This idea, that compensation should be in dollar terms, brings us to the next section.

8. Consumer's Surplus – Money Metrics

Since the time of Alfred Marshall (*Principles of Economics*, 1890), if not before, economists have attempted to measure gains and losses to individuals whose consumption bundles are shifting in money terms. The advantage is obvious – a dollar is a dollar, whereas a utility unit for you is not a utility unit for me. If changes are measured in money units, we can add the changes up over all members of society, in a sensible fashion. With utility units we are reluctant to do so.

A consumer's gain from a shift, in dollars, is traditionally measured in the following way: Suppose person i starts at x_i and ends up at y_i. Suppose $u_i(y_i) > u_i(x_i)$. When he is consuming y_i, he is choosing that bundle to maximize his utility subject to some budget constraint. To measure his gain in dollars we might ask this question: How much money could we subtract from the right hand side of his budget constraint, and still have him end up as well of as he was at the original bundle x_i?

(We are now thinking about a bank balance transfer, similar to the T_i's of Chapter 3.) In short, how much money could we take away, and leave him with his original utility level?

The answer to this question gives us a measure of his gain from the x_i to y_i transition, in terms of money. This is called the *compensating variation* measure, and was formalized by Sir John Hicks (1942). The idea is illustrated in Figure 9.5 below. In the figure we assume for simplicity that good 2 is measured in dollar units, that is, $p_2 = 1$.

In Figure 9.5, person i is starting at x_i. He is choosing x_i because his initial budget line is b_x, and given that budget line, he maximizes utility at x_i.

Then a change occurs. As a result of that change, person i has a new budget line given by b_y in the figure. (This shift in budget lines might be a consequence, for example, of i's having his initial bundle on the vertical axis, and of the price of good 1 dropping by half.) Given the shift, person i chooses a new utility-maximizing point y_i.

To measure i's gain in money terms, we ask this question: how much money could we take away from i, *given the new point* y_i *and the new relative prices associated with it*, so as to leave him exactly as well off

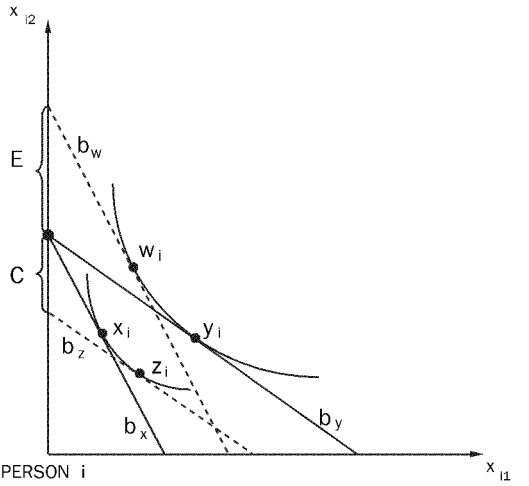


Figure 9.5.

as he was originally? Taking away money given the new prices means shifting b_y down in a parallel fashion; doing so is a way that leaves him as well off as he originally was at x_i, means shifting down to the (hypothetical) budget line b_z. Since p_2 is 1, the dollar budget shift is exactly the same as the shift on the vertical axis, shown as C. In short the compensating variation measure of i's gain is exactly C in Figure 9.5.

Note that this figure also shows the standard microeconomics textbook notions of substitution and income effect: For person i, the shift in consumption bundles from x_i to y_i can be viewed as a composite of the substitution effect (x_i to z_i, with utility constant but relative prices changing) and the income effect (z_i to y_i, with relative prices constant and income changing).

Hicks realized there is a similar and equally logical way to measure person i's gain. We might ask this question: Person i starts at x_i and ends up at y_i. Given the old point x_i and the relative prices associated with it, how much money would we have to give i to make him exactly as well off as he is at the new point y_i? The answer to the question is the *equivalent variation* measure. Giving him money but holding the old prices constant means shifting b_x up in a parallel fashion; doing so in a way that makes him as well off as he is at y_i, means shifting up to the (hypothetical) budget line b_w. Since p_2 is 1, the dollar budget shift is exactly the same as the shift on the vertical axis, shown as E. In short, the equivalent variation measure of i's gain is exactly E in Figure 9.5.

Note that C is *not* equal to E.

Mulling over these definitions leads to the following: For both compensating variation and equivalent variation, we are comparing budget lines that are tangent to the old indifference curve, and to the new indifference curve. Our motivation for wanting to compare budget line with budget line is to avoid comparing utility level with utility level; we want a dollar measure of person i's gain, rather than a utility measure. However, if the budget lines we compare are not parallel, there is an unavoidable ambiguity about the distance between them (and therefore person i's money gain.) If they are parallel, there is no such ambiguity. Figure 9.5 shows the two ways to construct parallel budget lines tangent to the old and new indifference curves, with, in each case, one of the parallel lines being the true (pre or post change) budget line. It is then crucial, in comparing C and E, that the C measure is based on the "new" or after-move-from-x_i-to-y_i prices, whereas the E measure is based on the "old" or before-move-from-x_i-to-y_i prices.

Now consider this: Suppose society makes 2 policy changes. At the very beginning, person i is at x_i. The first policy change moves him to y_i. By the compensating variation measure, his gain is C in Figure 9.5. The next policy change starts with him at y_i. It then switches him back to x_i. Person i is hurt by this change. To measure the amount of his loss, in compensating variation money terms, we ask what amount of money would we have to give him, based on the (now) new prices at x_i, to make him exactly as well off as he used to be at y_i? The answer is E. That is, the compensating variation measure of his loss from the return trip from y_i to x_i is E.

The next obvious question is this: If we were to measure the net result of his having started at x_i, gone to y_i, and then returned to x_i, using the compensating variation measure, what would the result be? The answer is $C - E$, which is *negative*.

This is a paradox, because he has ended up exactly where he started!

9. Consumers' Surplus - Boadway Paradox

We now turn to an interesting example discovered by Robin Boadway. In this example we have 2 people, and we will analyze a move from one Pareto optimal point x to another Pareto optimal point y, in one Edgeworth box diagram, Figure 9.6.

We will measure the aggregate gain, in terms of compensating variation consumer's surplus for the two individuals. Since we are adding together *consumer's* surplus for person 1 and *consumer's* surplus for person 2, we have titled this section *consumers'* surplus.

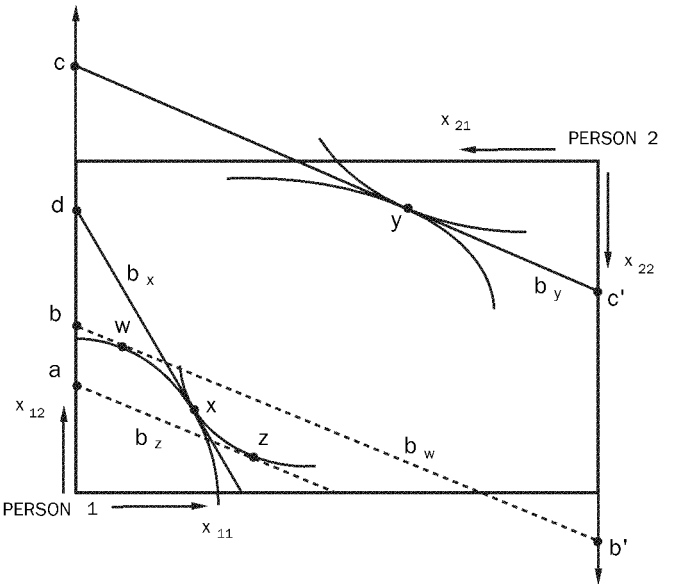


Figure 9.6.

The only crucial assumption for this construction is the assumption that the slope of the tangent indifference curves at point x is different than at point y.

Consider, then, the move from x to y in Figure 9.6. Person 1 is the gainer and 2 is the looser. We assume that prices are normalized so that a unit of good 2 costs a dollar. The old point is x and the new point is y. Therefore, since we are using the compensating variation measure, which relies on the new prices, we need to construct a hypothetical budget line with the same slope as the budget line b_y through y.

Person 1's gain is calculated much the same as in Figure 9.5. We project the budget line through y, b_y, to person 1's good 2 axis (which has been extended beyond the normal limits of the Edgeworth box diagram for these purposes.) We shift the budget line b_y downward in a parallel fashion until it is tangent to 1's old indifference curve, at the point z. Person 1's gain can now be read off the vertical axis as the difference $c - a$. (This corresponds to the magnitude C in Figure 9.5.)

Person 2's loss is calculated in a similar fashion. He starts at x and ends up, worse off, at the new point y. Based on the new prices, we can estimate his loss by asking this question: how much money would we have to give him, starting from the budget line b_y, to make him as well off as he was originally at x? The answer is that we would have to give him enough to shift b_y, in a parallel fashion, so that he can get to his

original indifference curve. This means we have to shift his budget line to b_w, and this implies we need to give him the amount $b' - c'$. That is, the compensating variation measure of person 2's loss equals $b' - c'$ on 2's vertical axis, which of course equals $c - b$ on 1's vertical axis.

We are now able to figure the total net change in compensating variation as society moves from x to y. It is 1's gain minus 2's loss, or $c - a - (c - b) = b - a > 0$. That is, compensating variation indicates that society has gained in the move from x to y. But this is nonsensical, since x and y are two equally attractive Pareto optimal allocations, and there is no plausible reason to favor one over the other! This is the Boadway paradox.

10. Quasilinear Utility

The paradoxes of the 2 preceeding sections don't arise if utility functions have a special form. Let's think for a moment in terms of a general exchange economy model, with a set of n individuals, consuming m goods. Person i's utility is a function $u_i(x_i)$, where $x_i = (x_{i1}, x_{i2}, \ldots, x_{im-1}, x_{im})$ is his bundle of goods.

Assume that, for every individual i, utility can be written in the following way:

$$u_i(x_i) = v_i(x_{i1}, x_{i2}, \ldots, x_{im-1}) + x_{im}.$$

We then say that the utility functions $u_1, u_2, \ldots, u_n$ satisfy the *quasilinearity assumption*, or, for short, that they are *quasilinear*.

Note that we are really assuming 2 things here: First, that each person i figures his utility as the sum of his utility from all the goods save one special good, plus the amount of that one special good he consumes. (The special good is good m in our definition.) Second, all the individuals agree on the identity of that one special good (the m*th*), which enters additively in the utility function of each person. That one special good can be thought of as "money" or "wealth."

In a two-person two-goods example, the quasilinear utility functions assumption means that the utility functions can be written

$$u_1(x_1) = v_1(x_{11}) + x_{12},$$
$$u_2(x_2) = v_2(x_{21}) + x_{22}.$$

When we analyze the two-goods case, we will assume, for simplicity, that the $v_i(x_{i1})$ part of the function shows "diminishing marginal utility," that is, $v_i'(x_{i1})$ declines as x_{i1} increases.

The assumption of quasilinearity may seem familiar, because we made essentially the same assumption in Chapter 8. Recall that in our public

goods model, we assumed person i's utility function was

$$u_i = v_i(x) + y_i,$$

where x = the quantity of the public good and y_i = i's consumption of the private good. We were therefore assuming that one good (the private good) entered additively in every person's utility function. That is, we were assuming quasilinear utility functions.

It turns out that the assumption of quasilinear utility functions changes Figure 9.5, and our discussion of compensating variation C and equivalent variation E, in a fundamental way. In fact, quasilinearity forces $C = E$, and eliminates the Figure 9.5 paradox. (It also eliminates the Figure 9.6 paradox.)

Recall the peculiar result from Figure 9.5: The shift from x_i to y_i, and then back from y_i to x_i, appears to leave person i worse off, because his gain on the out trip is C (compensating variation measure), and his loss on the return trip is E (compensating variation measure), and, since E is clearly greater than C, he seems to have a net loss from the round trip. But this is impossible, since he ends up exactly where he started!

But what if we modified Figure 9.5 in the following way: Place w_i directly above x_i, and z_i directly below y_i. Moreover, make the vertical gap at w_i/x_i exactly equal to the vertical gap at y_i/z_i. Figure 9.7 is constructed in this way.

In Figure 9.7, by elementary geometry E is equal to the w_i/x_i gap, C is equal to the y_i/z_i gap, and so $C = E$.

The key to Figure 9.7 is that the pair of indifference curves shown are "parallel," in the sense that (1) the slope agrees for any two points on the two indifference curves if and only if those points lie directly above/below one another, and (2) the vertical gap between the 2 indifference curves is the same, no matter what the x_{i1}. With properties (1) and (2) imposed, the w_i constructed in Figure 9.5 would be forced to lie directly above x_i, because MRS for person i is the same at x_i and w_i, and similarly for y_i/z_i. Moreover, the vertical gap at x_i/w_i would be forced to equal the vertical gap at y_i/z_i. This would transform Figure 9.5 into Figure 9.7, where $C = E$, and the paradox would disappear.

We now show that the assumption of quasilinear utility functions gives properties (1) and (2) of the previous paragraph, and therefore forces $C = E$. We do this in the two-goods case:

First, note that if

$$u_i(x_i) = v_i(x_{i1}) + x_{i2},$$

then at any point on any indifference curve,

$$\text{MRS for person } i = \frac{v_i'(x_{i1})}{1} = v_i'(x_{i1}).$$

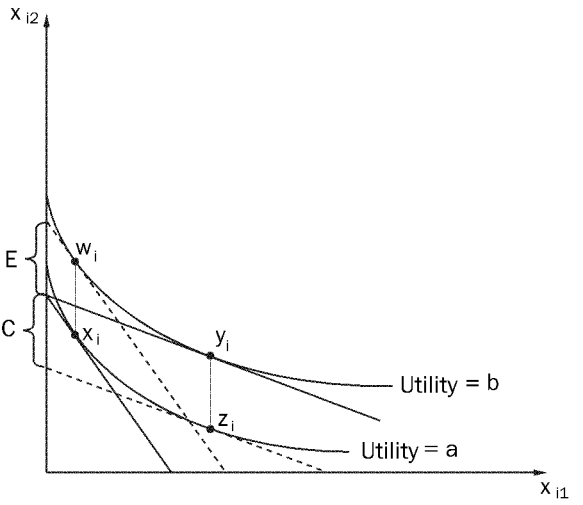


Figure 9.7.

Since we are assuming the $v_i'(x_i)$ function is monotonically decreasing, 2 points on 2 separate indifference curves, such as x_i and w_i in Figure 9.7, would have the same MRS if and only if they had the same x_{i1} coordinate, that is, if and only if they lie directly above/below one another. This gives (1).

For (2) we want to establish that the vertical gap between the indifference curve identified as "utility = a," and the "utility = b" indifference curve, is the same at w_i/x_i as it is at y_i/z_i. To do this, we write down the utility functions and utility levels at the 4 points.

$$\begin{aligned} &\text{At } w_i: && v_i(w_{i1}) + w_{i2} = b \\ &\text{At } x_i: && v_i(x_{i1}) + x_{i2} = a \\ &\text{At } y_i: && v_i(y_{i1}) + y_{i2} = b \\ &\text{At } z_i: && v_i(z_{i1}) + z_{i2} = a. \end{aligned}$$

Subtract the second equation from the first and you get $w_{i2} - x_{i2} = b - a$. Subtract the fourth equation from the third and you get $y_{i2} - z_{i2} = b - a$. Therefore

$$w_{i2} - x_{i2} = y_{i2} - z_{i2},$$

proving part (2). We have now shown that the assumption of quasilinear utility functions forces $C = E$ and precludes the paradox of Figure 9.5. This is the theoretical basis for the use of Marshallian consumers' surplus in applied studies.

To use the quasilinear utility functions in the general n-person m-goods exchange economy model, it must be the case that all the utility

functions separate into 2 parts, as indicated above, and the special good be the same good for everybody, also as indicated above. In other words, there must be unanimous agreement about the identity of the good that enters in a linear fashion in everyone's utility function. Moreover, calculations of compensating variation or equivalent variation must be made in terms of that one special good, not in terms of (in our notation) goods $1, 2, \ldots, m-1$.

11. Cost Benefit Analysis with One Good

The paradoxes of consumers' surplus disappear if the assumption of quasilinear utility holds. It is also true that there are no paradoxes if relative prices don't change, when the economy moves from one situation to another. The source of the odd results in the preceeding sections is the fact that there are 2 (or more) goods, and that relative prices change.

In an exchange economy with just one good, and with a fixed initial endowment $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$, any alternative is simply a redistribution of the total. Under the standard assumptions (of self-interestedness and monotonicity), every redistribution x is Pareto optimal, there are no Pareto moves from x to y (that is, no instances of Pareto improvements), nothing is Kaldor superior to anything else, and so on.

Let us now relax the assumption that the total is fixed.

Now if $x = (x_1, x_2, \ldots, x_n)$ is an alternative, the associated total is

$$\sum_{i=1}^{n} x_i,$$

and if $y = (y_1, y_2, \ldots, y_n)$ is another alternative, we might have

$$\sum_{i=1}^{n} y_i \neq \sum_{i=1}^{n} x_i.$$

We will say that x is superior to y by the *cost-benefit criterion* if the "size of the pie" under x is bigger than the "size of the pie" under y, that is, if

$$\sum_{i=1}^{n} x_i > \sum_{i=1}^{n} y_i.$$

We naturally define $S(x)$, the set of alternatives available from x, as that set of allocations that have the same total as x. In such a world everything collapses to comparisons of the aggregates, and paradoxes disappear. We will leave the details and the proofs as exercises for the reader.

12. Exercises

1. The cost-benefit criterion is a widely-used tool of policymakers, and it derives from the compensation criteria of Kaldor, Scitovsky and Samuelson. To analyze it in a truly simple (and simplistic) way, suppose, as in the last section of the chapter, that there is only one good in the economy, which everyone always wants more of, and suppose an alternative $x = (x_1, x_2, \ldots, x_n)$ is just a distribution or allocation of the one good among the people in the economy. The total amount of the good is

$$\sum_{i=1}^{n} x_i$$

and we shall assume for this problem that this quantity need not be fixed. That is, if y is another alternative, it is possible to have

$$\sum_{i=1}^{n} y_i \neq \sum_{i=1}^{n} x_i.$$

The sum

$$\sum_{i=1}^{n} x_i$$

can be regarded as the size of the total pie under x; while x_i is, of course, the size of person i's share of the pie under x.

In this ultra-simple model, the cost-benefit criterion says x is superior to y, which we can write $x\mathcal{C}y$, if the size of the pie under x exceeds the size of the pie under y, i.e., if

$$\sum_{i=1}^{n} x_i > \sum_{i=1}^{n} y_i.$$

We can interpret $S(x)$ as the set of possible distributions of a pie of size

$$\sum_{i=1}^{n} x_i.$$

Show that the Kaldor criterion, the Scitovsky criterion, the Samuelson criterion, and the cost-benefit criterion are all equivalent in this one-good model. That is, show that $x\mathcal{K}y$ implies $x\mathcal{S}y$, that $x\mathcal{S}y$ implies $x\mathcal{N}y$, that $x\mathcal{N}y$ implies $x\mathcal{C}y$, and that $x\mathcal{C}y$ implies $x\mathcal{K}y$.

2. Show with an Edgeworth box diagram that the assumption of quasi-linear utility makes the Boadway paradox impossible. (Assume for simplicity that $v_i'(x_{i1})$ declines as x_{i1} increases for $i = 1, 2$.)

13. Selected References

1. R. Boadway, "The Welfare Foundations of Cost-Benefit Analysis," *Economic Journal*, V. 84, 1974, pp. 926-939.

2. A.M. Feldman, "Kaldor-Hicks Compensation," in P. Newman (ed.) *The New Palgrave Dictionary of Economics and the Law*, McMillan ltd., London, V. 3, 1998, pp. 417-421.

3. J.R. Hicks, "The Foundations of Welfare Economics," *Economic Journal*, v. 49, 1939, pp. 696-712.

4. J.R. Hicks, "Consumers' Surplus and Index-Numbers," *Review of Economic Studies*, V. 9, 1942, pp. 126-137.

5. N. Kaldor, "Welfare Propositions of Economics and Interpersonal Comparisons of Utility," *Economic Journal*, V. 49, 1939, pp. 549-552.

   This is a short note about the classical arguments for free trade. Kaldor observes that the repeal of the English Corn Laws could be justified on these grounds: it would be possible for the government to compensate landlords for their losses from repeal, leaving everyone better off than before. So repeal is Kaldor superior to non-repeal. In general, welfare economists should be in favor of policies that increase aggregate production, according to Kaldor.

6. P.A. Samuelson, "Evaluation of Real National Income," *The Oxford Economic Papers*, V. 2, 1950, pp. 1-29.

   This paper starts with a discussion of the question "When does an increase in real national income correspond to an increase in social welfare?" Samuelson shows that if relative prices change in a move from situation 1 to situation 2, comparing real national income at the two situations will generally give inconclusive results. This result is a fatal blow for a line of theoretical welfare economics that starts with Pigou and continues through Kaldor, Hicks, and others. (However, theoretical rigor mortis does not imply practical rigor mortis; most of today's economists, including Samuelson, are very much concerned with growth in real national income. And this concern probably makes lots of sense, in spite of negative theoretical results.)

In the second half of the paper Samuelson analyzes production and utility frontiers (which he calls production and utility possibility functions). He applies the utility frontier analysis to the criteria of Kaldor and Scitovsky, both of which he criticizes. One of his conclusions is that "the only consistent and ethics-free definition of an increase in potential real income of a group is that based upon a uniform shift of the utility possibility function of the group."

7. A.K. Sen, *Collective Choice and Social Welfare*, Holden-Day, Inc., San Francisco, 1970, Chapters 2, 2*.

   Chapter 2 provides a short, clear, intuitive discussion of the Pareto criterion and related approaches. Chapter 2*, which is mathematical in style and notation, has a section on the Kaldor and Scitovsky criteria. Sen criticizes the Scitovsky criterion on the grounds that it might be nontransitive: it is possible to have xSy, ySz, but not xSz.

8. T. Scitovsky, "A Note on Welfare Propositions in Economics," *The Review of Economic Studies*, V. 9, 1941, pp. 77-88.

   This careful article elaborates on the criterion of Kaldor (and of J.R. Hicks). Scitovsky expresses his own criterion so clearly that it's worth quoting here:

   "We propose, therefore, to make welfare propositions on the following principle. We must first see whether it is possible in the new situation so to redistribute income as to make everybody better off than he was in the initial situation; secondly, we must see whether starting from the initial situation it is not possible by a mere redistribution of income to reach a position superior to the new situation, again from everybody's point of view. If the first is possible and the second impossible, we shall say that the new situation is better than the old was. If the first is impossible but the second possible, we shall say that the new situation is worse; whereas if both are possible or both are impossible, we shall refrain from making a welfare proposition."

Chapter 10

FAIRNESS AND THE RAWLS CRITERION

1. Introduction

The previous chapter described some of the properties and some of the shortcomings of the social improvement criteria of Pareto, Kaldor, Scitovsky and Samuelson. The question of when one alternative ought to be considered socially preferred to another remains largely unresolved. In this chapter we shall examine two new criteria for judging social states, both of which are explicitly egalitarian in their viewpoints. The first is the criterion of fairness in the sense of non-envy. Under some distributions of goods, no person would prefer any other's bundle of goods to his own. Under other distributions, some people would prefer others' bundles to their own. We call the first kind of distribution "fair" and the second "unfair." Can this distinction be fruitfully used to indicate when one distribution is socially preferable to another? The second is the criterion of justice, developed by John Rawls. According to Rawls, alternative x is more just than alternative y if the worst-off person under x is better off than the worst-off person under y. Does this criterion provide a good basis for judging between alternatives for society?

In the section below on fairness we shall revert to the exchange theory model, and (temporarily) put aside the preference relation approach of the last chapter. This discussion will also abstract from questions of production, although production could be explicitly introduced into the fairness question. The section on the Rawls criterion, however, will mainly be in the style of the last chapter, with abstract alternatives x, y, z that might or might not be allocations of goods in an n-person, m-good exchange economy.

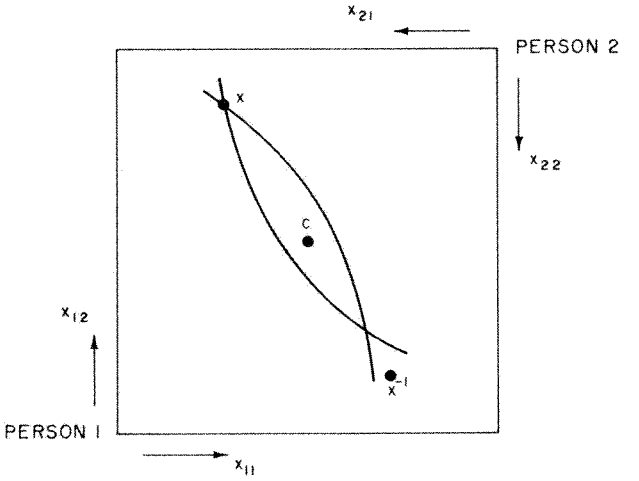


Figure 10.1.

Now let's examine the idea of fairness in more detail, and show how fair allocations are related to competitive equilibrium allocations and Pareto optimal allocations. Then we will discuss the usefulness of the fairness criterion as a guide to which allocations are socially preferable to which.

2. Fairness

Suppose $x = (x_1, x_2, \ldots, x_n)$ is an allocation of m goods among n people. Assume for simplicity that there are no externalities; so person i's utility depends only on his bundle of goods x_i. If $u_i(x_i) \geq u_i(x_j)$, then i likes his own bundle as well as he would like j's, and we say i does not envy j. If $u_i(x_i) \geq u_i(x_j)$ for every pair of people $\{i, j\}$, we say the allocation x is *fair*. (Other authors have used the terms *envy-free* and *equitable* for what we are calling *fair*.) A fair allocation x is illustrated in Figure 10.1.

In this figure, the allocation $x^{-1} = (x_2, x_1)$ reverses the bundles assigned by the allocation $x = (x_1, x_2)$. That is, x^{-1} gives person 1 person 2's bundle, and it gives person 2 person 1's bundle. Geometrically x^{-1} is the mirror image of x through the center of the box. So x, c, and x^{-1} all lie on a straight line, and the distance from x to c equals the distance from x^{-1} to c.

Observe that 1's indifference curve through x passes above the allocation x^{-1}. This means that he prefers x to x^{-1}, and, consequently, $u_1(x_1) > u_1(x_2)$. Similarly, 2's indifference curve through x passes above

x^{-1} ("above" in terms of 2's origin). Therefore, $u_2(x_2) > u_2(x_1)$. In short, at x neither individual would prefer the other's bundle to his own, so x is fair. Since x is obviously not Pareto optimal, the picture shows that fair allocations are not generally optimal allocations. The converse is also rather obvious: optimal allocations are not generally fair.

What then are the connections between fairness, on the one hand, and optimality on the other hand? The principal connection is indicated by the following:

Proposition 1. Suppose all people have monotonic self-interested utility functions. Let the initial allocation ω be the equal allocation, that is, suppose $\omega_1 = \omega_2 = \ldots = \omega_n$. Suppose $(\hat{x}, p)$ is a competitive equilibrium.

Then $\hat{x}$ is Pareto optimal, and it is fair.

Proof. $\hat{x}$ is Pareto optimal by the First Fundamental Theorem of Welfare Economics. To see that it must be fair, recall from the definition of a competitive equilibrium that $\hat{x}_i$ must maximize $u_i(x_i)$ subject to $p \cdot x_i \leq p \cdot \omega_i$, for all i. However, the righthand side of the inequality is the same for everyone, since ω_i is the same for all i. In short, each person attempts to maximize his utility subject to the same budget constraint. Since this is true, it is impossible for i to end up with a bundle which he finds inferior to j's. Therefore, for all i and j, $u_i(\hat{x}_i) \geq u_i(\hat{x}_j)$, or the allocation $\hat{x}$ is fair. Q.E.D.

The proposition indicates a way to achieve both Pareto optimality and fairness, at the same time. Start at the equal allocation, and allow a move to a competitive equilibrium allocation, which will be optimal and fair. And consequently, it suggests one way to choose among the Pareto optimal allocations–one way to decide when one Pareto optimal allocation is better than another. The proposition says that there exist allocations that are Pareto optimal and fair. If the choice is between an allocation that is Pareto optimal but not fair, and an allocation that is Pareto optimal and fair, choose the fair one.

Unfortunately, however, if we push the idea of fairness a little, if we pursue some of its implications, if we try to broaden Proposition 1, we start to uncover contradictions and dilemmas. Let's illustrate these difficulties with a few examples.

The first example is a three person, three good exchange economy, in which, for purposes of mathematical simplicity, each person's utility is a linear function of the amounts of the three goods he consumes. All the details are in the following table.

	Utility Function	ω_i	$u_i(\omega_i)$	x_i	$u_i(x_i)$
Person 1	$u_1(x_1) = 3x_{11} + 2x_{12} + x_{13}$	(1,1,1)	6	(3,2/3,0)	10 1/3
Person 2	$u_2(x_2) = 2x_{21} + x_{22} + 3x_{23}$	(1,1,1)	6	(0,0,2)	6
Person 3	$u_3(x_3) = x_{31} + 3x_{32} + 2x_{33}$	(1,1,1)	6	(0,7/3,1)	9

The initial allocation $\omega = (\omega_1, \omega_2, \omega_3)$ is the equal allocation, which must be fair. Although it is not a competitive equilibrium allocation based on ω, the final allocation $x = (x_1, x_2, x_3)$ is in the core. That is, no coalition or group of people can block it. The allocation x is also (necessarily) Pareto optimal. But is x fair? No it is not, since $u_2(x_1) = 6\ 2/3 > u_2(x_2) = 6$; that is, 2 envies 1 under x. Consequently, the example shows that even if we start at the equal allocation (the fairest of the fair), a move to the core, rather than to a competitive equilibrium, can destroy fairness.

It is true, of course, that this destruction of fairness, which subverts the spirit of Proposition 1, might be blamed on the nonprice trading in the example. Suppose we try to expand the scope of Proposition 1 by starting at a fair but not equal allocation, and then allow a move to a competitive equilibrium allocation. Is fairness preserved by such a move? The Edgeworth box diagram in Figure 10.2 shows it might not be.

In the figure, person 1's indifference curve through $\omega = (\omega_1, \omega_2)$ passes above the allocation $\omega^{-1} = (\omega_2, \omega_1)$; so he prefers his bundle under ω to person 2's. Similarly, person 2 prefers his own bundle under ω to person 1's. Consequently, ω is fair; neither one envies the other. The point $\hat{x}$ is a competitive equilibrium allocation based on ω. Person 2 prefers his own bundle at $\hat{x}$ to person 1's, since his indifference curve through $\hat{x}$ passes above $\hat{x}^{-1}$. However, person 1's indifference curve through $\hat{x}$ passes below $\hat{x}^{-1}$ and therefore 1 envies 2 at $\hat{x}$, and $\hat{x}$ is not fair. The move from ω to $\hat{x}$ destroys fairness.

The first example showed that a move from the equal allocation to the core can destroy fairness. The second example showed that a move from a fair allocation to a competitive equilibrium can destroy fairness. The examples both indicate that the clever idea of Proposition 1 (starting at the equal allocation and allowing the competitive market mechanism to bring the economy to a fair and Pareto optimal allocation) can't be easily broadened. And they also show that fairness will often be destroyed by the usual economic transactions, even transactions that make everyone better off.

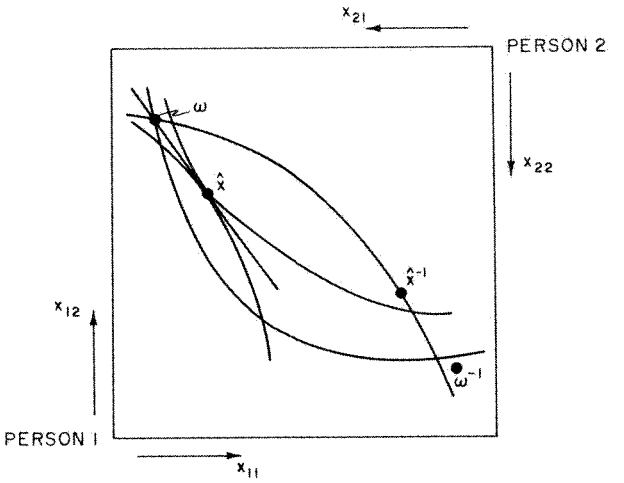


Figure 10.2.

This last point should be emphasized. In the three-person, three-good example, the core allocation x is Pareto superior to the initial allocation ω. But ω is fair, while x is not. In the example of Figure 10.2, the competitive equilibrium allocation $\hat{x}$ is Pareto superior to the initial allocation ω. But again, ω is fair, while $\hat{x}$ is not. In both cases, the criterion of fairness is contrary to the Pareto criterion. If the choice is between ω and x in the first example, or between ω and $\hat{x}$ in the second, which is socially preferable to which? Fairness is clearly a poor guide in these cases.

In short, there is considerable tension between fairness as a social improvement criterion and the usual criteria of welfare economics. Few economists would say that fairness is, by itself, a good guide to what alternatives are better than what alternatives. Fewer still would choose fairness over the Pareto criterion when the two criteria disagree.

3. The Rawls Criterion

No discussion of fairness would be complete without a treatment of the notion of justice developed in the 1960's by the philosopher John Rawls. In this section, we shall give a very brief (and therefore possibly "unfair") characterization of Rawls' notion.

Let us imagine that a group of individuals, a society, is choosing among social alternatives, or social states. Now these alternatives, these states, are for Rawls characterized by degrees of liberty as well as by

distributions of wealth or goods, but we shall focus only on the distributions of wealth or goods. Let us also suppose that the individuals who are considering the alternative states do not know what positions, what roles, they would occupy in those states. That is, person i might know that in state x there is one millionaire and one pauper, but he does not know whether he would be the millionaire or the pauper. He does not know what position he would occupy. The people in this society are making choices among the states from behind a veil of ignorance.

Now what state would they choose? To illustrate the Rawls criterion with a sensible concrete example, we assume here that there are two people and only one good (say wealth, or income). Let x be a social state in which one person's wealth is \$1,000,000 and the other person's wealth is \$0. Let y be a social state in which the wealth of both is \$1,000. Rawls argues that the social choice between x and y, made from behind a veil of ignorance, ought to be y. Not knowing whether he would be the millionaire or the pauper under x, each person ought to prefer y, because under y each is guaranteed a tolerable level of wealth. That is, the Rawls criterion is to opt for the state that maximizes the utility of the person in the worst position, or to maximize the minimum utility. It is therefore called a *maximin criterion.*

Formally, let $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$ be distributions of one good among n people.

Here $\sum_{i=1}^{n} x_i$ need not equal $\sum_{i=1}^{n} y_i$.

We say x is *Rawls superior* to y, written $x\mathcal{R}y$, if

$$\min\{x_1, x_2, \ldots, x_n\} > \min\{y_1, y_2, \ldots, y_n\}.$$

If it should happen that $\min\{x_1, x_2, \ldots, x_n\}$ equals $\min\{y_1, y_2, \ldots, y_n\}$, the logical thing to do would be to look at the second smallest of the x_j's and the second smallest of the y_j's. We won't worry too much about this nicety, however; we'll assume that the minimum of the x_j's is different from the minimum of the y_j's.

What are the properties of the Rawls criterion? First, it has clear egalitarian implications. If society is choosing among distributions of a fixed total of wealth, that is, among a set of x's for which

$$\sum_{i=1}^{n} x_i = C$$

and C is a constant, then the single distribution that is Rawls superior to all the others is the equal distribution $(C/n, C/n, \ldots, C/n)$.

Second, if choices are being made from among distributions which don't have fixed totals, the Rawls criterion can be perversely oversolicitous of the worst-off person. For instance, let

$$x = (50, 100, 150)$$
$$y = (90, 90, 90)$$
$$z = (80, 250, 250).$$

Then the Rawls criterion says y is better than x, which is plausible; going from x to y sacrifices some total wealth

$$\left(\sum_{i=1}^{3} x_i = 300 \text{ while } \sum_{i=1}^{3} y_i = 270\right)$$

but gains lots of equality. However, the Rawls criterion also says y is better than z, and in this case it sacrifices a great deal of total wealth

$$\left(\sum_{i=1}^{3} z_i = 580\right)$$

for the sake of some increase in equality. In fact, if real people named 1, 2 and 3 were choosing between y and z from behind a veil of ignorance, and if, like most real people, they were willing to take small risks for large potential gains, they would probably choose z, in spite of Rawls' advice.

The reader should note that the criteria of Kaldor, Scitovsky and Samuelson can be easily applied in this simple one good Rawlsian example. The set of states accessible for x, that is $S(x)$, is the set of all possible distributions of the total

$$\sum_{i=1}^{n} x_i.$$

$S(y)$ and $S(z)$ are defined analogously. Now according to the Kaldor, Scitovsky and Samuelson criteria, z is better than y, contrary to Rawls. That is, in a move from y to z, the gainers (persons 2 and 3) could easily compensate the loser (person 1), but not vice versa, and the utility frontier based on z lies entirely outside of the utility frontier for y. So the Rawls criterion is inconsistent with the usual economic compensation criteria, at least in this case.

Third, in a world in which there is more than one good, and in which individuals' tastes differ, the Rawls criterion may be impossible to apply.

Suppose, for instance, that x and y are alternative distributions of m goods, where $m > 1$. Define

$$\min_j u_i(x_j) = \text{ minimum}\{u_i(x_1), u_i(x_2), \ldots, u_i(x_n)\}.$$

Now person 1 might feel that

$$\min_j u_1(x_j) > \min_j u_1(y_j).$$

That is, he might think that the worst that could happen to him under x, if he were assigned to any of the n positions or given any one of the n bundles, is better than the worst that could happen to him under y. However, person 2 might find that

$$\min_j u_2(x_j) < \min_j u_2(y_j).$$

That is, the worst under y for person 2 might be better than the worst under x. In this case, the two individuals differ about which alternative distribution has a better worst position. If there is disagreement like this behind the veil of ignorance, which distribution should be chosen? The Rawls criterion can't say.

To make this possibility clear, let's construct a simple and slightly silly example. Let x be a distribution that gives each of three people a red hat; let y be a distribution that gives each of the three a blue hat; and let z be a distribution that gives each of three a green hat. Assume that x, y, and z are identical aside from hats. Suppose person 1 likes red better than blue and blue better than green. Suppose person 2 likes blue better than green and green better than red. Suppose person 3 likes green better than red and red better than blue. Then person 1 will rank the distributions x first, y second, and z third, even from behind a veil of ignorance. Person 2 will rank them y first, z second, and x third, from behind a veil of ignorance, while person 3 will rank them z first, x second and y third, from behind a veil of ignorance. There is a total disagreement about which alternative is best for the worst-off person!

In sum, the Rawls criterion, like the fairness criterion, doesn't mesh well with the usual economic criteria. Just as the fairness criterion can contradict the more fundamental Pareto criterion, the Rawls criterion can contradict the criteria of Kaldor, Scitovsky and Samuelson. Moreover, the Rawls criterion is based on a maximin test that is probably contradicted by the economic behavior of most people. That is, when faced with the choice between $y = (90, 90, 90)$ and $z = (80, 250, 250)$, most of us would, from behind a veil of ignorance, place out bets on z.

This in spite of Rawls' preference for y. Finally, if there are many goods and tastes differ, the Rawls criterion might simply be inapplicable.

For these reasons, the Rawls criterion, like the fairness criterion, cannot be the final answer. The question When is alternative x socially preferable to alternative y? remains open.

4. Exercises

1. a. For the exchange economy given by

$$\begin{aligned} u_1 &= x_{11}x_{12} & \omega_1 &= (1,0) \\ u_2 &= 2x_{21} + x_{22} & \omega_2 &= (0,1) \end{aligned}$$

find an allocation that is both Pareto optimal and fair. (Note: This is the same economy as in Exercise 5, Chapter 3.)

b. For this economy, find a specific pair of points x and y such that x is fair and y is not, and y is Pareto superior to x.

2. In an m-goods exchange model, the Rawls criterion can be formulated as follows: For each possible allocation x, each person i calculates

$$\min_j u_i(x_j).$$

This is the lowest utility he can imagine under x, the level he would get if he were assigned what he views as the worst possible bundle under x. Then person i prefers allocation x to allocation y, from behind a veil of ignorance, if

$$\min_j u_i(x_j) > \min_j u_i(y_j).$$

If all the individuals prefer x to y from behind a veil of ignorance, the Rawls criterion can be applied without ambiguity.

Now consider the exchange economy in which

$$u_1 = x_{11} + 2x_{12}$$

and

$$u_2 = 2x_{21} + x_{22}.$$

Let $x_1 = (1/2, 1/2)$, $x_2 = (1/2, 1/2)$, and let $y_1 = (0,1)$, $y_2 = (1,0)$. What are person 1's behind-a-veil-of-ignorance preferences regarding allocations x and y? Person 2's? Which allocation is Rawls superior to which?

Now remove the veil of ignorance, so person 1 knows he gets bundle 1, etc. Which allocation is Pareto superior to which? How does the Rawls criterion compare with the Pareto criterion?

5. Selected References

(Items marked with an asterisk (*) are mathematically difficult.)

1. A. Feldman and A. Kirman, "Fairness and Envy," *American Economic Review*, V. 64, 1974, pp. 995-1005.

   This paper criticizes the notion of fairness, on the grounds that a move that makes everyone better off from a fair allocation might end up at an unfair allocation. That is, the fairness criterion might be contrary to the Pareto criterion. The paper also discusses the question of minimizing the degree of unfairness in an economy, and shows that this minimization problem can also lead to paradoxical results.

2. D. Foley, "Resource Allocation in the Public Sector," *Yale Economic Essays*, V. 7, 1967, pp. 45-98.

   Section IV of Foley's paper is on "equity," and in it he introduces the fairness as non-envy concept used here: "This [discussion about inequality] suggests a new way to define equality even when preferences are diverse: an allocation is equitable if and only if each person in the society prefers his consumption bundle to the consumption bundle of every other person in the society."

3. E. Pazner, "Recent Thinking on Economic Justice," *Journal of Peace Science*, V. 2, 1976, pp. 143-154.

   This very readable short survey of topics in the theory of fairness starts with a discussion of fair and Pareto optimal allocations in exchange economies, and in economies with production and exchange. Production makes the analysis of fairness a little bit more complicated. Some people are born with very valuable productive resources — e.g., Luciano Pavarotti, Reggie Jackson, Michael Jordan, Tiger Woods — and if the ownership of these resources is taxed away in the interest of fairness, productive efficiency may disappear as well. So the idea of fairness may need modification to preserve Proposition 1.

   In the second half of the paper, Pazner discusses the idea of *egalitarian-equivalence*. There is a very brief but useful bibliographical section at the end of the paper.

4. E. Phelps, "Recent Developments in Welfare Economics: Justice et Equité," *Frontiers in Quantitative Economics*, Vol. IIIB (M. Intriligator, ed.), North Holland, 1976, pp. 703-730.

This is a sophisticated survey of many areas in economics, including justice and equity. It covers compensation criteria, Arrow's theorem, utilitarianism, and public finance models, in addition to the idea of fairness as non-envy, and the Rawlsian position.

5. J. Rawls, "Constitutional Liberty and the Concept of Justice," C.J. Friedrich and J. Chapman, eds., *Justice: Nomos 8*, Atherton Press, New York, 1963.

6. J. Rawls, *A Theory of Justice*, Harvard University Press, Cambridge, 1971.

7. A.K. Sen, *Collective Choice and Social Welfare*, Holden-Day Inc., San Francisco, 1970, Chapter 9.

   This is a good discussion of a number of topics related to the Rawls criterion.

8. W. Thomson, "The Consistency Principle," in T. Ichiishi et al (eds.) *Game Theory and Applications*, Academic Press, San Diego, 1990.

   This excellent survey covers numerous uses of the consistency axiom to characterize different allocation rules in different models. One of these characterizations concerns the competitive equilibrium from equal division of the endowments.

9. H.R. Varian, "Equity, Envy and Efficiency," *Journal of Economic Theory*, V. 9, 1974, pp. 63-91.

   Varian discusses fairness in an economy with production, as well as income-fairness, and coalition-fairness.

Chapter 11

LIFE AND DEATH CHOICES

1. Introduction

All of the models we have considered to this point, and many we will turn to later, have taken the population, the set of decision makers, the society, as given. The exchange economy model is based on a fixed population of traders. The Pareto criterion says x is superior to y if all the people in a given population like x at least as well as y, and some like it better. The Kaldor criterion is based on a fixed population. Majority voting, which we will study in the next chapter, says x is as good as y if the number of people in a given population who prefer x to y is greater than or equal to the number of people in the same population who prefer y to x. An Arrow social welfare function, which we will analyze in Chapter 13, takes the preferences of each member of a given society and transforms those preferences into a social preference relation.

But what if the population changes? For instance, what if a set of individuals $\{1, 2, \ldots, n\}$ is attempting to choose between alternatives x and y, but x will kill off some of the people, and y will add additional people?

In fact, this is an extremely common question that policy makers and economists face almost every day. For instance: Should a state spend \$5 million replacing a highway if those repairs will likely result in 1 less traffic fatality in the next year? Should a government spend \$10 billion on AIDS drugs if those drugs will prevent 1,000 deaths? Should a government prohibit a sport or leisure activity if that sport creates a 1/6 probability of death per play (e.g., Russian roulette with a 6-chamber revolver)? Should it prohibit a sport or leisure activity if that

sport creates a 1/1,000,000 probability of death per day (e.g., downhill skiing)?

Is it better for a country to have a higher population or lower? If it is better to have more people, should this be done by encouraging births, or increasing life expectancy? If it is better to have fewer people, is it better to reduce birth rates or increase deaths?

In this chapter we will look at some partial answers to questions like these. We will start with the standard economic approach to life/death issues, and then we will turn to the more abstract, philosophical approaches.

2. Economic Model – The Money Value of a Life

Placing a money value on a life in legal disputes is an ancient practice. For instance, in the Code of Hammurabi (circa 1750 B.C.) there is a paragraph that reads, in part, "if a citizen has struck a citizen in a quarrel, and has inflicted on him a wound, [and] if he has died as a consequence..., he shall swear, [he struck him unwittingly], ... and ... he shall pay a half-mina of silver." Some lines of *Exodus* suggest money damages for accidental killings in limited circumstances. The Justinian Code provided for money payments to be made in case of accidental killing of slaves. Islamic law distinguishes between intentional and accidental killing, and provides for paying of *diyah* "blood-wit" in the case of accidental killing. Similarly, early custom in pre-Norman Britain put a compensating price (*wirgild*) on deaths.

The modern Anglo-American legal treatment of accidental killing, which started in the mid 19th century, typically provides that dependents of a deceased person may recover for pecuniary losses they suffer, especially lost wages the deceased would have provided. The deceased is primarily viewed as a money making machine. The value of his life is mainly given by lifetime income or earnings, possibly net of expenses needed to maintain the machine (e.g., food, clothing, etc.), possibly discounted to present value, and possibly augmented by the value of non-paid services provided. This can be called the *human capital* approach: the person is valued as a (human) money making machine.

The human capital approach to valuing lives, however, ignores how much the deceased himself would value *being alive*.

How can being alive be valued? Consider the question: "How much is your life worth, in dollars?" What does it mean? It might be a threat: a robber has a gun pointed at you; you have $1,000 in your pocket, and he asks the question. You hand over the $1,000 and he goes away. You have then revealed that you are willing to pay $1,000 (or more) to preserve your life. Or, it might be an opportunity: a benefactor with great wealth

approaches you (and your attorney) with an offer: if you sacrifice your life, he will pay $10 billion to your estate, which will then be distributed as your will provides. You accept the offer, sacrifice your life, and your spouse and children are wealthy as a result. You have now revealed that you are willing to accept $10,000,000,000 (or less) to sacrifice your life.

Obviously exercises like these will tend to produce wildly disparate numbers. The standard economic measure of the value of a life, in contrast, looks at willingness-to-pay (or willingness-to-accept) for small changes in the probability of death.

The *willingness-to-pay* approach to the value of life works as follows: Suppose one person has an opportunity to reduce his probability of dying by ε, if he participates in some government program, uses some medical procedure, buys some safety enhancement for his car, and so on. Let c be the maximum he would be willing to pay for the given ε. Then we say the willingness-to-pay value of his whole life is c/ε.

Alternatively, a person may be faced with an increase in the probability of dying of ε, because of a riskier job, a hazardous trip, etc. Let c be the minimum he would be willing to accept to compensate him for the given ε. Then we say the willingness-to-accept value of his whole life is c/ε.

(This type of analysis is largely due to Thomas Schelling (1968), E.J. Mishan (1971) and M.W. Jones-Lee (1974). A good survey can be found in W. Kip Viscusi (1993).)

3. A Formal Version of the Economic Model

We will now develop a relatively simple model to show how one individual "computes" the value of his life.

In this model there is just one person, so we will dispense with an identifying subscript. There are two time periods. In period 1, the planning or ex-ante period, he decides on how to allocate his spending. He can spend on consumption, on precaution, or on insurance. Between period 1 and period 2, the ex-post period, events unfold, which leave him either alive, or dead. The probability that he ends up alive in period 2 depends on how much he spends on precaution in period 1. If he is alive, he consumes the amount he chose in period 1. If he is dead, the amount he would have consumed, plus the value of any insurance policy he bought, is bequeathed to his heirs.

We use the following notation:

$$\begin{aligned} x &= \text{ consumption in period 2 (or part of bequest, if he is dead)} \\ y &= \text{precaution expenditure} \\ z &= \text{insurance expenditure} \\ w &= x + y + z = \text{ initial cash endowment} \\ q(y) &= \text{probability he is alive in period 2} \\ V &= \text{face value of any life insurance policy he buys} \end{aligned}$$

We assume the $q(y)$ function is nicely behaved: $0 < q(y) < 1$ for all y, $q(y)$ increasing in y, concave, and smooth.

We assume that the cost of life insurance would reflect the actual odds that he will die, so that $z = V \cdot (1 - q(y))$. That is, the price of insurance is "actuarially fair."

We assume our individual reasons as follows: He recognizes there are 2 states of the world for him in period 2, alive, or dead. He has a state-contingent utility function:

$$\begin{cases} f(x) = x^{\alpha} \text{ if alive} \\ g(x + V) = (x + V)^{\alpha} - K \text{ if dead.} \end{cases}$$

Note that α is some fixed parameter, with $0 < \alpha < 1$, and K is some constant.

Let's pause to consider this assumption. The "if alive" part is fairly reasonable, as consumption x increases, he is happier, which makes sense. The particular power function form of the utility function implies that he has diminishing marginal utility from consumption.

The utility "if dead" part requires several comments. First, the $(x + V)^{\alpha}$ part of the function means that our individual contemplates the bequest of what he has not lived to consume (x) plus any life insurance policy (V), just as he would contemplate it if he himself were doing the consuming. That is, we are assuming the same power utility function here. So our person presumably has dependents in mind, or a charity, or some other bequest motive. He cares about consumption by, e.g., his widow and orphans, in much the same way he cares about his own consumption. Second, he (probably) views the dead state as undesirable, with the degree of undesirability captured by the constant K. If K is a large (positive) number, he thinks the dead state is very bad. On the other hand, if it is zero, he views the dead state as similar to the live state, and if K is a negative number, he wants to be a martyr. The position of K in the g function is important. It is outside the power function $(\cdot)^{\alpha}$ and so affects utility in a way that is different than the way

money affects utility. This limits the applicability of what is called the complete insurance theorem.

Third, the notion of utility "if dead" is obviously odd: We aren't claiming this is utility *in period 2*. It is not. Once dead, our individual has no utility. However, in period 1, when he is planning, he can think about alternatives (e.g., should I buy a big insurance policy?) contingent on the dead state. The $g(\cdot)$ function is designed to allow rational analysis of such planning.

At this point we combine the utility if alive function $f(x)$ and the utility if dead function $g(x+V)$ into a von Neumann-Morgenstern expected utility function (recall the relevant section of Chapter 1). We assume that in the planning period, our individual maximizes expected utility, which we will call $u(x,y,z)$. Expected utility is the probability of being alive times utility if alive, plus the probability of being dead, times utility if dead. We now have

$$u(x,y,z) = q(y)x^{\alpha} + (1-q(y))[(x+V)^{\alpha} - K].$$

Substituting for V gives:

$$u(x,y,z) = q(y)x^{\alpha} + (1-q(y))\left[\left(x + \frac{z}{1-q(y)}\right)^{\alpha} - K\right].$$

Our rational planner chooses x, y and z to maximize this function. He is subject to the budget constraint $x+y+z=w$. Also, the variables x and y are constrained to be non-negative, but z need not be. (That is, a person could have a "negative" insurance policy that provides greater consumption if he is alive, but reduces his bequest if he is dead.)

Maximizing this function of 3 variables, subject to the budget constraint, is a somewhat complex exercise. We will not go through all the detailed steps. For interested readers, the procedure is as follows: First, take partial derivatives of $u(x,y,z)$ with respect to the 3 variables. This gives $\frac{\partial u}{\partial x}$, $\frac{\partial u}{\partial y}$ and $\frac{\partial u}{\partial z}$. Set these partial derivatives equal to each other, to get "first-order" conditions for the maximum. Use the first-order conditions, and the budget constraint, to characterize the solution. (Note that solving for specific values for x and y would require assuming a specific functional form for $q(y)$ which we will not do.)

As it turns out, given the assumptions listed above, the utility maximizing choices of z and V (i.e., the amount spent on life insurance, and the value of the policy) are both zero. So z and V drop out, and expected utility can therefore be rewritten as

$$u(x,y) = q(y)x^{\alpha} + (1-q(y))[x^{\alpha} - K].$$

It also turns out that the first-order conditions produce the following equation relating $q'(y)$, the derivative of the $q(y)$ function, x, and y:

$$q'(y) = \frac{\alpha}{K} x^{\alpha-1}.$$

Now recall the modern economic measure of willingness-to-pay value of life (or VOL for short), as described earlier. VOL is the highest c a person is willing to pay, per incremental reduction in his probability of dying ε. That is, VOL $= c/\varepsilon$. But the ratio c/ε has units (change in dollars)/(change in probability). Now note that for our model, the derivative $q'(y)$ has units (change in probability)/(change in dollars). Moreover when our individual is maximizing expected utility subject to a budget constraint, he is increasing y to the point where he is indifferent about that last dollar spent on precaution. That is, he is paying the maximum he would be willing to pay for that last reduction in the probability of dying.

In short, our model produces the modern economic measure of VOL, and it is simply $1/q'(y)$. Therefore, we have

$$\text{VOL} = \frac{K}{\alpha} x^{1-\alpha}.$$

For instance, if $\alpha = 1/2$ (square root utility), if $x = 25,000$ (the right order of magnitude for annual income of a wage-earner in the United States) and $K = 3162$, then VOL $= 1,000,000$ (the right order of magnitude for late $20th$ century empirical studies on VOL in the U.S.)

4. The Broome Paradox

John Broome (1978) attacks the standard economic approach to valuing life in the following way: Suppose the government plans a project that will put some lives at risk. (For instance, a long highway tunnel, or a large bridge, or a military action.) Suppose that, based on past experience, it anticipates 5 deaths from this project. Assuming the formula for VOL developed above, it figures a cost in terms of lives lost of 5 million dollars. Suppose further that, other than the lives lost, the project will cost 20 million, but will produce benefits valued at 50 million.

Cost-benefit analysis then indicates the net project benefit at $50-20-5 = +25$ million, and the project is worthwhile. Suppose the government proceeds with the project, and, as anticipated, 5 workers die. Assume they are Adam, Ben, Charles, Dave and Edward. Consider Adam. He was hired to work on the project, and may have been paid for the additional risk he incurred. However, the additional risk was small, as was his payment to accept it. But we know, ex post, that he did die. We might wonder about the following hypothetical question:

How much cash, say c, would he have required to make him indifferent between the live state and the dead state? Recall that in our model, where individuals are choosing not to buy insurance, the live state utility is x^α and the dead state utility is $x^\alpha - K$. Add c to the dead state x, and set the live state utility equal to the dead state utility. This gives:

$$x^\alpha = (x + c)^\alpha - K.$$

It follows that

$$c = (x^\alpha + K)^{1/\alpha} - x.$$

Using the numbers given above implies

$$c = (25,000^{1/2} + 3162)^2 - 25,000 = 11 \text{ million}.$$

If the cost benefit analysis were redone, knowing that each of the 5 victims would require 11 million to be compensated for his certain death, net project benefit, in millions, would be

$$50 - 20 - 11 \times 5 = -25.$$

So, if we knew which 5 individuals were to die, and if we were to compensate them for the certain losses of their lives, the net benefit of the project would change from +25 million to -25 million.

In short, there is a glaring inconsistency in the valuation of this project, between the ex-ante evaluation (5 statistical lives lost, +25 million net value) and the ex-post valuation (Adam, Ben, Charles, Dave and Edward dead, -25 million net value.)

5. Ex-Ante and Ex-Post

One response to the Broome paradox is to say that it does not matter: Often society must make life or death decisions before knowing *who* will live or die, and in those circumstances it has no choice but to use the ex-ante method. The ex-post valuation is interesting but irrelevant.

Another response is to ask whether or not there exist any approaches to valuing lives that give the same answers, ex-ante and ex-post.

To illustrate, let's consider an example with 100 identical people, all with the utility and VOL functions described above. That is, for utility,

$$\begin{aligned} u(x, y) &= q(y)x^\alpha + (1 - q(y))[x^\alpha - K] \\ &= x^\alpha - (1 - q(y))K. \end{aligned}$$

Assume as before that $x = 25,000$, and $K = 3162$, and so VOL = 1 million. We do not want to have to assume a particular $q(y)$ function.

Instead, let us simply assume that $q(y)$ starts at 0.99. That is, one person is expected to die.

Now assume that the government is deciding between the status quo, which we will call d_1, or, alternatively, a project that will put 15,000 dollars in the pockets of each of the one hundred original people, but will kill off an extra person. We call this alternative d_2.

The money needed to compensate any of our 100 people for this additional 1 percent risk of death is $0.01 \times \text{VOL} = 10,000$. Therefore the 15,000 more than compensates. That is, ex-ante, d_2 is superior to d_1, based on the dollar amounts. But if we knew *which* additional person was going to die as a result of society's choice of d_2, the necessary compensation for that person would be 11 million, which far exceeds the $15,000 \times 100 = 1.5$ million total project benefit. In short, ex-post d_2 is inferior to d_1.

That is, the money test of d_1 against d_2 is inconsistent. Hello again to the Broome paradox.

But now let's try a similar test with *utilities* rather than *dollars*. In other words, let us become utilitarians, for a change. We will now add together the (identical) von Neumann-Morgenstern utility function of all the people in this society.

Under d_1, ex ante, there are 100 people, all with utility levels

$$.99(25,000)^{1/2} + .01[25,000^{1/2} - 3162] = 126.49.$$

Multiply by 100 gives total utility for society of 12,649. Under d_1, ex-post, there are 99 living people and 1 dead person. The utility levels are

$$(25,000)^{1/2} = 158.1 \text{ for the living,}$$

and

$$(25,000)^{1/2} - 3162 = -3004.2 \text{ for the dead.}$$

Multiplying the first by 99 and the second by 1, and adding all together gives total utility for society of 12,649. That is, using summed utility as the social metric, d_1 is precisely as good ex-ante as ex-post.

Next, consider d_2. Ex-ante, we have 100 people, each with utility level of:

$$.98(25,000 + 15,000)^{1/2} + .02((25,000 + 15,000)^{1/2} - 3162) = 136.76.$$

Multiplying by 100 gives total utility for society of 13,676. Under d_2, ex-post, there are 98 living people and 2 dead. They have utility levels of:

$$(25,000 + 15,000)^{1/2} = 200.0 \text{ for the living, and}$$

$$(25,000 + 15,000)^{1/2} - 3162 = -2962 \text{ for the dead.}$$

Multiplying the first by 98 and the second by 2 gives a total utility for society of 13,676. That is, using summed utility as the metric, d_2 is precisely as good ex-ante as ex-post.

But this means d_1 and d_2 can be unambiguously compared. By the total utility test, d_2 is clearly superior to d_1. No Broome paradox is possible because there is no difference between the ex-ante and ex-post measures.

This example suggests one way to avoid Broome paradox inconsistencies: Use a "utilitarian metric" for measuring the effects of life-death choices, rather than a "money metric."

This possibility is mentioned by Alistair Ulph. The formalization is based on a theorem of Peter Hammond (1981):

Assume there are n people in society, denoted by $i = 1, 2, \ldots, n$. Society can make alternative decisions, $d = 1, \ldots, D$. There are several alternative states of the world, $s = 1, 2, \ldots, S$, whose probabilities depend on the social decision that is taken. We will assume that the decision taken only affects the probabilities of the various states; it does not affect an individual's ex-post utility in a particular state. Let the probability of state s when decision d is taken be $q_s(d)$.

Person i has state-contingent ex-post utility u_{is}. His von Neumann-Morgenstern expected utility function is

$$u_i(d) = \sum_s q_s(d) u_{is}.$$

If society measures welfare using a utilitarian approach, it calculates social welfare as a weighted sum of the $u_i(d)$ terms. (Taking a weighted sum allows counting person i's utility more or less heavily than person j's. If the weights are equal, then $u_i(d)$ and $u_j(d)$ are counted equally.) Let α_i be the weight attached to person i. Let $SW(d)$ represent social welfare. Then, ex ante,

$$SW(d) = \sum_{i=1}^{n} \alpha_i u_i(d) = \sum_{i=1}^{n} \alpha_i \sum_{s=1}^{S} q_s(d) u_{is}.$$

When social welfare is being measured ex-ante, we know which decision d society is taking, but we do not yet know which state of the world will occur.

Ex-post, we do know which state has occured. If state s has occured, ex-post social welfare is

$$\sum_{i=1}^{n} \alpha_i u_{is}.$$

But this state occurred with probability $q_s(d)$. To get an *average* ex-post measure of social welfare, we would have to take a weighted sum over all the states, using the state probabilities as weights. This gives an average ex-post measure:

$$SW(d) = \sum_{s=1}^{S} q_s(d) \sum_{i=1}^{n} \alpha_i u_{is}.$$

Since

$$\sum_{i=1}^{n} \alpha_i \sum_{s=1}^{S} q_s(d) u_{is} = \sum_{s=1}^{S} q_s(d) \sum_{i=1}^{n} \alpha_i u_{is},$$

the ex-ante utilitarian social welfare measure agrees with the ex-post utilitarian social welfare measure. And this is true for *any* ex-post utility functions.

Therefore the cure for Broome-type paradoxes in evaluating life and death choices may be to replace money metrics with utility metrics. But, then again, maybe not, as the next section will show.

6. Problems with Utilitarian Measures of Life/Death Choices

At this point we will drop the complications created by randomness. We assume society can choose among alternative policies $a, b, c, \ldots$. The choice may be made at a particular point in time, at which point there is a population of persons alive, whom we call, as usual, $\{1, 2, \ldots, n\}$. Such a point in time will be called time zero.

Any alternative choice by society, such as a, has implications for the utilities of everyone alive at time zero, and also for the utilities and *the existence* of persons who are not alive at time zero. (For example, the grandchildren and great grandchildren of those who are, at time zero, 10 years old.)

The *timeless population* for a is that set of all people who have been alive, who are alive, and who will be alive, under policy a. We denote that set of people N_a. Obviously it contains $\{1, 2, \ldots, n\}$ as well as many others. The number of people in N_a is called n_a.

Analysis of the choice between alternative a and b can be done in a way which focuses on the time dimension, or in a way that abstracts from it. Our discussion of ex-ante and ex-post, for example, focuses on the time dimension. We will now analyze the choice between alternatives a and b in a way which mainly abstracts from the time dimension. (We do not abstract from time entirely, since we recognize a time zero, and a given population as of that time.) This type of analysis, which deemphasizes the time dimension, is called timeless utilitarianism.

We will now look closely at some standard utilitarian principles, and critique them following the arguments of John Broome (1985).

(1) The total principle. Consider alternative a. For each person in the timeless population for a, measure the utility of his life. Call it $u_i(a)$. Sum these utility numbers for all such people.

This gives

$$\sum_{i \in N_a} u_i(a).$$

Consider an alternative choice b. Do the same. Then compare

$$\sum_{i \in N_a} u_i(a) \text{ and } \sum_{i \in N_b} u_i(b).$$

The total principle holds that the alternative with the higher total is the better alternative.

When writing utility of life numbers like $u_i(a)$, we will follow this standard convention: if a life is worth living, it has a *positive* utility; if it is not worth living, it has a *negative* utility. So $u_i(a) > 0$ means that, under alternative a, person i has a worthwhile life; $u_i(a) < 0$ means that, under alternative a, person i has such an awful life he would be better off dead, and $u_i(a) = 0$ means that, under alternative a, person i is alive, but with a "neutral" life, not so good as to be worth living, not so bad as to be worth dying. Further, if individual i does not *exist* in alternative a, we will use a dash "$-$" to represent his utility, if we need to represent it in an array of utility numbers.

There are two fatal objections to the total principle.

First is the so-called "repugnant conclusion" of Derek Parfit, discussed in his 1984 book *Reasons and Persons*. Consider an alternative a, with a small or moderate sized (timeless) population of persons, all with high levels of utility. Generally there will be an alternative b with a very large (timeless) population of persons with low (but positive) levels of utility, such that b is superior to a. That is, the total principle opts for a huge number of miserable persons (persons with utility barely above zero) over a moderate number of very happy persons.

Second, the total principle attaches the same weights to the person already dead at time 0, and to the person not-yet-born at time 0, as it does to the person alive here on earth. It places no special importance on the *continued* existence of a person alive at time 0. Many people would reject an ethical principal that treats a currently living person the same as a person who is not yet conceived.

(2) The person-restricted principle. This is meant to escape the "repugnant conclusion." We focus on those persons, and only those persons,

who exist under both a and b. That is, let I be the intersection of the sets N_a and N_b. When weighing alternative a against alternative b, compare

$$\sum_{i \in I} u_i(a) \text{ and } \sum_{i \in I} u_i(b).$$

The person-restricted principle opts for the alternative with the higher (person-restricted) total.

This obviously eliminates the possibility that an alternative a with a small number of very happy people will be found inferior to an alternative b with a huge number of miserable people.

However, like the total principle, it attaches the same weight to a currently living person as to a not-yet-born person (who would be born under both a and b).

Moreover, it creates anomalies like the following example, taken from Broome (1985):

Alternative a: An infant is born with severe disabilities. One million dollars might be spent to save his life, but is not spent, and he dies after 3.65 days. The 1 million is spent instead to save the life of a 20-year-old woman, who then lives happily to age 80. The infant's parents conceive another child after its death; the second baby lives happily for 80 years.

Alternative b: An infant is born with severe disabilities. One million dollars is spent to save his life, and he lives happily to age 80. The 20-year old women died at age 20. The infants's parents do not conceive another baby.

We focus on 3 people: the original baby, the replacement baby, and the young woman. We assume that everyone else's utility is the same under a and b. Let's measure utility by simply counting (assumed happy) years of life.

Under the total principle, the comparison is:

	$\underline{a}$	$\underline{b}$
Original baby	.01	80
Replacement baby	80	–
Woman	80	20
Totals:	160.01	100

So a is a better choice by the total principle.

Under the person-restricted principle, the replacement baby is stricken from the calculation, since he does not exists in both a and b. The result is:

	$\underline{a}$	$\underline{b}$
Original baby	.01	80
Woman	80	20
Totals:	80.01	100

So b is better by the person-restricted principle. But many people would reject this result. That is, the person-restricted principle seems wrong in this example.

Another objection to the person-restricted principle is this: Again, let I represent the intersection of the sets N_a and N_b. Suppose all the people in I are equally happy with either a or b. Then the person-restricted principle says a and b are equally good. But suppose the people in $N_a \setminus N_b$ (i.e., who exist under a but not under b) are all happy, whereas the people in $N_b \setminus N_a$ (i.e., who exist under b but not under a) are all miserable. Assume the numbers of people in $N_a \setminus N_b$ and $N_b \setminus N_a$ are the same. Then most people would opt for a over b, contrary to the person-restricted principle result.

(3) The average principle. One of the unpalatable things about utilitarianism is its strong bias towards population increase. An obvious way to eliminate that bias is to calculate average utilities, rather than total utilities. Again, consider a and b. Recall that n_a = the number of persons in the timeless population N_a, and n_b = the number of people in N_b.

When weighing alternative a against alternative b, compare

$$\frac{1}{n_a}\sum_{i\in N_a} u_i(a) \text{ and } \frac{1}{n_b}\sum_{i\in N_b} u_i(b).$$

Choose the alternative with the higher average utility level.

Applied to our original baby, replacement baby and woman example, average utilities under the 2 alternatives are:

$$160.01/3 = 53.34 \text{ for } a, \text{ and } 100/2 = 50.00 \text{ for } b.$$

Therefore a is better.

Although this is a logically neat principle, it has flaws. First, it violates a rather natural and appealing axiom called utility independence. (See Charles Blackorby, Walter Bossert and David Donaldson (2003) for a discussion of utility independence and related independence axioms.) *Utility independence* requires that if there are other people in the timeless populations under alternatives a and b, but their utilities are the same in both, then the comparison of a and b should not depend on the utility levels of those other people.

Let's now add one "other person" to modify the example, as follows:

	$\underline{a}$	$\underline{b}$
Original baby	.01	80
Replacement baby	80	–
Woman	80	20
"other person"	x	x
Totals:	$160.01 + x$	$100 + x$

Now the average utilities are:

$$\frac{160.01 + x}{4} = 40.00 + \frac{x}{4} \text{ for } a, \text{ and}$$
$$\frac{100 + x}{3} = 33.33 + \frac{x}{3} \text{ for } b.$$

It follows that average utility under a is greater than average utility under b if and only if $80.04 > x$. That is, if the "other person" has utility of, say, 60 in both a and b, a is better: but if he has utility of, say, 100 in both a and b, then b is better! This seems a glaring inconsistency.

But second, it gets even worse : Since these are timeless utilities, the other person may have lived, and died, 2000 years before our babies and woman! (Blackorby et al. call him Euclid, the inventor of geometry, who lived around the 3*rd* century B.C.)

In short, average utilitarianism makes the judgment between a and b depend on utility levels of people who are indifferent between a and b, and those people might have died years ago!

Third, under the average utility principle, it may be an improvement to get rid of relatively unhappy people. Kill the unhappy, or, more consistent with the timeless utilitarian approach, don't let them be born. This is an unpalatable ethical conclusion for some.

(4) The critical level principle. As an alternative to average utilitarianism, consider the following approach, developed in the 1980's by Charles Blackorby and others: To evaluate an alternative a, first figure $u_i(a)$ for all i in N_a, as before. Let $\bar{u}$ be some positive constant. When weighting a against b compare

$$\sum_{i \in N_a} (u_i(a) - \bar{u}) \text{ and } \sum_{i \in N_b} (u_i(b) - \bar{u}).$$

The summation of the terms $(u_i(a) - \bar{u})$ can be viewed intuitively as a calculation of utility *surplus* under a. Critical level utilitarianism says we should choose the alternative with the higher surplus.

Critical level utilitarianism does satisfy utility independence, unlike average utilitarianism.

However, like the average principle it can be criticized because it implies that the unhappy (those with $u_i(a) < \bar{u}$), should be killed, or at least prevented from being born. This is again an unpalatable ethical position for some.

The total principle, the average principle, and the critical level principle can be summarized and compared in the following way:

Let's use TU(a) as shorthand for total utility under a, i.e.,

$$\mathrm{TU}(a) = \sum_{i \in N_a} u_i(a).$$

Then the goal of total principle is to maximize TU(a). The goal of the average principle is to maximize $\frac{1}{n_a}\mathrm{TU}(a)$. The goal of the critical level principle is to maximize $\mathrm{TU}(a) - n_a\bar{u}$. Both the average principle and the critical level principle escape the "repugnant conclusion" of the total principle (of prescribing a huge population of miserably poor people). Average utilitarianism escapes it by looking at the average, rather than the total of the utilities. Critical level utilitarianism escapes it by requiring that a person being added to the population should be viewed as a "good," only if that person's utility exceeds the critical level.

7. The Pareto Principle and Extended Pareto Principles

In judging among alternatives, when the population is fixed and all persons are alive in all alternative states, we can construct a matrix of utilities to help see what is better than what, in terms of the Pareto criterion, the total utility principle, the average utility principle, and so on. We now list in Table 11.1 the people (by number) in the column headings; each row represents a vector of utility levels, one for each person, resulting from a choice by society.

Table 11.1.

	1	2	3	Total	Average
a	1	8	5	14	4.66
b	2	8	5	15	5
c	1	1	1	3	1

Thus b is Pareto superior to a, b is superior to a in terms of total utility and average utility, and c is inferior to a and b in terms of the Pareto criterion, and the utilitarian measures.

We can easily extend this matrix format to accommodate varying populations in a timeless utilitarian framework. We need only to expand the number of columns so as to list every person alive under any alternative. On the line for each alternative, we enter positive utilities for those persons who are alive in that alternative and have lives worth living. Recall that a negative entry would indicate that i is alive but has a life worth avoiding, and a zero would indicate i exists but has a neutral life. A

Table 11.2.

	1	2	3	4	5	...	1000	n_x	Total	Average
a	1	8	5	–	–	...	–	3	14	4.66
b	2	8	5	–	–	...	–	3	15	5.0
c	–	8	5	–	–	...	–	2	13	6.5
d	.02	.02	.02	.02	.02	...	.02	1000	20	.02

dash indicates that person i does not exist under the given alternative. As before, N_x is the population who exist under alternative x, and n_x is the number of persons in N_x. Now our matrix of alternatives may look like Table 11.2 above.

Note that we have changed c and added a d. As before, alternative b is superior to alternative a by Pareto, total utility, average utility, and so on.

Now a comparison of a and d in Table 11.2 shows the intuition of the Derek Parfit's "repugnant conclusion": Total utilitarianism rates d as better than a, even though the people alive in a have very high average utility (4.66) and those alive in d have very low average utility (.02). (Similar comparisons can be made between b and d, and between c and d.)

A comparison of a and c is very useful. Note first that there is not a "transition" from a to c that involves killing person 1. For person 1, alternative c shows a dash. So in c, person 1 has not been killed, rather he was never born.

Now according to some people, there is a list of souls in heaven, and each soul might come to be born on earth, or might not. (If a soul is never born, it would appear in our matrix as a dash.) If a soul is born, it is a good thing, according to some. If an additional person is born, with a positive lifetime utility, and no existing persons *are made worse off*, we will call the change an improvement by the *weak list of souls principle.* If an additional person is born, with a positive lifetime utility and no existing persons are *erased*, we will call the change an improvement by the *strong list of souls principle.*

By the strong list of souls principle, d is best in Table 11.2, and both a and b are superior to c. However, enlightened people generally reject the strong list of souls principle, first, because it leads directly to the "repugnant conclusion," and second, even worse, because it implies that 1000 happy people should be made miserable to allow one miserable person to be born.

The weak list of souls principle is another matter. In a choice between c and a, many would opt for a on the grounds that $c \rightarrow a$ is a "Pareto-like" move. Following Blackorby et al. (1984) and others, we will say the following: If a given alternative is modified by adding one person with a positive lifetime utility, while leaving the original population's utility levels unchanged, the change is a *Pareto population improvement.* If the given alternative is modified by adding one person with lifetime utility greater than or equal to the average utility of the original population, while leaving the original population's utility levels unchanged, the change is an *average-utility Pareto population improvement.* If the given alternative is modified by adding one person with lifetime utility greater than or equal to a constant $\bar{u} > 0$, the change is a *critical-level Pareto population improvement.* Note that all these Pareto population improvement criteria are improvements by the weak list of souls principle.

The move from c to a in Table 11.2 is a Pareto population improvement, and a critical-level Pareto population improvement for $\bar{u} \leq 1$. However it fails as an average-utility Pareto population improvement.

The next matrix reveals that both the Pareto population improvement criterion and the critical-level Pareto population improvement criterion are liable to Derek Parfit-like conundrums. For this example, let $\bar{u} = 2$.

Table 11.3.

	1	2	3	4	5	...	1000	n_x	Total	Average
a	1	8	5	–	–	...	–	3	14	4.66
c	–	8	5	–	–	...	–	2	15	6.5
e	.01	8	5	.01	.01	...	.01	1000	22.97	.023
f	2.01	8	5	2.01	2.01	...	2.01	1000	2017	2.02

In Table 11.3, we first compare c and e: By the Pareto population improvement criterion, e is superior to c. But it creates a crowd (997) of miserable people. We conclude from this example that the Pareto population improvement criteria may be unreliable. Now we compare c and f: By the critical-level Pareto population improvement criteria, f is superior to c. The point of this comparison is that this principle will not endorse a crowd of miserable people (people with utility barely above zero), but it will endorse a crowd of people with utilities marginally above the critical level.

8. What's Repugnant About the Repugnant Conclusion?

We should not leave the topic of utilitarian measures over differing populations without a few comments about real-world issues of population size and the resources of the earth.

When the first edition of this book was published, the world population (of humans) was around 4.5 billion. As this second edition is being prepared, it is around 6.5 billion. It is projected to be around 9.0 billion in the year 2050.

Many of the economic resources that are used by humans on earth are obviously physically limited. The land is limited, the seas are limited, fresh water is limited, easily-recoverable oil, natural gas, iron, coal, etc. are limited. It follows that the larger the population, the smaller is the amount of land, sea, fresh water, oil, natural gas, iron, coal, etc. etc., per person. It seems a truism that a great enough population, and small enough per-capita resources, will cause us to suffer someday.

Of course economists have been saying things like these since Thomas Malthus, and we are materially much better off now than we were when, in 1798, he wrote his *Essay on the Principle of Population.* Nonetheless, it is clear that at some level of the human population, human living standards must drop as population rises, because of the finite resources of the planet. (The belief of some that we can export our extra population to other planets or mine other planets for resources, remains a pipe dream.)

But even if we humans can be 9 billion, or 12 billion, or 15 billion strong without threatening our own living standards, there are the non-human living things on earth to consider. And, according to many experts, non-human life has suffered greatly from our growth.

According to biologists such as Edward O. Wilson, humans have precipitated a massive extinction of non-human species, which may become comparable to the great extinction catastrophes of geologic times. We probably hunted to extinction much of the megafauna of North and South America; we hunted to extinction the large flightless birds of New Zealand, we apparently hunted to extinction the megafauna of Madagascar and Australia. We almost hunted the American bison to extinction, and we hunted the passenger pigeon to extinction. Wilson indicates that one fifth of bird species have disappeared in the last 2000 years, because of human action.

But hunting is only one way humans have devastated other species. We have also damaged other species, in isolated environments, particularly islands, by introducing alien species like rats, pigs and snakes. And

most important, we have destroyed habitant. When we cut the forest for timber, for farming, or for cattle raising, we often have a massive impact on the plants and animals in those forests, particularly in tropical rain forests. We have already reduced rain forests to around half their prehistoric size, and we are currently cutting them at a rate of 1-2 percent per year. Wilson estimates that at current rates of deforestation, 10 to 25 percent of rain forest species of plants and animals will disappear in 30 years. There is of course a natural background extinction rate for species, but, according to Wilson, "human activity has increased extinction between 1,000 and 10,000 times [the background rate] in the rain forest by reduction in area alone. Clearly we are in the midst of one of the great extinction spasms of geological history." (Wilson, p. 280).

In our discussion of the principles of total utility and average utility above, we noted that the total utility measure might opt for a huge (human) population of miserable people over a small (human) population of happy people, which is the "repugnant conclusion."

For us, one underlying reason such a conclusion is repugnant is that the huge human population may have a huge negative effect on other living things.

In other words, the utility measures we have discussed only account for human utility, and that makes us wary of pushing them too far.

9. Conclusions About Life and Death Choices

In this chapter we have outlined a basic economic value of life model, in which an individual makes decisions about how much to spend to reduce his probability of dying. It is a rather simplistic 2-state uncertainty model, but it does allow some computations of value of life (VOL) numbers.

The economic model raises difficult philosophical questions, however. We have suggested that some of the philosophical objections to the money metric value of life model, particularly the inconsistency between ex-ante and ex-post evaluations, could be met by a utilitarian metric rather than a money metric. But the utilitarian analysis is also full of problems.

As we will see later in this book, there are very serious objections to *almost any* procedure for aggregating the preferences (or ordinal utilities) of a given set of n persons. If it is difficult or impossible to aggregate the preferences of a given population $\{1, 2, \ldots, n\}$, we should probably expect it to be difficult or impossible to aggregate the utilities of arbitrary populations of arbitrary compositions and arbitrary sizes. And so it is.

10. Exercises

1. Suppose an individual spends x on consumption and y on precaution. (There is no life insurance available.) Let the state contingent utility function be

$$\begin{cases} f(x) = x^{\alpha} \text{ if alive} \\ g(x) = x^{\alpha} - K \text{ if dead} \end{cases}$$

   Assume $q(y)$ = probability the individual is alive in period 2.

   Derive an equation for VOL.

2. Suppose the state contingent utility function is

$$\begin{cases} f(x) = \ln x \text{ if alive} \\ g(x) = \ln x - K \text{ if dead} \end{cases}$$

   Assume $q(y)$ = probability the individual is alive in period 2.

   Derive an equation for VOL.

3. Consider the example presented in the discussion of the person-restricted principle:

	$\underline{a}$	$\underline{b}$
Original baby	.01	80
Replacement baby	80	–
Woman	80	20
Totals:	160.01	100

   Indicate which is better, and why, by the total principle, the average principle, and the critical level principles with $\bar{u} = 20$.

   For what critical level $\bar{u}$ would a and b be equally good?

4. Show that critical level utilitarianism satisfies utility independence.

11. Selected References

(Items marked with an asterisk (*) are mathematically difficult.)

*1. C. Blackorby and D. Donaldson, "Social Criteria for Evaluating Population Change," *Journal of Public Economics*, V. 25, 1984, pp. 13-33.

   As the title suggests, this paper explores ways to evaluate population change. The authors show that a utilitarian social welfare function satisfying certain plausible assumptions must be a (generalized) critical level utility function.

2. C. Blackorby, W. Bossert and D. Donaldson, "The Axiomatic Approach to Population Ethics," *Politics, Philosophy and Economics*, V. 2, 2003, pp. 342-381.

   This is an excellent non-technical survey of the theory of population ethics. The authors lay out axioms of population change, provide clear examples, and review the major principles: classical utilitarianism, critical level utilitarianism, and average utilitarianism. Many variants of these principles are also covered. Blackorby et al. conclude that there are no principles that satisfy all of a set of reasonable axioms, and they discuss the trade offs among the axioms.

   The paper confines mathematical notation to an appendix.

3. J. Broome, "Trying to Value a Life," *Journal of Public Economics*, V. 9, 1978, pp. 91-100.

   Broome lays out his objection to the economic value of life measurement with clarity and elegance.

4. J. Broome, "The Economic Value of Life," *Economica*, V. 52, 1985, pp. 281-294.

   This is another clearly written piece by Broome, in which he describes and criticizes various utilitarian principles for weighing alternatives with life/death implications. It turns out that, according to Broome "none of the principles I considered seems acceptable." Broome opines that the usual practice for valuing life "has no sound basis," and admits that he has no sound alternative to offer in its place. Much of the "Problems with Utilitarian Measures..." section of this chapter comes from this Broome paper.

5. P. Dasgupta, *Human Well-Being and the Natural Environment*, Oxford University Press, 2001.

   This is a deep and complex book on economics and philosophy, touching on human well-being, population principles, sustainable development, and the natural environment. Dasgupta argues that states, particularly poor ones, have mismanaged the natural environment, and that economists have overestimated growth in human well-being by using measures that ignore environmental degradation.

   According to Dasgupta, wealth includes all of manufactured capital, human capital, knowledge, and natural capital, and sustainable development means each generation should bequeath to its successor generation as much capital as it inherited. Dasgupta also revisits

classical (total) utilitarianism and average utilitarianism. He categorically rejects the idea of weighing the well-being of a not-yet-conceived person the same as the well-being of a person now alive.

6. P.J. Hammond, "Ex-Ante and Ex-Post Welfare Optimality under Uncertainty," *Economica*, V. 48, 1981, pp. 235-250.

   Hammond proves that if a social welfare function has the property that ex-post efficiency implies ex-ante efficiency, then the function must be a weighted sum of von Neumann-Morgenstern expected utility functions.

7. M.W. Jones-Lee, "The Value of Changes in the Probability of Death or Injury," *Journal of Political Economy*, V. 82, 1974, pp. 835-847.

   Jones-Lee develops a simple and tractable expected utility model for risk of death.

8. E. J. Mishan, "Evaluation of Life and Limb: A Theoretical Approach," *Journal of Political Economy*, V. 79, 1971, pp. 687-705.

   Mishan surveys various methods for valuing loss of life or limb, and opts for compensating variation measures of changes in probabilities of death or injury. These methods should include external effects. Projects should be undertaken if and only if aggregate net benefits, including these measures, are positive.

9. A. Ulph, "The Role of Ex-Ante and Ex-Post Decision, in the Valuation of Life," *Journal of Public Economics*, V. 18, 1982, pp. 265-276.

   Ulph's paper follows in the wave of controversy created by Broome's paper. He notes the ex-ante/ex-post consistency of utilitarian measures of social welfare, but does not advocate utilitarianism. Instead he suggests that cost-benefit analysts "will need to think more carefully about how to capture both the ex-ante and ex-post distributional considerations ... ."

10. W. Kip Viscusi, "The Value of Risk to Life and Health," *Journal of Economic Literature*, V. 31, 1993, pp. 1912-1946.

    A useful survey of many empirical value of life studies. This summarizes labor-market studies on fatal and non-fatal injuries, as well as some non-labor market studies, e.g., on highway speed risks, cigarette smoking risks, and risks from fires. Viscusi also discusses some questionnaire-based studies.

11. E.O. Wilson, *The Diversity of Life*, W.W. Norton & Company, New York, 1992.

A beautifully written book by an eminent biologist on man's impact on non-human life on earth. Wilson has attempted in this and other books to develop an ethical outlook that incorporates what we know about human evolutionary origins. The ethical bottom line is "prudence." We should be aware of our origins as animals, and not destroy the biological world in which we were born.

Chapter 12

MAJORITY VOTING

1. Introduction

In Chapter 9 we saw that the social improvement criteria of Pareto, Kaldor, Scitovsky and Samuelson were all, in one way or another, unsatisfactory. We also saw that the usual economic welfare measure of consumers' surplus may produce nonsensical results. In Chapter 10 we saw that the criterion of fairness is in certain ways inconsistent with other utility-based criteria, and the Rawls criterion has several objectionable properties. Therefore, we are still faced with these questions: When is alternative x socially preferable to alternative y? How can the best social alternatives be chosen from among the good ones? An ancient and obvious mechanism for answering such questions is the mechanism of voting. This chapter examines the majority voting mechanism.

We assume throughout this chapter that person i votes for x over y if he prefers x to y, and abstains if he is indifferent. This type of honest voting precludes, for example, strategically voting for one's second choice over one's first when one knows that, given expected votes by other people, the first choice cannot win. Strategic voting complications will be introduced in succeeding chapters.

A few words are appropriate at this point about the social alternatives in this chapter and the chapters that follow. These alternatives might be allocations in an exchange economy, with or without externalities. Or they might be production plans, or production and consumption patterns in an economy with production. Or they might be levels of expenditure on a public good. Or they might be political candidates. In short, they might be just about anything that people might choose collectively. They are not, however, states viewed from behind a veil of

ignorance, as in Rawls, nor are they whole sets of possibilities, as in our abstract descriptions of Kaldor, Scitovsky and Samuelson. We assume here that people are quite clear about what will happen to them, as individuals, if x obtains rather than y. So it makes good sense to say something like "person i prefers x to y" or "$u_i(x) > u_i(y)$."

2. The Majority Voting Criterion

If majority rule is applied to a simple choice between two alternatives x and y, then x wins if it gets more votes than y, and they tie if they get the same number of votes. If x wins we say that x is superior to y according to the majority voting criterion, and if they tie we can say that x and y are equally good according to the majority voting criterion. We combine these two ideas with the following definition: x is as good as y according to the *majority voting criterion* if x gets at least as many votes as y. If x is as good as y according to the majority voting criterion, that is, if x beats or ties y, we write $x\mathcal{M}y$. Note that the relation $\mathcal{M}$, which allows social indifference, is analogous to an individual's R_i relation, rather than an individual's P_i relation.

The relation $\mathcal{M}$ has two cardinal virtues. First, unlike the Pareto criterion, it is always capable of judging between two alternatives. Formally, it is complete. For any two alternatives x and y, either $x\mathcal{M}y$ or $y\mathcal{M}x$. That is, for any x and y, either x beats y, or y beats x, or they tie.

Second, unlike the fairness criterion and the Rawls criterion, $\mathcal{M}$ is consistent with the Pareto criterion. If x is Pareto superior to y, then x must win a majority over y, which means $x\mathcal{M}y$ and not $y\mathcal{M}x$. For if x is Pareto superior to y, some people will vote for x over y (those for whom $u_i(x) > u_i(y)$) and the rest will abstain. No one will vote for y over x.

What are $\mathcal{M}$'s shortcomings? Its fundamental logical flaw, already revealed in Chapter 1, is its nontransitivity. Recall that in Chapter 1 we assumed an individual's preference relation R_i is complete and transitive. We make this assumption for two reasons: first, it is consistent with empirical research and with simple day-to-day observation; and second, if a person has complete and transitive preferences he can always make choices, according to Proposition 2 in Chapter 1. We saw in Proposition 3 in Chapter 1 that the transitivity assumption could be weakened to acyclicity without disastrous effect: if a person has complete and acyclic preferences he can always make choices. And we saw, in Proposition 5 in Chapter 1, that completeness and transitivity for a person's preferences implied the existence of a utility function for that person. In short, completeness and transitivity, or at least completeness and acyclicity,

make rational choice possible for an individual. If rational choice through the use of majority rule is to be possible for society, then $\mathcal{M}$ had better be complete and transitive, or at least complete and acyclic. We know $\mathcal{M}$ is complete. Is it transitive? Or is it at least acyclic?

Unfortunately, the answer to both questions is No. This has been known at least since the time of Marie Jean Antoine Nicolas Caritat, the Marquis de Condorcet, who provided examples of voting paradoxes in his *Essai sur l'Application de l'Analyse à la Probabilité des Decisions Rendues à la Pluralité des Voix*, in 1785. The following example of a Condorcet voting paradox, which was introduced in Chapter 1, involves three people, and three alternatives $\{x, y, z\}$. Suppose the preferences of the three are as follows:

$$u_1(x) > u_1(y) > u_1(z)$$
$$u_2(y) > u_2(z) > u_2(x)$$
$$u_3(z) > u_3(x) > u_3(y)$$

In a vote between x and y, persons 1 and 3 will vote for x over y, so x beats y. In a vote between y and z, persons 1 and 2 vote for y over z, so y beats z. At this stage, we note that if $\mathcal{M}$ were transitive, x would have to beat z. If $\mathcal{M}$ were merely acyclic, z would have to not beat x. However, in a vote between x and z, person 2 and 3 vote for z. That is, z beats x, which completes the cycle. Consequently, $\mathcal{M}$ is neither transitive nor acyclic.

What does this imply? Since x is socially preferred to y, according to majority rule, and y is socially preferred to z, and z is socially preferred to x, there is no best alternative among the three. Each one is inferior to one of the others. Also, there is no social utility function that represents these social preferences.

Some readers might think such cycling is unlikely, bizarre, artificial, implausible, and generally of no concern to real world folks. To show that it's not, we shall make the example a little more down-to-earth. Let the set of alternatives be distributions of \$100 among three people. Let x be the distribution (\$50, \$20, \$30); that is, under x, person 1 gets \$50, person 2 gets \$20, and person 3 gets \$30. Let y be (\$30, \$50, \$20); that is, under y, person 1 gets \$30, person 2 gets \$50, and person 3 gets \$20. Let z be (\$20, \$30, \$50); that is under z, person 1 gets \$20, person 2 gets \$30, and person 3 gets \$50. Then the preferences of the three individuals are

$$u_1(x) > u_1(y) > u_1(z)$$
$$u_2(y) > u_2(z) > u_2(x)$$
$$u_3(z) > u_3(x) > u_3(y)$$

These are the voting paradox preferences! But skeptics might now object that the particular distributions x, y, and z were improperly chosen. Surely there is some distribution that isn't defeated in a majority vote by some other distribution! For instance, what if we let x' =(\$33 1/3, \$33 1/3, \$33 1/3)? If we do, x' is beaten by y' =(\$50, \$50, 0). And this new y' is in turn beaten by some other distribution. In fact, every distribution of \$100 among three people is defeated in a majority vote by some other distribution of the \$100 among the three. Majority voting crumbles under this very simple distributional problem.

3. Majority Voting and Single-Peakedness

The thrust of the section above is that majority voting might cycle, or produce illogical results. It might make social choice impossible. Yet we feel that majority voting often works; that is, we feel that there are circumstances under which majority voting yields sensible transitive results. Are there then some conditions we can find which necessarily imply that voting is transitive? The answer is Yes.

There are circumstances under which majority voting is transitive, and under which it does allow for the choice of best alternatives, and does define a social utility function. Those circumstances hinge on a certain property of the traders' tastes which we call *single-peakedness*. To explain single-peakedness, it is easiest to start with the Condorcet voting paradox example, in which tastes are not single-peaked.

Suppose that x, y, and z are lined up on an axis of alternatives, with x on the left, y on the middle, and z on the right, as in Figure 12.1a. Then when person 1's utility function u_1 is plotted about the axis of alternatives, it has a peak at x, and drops monotonically to the right of that peak, because $u_1(x) > u_1(y) > u_1(z)$. Similarly, when person 2's utility function u_2 is plotted above the axis of alternatives, it has a peak at the middle alternative y, and it drops monotonically to the left and to the right of that peak, because $u_2(y) > u_2(z) > u_2(x)$. Finally, when person 3's utility function u_3 is plotted above the axis of alternatives, it has a low peak at x, it drops at y, and it has a high peak at z, because $u_3(z) > u_3(x) > u_3(y)$. In other words, u_3 has two peaks. The important thing about Figure 12.1a is not the spacing between the alternatives, nor the absolute heights of the dots, nor the slopes of the line segments. What is important here is that the u_3 function has two peaks, while the u_1 and u_2 functions have only one.

Given that we have placed the alternatives in the order x, y, z, person 1's utility function has one peak, person 2's utility function has one peak and person 3's utility function has 2 peaks. What if we were to change the ordering of alternatives along the horizontal axis? The reader can

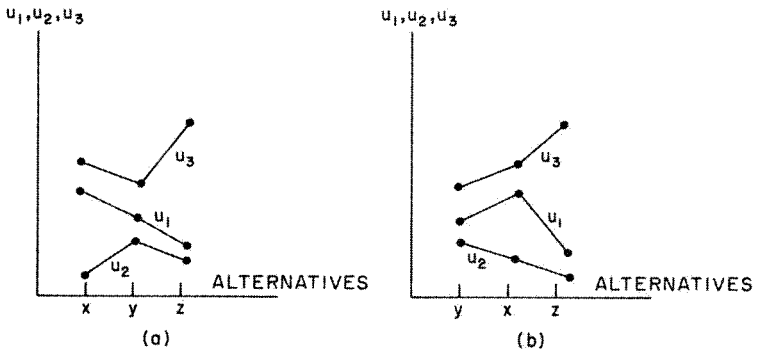


Figure 12.1.

verify that no matter how we permute the alternatives $\{x, y, z\}$, one trader's utility function will always show a double peak. There is no permutation of the alternatives under which all traders utility functions have a single peak. We therefore say that these preferences do not satisfy the single-peakedness property, which we define formally as follows:

Suppose that there exists an ordering of the alternatives along the horizontal axis, such that, given that ordering, the graph of each person's utility function has a peak at a single point; it rises monotonically to the left of the peak, providing the peak is not at the farthest-left point, and it drops monotonically to the right of the peak, providing the peak is not at the farthest-right point. The particular ordering of the alternatives for which this happens is called the *single-peakedness ordering.* If there does exist such a single-peakedness ordering, the preferences $R_1, R_2, \ldots, R_n$ or the utility functions $u_1, u_2, \ldots, u_n$ are said to satisfy the *single-peakedness property.*

To make the definition clear, let's take the voting paradox preferences and change u_2. Let the new u_2 function be one for which

$$u_2(y) > u_2(x) > u_2(z).$$

Now if we plot the old u_1, the new u_2, and the old u_3 functions in a picture like Figure 12.1a, that is, if we order the alternatives x, y, z and then plot the points, we still get a double peak in u_3, since by assumption person 3's utility function has not changed. However, if we reorder the alternatives, if we list them y, x, z from left to right, we get the picture in Figure 12.1b. With this single-peakedness ordering, every u_i function has one peak. Consequently this modified set of utility functions u_1, u_2, u_3 does satisfy the single-peakedness property.

Now consider what happens when majority voting is applied to the modified preferences. In a vote between x and y, person 1 and 3 vote for x, so x beats y, as before. In a vote between y and z, persons 1 and 2 vote for y, so y beats z, as before. However, in a vote between x and z, persons 1 and 2 vote for x, so x now beats z, and the voting cycle disappears!

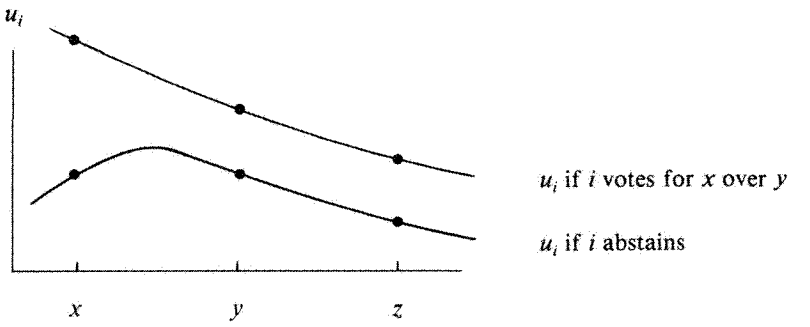
Forcing the preferences into the single-peaked mold has abolished the voting paradox, at least in this example. In fact, the example illustrates a general result, an important theorem that provides conditions under which the majority voting relation $\mathcal{M}$ is complete and transitive, just like an individual's preference relation. This theorem, along with the idea of single-peakedness, was developed by Duncan Black in the 1940's. We now turn to a formal statement and proof of one version of the theorem.

Black's Theorem 1. Suppose that the number of people n is odd. If the single-peakedness property is satisfied, the majority voting relation $\mathcal{M}$ is transitive. That is, for any three alternatives x, y, and z, if x beats or ties y, and y beats or ties z, x must beat or tie z.

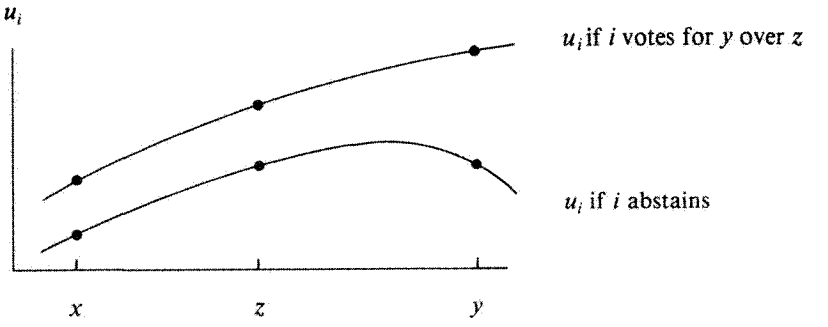
Proof. Take three alternatives x, y, and z. They will be placed, in some order, along the horizontal axis. To avoid trivial cases, assume that they are distinct. Suppose $x\mathcal{M}y$ and $y\mathcal{M}z$. We must show $x\mathcal{M}z$.

The proof will hinge on how x, y, and z appear in the single-peakedness ordering. Since there are six possible orderings of these three alternatives, there are six cases to consider. In each case we will use one of the following two observations, which hold for any n. (1) If alternative a beats or ties alternative b, then the number of people who vote for a or abstain in the a vs. b contest, is greater than or equal to $n/2$. (2) If alternative a beats b, then the number of people who vote for a or abstain in the a vs. b contest, is greater than $n/2$. (Actually if n is odd, these two cases reduce to one, but we state them here because they are useful in exercise 1.)

Case 1. The single-peakedness ordering is x, y, z. That is, x is to the left, y is in the middle, and z is to the right. Now if i votes for x over y, that is, if $u_i(x) > u_i(y)$, then $u_i(x) > u_i(z)$ must hold for u_i to have only one peak. (See the illustration following.) Similarly, if i abstains in the x vs. y vote, that is, if $u_i(x) = u_i(y)$, then $u_i(x) > u_i(z)$ must hold. Since $x\mathcal{M}y$, the number of people who vote for x or abstain in the x vs. y contest is greater than or equal to $n/2$. Since all these people must vote for x over z, $x\mathcal{M}z$.

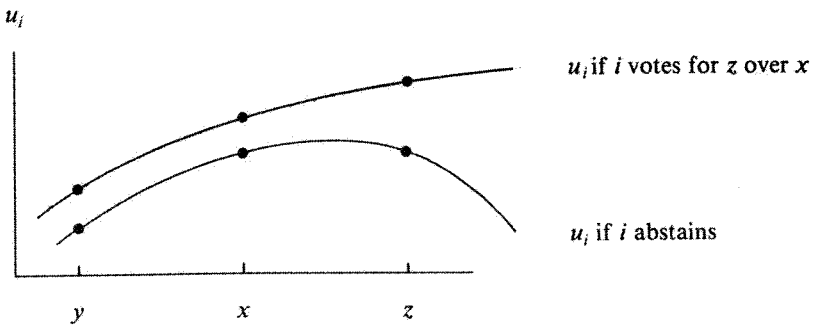


Case 2. The single-peakedness ordering is x, z, y. Now if i votes for y over z, that is, if $u_i(y) > u_i(z)$, then $u_i(y) > u_i(x)$ must hold, for u_i to have only one peak. (See the illustration below.) Similarly, if i abstains in the y vs. z vote, that is, if $u_i(y) = u_i(z)$, then $u_i(y) > u_i(x)$ must hold. Since $y\mathcal{M}z$, the number of people who vote for y or abstain in the y vs. z contest is greater than or equal to $n/2$. Since n is odd, this number is greater than $n/2$. Since all these people must vote for y over x, y must beat x. But this contradicts the assumption that $x\mathcal{M}y$. Therefore, the single-peakedness ordering cannot be x, z, y.



Case 3. The single-peakedness ordering is y, x, z. We want to show that $x\mathcal{M}z$. Suppose to the contrary that z beats x. Now if i votes for z over x, that is, if $u_i(z) > u_i(x)$, then $u_i(z) > u_i(y)$ must hold, for u_i to have only one peak. (See the illustration on the next page.) Similarly, if i abstains in the x vs. z vote, that is, if $u_i(x) = u_i(z)$ then $u_i(z) > u_i(y)$ must hold. Since z beats x, the number of people who vote for z or abstain in the x vs. z contest is greater than $n/2$. Since all these people must vote for z over y, z must defeat y. But

this contradicts the assumption that $y\mathcal{M}z$. Therefore, z cannot beat x, and so $x\mathcal{M}z$.



Cases 4,5, and 6. The single-peakedness orderings are y, z, x; z, x, y; and z, y, x, respectively. These cases are mirror images of cases 2,3, and 1, respectively, and the corresponding arguments apply. Q.E.D.

What happens to this theorem if the number of people n is even? What happens if there are say 1000 voters rather than 999? Then the theorem must be slightly modified.

To see the minor problem with single-peakedness when n is even, we consider a very simple example. Suppose there are two people and three alternatives x, y, and z. Suppose the utility functions are as follows:

$$u_1(z) > u_1(x) > u_1(y)$$
$$u_2(y) > u_2(z) > u_2(x)$$

You can check that these preferences satisfy the single-peakedness property; the single-peakedness ordering is x, z, y. Now what happens when votes are taken? In a vote between x and y, there is a tie. So x and y are socially indifferent. In a vote between y and z there is another tie. So y and z are socially indifferent. But in a vote between x and z, z defeats x. Consequently, $\mathcal{M}$ is not transitive; it is only quasi-transitive. This brings us to our second version of Black's Theorem.

Black's Theorem 2. If the single-peakedness property is satisfied, the majority voting relation $\mathcal{M}$ is quasi-transitive. That is, for any three alternatives x, y, and z, if x beats y and y beats z, then x beats z.

The proof of Black's Theorem 2 is very similar to the proof of the first version, and is left as an exercise for the reader.

Before turning to our next results, we need to define an important concept. Recall the idea of a "best set" or "choice set" from Chapter

1: an alternative x is in a person i's choice set if, for all alternatives y, xR_iy. For an alternative to be in i's choice set, he must like it as well as, or better than, any alternative.

There is a "best set" or "choice set" for the majority voting relation $\mathcal{M}$. For an alternative x to be in the choice set for $\mathcal{M}$, it must beat or tie any alternative. That is, the choice set for $\mathcal{M}$ is $\{x|x\mathcal{M}y \text{ for all } y\}$. An x that is in this choice set is called a *majority voting equilibrium.*

In short, a majority voting equilibrium beats or ties any alternative put against it in a majority vote.

Now we can apply the propositions of Chapter 1 to get two corollaries.

Corollary 1. Suppose the set of alternatives is finite, and the single-peakedness property is satisfied.

Then there exists at least one majority voting equilibrium.

We won't give a formal proof; the proof follows easily from Black's Theorem 2 above, and Propositions 1 and 3 in Chapter 1.

Corollary 2. Suppose the set of alternatives is finite, the number of people n is odd, and the single-peakedness property is satisfied.

1. Then there exists a unique majority voting equilibrium.
2. Moreover, there exists a social utility function U that exactly reflects the social preferences defined by majority voting. That is, for all alternatives x and y, $x\mathcal{M}y$ if and only if $U(x) \geq U(y)$.

We won't give a formal proof of this corollary either. Part 1 can be proved using Corollary 1 and assuming, contrary to what is to be proved, that there are two distinct alternatives, say x and y, each of which beats or ties all the alternatives. Using the fact that n is odd, it can be established that there is an alternative z that lies between x and y in the single-peakedness ordering, and that beats x and y. This gives a contradiction. The details are left as an exercise. Part 2 follows easily from Black's Theorem 1 above, and Proposition 5 in Chapter 1.

Let's pause to think about these results. Their basic thrust is that single-peakedness makes the majority voting relation $\mathcal{M}$ as sensible, as rational, as an individual's preference relation. If preferences are single-peaked, there are no paradoxes and no surprises inherent in majority voting; it is a perfectly logical way to make social choices. Majority voting becomes an acceptable vehicle for deciding when one alternative is socially preferred to another.

We have already seen, however, that the simple problem of distributing \$100 among three people gives rise to voting paradoxes. How useful

is single-peakedness? Is it an assumption that might actually be met? Or is it simply an empty theoretical construction? To get a feel for the answers to those questions, it's appropriate to examine a few more examples.

One example has already been introduced in Chapter 8. Suppose a government is deciding on the output of a single public good. Suppose also that person i's utility function is

$$u_i = v_i(x) - y_i$$

where v_i is the type of function illustrated in Figure 8.1, x is the quantity of the public good, and y_i is i's consumption of the private good. Suppose finally that i has a fixed tax share t_i; that is, his tax bill is $t_i x$. Then i's preferences for various levels of x are given by his utility-from-the-public-good function net of taxes, or $v_i(x) - t_i x$. If this function is plotted against x, it rises monotonically to a maximum, and then declines. And this is true for every i, although different people will generally have different maxima, and functions of slightly different shapes. Therefore, preferences for the provision of one public good, financed by fixed tax shares, satisfy the single-peakedness property. Note that the single-peakedness ordering is the natural ordering — x goes from 0 at the left to $+\infty$ at the right.

Black's Theorems indicate that there will always exist a level of expenditures x that beats or ties all other proposed levels of expenditure. If levels of expenditure are subject to votes, there will be no voting cycles, no voting paradoxes. And therefore majority voting will provide a logically acceptable way to make decisions about the provision of the public good. (However we should note the winning level of expenditure might not satisfy the Samuelson optimality condition. See Chapter 8.)

Another common single-peakedness example comes from political theory. We occasionally read about a left-right political spectrum, along which each of us has a different favorite position, and such that when we move away from our favorite position each of us is more and more unhappy:

Radical-Liberal-Moderate-Conservative-Reactionary.

The moderates love the moderate position, dislike liberal and conservative positions, and dislike radical and reactionary positions even more. The reactionaries love the reactionary position, are lukewarm about the conservative position, indifferent about the moderate position, and so on. If there is such a spectrum, then our preferences regarding political positions satisfy the single-peakedness property. Consequently, majority voting among referenda or candidates on this spectrum will be transi-

tive, or at least quasi-transitive; there will be no voting cycles and no voting paradoxes. And there will be a position that beats or ties any other position. Clearly that winning position will be somewhere near the middle of the spectrum, and this is why, according to some political theorists, candidates in U.S. elections tend toward centrist positions. Single-peakedness has another application here.

The conclusion of this section must be that single-peakedness makes some sense, that there are interesting cases in which preferences are indeed single-peaked, and in which, by Black's Theorems, majority voting works well. But the discussion of majority voting and single-peakedness cannot stop here. In the 1960s and 1970s several economists and political scientists, particularly Charles Plott and Gerald Kramer, explored the problem of single-peakedness and voting over multidimensional sets of alternatives. They added new, disturbing results, results that we shall discuss in the next section.

4. Majority Voting and Single-Peakedness: The Multidimensional Case

In Black's Theorems, the ordering of the alternatives along a single dimension plays a crucial role. A left-right political spectrum is one such unidimensional ordering. The amount spent by a town on maintaining a particular road is a unidimensional set of alternatives that is likely associated with single-peaked preferences for the town's citizens. Each person is likely to have a most preferred expenditure level, and is likely to be less and less happy with expenditure levels further and further from his most preferred level. Similarly, the amount spent by a state on public schools is a unidimensional set of alternatives likely associated with single-peaked preferences, as is the amount spent by a nation on defense. In all these single-peaked cases, Black's Theorems indicate that majority voting should give rise to sensible — i.e., transitive or quasi-transitive — results.

However, we often make choices in multidimensional sets of alternatives. For example, political candidates might have to adopt positions with regard to several unconnected issues — like gun control, abortion, unemployment, inflation, and crime — rather than to adopt one position on the left-right spectrum. A town meeting might decide simultaneously on expenditures for road maintenance and expenditures for schools — in which case the set of alternatives is most naturally two-dimensional. Or, finally, policy makers in the national government might choose among multidimensional policy alternatives in which a single alternative may be characterized by a level of unemployment, an interest rate, a rate of inflation, and an exchange rate between the yen and the dollar.

The possibility of two or more issues being decided simultaneously gives rise to new questions: Are the single-peakedness results applicable when there are multidimensional sets of alternatives? If alternatives are two-dimensional (or of higher dimensionality), and if preferences of voters are similar, in some sense, to one-dimensional single-peaked preferences, is majority voting transitive? Or do Black's Theorems collapse when the dimensionality of the set of alternatives goes from one to two or more?

The first step in answering these questions is to provide a multidimensional analog for single-peakedness. To keep the exposition simple we shall assume that social alternatives are two-dimensional vectors. Thus in this section an alternative $x = (x_1, x_2)$ specifies the levels of two variables. For example, x_1 might be the dollar amount spent on schools, a number that can range between 0 and ∞, while x_2 might be the dollar amount spent on police. Graphically, the set of alternatives is represented by the first quadrant of a plane, rather than by a line. Let us say that preferences are *regular* if each person has a single most preferred point on the plane, and if each person's indifference curves are roughly circular around his most preferred point. More precisely, indifference curves must be lines (they cannot be "fat"), they must have no flat segments or kinks, and they must enclose convex sets. The assumption of a most preferred point for each person is especially natural in the analysis of the provision of public services. No one wants infinite expenditures on schools or police, since those expenditures are financed through taxes.

Figure 12.2 illustrates regular preferences for three people. The set of alternatives is the set of points on the plane.

In the figure the point a represents the favorite alternative of person 1. Person 1 likes the levels of x_1 and x_2 given by the coordinates of a best. His indifference curves encircle a. This means, for example, that 1 gets the same level of utility from alternative x as he does from alternative y. Point b is the favorite of person 2, and his indifference curves are the ones encircling it; and c is the favorite of person 3, while his indifference curves are the ones encircling it.

This preference configuration provides the two-dimensional analog of single-peaked preferences: Each individual has a preference peak, and each individual's utility declines monotonically as one moves (in a straight line) from his preference peak. Moreover, if one draws a straight line (like L) anywhere through the alternative set, and looks at individual's utility levels on that line, those utility levels are single-peaked in the Black sense. Figure 12.3 illustrates the utility levels for 1, 2 and 3 *along* L. This figure is constructed by referring to Figure 12.2 and reading off the utility levels of the three individuals along L.

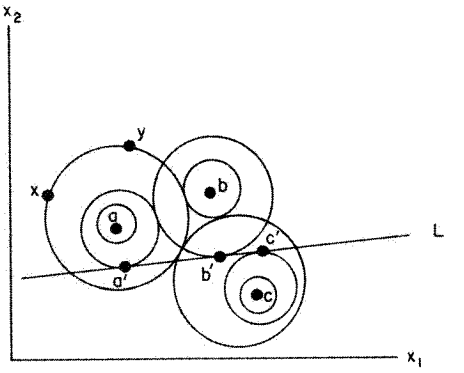


Figure 12.2.

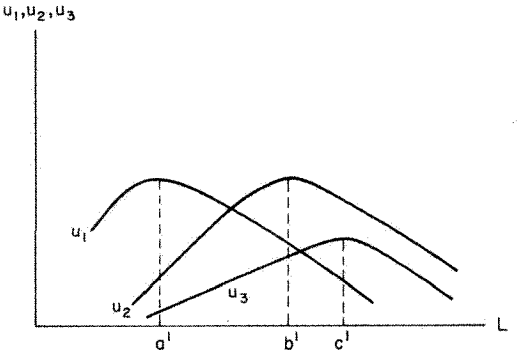


Figure 12.3.

Note that a' is the alternative on L that person 1 likes best, and that 1's utility declines monotonically as one moves away from a' along L; similarly b' is the alternative on L that person 2 likes best, and c' is the alternative on L that 3 likes best.

Now we consider how majority voting works on a two-dimensional alternative set. The first thing to observe is that if the only alternatives considered lie on a line like L, the single-peakedness theorem holds. In this case, majority voting is transitive, or at worst, if n were even, quasi-transitive. We know by Black's Theorems that there must be an alternative that beats or ties all the rest. In the particular example of Figure 12.3, majority voting leads to the most preferred alternative of the middle, or median voter, namely person 2. That is, majority voting

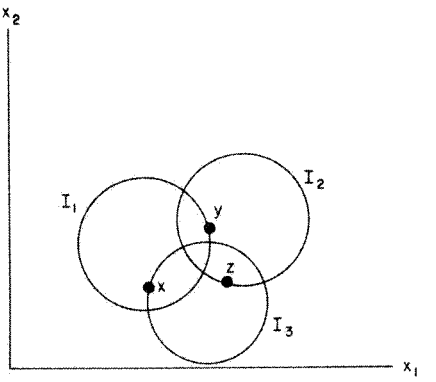


Figure 12.4.

on L leads to b'. But constraining the voting to a line like L forces this two-dimensional problem into a one-dimensional mold. What if all alternatives are permitted on the agenda? That is, what if votes can be taken on any pair of points on the plane?

In Figure 12.4 we show three indifference curves for persons 1, 2 and 3, labeled I_1, I_2, I_3 respectively, and three alternatives x, y, and z. Note that each voter prefers any alternative on his indifference curve to any outside it, and prefers any alternative inside his indifference curve to any on it. Thus person 1 prefers x to y to z; person 2 prefers y to z to x, and person 3 prefers z to x to y.

In other words, the utility levels of the three individuals for the three alternatives are as follows:

$$u_1(x) > u_1(y) > u_1(z)$$
$$u_2(y) > u_2(z) > u_2(x)$$
$$u_3(z) > u_3(x) > u_3(y)$$

But these are precisely the utilities that give a voting paradox! In votes between pairs of alternatives, x beats y, y beats z, and z beats x, completing the cycle.

You can check to see that, in this three-person, two-dimensional case, every alternative, including every Pareto optimal one, is defeated by some other alternative in a majority vote. Majority voting is riddled with cycles, and leads to no socially best alternatives.

Nor is this a contrived example. In general, for a multidimensional alternative set, even if people's preferences are regular, unconstrained

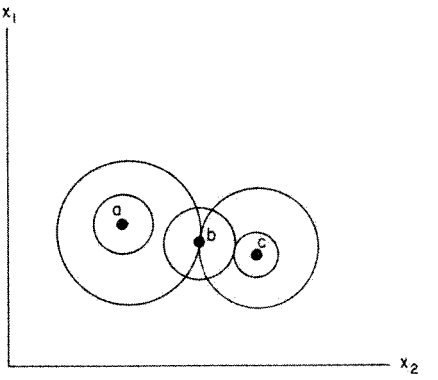


Figure 12.5.

majority voting will be cyclic. It will almost always fail as a guide for choosing best alternatives.

There is, however, one special circumstance under which majority voting over a multidimensional set of alternatives has an equilibrium; that is, an alternative that defeats all other alternatives. Majority voting in this circumstance can indicate a socially best choice, namely the one that beats or ties all the others. (There may, however, still be cycles among inferior alternatives. This is not ruled out by Proposition 3 in Chapter 1.)

The special circumstance is illustrated in Figure 12.5:

In this figure, a is again 1's favorite, b is 2's favorite, and c is 3's favorite. What is crucial here is that person 2's favorite is on the locus of tangencies of 1's and 3's indifference curves. In this case, alternative b wins a majority over any other alternative.

The peculiarity of this case is striking, however, because the a priori likelihood that person 2's favorite is on the thin line of tangencies connecting points a and c must be zero! Moreover, this general configuration is the only regular preferences configuration that gives rise to the existence of a majority voting equilibrium, in the three person two dimensional case. A voting equilibrium exists if and only if one person's favorite alternative lies on the locus of tangencies connecting the favorites of the other two people. This is surely a rarity.

We must conclude that the connection between single-peakedness and transitivity or quasi-transitivity for majority voting depends crucially on there being a unidimensional set of alternatives. If the alternatives can be lined up, and utility functions have single peaks given that line

up, majority voting is logically acceptable. If, however, the alternatives lie on a plane, and utility functions have single peaks over the plane, then majority voting will almost surely be logically unacceptable. It will almost surely be riddled with cycles. And there will almost surely be no alternative x such that $x\mathcal{M}y$ for all y. It will almost surely be the case that every alternative is, paradoxically, inferior to some other alternative.

5. Multidimensional Voting – Are the Cycles That Bad?

It is possible that voting cycles in multidimensional sets are not as bad as the preceeding section suggests. Plott's result, that a majority voting equilibrium exists only if (in the 3 person case) 1 person's favorite point lies on the tangency locus for the other 2 people, is certainly very restrictive.

But it is possible that, in the general case where there is no majority voting equilibria, voting will nonetheless tend to lead into some more-or-less stable solution set.

Let us be more specific with an example. Suppose we have two dimensions, as before, and 3 individuals. Again, we will let a = 1's favorite (x_1, x_2) combination, b = 2's favorite, and c = 3's favorite. We will again assume circular indifference curves around the favorite points. For the sake of simplicity in Figure 12.6 below, in fact, we will assume the indifference curves are exactly concentric circles around favorite points. (This means that when evaluating 2 alternatives, say x and y, person i prefers the one that is *a shorter distance from his favorite point.*)

In the figure we have drawn straight lines connecting the favorite points a, b, and c. These lines are in fact tangency points of indifference curves of pairs of individuals; for instance the line connecting a and b is comprised of tangencies of indifference curves of persons 1 and 2. (The tangency loci are straight lines because we are assuming here that the indifference curves are concentric circles around the favorite points.)

Now a quick examination of Figure 12.6 should convince the reader that every point outside of the a, b, c triangle (such as x), is not Pareto optimal. For instance, x is Pareto dominated by y; that is, the move from x to y would make person 1 (whose favorite is a) better off, while leaving persons 2 and 3 as well off. And another moment of thought should convince the reader that points that are on the boundary of or inside the a, b, c triangle (such as y) *are* Pareto optimal. At a point like y, for instance, any shift must make someone worse off.

A casual examination of Figure 12.6 leads to the following speculation: From points outside the Pareto optimal triangle (like x), it seems that

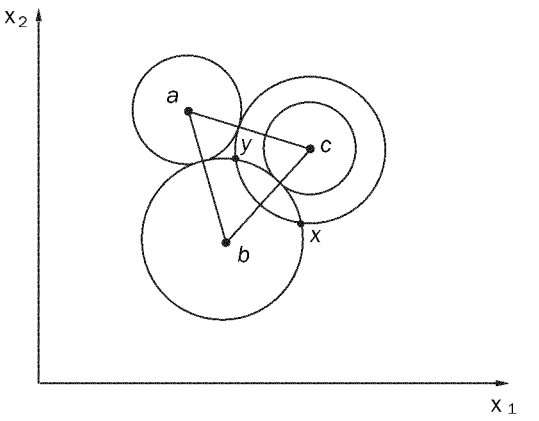


Figure 12.6.

majority voting would quickly and easily move society into the Pareto optimal triangle. From points inside, it seems that majority voting would not be likely to lead outside.

So, although there may be no majority voting equilibrium, strictly speaking, and although there might be voting cycles in this 2-dimensional model, perhaps voting tends to move into the triangle, and perhaps it tends to stay there. Call this the comforting conjecture on multidimensional voting.

Unfortunately, the comforting conjecture has the following discomforting rejoinder, proved by Richard McKelvey (1976):

> *McKelvey's Theorem.* Suppose there are 3 or more voters voting on alternatives in an issue space of 2 or more dimensions. Suppose their indifference curves are concentric circles (or spheres) around their favorite alternatives. Suppose there is no majority voting equilibrium.
>
> Let x and y be any two alternatives. Then a sequence of points $x_1, \ldots, x_k$ can be found such that: x_1 beats x in a majority vote, x_2 beats $x_1, \ldots, x_k$ beats x_{k-1}, and y beats x_k.
>
> That is, by a sequence of majority votes, we can move from x to y. By the same token, by a (different) sequence of votes, we can move from y back to x.
>
> In other words, all pairs of alternatives are connected in voting cycles.

We dispense with a rigorous proof. However the interested reader can easily convince himself that the theorem must be true, with the following

exercise: Make a big copy of Figure 12.6. It's pretty easy to start outside the set of Pareto optima and construct a sequence of votes to get in: in fact, it can always be done in one step. For instance, for x and y as illustrated in Figure 12.6; start at the non-optimal x and move to the optimal y in just 1 vote: Persons 2 and 3 abstain, and person 1 votes for y over x.

The more challenging thing is to start inside the set of Pareto optima and construct a sequence of votes to get out. In Figure 12.6, for example, we can start at the Pareto optimal y and construct intermediate points x_1, x_2, etc. to get to the non-optimal x. The trick at each stage of this process is to get a majority to support a proposed point over the current point. For instance, starting with y as the current point, move *beyond* the a/c straight line segment to an x_1 point that wins the votes of persons 1 and 3 over y. Given our assumption of concentric circle indifference curves, this means x_1 must be closer to the favorite points a and c than y is. Then, with x_1 as the current point, move to an x_2 that is beyond the a/b straight line segment.

Continue in this fashion, with bigger and bigger swings from the current to the proposed point, until you can swing to x!

This process of moves from x to x_1 to x_2, etc. can evidently require large jumps in a figure like Figure 12.6, and those jumps might be large enough to require *negative* coordinates. In fact, McKelvey's Theorem takes the choice space as the entire m-dimensional space, rather than the m-dimensional non-negative orthant. So this creates a possible restriction on the scope of McKelvey's theorem: if negative quantities are not possible, the McKelvey cycling possibility may be restricted. Another possible restriction has to do with the assumption of concentric circular indifference curves, but a few years after his first paper McKelvey (1979) provided a generalization of his theorem that dispenses with this restrictive assumption.

6. Exercises

1. Prove Black's Theorem 2.

2. Prove Corollary 2 to Black's Theorems.

3. Suppose there are n voters, where n is an odd number, and a set of distinct alternatives $\{x_1, x_2, \ldots, x_n\}$. Suppose the voters' preferences are single-peaked, and the single-peakedness ordering is $x_1, x_2, \ldots, x_n$. Assume that x_1 is person 1's favorite alternative; x_2 is person 2's, and so on down through n. Let $m = (n+1)/2$, so m is the median of the n numbers, person m is the median voter, and x_m is the favorite of the median voter.

Show that x_m will win a majority over any other alternative. (Hint: Draw a picture of the utility functions over $x_1, x_2, \ldots, x_n$. Then figure out who will vote for x_m and who will vote for x_i, for an arbitrarily chosen x_i other than x_m.)

This result is essentially Black's original version of the single-peakedness theorem.

4. Show with a graph why McKelvey's argument fails if variables are constrained to be nonnegative and if the a, b, c triangle is located close to the origin.

7. Selected References

(Items marked with an asterisk (*) are mathematically difficult.)

1. D. Black, "On the Rationale of Group Decision Making," *Journal of Political Economy*, V. 56, 1948, pp. 23-34.

   In this straightforward article Black explains the idea of single-peakedness, and shows that one motion exists that wins a simple majority over every other motion.

2. D. Black, *The Theory of Committees and Elections*, Cambridge University Press, London, 1958.

   This excellent book covers lots of topics in the theory of elections. We name a few of the topics in Part I. Chapter IV discusses single-peakedness. Chapters V, VI and VII cover non-single-peaked preferences, and cycles in majority voting. Chapter X covers some election procedures other than majority voting. Part II of the book provides a history of analysis of voting, with excellent treatments of the contributions of de Borda, Condorcet, C.L. Dodgson (Lewis Carroll) as well as others. This part is essential for anyone seriously interested in the theory of elections.

*3. P. Dasgupta and E. Maskin, "On the Robustness of Majority Rule and Unanimity Rule," Economics Working Paper 36, Institute for Advanced Study, Princeton, 2003.

   Assuming a continuum of voters, Dasgupta and Maskin show that the failures of majority voting to satisfy transitivity, i.e. the Condorcet cycles, are "rare." For most preference profiles, majority rule is transitive. This property is termed "generic transitivity." They also show that if, over a given domain of preferences, a voting rule satisfies Pareto consistency, anonymity (equal treatment of voters), neutrality (equal treatment of alternatives) and generic transitivity,

so does majority voting over the same domain. In this sense, majority is "at least as good as" or "better than" any other voting rule.

4. O. Davis, M.J. Hinich, P.C. Ordeshook, "An Expository Development of a Mathematical Model of the Electoral Process," *American Political Science Review*, V. 64, 1970, pp. 426-448.

   Davis, Hinich and Ordeshook develop a model in which candidates in an election adopt positions in a multidimensional issues space. Each voter has preferences over the issues space, and a most preferred alternative; that is, a favorite position on all the issues. These preferences for individual i give rise to a *loss function* that essentially measures how much individual i dislikes candidate j. The loss is zero if candidate j's position is identical to voter i's most preferred alternative, and the loss increases as the difference between the voter's most preferred alternative and candidate j's position increases. The authors then go on to characterize some conditions under which there will exist an equilibrium position; i.e., a position that, if adopted by candidate j, will guarantee j's beating or tying any other candidate in a majority vote. It turns out that the existence of such a position for every candidate is unlikely.

5. P. Fishburn, "Paradoxes of Voting," *American Political Science Review*, V. 67, 1974, pp. 537-546.

   This interesting article describes several voting paradoxes other than Condorcet's. These are paradoxes that arise for voting procedures that are different from straightforward majority rule, particularly sequential-elimination simple-majority voting, the de Borda procedure, and truncated variants of the de Borda procedure. Fishburn also presents data from computer simulations to suggest how likely (or unlikely) the paradoxes might be.

*6. G.H. Kramer, "On a class of Equilibrium Conditions for Majority Rule," *Econometrica*, V. 41, 1973, pp. 285-297.

   Kramer examines voting over multidimensional alternatives, in a model where people have utility functions that satisfy standard properties of convexity and smoothness. He examines many "exclusionary" properties — properties which, like single-peakedness, rule out or exclude certain configurations of preferences. These properties were developed to avoid the quirks of majority rule. Kramer shows that in a standard utility-function model these exclusionary properties are extremely restrictive, "in the sense that they are incompatible with even a very modest degree of heterogeneity of tastes." Voting cy-

cles are, therefore, the norm rather than the exception in "problems involving voting over multidimensional choice spaces."

7. K. May, "A Set of Independent Necessary and Sufficient Conditions in Simple Majority Decisions," *Econometrica*, V. 20, 1952, pp. 680-684.

   In this early paper, May analyzes "group decision functions" limited to 2 alternatives x and y. He shows that if a group decision function satisfies 4 reasonable conditions ((i) that it is defined and single-valued; (ii) that it treats different individuals the same, (iii) that it is neutral between x and y, and (iv) that it satisfies a mild monotonicity assumption) then it must be the majority voting rule. He also shows conversely, that majority voting satisfies conditions (i) through (iv).

*8. R. McKelvey, "Intransitivities in Multidimensional Voting Models and Some Implications for Agenda Control," *Journal of Economic Theory*, V. 12, 1976, pp. 472-482.

*9. R. McKelvey, "Generalized Conditions for Global Intransitivities in Formal Voting Models," *Econometrica*, V. 47, 1979, pp. 1085-1112.

*10. C.R. Plott, "A Notion of Equilibrium and its Possibility Under Majority Rule," *American Economic Review*, V. 57, 1967, pp. 787-806.

    In a model with a multidimensional alternative set, Plott characterizes a voting equilibrium, that is, an alternative that beats or ties any other in a majority vote. In the case where the number of people n is odd, an equilibrium point must satisfy these conditions: (1) It must be a favorite alternative for at least one person. (2) "All the individuals for which the point is not a maximum can be divided into pairs whose interests are diametrically opposed." Since it is unlikely that an arbitrary set of preferences will satisfy these conditions, it is unlikely that an equilibrium will exist. This article by Plott is the basis for our discussion surrounding Figure 12.5.

11. A.K. Sen, *Collective Choice and Social Welfare*, Holden-Day, Inc., San Francisco, 1970, Chapter 10, 10*.

    Chapter 10, on majority rule, provides a nice discussion of literature on majority voting. One branch of this literature is about the likelihood of voting cycles — in an "average" "randomly selected committee," is a voting cycle a real possibility? The second branch is about extensions of the approach of Black. These extensions impose conditions (like single-peakedness) on preference relations, conditions which iron out some of the quirks in majority voting.

Chapter 10*, which is rather mathematical, has a set of theorems on majority rule under various preference relation similarity conditions.

Chapter 13

ARROW'S IMPOSSIBILITY THEOREM

1. Introduction

We have now looked at several ways to answer the question "When is one alternative socially preferred to another?" Each of the answers has been somehow disappointing. The Pareto criterion is incomplete; the Kaldor criterion is possibly inconsistent. The Scitovsky criterion reduces to a question of optimality vs. non-optimality. The Samuelson criterion might be completely devoid of content. The fairness and Rawls criteria might be inconsistent with the Pareto criterion. Majority voting might generate cycles, and the single-peakedness condition, which forces transitivity, is quite restrictive, and almost useless when the issue space has two or more dimensions.

Our goal throughout has been to discover an unerring rule for generating rational social preferences, rational in the sense that an individual's preferences are rational. That is, we have been looking for a rule that generates complete and transitive social preferences, or, at least, complete and acyclic social preferences. But our series of disappointments raises some questions: Does a foolproof method exist for constructing complete and transitive social preference relations? Does a foolproof method exist for constructing complete and acyclic social preference relations? Does a foolproof method exist for finding best social alternatives? In this chapter we address the first question, and we briefly discuss the answer to the second question. We will provide partial answers to the third question in Chapters 14-16.

Does a foolproof method exist for constructing a complete and transitive social preference relation? The answer to this question clearly depends on what we mean by foolproof. We will impose formal require-

ments on the method for constructing social preferences, requirements that allow a definite answer. The list of requirements, and the answer to the question, were developed by Kenneth Arrow around 1950, and the answer, to which we shall soon turn, is call Arrow's Impossibility Theorem.

2. A Simple Model for Arrow's Impossibility Theorem

We now assume for the sake of simplicity that there are only two individuals, and three social alternatives x, y, and z. Also, we suppose that no individual is ever indifferent between any two alternatives. As usual, xP_iy means i prefers x to y. Individual i's preference ordering is assumed to be complete and transitive. Since there are only three alternatives and indifference is disallowed, there are only six ways individual 1 can order the alternatives. He can prefer x to y to z, or he can prefer x to z to y, or he can prefer y to x to z, and so on. The same is true of individual 2. Therefore, if any preference ordering for 1 or 2 is allowed, there are exactly $6 \times 6 = 36$ different constellations of individual preferences, or *preference profiles*, possible in this small society. Table 13.1 includes them all.

Each cell in this table shows a possible pair of rankings of the three alternatives by individuals 1 and 2. On the left side the alternatives are ordered, from top to bottom, according to person 1's preferences, and on the right side they are ordered according to 2's preferences. For example, the first row, second column cell has 1 preferring x to y to z, and 2 preferring x to z to y.

Our concern here is whether or not there is a foolproof rule to transform any cell in Table 13.1 into a social preference relation. Such a rule is called an *Arrow social welfare function.* An Arrow social welfare function takes preference profiles and produces social preferences.

Let R stand for a social preference relation, so xRy means x is socially at least as good as y. P is the corresponding strict social preference relation: xPy means x is socially preferred to y; i.e., xRy and not yRx. (Don't confuse this R with the Rawls criterion, or this P with the Pareto criterion.) Finally, I is the social indifference relation: xIy means x and y are socially indifferent; i.e., xRy and yRx.

In the next section we will list five plausible requirements that will be imposed on the Arrow social welfare function. Taken together, these five requirements define what we mean by foolproof in our simple Arrow's theorem model.

Table 13.1.

	Individuals											
Choices	1	2	1	2	1	2	1	2	1	2	1	2
1st	x	x	x	x	x	y	x	y	x	z	x	z
2nd	y	y	y	z	y	x	y	z	y	x	y	y
3rd	z	z	z	y	z	z	z	x	z	y	z	x
1st	x	x	x	x	x	y	x	y	x	z	x	z
2nd	z	y	z	z	z	x	z	z	z	x	z	y
3rd	y	z	y	y	y	z	y	x	y	y	y	x
1st	y	x	y	x	y	y	y	y	y	z	y	z
2nd	x	y	x	z	x	x	x	z	x	x	x	y
3rd	z	z	z	y	z	z	z	x	z	y	z	x
1st	y	x	y	x	y	y	y	y	y	z	y	z
2nd	z	y	z	z	z	z	x	z	z	x	z	y
3rd	x	x	x	y	x	x	z	x	x	y	x	x
1st	z	x	z	x	z	y	z	y	z	z	z	z
2nd	x	y	x	z	x	x	x	z	x	x	x	y
3rd	y	z	y	y	y	z	y	x	y	y	y	x
1st	z	x	z	x	z	y	z	y	z	z	z	z
2nd	y	y	y	z	y	x	y	z	y	x	y	y
3rd	x	z	x	y	x	z	x	x	x	y	x	x

3. Requirements on the Arrow Social Welfare Function

1. *Completeness and transitivity.* The social preference relations generated by an Arrow social welfare function must be complete and transitive. If some preference profile is transformed into a particular R, then for any pair of alternatives x and y, either xRy or yRx must hold, and for any triple x, y and z, xRy and yRz must imply xRz. The requirement says that an Arrow social welfare function must always permit compar-

isons between two alternatives, and that social preferences must have the nice transitivity property assumed for an individual's preferences.

We have examined Arrow social welfare functions that don't generate complete and transitive social preference relations. The Pareto criterion gives incomplete social rankings: if 1 prefers x to y and 2 prefers y to x, the two alternatives are Pareto noncomparable. Majority voting gives nontransitive social rankings. As an example in a two-person case (where the voting cycle of Condorcet cannot be generated), suppose 1 prefers x to z to y, and 2 prefers y to x to z. If a vote is taken between x and y, there is a tie (1 votes for x, and 2 votes for y). If a vote is taken between y and z, there is again a tie. According to majority voting, x and y are socially indifferent, and y and z are socially indifferent. Transitivity would require that x and z be socially indifferent. But in a vote between x and z, x gets 2 votes and z none; so x beats z, and transitivity fails.

2. *Universality.* An Arrow social welfare function should work no matter what individual preferences happen to be. This means that the rule should give us a social preference ordering for every cell in Table 13.1, not just for the easy ones, like the ones where there is unanimous agreement.

Universality is a significant requirement. It precludes the assumption of single-peakedness, since it says that the Arrow social welfare function must work for all preference profiles, not just the ones where the utility functions have single peaks. Why should it be imposed?

First, it is difficult to see where to draw the line between permissible and impermissible individual preferences. Which cells in Table 13.1 should be disallowed or ignored? How much diversity can be expected in society? When is there so much conflict that the very idea of social welfare becomes implausible? There are no easy answers to these questions. Second, the theorem we will prove remains valid even when the universality requirement is substantially weakened, and we will indicate how much it can be weakened in a subsequent section.

3. *Pareto consistency.* An Arrow social welfare function should be consistent with the Pareto criterion. For any pair of alternatives x and y, if both individuals prefer x to y, x must be socially preferred to y.

Pareto consistency is a very mild requirement for an Arrow social welfare function. One would not expect it to hold in societies that are ruled by external forces; in which, for example, everyone prefers lust and gambling, on the one hand, to chastity and frugality on the other; but where, according to a Holy Book, the social state of chastity and frugality is preferable to the social state of lust and gambling. Economists naturally would recommend lust and gambling.

On a more serious note, let's recall that the fairness and Rawls criteria could produce results contrary to the Pareto criterion. In our view, this fact is an indictment of fairness and Rawls, not of Pareto. We take Pareto consistency to be fundamental.

4. *Non-dictatorship.* An Arrow social welfare function must make no one a dictator. Individual i is said to be a *dictator* if his wishes prevail, no matter how j feels; that is, if xP_iy implies xPy for all x and y, irrespective of P_j. Ruling out dictatorship does not mean that it is never possible to have xP_iy implying xPy for all x and y. Obviously, if both people agree on the rankings of all alternatives (so that $P_1 = P_2$, as in the diagonal cells of Table 13.1), then it is perfectly reasonable to have the social preference relation agreeing with 1's (and 2's) preference relation, and in fact, the Pareto consistency requirement makes such agreement necessary. Nondictatorship simply says that 1 (or 2) must not always prevail, no matter how 2 (or 1) happens to feel.

5. *Independence of irrelevant alternatives.* If people's feelings change about some set of irrelevant alternatives, but do not change about the pair of alternatives x and y, then an Arrow social welfare function must preserve the social ordering of x and y. The social preference between x and y must be independent of individual orderings on other pairs of alternatives. (We should note that this formulation of independence differs slightly from Arrow's original formulation.)

Independence is the subtlest of the five requirements, and it takes some explanation. Suppose society prefers x to y when z is a third alternative lurking in the wings. Next suppose everyone suddenly changes his mind about the desirability of z, but no one changes his mind about x vs. y. The independence requirement says that, if society is deciding on the relative merits of x and y, and only those two, it must still prefer x to y.

The standard example of an otherwise-nice Arrow social welfare function that violates independence is *weighted voting.* This type of rule was first analyzed in 1781 by Jean-Charles de Borda, in his *Mémoire sur les Élections au Scrutin*, and it is consequently called *de Borda voting.* It works as follows. Each person reports his preference relation, his rank ordering. A first place in a rank ordering is assigned a certain fixed weight, a second place is assigned a (smaller) fixed weight, a third place is assigned a (yet smaller) fixed weight, and so on. (In the two-person, three-alternative current model we have no ties, no cases of indifference, to worry about.) The weights that each alternative gets from each person are summed, and the social preference relation is derived from the sums of the weights.

For instance, suppose person 1 prefers z to x to y, while person 2 prefers y to x to z. Suppose a person's first choice gets a weight of five points, a second choice gets four points, and a third choice gets one point. (The weights are obviously arbitrary. It is common to use equally spaced weights, like 3, 2, and 1, but there is no logically compelling reason to do so. You may construct an example similar to this one using the common weighting scheme.) Now alternative x gets $4+4 = 8$ points, alternative y gets $1+5 = 6$ points, and alternative z gets $5+1 = 6$ points. Therefore, for this preference profile, x is socially preferred to y according to the weighted voting rule.

However, suppose person 1 becomes disillusioned with alternative z, and his preference ordering changes to x over y over z. If the voting is repeated, x gets $5 + 4 = 9$ points, y gets $4 + 5 = 9$ points, and z gets $1 + 1 = 2$ points. Therefore, given this new preference profile, x is socially indifferent to y. Society has become indifferent between x and y, even though neither person has changed his feelings about x and y! Consequently, weighted voting violates the independence requirement.

4. Applying the Requirements

At this stage we shall apply requirements 1, 2, 3 and 5 to Table 13.1. The applications should clarify the meanings of the four requirements. They will also lay the groundwork for the proof of Arrow's Impossibility Theorem for this two-person three-alternatives case.

The completeness/transitivity and the universality requirements say that, when applied to any cell of Table 13.1, an Arrow social welfare function must generate a complete and transitive social ordering.

The Pareto consistency requirement says an Arrow social welfare function must respect unanimous opinion: If both 1 and 2 prefer one alternative to another, then so must society. For example, given the preference profile of the first row, second column cell of Table 13.1, the Pareto requirement says x must be socially preferred to y and x must be socially preferred to z. We must have xPy and xPz. Application of Pareto consistency over the entirety of Table 13.1 gives rise to Table 13.2.

Each cell of this table is produced by applying Pareto consistency to the corresponding cell of Table 13.1, and therefore, any rule for generating social preferences must be entirely consistent with Table 13.2.

Now we turn to the independence requirement. Suppose that when person 1 prefers x to y to z and person 2 prefers y to x to z (the first row, third column cell of Table 13.1) an Arrow social welfare function declares x is socially preferred to y, or xPy. Then independence requires that xPy hold whenever xP_1y and yP_2x, no matter how 1 and 2 rank alternative z. Similarly, if yPx (or xIy) holds when person 1 prefers

Table 13.2.

xPy xPz yPz	xPy xPz	xPz yPz	yPz	xPy	
xPy xPz	xPy xPz zPy	xPz		xPy zPy	zPy
xPz yPz	xPz	xPz yPx yPz	yPx yPz		yPx
yPz		yPx yPz	yPx yPz zPx	zPx	yPx zPx
xPy	xPy zPy		zPx	xPy zPx zPy	zPx zPy
	zPy	yPx	yPx zPx	zPx zPy	yPx zPx zPy

x to y to z and person 2 prefers y to x to z, then yPx (or xIy) must hold whenever xP_1y and yP_2x. In short, the independence requirement forces an Arrow social welfare function to give rise to social preferences that agree over certain preference profiles.

Let's be specific about those areas of agreement. Independence requires that all the cells in Table 13.1 where xP_1y and yP_2x must yield identical social rankings of x and y. Similarly, all the cells where yP_1x and xP_2y must yield identical social rankings of x and y. There is no presumption, however, that the social ranking of x and y on the xP_1y and yP_2x cells need be the same as the social ranking of x and y on the yP_1x and xP_2y cells. Such an anonymity condition is unnecessary for the

proof of the Impossibility Theorem, although it is intuitively appealing and useful in other contexts.

Independence also implies these areas of agreement: All the cells of Table 13.1 where xP_1z and zP_2x must give rise to identical social rankings of x and z; all the cells where zP_1x and xP_2z must give rise to identical social rankings of x and z; all the cells where yP_1z and zP_2y must give rise to identical social rankings of y and z; and, finally, all the cells where zP_1y and yP_2z must give rise to identical social ranking of y and z. All of this information can be incorporated in a third table. Table 13.3a indicates where the social rankings of x and y must agree because xP_1y and yP_2x in all the X'd cells and where the social rankings of x and y must agree because yP_1x and xP_2y in the O'd cells. Tables 13.3b and 13.3c show the areas of agreement which arise from applications of the independence requirement to the social preferences between x and z, and y and z, respectively.

With these preliminaries out of the way, we can turn to a truly remarkable theorem.

5. Arrow's Impossibility Theorem

At least since the time of Condorcet and de Borda in the eighteenth century, people have been concerned with the properties of rules for making social choices, election rules in practice, Arrow social welfare functions in theory. Does there exist a foolproof rule for discovering, or for defining, social preferences? Arrow showed that, if foolproof means consistent with the five requirements above, the answer is No.

We now turn to a formal statement and proof of the theorem for the case of two persons and three alternatives that we are considering so far. In a subsequent section we shall provide a proof for another version of the theorem, in a more general case.

Arrow's Impossibility Theorem, Version 1. Any Arrow social welfare function which is consistent with the requirements of (1) completeness and transitivity, (2) universality, (3) Pareto consistency, and (5) independence of irrelevant alternatives, makes one person a dictator. Therefore, there is no rule which satisfies all five requirements.

Proof. We start by looking at the preference profile of the first row, second column cell of Table 13.1. For these preferences Pareto consistency requires xPy and xPz (Table 13.2). There are three and only three complete and transitive social preference orderings which satisfy xPy and xPz. They are:

Table 13.3.

		X	X		X
		X	X		X
0	0			0	
0	0			0	
		X	X		X
0	0			0	

(a) The crossed cells all produce the same x-y social rankings. The circled cells all produce the same x-y social rankings (which need not be the same as in the crossed cells).

			X	X	X
			X	X	X
			X	X	X
0	0	0			
0	0	0			
0	0	0			

(b) The crossed cells all produce the same x-z social rankings. The circled cells all produce the same x-z social rankings (which need not be the same as in the crossed cells).

	X			X	X
0		0	0		
	X			X	X
	X			X	X
0		0	0		
0		0	0		

(c) The crossed cells all produce the same y-z social rankings. The circled cells all produce the same y-z social ranking (which need not be the same as in the crossed cells).

1. xPy, xPz and yPz
2. xPy, xPz and zPy
3. xPy, xPz and yIz

Each of these three possibilities will be considered in turn.

Case 1: yPz. First a word about strategy. The Pareto consistency requirement tells a lot about what social preferences must be, but it leaves a lot unsaid. Table 13.2 is full of blank and partially blank spaces. We will now show how all the blanks can be filled in by repeatedly applying the independence and transitivity requirements.

If yPz holds in the first row, second column cell, then independence (Table 13.3c) requires that y be socially preferred to z whenever individual preferences about y and z are the same as they are in that cell. Therefore yPz holds in all the cells indicated in Table 13.4a. (The cells that provide crucial steps in the proof are numbered 1 and 2.)

Now consider the first row, fifth column cell, or cell number 2 in Table 13.4a. Pareto consistency (Table 13.2) requires that xPy here, but xPy and yPz implies xPz, by transitivity. So in this cell we must also have xPz.

But if xPz holds in cell number 2, then independence (Table 13.3b) requires that x be socially preferred to z whenever individual preferences about x and z are the same as they are in that cell. Therefore, xPz holds in all the cells indicated in Table 13.4b. (The cells that provide crucial steps in the proof are numbered 2 and 3.)

Now we have xPz in cell 3. We again invoke Pareto consistency and transitivity to conclude that xPy must hold in cell 3 as well. But this allows us to use independence again, (Table 13.3a), to fill in eight more bits of information.

The filling-in process is repeated four more times. You can complete this part of the argument using the sequence of crucial cells indicated in Table 13.4c.

This filling-in process produces six diagrams like Tables 13.4a and 13.4b. Each one shows nine identical social preferences. If the information contained in all these diagrams is added to the information of Table 13.2, the result is the pattern of social preferences of Table 13.5.

But the social preferences shown in Table 13.5 are identical to person 1's preferences. Therefore, in Case 1, 1 is a dictator. He gets his way, no matter how 2 feels.

Case 2: zPy. If zPy holds in the first row, second column cell, an argument analogous to the one above establishes that 2 is a dictator. The argument is left as an exercise.

Table 13.4.

	1. yPz			2. yPz	yPz
	yPz			yPz	yPz
	yPz			yPz	yPz

(a)

			xPz	2. xPz	xPz
			xPz	xPz	3. xPz
			xPz	xPz	xPz

(b)

	1.			2.	
					3.
6.					
			4.		
		5.			

(c)

Table 13.5.

x y z	x y z	x y z	x y z	x y z	x y z
x z y	x z y	x z y	x z y	x z y	x z y
y x z	y x z	y x z	y x z	y x z	y x z
y z x	y z x	y z x	y z x	y z x	y z x
z x y	z x y	z x y	z x y	z x y	z x y
z y x	z y x	z y x	z y x	z y x	z y x

Case 3: yIz. Suppose yIz holds in the first row, second column cell. Then by independence of irrelevant alternatives (Table 13.3c) yIz must also hold in the third row, second column cell, as well as the fourth row, fifth column cell. By Pareto consistency (Table 13.2), z must be socially preferred to x in the latter cell. Now by transitivity, yIz and zPx implies yPx for the fourth row, fifth column cell. By independence again (Table 13.3a), yPz in the fourth row, fifth column cell implies yPz in the third row, second column cell. Using transitivity again, yIz and yPx implies zPx in this cell. However, this contradicts Pareto consistency (Table 13.2), which says that xPz holds here. Therefore, Case 3 is impossible. (John Weymark has

pointed out to us (private communication, 2005) that case 3 follows from a general result: If an Arrow social welfare function satisfies requirements 1, 2, 3 and 5 (that is, all the Arrow requirements but non-dictatorship), and if society can be divided into two groups with opposite strict preferences over some pair of alternatives, then the social preference between this pair must be strict. That is, if no person is indifferent between a pair of alternatives x and y, society cannot be indifferent between x and y.)

The proof of the theorem for the two-person three-alternatives case analyzed here is now complete, for it has been shown that requirements 1, 2, 3, and 5 together imply that either

i. 1 is a dictator, or

ii. 2 is a dictator. Q.E.D.

6. Relaxing the Universality Requirement

We said near the beginning of this chapter that the universality requirement, which demands that an Arrow social welfare function work for any preference profile, is overly strong. This section will indicate why. In fact, the construction of a sensible Arrow social welfare function remains impossible even if a large number of possible preference profiles are excluded.

To show that lots of preference profiles might be dispensed with, without affecting the theorem, we will just count the cells that are crucial in our proof. In case 1, we use the cells numbered 1 through 6 in Table 13.4c. In Case 2, we need another six crucial cells, but the first crucial cell is again the first row, second column cell, which we have already counted. Thus far, we have $6+5$ or 11 cells. In Case 3, we use three crucial cells, but one is the familiar first row, second column cell, which has been counted. Therefore, the total number of crucial cells for the purposes of our proof is $11+2=13$. And the Arrow Impossibility Theorem holds even if any or all of the remaining cells are discarded. Incidentally, the crucial cells we have used are not the only ones which can establish the theorem—other sequences of steps can be used to prove it. But they are a full set, in the sense that they will do the job, and so long as they are all retained, any or all of the other cells are disposable.

In short, with a properly chosen set of thirteen preference profiles, out of the total set of thirty-six, Arrow's Theorem can be established. So universality is really a much stronger assumption than is needed to prove the theorem.

7. Reactions to Arrow's Impossibility Theorem

Since Arrow published his seminal paper in 1950, a vast literature has grown on the Impossibility Theorem and related topics. It's interesting to think about why. The theorem provides an unambiguous answer to the question "Is there a foolproof way to derive complete and transitive social preference relations?" The answer is No. This clearly negative result casts doubts an all assertions that there is a "general will," a "social contract," a "social good," a "will of the people," a "people's government," a "people's voice," a "social benefit," and so on and so forth. That is, it casts doubts on all notions that explicitly or implicitly attribute preferences to society that are comparable to preferences of an individual. Therefore, it undermines vast areas of twentieth century social thought. Of course any theorem that casts so much doubt will generate a lot of responses.

There are many possible reactions to Arrow's Theorem. The first, and perhaps most obvious reaction is this. It is quite silly in the first place to think that there might be social preferences that are analogous to individual preferences. It is nonsense to talk about social preferences since society itself is nothing more than a collection of individuals, each with his own interests. The idea that a motley collection of individuals should have social preferences that are like an individual's preferences is just an example of illegitimate reasoning by analogy. To attribute the characteristics of an individual to a society commits the logical error of personification. Arguments like these have been made by James Buchanan, Charles Plott, and others.

Now this line of reasoning could be pursued further. For instance, the idea that anyone should be interested in Pareto optimality is also silly, since each person i just wants to maximize his own utility u_i. Person i couldn't care less about optimality. Further, the idea that a government official might be interested in pursuing the public good is equally silly, first because the public good is an empty idea, and second because even if there were a public good to be pursued, public official i just wants to maximize his utility function u_i, and what's good for the public will often differ from what's good for the public official. (See Anthony Downs' *An Economic Theory of Democracy.*) Further, the idea that an economist or a political scientist might be interested in instructing the citizenry or electorate, or reforming public officials, is also silly, first because there is no public good, second because the citizens and/or the officials are only trying to maximize their own utility functions, and third because the economist or the political scientist is also simply interested in maximizing his u_i! And why should anyone listen to such an obviously self-serving advisor or reformer?

So this first reaction to Arrow's Theorem is logically attractive, but it can lead to varieties of nihilism that are unappealing to some people, including us. (We wrote this book because we are idealists, not nihilists!)

The second type of reaction to Arrow's Theorem accepts the legitimacy of the basic idea of social preferences, and attacks one or more of Arrow's five requirements. Let's very briefly touch on some of these lines of attack.

First, the completeness requirement might be jettisoned. For instance, some people are quite satisfied with the Pareto criterion alone, and hold that if x and y are Pareto noncomparable, only an act of God can decide between them. There is no way that a reasonable person, or government, can decide, and reasonable people or governments have no business trying to do so. One implication of this line of reasoning is that a Pareto optimal status quo, brought about, perhaps, by a competitive market mechanism, should be left untouched—government officials should not waste their time (and tax dollars) scheming about how to redistribute wealth. Nor should economists and political scientists.

Second, the transitivity requirement might be dropped, or weakened to quasi-transitivity, or acyclicity. If transitivity is dropped entirely, majority voting becomes acceptable, and its advocates simply hope embarrassing voting cycles do not arise. If cycles are potentially there, they might be suppressed by clever committee chairmen, by agenda rules, or by some other *deus ex machina*.

Or, transitivity might be weakened to quasi-transitivity. If this is done, however, a new version of Arrow's Impossibility Theorem rears its head. In this version, due to Allan Gibbard, the requirements of completeness and quasi-transitivity, universality, Pareto consistency, and independence together imply that there must exist an *oligarchy*, rather than a dictatorship. That is, there must exist a group of people G such that (1) if xP_iy for all i in G, then xPy, and (2) if xP_iy for some i in G, then xRy. Acting together, members of the oligarchy can force a social preference for x over y, and, acting apart, each member of the oligarchy has *veto power* over a y which he regards as inferior to x. If G has only one member, then the oligarchy is a dictatorship. If G has two members, it is a duumvirate; if it has three members, it is a triumvirate, and so on. (G might include everyone. Consider, for instance, Amartya Sen's *Pareto-extension* rule, which is defined as follows. Let xRy if y is not Pareto superior to x. This rule satisfies completeness and quasi-transitivity, universality, Pareto consistency, and independence. It makes the whole group an oligarchy. And, we should note, it doesn't provide much information. It says, for example, that all Pareto optimal allocations are equally good.)

If transitivity is weakened to acyclicity, there arises yet another version of Arrow's Impossibility Theorem, a version due to Donald Brown. This result says the requirements of completeness and acyclicity, universality, Pareto consistency, and independence together imply that there must exist what Brown calls a *collegial polity*, rather than an oligarchy or a dictatorship. In a collegial polity, there may be several groups G who can force a social preference for x over y, that is, several G's such that if xP_iy for all i in G, then xPy. These groups have a nonempty intersection, and this intersection, this set of people who belong to all the powerful G-groups is called a *collegium*, which we abbreviate C. Now C need not be able to force the social preference of x over y by itself, but any group that can force that preference must include C. C is thus the elite of the powerful. As an example, suppose there are five people, and the Arrow social welfare function defines xPy if persons 1 and 2 plus any one other person prefer x to y. Otherwise, yRx. Then the powerful G-groups are {1,2,3,}, {1,2,4} and {1,2,5}. That is, any of these groups can force a social preference for x over y. The elite among the powerful, the collegium, is {1,2}. However, the collegium {1,2}, acting alone cannot force a social preference for x over y.

A third reaction to Arrow's Theorem is to drop the universality requirement. This is the approach of the single-peakedness mode of analysis, and similar analyses of restrictions of allowable preference profiles. We have already seen in Chapter 12 and this chapter that this tack is disappointing.

A fourth, and for our purposes final, way is to drop the independence requirement. There are at least two reasons why one might want to move in this direction. The first is that the independence requirement is, of all the Arrow requirements, the least intuitive, the least compelling, the least understandable, and these weaknesses suggest it should be sacrificed. The second is that the independence requirement depends on there being more than one preference profile; it depends on changes in individuals' preferences. Why not constrain the entire discussion to a fixed preference profile? Why not say: "Look, we have so many people with particular preferences that are given. How might we aggregate these given preferences? What might or might not happen if preferences change is of no particular interest, because we want to aggregate the fixed preferences of our given population."

What is wrong with this position? The essential problem here is that with the independence requirement we are forced to admit only one type of Arrow social welfare function - dictatorship - but without it we have an unlimited set of admissible Arrow social welfare functions, a real embarrassment of riches. For instance, suppose we forget about

independence, suppose for simplicity that no one is ever indifferent between two alternatives, and suppose the total number of alternatives is k. Let $(a_1, a_2, \ldots, a_k)$ be any k numbers satisfying $a_1 > a_2 > \ldots > a_k$. Consider the generalized weighted voting rule, or generalized de Borda rule, in which the weight assigned to a first choice is a_1, the weight assigned to a second choice is a_2, and so on. Then this Arrow social welfare function is perfectly well behaved; it satisfies completeness and transitivity, universality (which might be moot if we are really concerned with a fixed preference profile) and Pareto consistency. Independence has been jettisoned. And the rule is most certainly nondictatorial, nonoligarchic, and noncollegial. But the problem here is that the outcome, the social preference relation, depends on the actual magnitudes of the weights $(a_1, a_2, \ldots, a_k)$, and that a different set of weights will generally give a different social preference relation. And there are infinitely many ways to choose the weights! So the resulting social preference relation is arbitrary, insofar as the particular weights are arbitrary.

How is a (social) choice of a set of weights going to be made? How can we decide if one set of weights $(a_1, a_2, \ldots, a_k)$ is socially preferable to another set of weights $(b_1, b_2, \ldots, b_k)$? The problem of social preferences has not been solved in this case, it has only been thrust back onto the choice of the weights.

8. Another Version of Arrow's Theorem

There are many different versions of Arrow's Theorem. We have discussed one simple version, in a special model, in a section above. We will now turn to a general version, that is not restricted to 2 people and 3 alternatives, but which is elegant and easy to prove. Its elegance and ease of proof result from its use of an enhanced and strengthened version of the independence of irrelevant alternatives assumption.

The stronger version of independence that we shall use is called *neutrality-independence-monotonicity*, or NIM for short. NIM was developed by Julian Blau and Rajat Deb, who call it "full neutrality and monotonicity," and further discussed by Amartya Sen, who calls it "condition NIM," short for "weak neutrality, independence, and monotonicity." With NIM we will prove Arrow's Theorem in a general framework, with n individuals, numbered $i = 1, 2, \ldots, n$, and an arbitrary set of 3 or more alternatives. We no longer require that preferences be strict, so person i can prefer x to y (xP_iy), prefer y to x (yP_ix), or be indifferent between them (xI_iy). A *preference profile* is again a specification of everyone's preferences; we write $(R_1, R_2, \ldots, R_n)$ for a profile. As in the previous sections, the *Arrow social welfare function* aggregates in-

dividual preferences into social preferences, and we use R for the social preference relation.

This is the idea of NIM: Suppose $(R_1, R_2, \ldots, R_n)$ and $(R'_1, R'_2, \ldots, R'_n)$ are two preference profiles, and the Arrow social welfare function transforms the former into the social preference R, and the latter into the social preference relation R'. We will call the unprimed profile and social preference relation "old," and the primed "new." Suppose that x, y, z, w are any 4 alternatives (not necessarily distinct). Suppose that given the old profile $(R_1, R_2, \ldots, R_n)$, it turns out that xPy; x is socially preferred to y.

Now compare old preferences with new. Suppose that the support for w over z in the new situation is *equal to or stronger* than the support for x over y under the old, in the sense that all the people that preferred x over y in the old situation now prefer w over z under the new. (That is, for all i, xP_iy implies wP'_iz.)

Suppose further that the support for z over w under the new is *equal to or weaker* than the support for y over x under the old, in the sense that all the people that prefer z over w in the new did prefer y over x in the old. (That is, for all i, zP'_iw implies yP_ix.)

To put it briefly: The people who favored x over y before now favor w over z. The people who now favor z over w were previously in favor of y over x.

Then, since society preferred x over y, it now must prefer w over z. That is, $wP'z$ must hold.

Note that NIM is a "neutrality" assumption in the sense that it says the Arrow social welfare function should be neutral among alternatives, but not necessarily neutral among people. It is an "independence" assumption in the sense that the choices among x, y, w, z are independent of preferences over alternatives other than these 4. It is a "monotonicity" assumption in the sense that the support for w over z might grow compared to the one for x over y; with such growth in support, xPy should imply $wP'z$.

Before turning to the main result, we must introduce another definition, and establish a preliminary result.

Let V represent any non-empty set of individuals, chosen from among persons $1, 2, \ldots, n$. Suppose that, for any pair of alternatives x and y, if all the people in V prefer x to y, then the social preference must be xPy, no matter how others feel. That is, for any x and y, if xP_iy for all the i in V, then xPy. If this is the case, we say that V is *decisive*, or a *decisive set*. (Note that an oligarchy, as defined earlier, must be a decisive set, but not conversely.)

We described the Pareto consistency requirement earlier. The generalization to the case of n individuals is straightforward. We will say an Arrow social welfare function satisfies *Pareto consistency* if xP_iy for all $i = 1, 2, \ldots, n$ implies xPy. That is, if everybody prefers x to y, the social preference must be x over y. Note that for any rule satisfying Pareto consistency, the set of all individuals $\{1, 2, \ldots, n\}$ is decisive.

We described dictatorship earlier, and the definition remains the same: person i is a *dictator* if for any pair of alternatives x and y, xP_iy implies xPy, no matter what are the preferences of the other people. Note that if i is a dictator, then the set $\{i\}$ – that is, i himself – is a decisive set.

Now to the preliminary result:

Proposition 1. Suppose an Arrow social welfare function satisfies NIM. Suppose there is a group of people V, a pair of alternatives x and y, and some preference profile $(R_1, R_2, \ldots, R_n)$, such that xP_iy for all i in V, and yP_ix for all i outside of V. Suppose xPy.

Then V is decisive.

Proof. It is necessary to show that for any preference profile, say $(R'_1, R'_2, \ldots, R'_n)$, if wP'_iz for all the people in V, then $wP'z$. This follows immediately from NIM.

We can now turn to our alternative version of Arrow's Theorem. We are looking for an Arrow social welfare function that transforms preference profiles for n individuals into social preference relations for society. We require, as before, that the social preference relation always be complete and transitive. We require, as before, that the rule satisfy universality: it should work no matter what individual preferences happen to be. We require Pareto consistency: if all n individuals prefer x to y, then the social preference should be x over y. We require non-dictatorship: there should not be one person i who always gets his way: xP_iy always implies xPy. Finally, we require NIM. The result is – we require too much!

Arrow's Impossibility Theorem, Version 2. Any Arrow social welfare function which satisfies (1) completeness and transitivity, (2) universality, (3) Pareto consistency and (4) NIM, makes one person a dictator. Therefore, there is no rule which satisfies all five requirements.

Proof. By Pareto consistency, the set of all individuals $\{1, 2, \ldots, n\}$ is decisive. This establishes that decisive sets exist.

Let V be a decisive set of minimal size, i.e., V is decisive and there does not exist $V' \subseteq V$, $V' \neq V$ that is also decisive. We will

show that there is only 1 person in V, which will make that person a dictator. This will establish Arrow's Theorem.

Suppose to the contrary that V has 2 or more members. Divide V into non-empty subsets V_1 and V_2, and let V_3 represent everyone in society, not in V itself. (Note that V_3 might be empty.)

Consider a preference profile for which:

$$\begin{array}{ll} \text{For } i \text{ in } V_1\text{:} & xP_iyP_iz \\ \text{For } i \text{ in } V_2\text{:} & yP_izP_ix \\ \text{For } i \text{ in } V_3\text{:} & zP_ixP_iy \end{array}$$

(Hello again Condorcet's voting paradox!)

Now since V is decisive, it must be the case that y is socially preferred to z: yPz. We consider 2 cases:

Case 1. Suppose the social preference is xRy. Then xPz by transitivity. Then V_1 becomes decisive by proposition 1 above. But this is a contradiction, since we assumed V was a decisive set of minimal size.

Case 2. Suppose the social preference is not xRy. Then it must be yPx, since social preferences are complete. But if yPx then V_2 is getting its way in the face of opposition by everyone else. (That is, yP_ix for all i in V_2, but xP_iy for everyone else in society.) Then by proposition 1 above V_2 is decisive. This is again a contradiction since we assumed V was a decisive set of minimal size.

Since the assumption that V could be divided into non-empty subsets leads to a contradiction, V must have only 1 member, who is a dictator. Q.E.D.

9. Exercises

1. Suppose there are three people and four alternatives. Assume 1 prefers w to x to y to z, 2 prefers y to z to w to x, and 3 prefers z to x to y to w. Find weights (a_1, a_2, a_3, a_4) so that weighted voting indicates the social preference is y over z over x over w. That is, y's total weighted vote is the highest, z's the second highest, and so on. Next find weights (b_1, b_2, b_3, b_4) so the weighted voting indicates the social preference is z over x over y over w.

2. Suppose there are four people, and no one is ever indifferent between two alternatives. One of the alternatives is special, and is labeled x_0. Consider an Arrow social welfare function that works as follows. If

person 1 prefers x to y, and y isn't the special alternative, then x is socially preferred to y. That is, xP_1y implies xPy, if $y \neq x_0$. (So person 1 is a dictator except when the special alternative is at stake.) If persons 1 and 2, plus at least one other person, prefer x to x_0, then x is socially preferred to x_0. (So person 1 needs person 2 plus someone else to overrule the special alternative.) In all other cases, x_0Rx.

a. Show that this rule satisfies Pareto consistency and independence.

b. Show with specific examples of preferences that the rule is not quasi-transitive.

c. Show that it must be acyclic.

d. Identify the collegium.

e. If all members of the collegium prefer x to y, does it necessarily follow that xPy?

3. Prove Case 2 of Arrow's Impossibility Theorem, Version 1.

10. Selected References

(Items marked with an asterisk (*) are mathematically difficult.)

1. K.J. Arrow, "A Difficulty in the Concept of Social Welfare," *The Journal of Political Economy*, V. 58, 1950, pp. 328-346.

   Arrow's 1950 paper is the source of the enormous literature on the Impossibility Theorem. (By now there are many hundreds of published papers on the topic.) After briefly surveying dictatorship, rule by convention, majority rule, and compensation criteria, Arrow turns to the question: "Can we find other methods of aggregating individual tastes which imply rational behavior on the part of the community and which will be satisfactory in other ways?" The answer is No.

2. K.J. Arrow, *Social Choice and Individual Values.* 2nd edition, John Wiley and Sons, Inc. New York, 1963.

   Arrow's excellent monograph extends and supplements his original 1950 article. The monograph is as valuable for its treatment of related topics as it is for its proof of the Impossibility Theorem. For instance, Arrow discusses compensation criteria at great length; he has a chapter on preference similarity that deals with unanimity, and single-peakedness, and, in the last chapter, he discusses some of the issues raised in the literature on social choice since the publication of his 1950 article (and the first, 1951 edition of this monograph).

The last section of the last chapter is especially interesting, since it is about whether or not rationality is a property that ought to be attributed to society.

*3. K.J. Arrow, A.K. Sen and K. Suzumura (eds.) *Handbook of Social Choice and Welfare, Vol. 1*, Elsevier, New York, 2002.

This is an up-to-date presentation of different topics in the theory of social choice and welfare. The first part of the volume is concerned with Arrow's impossibility theorem.

4. J. Blau and R. Deb, "Social Decision Functions and the Veto," *Econometrica*, V. 45, 1977, pp. 871-879.

Blau and Deb introduce the NIM assumption, which they call "full neutrality and monotonicity." As far as we know it was renamed "condition NIM," short for "weak neutrality, independence, and monotonicity," by A. Sen in his 1977 survey referenced below.

Blau and Deb use NIM to prove a very simple and clever veto theorem.

5. D.J. Brown, "Aggregation of Preferences," *Quarterly Journal of Economics*, V. 89, 1975, pp. 456-469.

See the references section at the end of Chapter 1.

6. J.M. Buchanan, "Individual Choice in Voting and in the Market," *Journal of Political Economy*, V. 62, 1954, pp. 334-343.

*7. M. Le Breton and J. Weymark "Arrovian Social Choice Theory on Economic Domains," 2003 working paper, forthcoming in K.J. Arrow, A.K. Sen and K. Suzumura (eds.) *Handbook of Social Choice and Welfare, Vol. 2*, Elsevier, New York.

This survey presents the scope for obtaining possibility results when preferences and alternatives are required to satisfy the usual assumptions made when one studies economies. The assumption of universality is relaxed, and additional structure is imposed on the set of social alternatives.

8. C. Plott, "Axiomatic Social Choice Theory: An Overview and Interpretation," *American Journal of Political Science*, V. 20, 1976, pp. 511-596.

Plott provides an interesting and relatively nontechnical survey of Arrow Impossibility Theorem topics in the first part of this paper. In the second, he surveys axiomatic characterizations of Arrow social

welfare functions. These are statements of the form "if a rule has properties A, B and C, then it must be such-and-such."

9. A.K. Sen, *Collective Choice and Social Welfare*, Holden-Day, Inc., San Francisco, 1970, Chapters 3, 3*, 4, 4*.

   Chapter 3 is a literary treatment of Arrow's results, 3* is a mathematical treatment and a general proof. Chapters 4 and 4* are about weakening the transitivity requirement.

*10. A.K. Sen, "Social Choice Theory: A Re-examination," *Econometrica*, V. 45, 1977, pp. 53-90.

   This article surveys the technical literature on collective choice rules (Arrow social welfare functions for us) that generate binary social preference relations, and collective choice rules that generate choice functions. (A choice function indicates what alternatives are best from among any subset of alternatives.) Impossibility results are given for both types of rules.

Chapter 14

DOMINANT-STRATEGY IMPLEMENTATION

1. Introduction

In the last chapter the goal was to find a foolproof Arrow social welfare function, which would transform any preference profile into a complete and transitive social preference relation. It turned out that this was impossible. Evidently we were hoping for too much. Now we shall modify that goal. We shall try to find a foolproof rule that can transform any preference profile into a single best alternative, or a single winner.

Is it possible that it is less demanding for a rule to produce single winners, than to produce rankings of all the alternatives? Is it possible that there exist satisfactory rules which generate single winners?

What are the differences between rules that transform preference profiles into social preference relations, or what we call Arrow social welfare functions, and rules that transform preference profiles into single winners? We call the latter *social choice functions*, or SCF's for short. According to Proposition 2 of Chapter 1, if we have a complete and transitive social preference relation we can always find best social alternatives. That is, we can always find alternatives that are socially preferred or indifferent to all the rest. But by Proposition 3 of Chapter 1, even with a complete and acyclic social preference relation, we can always find best social alternatives. So completeness and transitivity for the social preference relation is more than enough to guarantee that socially best choices exist. Therefore, if we only require that a rule generate socially best alternatives, we are asking for less than if we require that a rule generate complete and transitive social preferences. And so the disappointment of Arrow's Impossibility Theorem might not extend to SCF's.

This is our first reason for analyzing rules that transform preference profiles into single winners. We might find a way to escape Arrow's dilemma. The second reason to analyze SCF's is more down-to-earth. Our concern with social preference relations per se is somewhat academic. We analyze social preferences in part because we have already analyzed individual preference relations; we are used to thinking in terms of preference relations. But there is no practical need to be able to judge between every pair of alternatives. From a practical point of view, it is quite sufficient to say: "Given this set of alternatives, this one is best." The practical question of social choice is not "What is the ranking of all the alternatives?" but "Which alternative is on top?" This is our second reason for analyzing SCF's.

The question now is "Is there a foolproof way to find single best alternatives, is there a foolproof SCF?" Once again, the answer depends on what we mean by foolproof.

Let's first consider the requirements that defined a foolproof Arrow social welfare function. (1) The completeness and transitivity requirement is not directly applicable to SCF's, and we will drop it. (2) Universality is applicable, and we will continue to require it. We will require that the rule always work, no matter what individual preferences might be. (3) Pareto consistency might easily be adapted to SCF's. For instance, one might insist that the chosen alternative, the best alternative, always be Pareto optimal. However, we will not insist on Pareto consistency as a requirement for foolproof SCF's; we will drop this requirement. (4) Nondictatorship is applicable here, and we will continue to require it. However, the definition of an SCF dictator will differ slightly from the definition of an Arrow social welfare function dictator. (5) Independence of irrelevant alternatives is not directly applicable to SCF's, and we will drop this requirement also.

So, of the Arrow requirements, we are left with universality and nondictatorship. Are there any new requirements to impose? What in fact makes a foolproof SCF? The answer to this question is closely connected to our analysis of public goods in Chapter 8, and leads to the approach taken in this and the next two chapters, the theory of implementation.

Recall that the Wicksell-Lindahl tax scheme, in which a person's tax share was equal to his marginal utility from the public good, has one fatal flaw: it induces people to misrepresent their preferences, to lie to the Public Good Board. We take this flaw very seriously in this and the remaining chapters of this book, and we incorporate the idea of misrepresentation into our definition of a foolproof SCF.

If a tax scheme causes people to behave strategically, that is to lie about their marginal utilities for a public good, then we say it isn't

foolproof. If a voting mechanism causes people to vote strategically, that is, to vote as if they prefer A over B over C when in fact they prefer C over B over A, then we say it isn't foolproof.

In short, in this chapter the new requirement for a foolproof rule is that it provide no incentives for strategic behavior in the sense of misrepresentation of preferences. The requirement that no one has an incentive to misrepresent his preferences, no matter what the preferences of the others may be, is called *strategy proofness*. It amounts to requiring that telling the truth to the Public Good Board, or more generally, to any Central Authority is always a dominant strategy.

If a way can be devised to get everyone to tell the truth to the Central Authority, no matter what the others are telling it, and to have the Central Authority then choose an outcome according to the SCF, we say the SCF is *dominant-strategy implementable*. Can we find SCF's that are dominant strategy implementable?

Note that the area of economic theory that studies the possibility of achieving desired social goals, when agents act strategically and the Central Authority does not have all the relevant information, is now called *implementation theory*. The term "implementation" was first used in this context by Eric Maskin, in a seminal paper written in 1977. (We shall describe in detail Maskin's contribution in Chapter 15.)

We now turn to examples of SCF's to further illustrate the idea of strategy proofness.

2. Examples of Strategic Manipulation

What follows are examples of SCF's that are *not* strategy proof, i.e., allow strategic manipulation.

Example 1 is a simplified version of a problem already analyzed in Chapter 8. There is a public good, a bridge, that might or might not be built. We'll assume here that the bridge can be only one size, so the choice is to build it or not. Consequently, there are just two alternatives.

There is one good other than bridges; this is the private good, which we now simply call money. Person i's initial quantity of the private good is ω_i, and we let T_i represent the tax that i pays to finance the bridge, if the bridge is built. We assume that i's utility function is separable, and can be written this way:

$$u_i = \begin{cases} v_i + \omega_i - T_i \text{ if the bridge is built} \\ \\ \omega_i \quad \text{if the bridge is not built} \end{cases}$$

The number v_i is then i's valuation of the bridge, and is analogous to the function v_i of Chapter 8. We assume that $v_i \geq 0$ for all i.

If the bridge is built, its cost is C, measured in units of the private good or money. As usual we assume that whatever the tax system might be, each person knows how all taxes are computed.

The SCF works as follows. Each person is asked his valuation v_i. The Bridge Board goes ahead and builds the bridge if

$$\sum_{i=1}^{n} v_i \geq C.$$

It does not build it if

$$\sum_{i=1}^{n} v_i < C.$$

Now let's turn to the tax rules. We shall examine two possibilities. First, suppose that the tax shares are fixed and equal; or $T_i = C/n$ for all i. Suppose for some particular i, $v_i - C/n > 0$. Then i is better off if the bridge is built than if it is not built. When the Bridge Board asks i for his v_i, i has a clear incentive to lie, to misrepresent his preferences. In particular, i will exaggerate his valuation of the bridge. Similarly, if $v_i - C/n < 0$, i has a clear incentive to claim his valuation is zero. With this tax scheme, the SCF encourages strategic manipulation, and any decision the Bridge Board reaches must be suspect.

Next suppose that i's tax is linked to his valuation, as it is in the Wicksell-Lindahl tax scheme. In particular, suppose T_i is set equal to v_i. There might be a budget surplus, which we know could destroy Pareto optimality, but let's ignore that problem and concentrate on the possibility of strategic manipulation. If $T_i = v_i$, it is clear that many people, including some who value the bridge rather highly, will claim their valuations are zero! They will try to be free riders. Again, this SCF elicits false preferences, and any decision it reaches is suspect.

In short, these two very plausible tax schemes give rise to a bridge-building SCF that rewards manipulation. And this elicitation of false preferences, in our view, means these tax schemes are not foolproof. (Of course there do exist demand-revealing taxes to finance this bridge, taxes which provide no incentives for false revelation of preferences. Examples are given in the exercise section below. It should be recalled, however, that the effectiveness of demand-revealing taxes depends crucially on the special assumption that individual i's valuation v_i is independent of ω_i and T_i.

In the next examples we turn away from tax expenditure schemes, and turn back to voting rules.

Example 2 is the weighted voting rule, or the de Borda rule, which we met in the last chapter. Here we analyze the common version of

de Borda's rule. Suppose there are k alternatives. Each voter ranks the alternatives according to his order of preference. (We assume that there is no indifference.) A voter's first choice is assigned k points, his second $k-1$ points, and so on, down to his last choice, which is assigned one point. The total vote for an alternative is the sum of the points assigned to it by the various voters, and the winner (ignoring ties) is the alternative with the highest sum.

Aside from the minor complication caused by the possibility of ties, the de Borda rule takes a collection of individual preferences, and, based on those preferences, produces a single best alternative, or a winner.

Let's see how it might work with a particular preference profile, when there are seven voters and five alternatives. Suppose the voters have the following preferences:

Type 1	Type 2	Type 3
x	y	z
y	z	x
z	x	y
u	u	u
v	v	v
(3 people)	(2 people)	(2 people)

This table indicates that three of the people, the ones labeled "type 1," prefer x to y to z to u to v. Two of the people, the ones labeled "type 2," prefer y to z to x to u to v; and two of the people, the ones labeled "type 3," prefer z to x to y to u to v.

In this case the de Borda votes are:

$$3 \times 5 + 2 \times 3 + 2 \times 4 = 29 \text{ points for } x,$$

$$3 \times 4 + 2 \times 5 + 2 \times 3 = 28 \text{ points for } y,$$

$$3 \times 3 + 2 \times 4 + 2 \times 5 = 27 \text{ points for } z,$$

$$3 \times 2 + 2 \times 2 + 2 \times 2 = 14 \text{ points for } u, \text{ and}$$

$$3 \times 1 + 2 \times 1 + 2 \times 1 = \ \ 7 \text{ points for } v.$$

Consequently, when all individuals honestly report their preferences, alternative x wins.

However, if one of the type 2 voters declares

$$\begin{matrix} y \\ z \\ u \\ v \\ x \end{matrix}$$

as his preferences ordering, the de Borda votes are 28 for y, 27 for z and 27 for x; so y wins.

Consequently, the de Borda rule may tempt people to vote as if their preferences are something other than what they really are. It might provide incentives for the declaration of false preferences and therefore the decisions it reaches are suspect. In short, it's not a foolproof SCF. (Incidentally, this kind of manipulation of the de Borda rule was forseen by de Borda himself, who is supposed to have said, "My scheme is only intended for honest men." See Duncan Black's *Theory of Committees and Elections*, Chapter XVIII-3.)

In Example 3, we return to majority voting, but majority voting modified by the introduction of an agenda. We have already seen that majority voting between pairs of alternatives may produce cycles of social preference, unless individuals' preferences are single-peaked. And cycles might make the choice of a winner impossible. However, cycles infrequently arise in the real world, partly because of single-peakedness, but also partly because of the effects of committee agendas. A typical agenda rules out many pairwise comparisons, and therefore makes the appearance of cycles less probable. Without cycles, majority voting does produce overall winners (barring ties), and is consequently a proper SCF. To clarify these points, we turn to our concrete example.

Again, suppose there are seven voters of three types. This time, however, assume there are three alternatives. One is the *status quo*, which we abbreviate SQ. The second is a *motion*, which we abbreviate M. And the third is an *amended version of the motion*, which we abbreviate AM. Assume the preferences are as follows:

Type 1	Type 2	Type 3
SQ	M	AM
M	AM	SQ
AM	SQ	M
(3 voters)	(2 voters)	(2 voters)

Majority voting with no agenda produces a cycle here. The status quo defeats the motion 5 to 2; the motion defeats the amended motion 5 to 2; but the amended motion defeats the status quo 4 to 3.

But what does a typical committee do in a case like this? Does it hold a vote between all three pairs? In fact, it doesn't. The standard procedure is to first vote on the motion (i.e., decide between M and AM), and then put the winner of that vote against the status quo. Only two votes are held, not three. If people vote their preferences honestly, M first defeats AM, or the amendment is defeated. The final winner is SQ.

But under these circumstances, the two type 2 voters can misrepresent their preferences, by voting as if they prefer the amended motion to the

motion, and the motion to the status quo. If they do so, AM defeats M in the first round, or the amendment is adopted. Then AM defeats SQ, or the amended motion is adopted. The final winner is AM, which type 2 voters prefer to SQ. (Note that this is an example of "group or coalitional manipulation", rather than manipulation by a single individual. A similar, but more complex example, could be constructed to illustrate manipulation by a single individual in a committee.)

In short, the method of majority rule with the usual agenda may provide opportunities for people to profitably misrepresent their preferences. Consequently, decisions arrived at with this very widely used procedure are suspect. This SCF isn't foolproof either.

Our idea of a foolproof social choice function should be clear by now. Each of the above SCF's is unsatisfactory because it produces incentives for individuals to misrepresent their preferences, and so the decisions reached are in each case suspect. We require of a foolproof SCF that it never provide incentives for declarations of false preferences. The SCF must be immune to manipulation or cheating by any dishonest individual.

As we noted above, we also require that a foolproof SCF work for all possible preference profiles (the universality requirement), and that it be nondictatorial.

There is one last minor requirement. Suppose the set of alternatives is $\{x, y, z, u, v, \ldots\}$, and the SCF chooses y as the winner, no matter what the preferences of the individuals might be. This is clearly an unsatisfactory SCF because it is trivial or degenerate. We shall also require nontriviality, or nondegeneracy, for our foolproof SCF.

We have seen several examples of SCF's that are not foolproof. Does there exist an SCF that meets the four requirements of universality, nontriviality, nondictatorship, and strategy proofness?

3. The Gibbard-Satterthwaite Impossibility Theorem

Does a foolproof SCF exist? The question is answered here for the simple model that was analyzed in the first part of the previous chapter. We assume that there are two people, and three alternatives, x, y, and z. We again suppose that no individual is ever indifferent between two alternatives. Each person can then have six possible rankings of the three alternatives, and there are $6 \times 6 = 36$ possible preference profiles, all of which are pictured in Table 13.1 of Chapter 13, which is reproduced in Table 14.1.

Each cell in this figure shows a preference ranking for each of the two people; so, for example, the cell in the first row, second column, indicates person 1 prefers x to y to z, and person 2 prefers x to z to y.

The universality requirement says a social choice function must work no matter what preference profile is given to it. Therefore, a social choice function for this little society is a rule which takes every cell of Table 14.1, or every preference profile, and transforms it into a winner, or a social choice. For each of the thirty-six preference profiles of Table 14.1, there are three possible social choices. Therefore, the number of conceivable SCF's is 3^{36}, or (approximately) 1.5×10^{17}, or a hundred and fifty thousand trillion. Our having only two people and three alternatives does not imply that only a few SCF's are available!

Any one of the many possible SCF's can be represented by another 6×6 matrix, whose entries are winners, or social choices, corresponding to the preference profiles of Table 14.1.

Table 14.2 represents one such SCF:

Each cell of Table 14.2 shows the social choice or winner, given the preference profile of the corresponding cell of Table 14.1. For example, given the preference profile in the first row, second column cell of Table 14.1, Table 14.2 says the winner is alternative x.

Now it should be obvious that Table 14.2 represents a very special SCF. For each choice in it is person 1's most preferred alternative. This is a *dictatorial* SCF: the social choice is always the favorite alternative of person 1. There is, of course, one other dictatorial SCF; it would be represented by the transpose of the Table 14.2 matrix, and it would make 2 a dictator.

But we require that a foolproof SCF must be nondictatorial. Therefore, the SCF represented in Table 14.2 is unacceptable, as is the SCF that makes 2 a dictator.

We also require that a foolproof SCF be *nontrivial* or *nondegenerate.* In terms of the simple model at hand, this means that there must be some preference profile that gives rise to the choice of x, that there must be some (other) preference profile that gives rise to the choice of y, and that there must be some (other) preference profile that give rise to the choice of z. In other words, none of the three alternatives is irrevocably excluded from choice. In formal terms, each of x, y and z is included in the range of the SCF. If one alternative were always excluded from choice, we would have a two-alternative model, or a degenerate, trivial version of the model at hand.

Now let's turn to the crucial requirement of nonmanipulability or *strategy proofness.* How can we represent the idea that no person should ever have an incentive to falsely reveal his preferences? How, in fact, can

Table 14.1.

	Individuals											
Choices	1	2	1	2	1	2	1	2	1	2	1	2
1st	x	x	x	x	x	y	x	y	x	z	x	z
2nd	y	y	y	z	y	x	y	z	y	x	y	y
3rd	z	z	z	y	z	z	z	x	z	y	z	x
1st	x	x	x	x	x	y	x	y	x	z	x	z
2nd	z	y	z	z	z	x	z	z	z	x	z	y
3rd	y	z	y	y	y	z	y	x	y	y	y	x
1st	y	x	y	x	y	y	y	y	y	z	y	z
2nd	x	y	x	z	x	x	x	z	x	x	x	y
3rd	z	z	z	y	z	z	z	x	z	y	z	x
1st	y	x	y	x	y	y	y	y	y	z	y	z
2nd	z	y	z	z	z	x	z	z	z	x	z	y
3rd	x	z	x	y	x	z	x	x	x	y	x	x
1st	z	x	z	x	z	y	z	y	z	z	z	z
2nd	x	y	x	z	x	x	x	z	x	x	x	y
3rd	y	z	y	y	y	z	y	x	y	y	y	x
1st	z	x	z	x	z	y	z	y	z	z	z	z
2nd	y	y	y	z	y	x	y	z	y	x	y	y
3rd	x	z	x	y	x	z	x	x	x	y	x	x

strategic behaviour, manipulation, or profitable lying, be represented in terms of these figures? We illustrate with an example.

Suppose we know some of the social choices for the preference profiles of row one of Table 14.1:

Social Choices

x	x	y	z	?	?

Table 14.2.

x	x	x	x	x	x
x	x	x	x	x	x
y	y	y	y	y	y
y	y	y	y	y	y
z	z	z	z	z	z
z	z	z	z	z	z

(This is just part of one possible SCF.) That is, for the preference profile of row one, column one, we know that x wins; for the preference profile of column two, x wins; for the preference profile of column three, y wins; for the preference profile of column four, z wins, and we know nothing more. In this case, person two has an opportunity to profitably misrepresent his preferences. Suppose his real preferences are

$$\begin{matrix} y \\ z \\ x \end{matrix} \quad \text{(column four in Table 14.1).}$$

(That is, he prefers y to z to x.) If he reports this honestly (and 1 is also honest), the SCF outcome is z. However, if he (falsely) claims his preferences are

$$\begin{matrix} y \\ x \\ z \end{matrix} \quad \text{(column three in Table 14.1),}$$

the SCF outcome is y, when he (truly) prefers y to z. In short, person 2 can profitably manipulate the SCF when person 1 prefers x to y to z, and he prefers y to z to x (that is, when the preference profile is the one in row one, column four of Table 14.1).

If there is any opportunity for 1 (or 2) to secure a preferred outcome by misrepresenting his preferences, the SCF is said to be *manipulable.* If it is never possible for 1 (or 2) to secure a preferred outcome by misrepresentation, the SCF is said to be *nonmanipulable*, or *strategy proof.*

Note that the definition of nonmanipulability used here only requires that an SCF be immune to lying by single individuals — not that it be immune to lying by arbitrary groups of individuals. In the majority voting with an agenda example of the last section, a group (both of the type 2 voters) profitably misrepresented its preferences. As noted, that was group-manipulation, or *coalition-manipulation.* One might require that an SCF be coalition-nonmanipulable. But we only require individual-nonmanipulability, and our version of nonmanipulability is a less stringent requirement.

The SCF partly illustrated above is evidently manipulable. What about the dictatorial SCF of Table 14.2? Clearly 2 cannot manipulate it since his preferences never affect the outcome. Misrepresenting them must be useless. Nor can 1 manipulate it, since he always gets his (true) first choice. He can never secure a preferred outcome by lying. Dictatorship is, therefore, nonmanipulable. However, it is unacceptable. We require that a foolproof SCF be nondictatorial.

Are there any nontrivial, nondictatorial and nonmanipulable SCF's?

Since there are 150 thousand trillion possible SCF's in this simple model, it is obviously impossible to systematically examine all of them to discover which, if any, are nondictatorial, nonmanipulable, and nondegenerate. Nonetheless, an unambiguous negative answer to the question can be reached, in a profound, inescapable "impossibility" theorem much like Arrow's theorem. The theorem, discovered independently by Allan Gibbard and Mark Allen Satterthwaite in the 1970's, says *there are no foolproof social choice functions.*

We shall turn to a formal statement and proof of the theorem below. The proof, however will depend on an intuitively appealing proposition, which we state here and prove in the Appendix to this chapter:

> *Proposition 1*: Suppose an SCF is nondegenerate and nonmanipulable. Suppose both people prefer one alternative to a second alternative. Then the social choice, the winner, cannot be the second alternative.

Now we can turn to the fundamental result:

> *The Gibbard-Satterthwaite Impossibility Theorem*: Any SCF which satisfies the universality, nondegeneracy, and nonmanipulability re-

quirements must be dictatorial. Therefore, there is no SCF which satisfies all requirements.

Proof: The proposition above implies that any nondegenerate nonmanipulable SCF must be consistent with Table 14.3.

For example, if person 1 prefers x to y to z and person 2 prefers x to z to y (the first row, second column cell of Table 14.1), the proposition says that neither y nor z can be the social choice, since both people prefer x to y, and both people prefer x to z. Therefore, the social choice must be x. Again, for the first row, third column preference profile of Table 14.1, where person 1 prefers x to y to z and person 2 prefers y to x to z, the social choice cannot be z, since both prefer x to z. This line of reasoning is applied to the entirety of Table 14.1, to get Table 14.3.

Table 14.3.

Social Choices

x	x	not z	not z	not y	
x	x	not z		not y	not y
not z	not z	y	y		not x
not z		y	y	not x	not x
not y	not y		not x	z	z
	not y	not x	not x	z	z

Now let us focus on the first row of Table 14.1:

1	2	1	2	1	2	1	2	1	2	1	2
x	x	x	x	x	y	x	y	x	z	x	z
y	y	y	z	y	x	y	z	y	x	y	y
z	z	z	y	z	z	z	x	z	y	z	x

Thus far, this much is known about the corresponding social choices:

Social Choices

x	x	not z	not z	not y	

To start the machinery cranking, an assumption must be made: Suppose that the social choice for the third column cell (which cannot be z) is x.

It follows that:

Social Choices

x	x	x	not z	not y	

Now if the social choice in column four, five or six were y, person 2 would have an opportunity to manipulate in column three. That is, he could force the choice of y instead of x, when his real preferences are

$$\begin{matrix} y \\ x, \\ z \end{matrix}$$

by pretending his preferences were as in four, five, or six. Therefore, for any nonmanipulable SCF, we must have:

Social Choices

x	x	x	x	not y	not y

Next, if the social choice in column five or six were z, person 2 would have an opportunity to manipulate in column four. That is, he could force the choice of z instead of x, when his real preferences are

$$\begin{matrix} y \\ z, \\ x \end{matrix}$$

by pretending his preferences were as in five or six. Therefore, for any nonmanipulable SCF, we must have:

Social Choices

x	x	x	x	x	x

Similar reasoning forces particular social choices as one drops down and fills in all thirty-six cells in Table 14.3. The rest of the filling in process is left as an exercise. When all thirty-six cells are filled in, the result is a table identical to Table 14.2, and consequently, *person 1 is a dictator.*

This outcome became inevitable when we assumed that the social choice for the first row, third column cell was x. Had we assumed y, person 2 could have been the dictator.

In either case, a nondegenerate nonmanipulable SCF must be dictatorial. Therefore, there is no nondegenerate, nondictatorial strategy proof SCF. Q.E.D.

Before leaving the theorem we should make some observations about the special case nature of the proof. Here there are two individuals, three alternatives and no indifference. Using more sophisticated tools the theorem is generalizable to two or more individuals, three or more alternatives, and indifference permissible. Proofs are in the articles referred to at the end of this chapter.

4. Significance of the Gibbard-Satterthwaite Theorem

It is worthwhile to briefly discuss the significance of the Gibbard-Satterthwaite Theorem. What does the theorem say, and what doesn't it say? First of all, it doesn't say that there are no useful social choice functions. Obviously we can and do live with imperfect decision rules. Majority voting, with an agenda to make cycles impossible, is for many people an acceptable SCF, in spite of the fact that it is not foolproof. It is certainly superior to dictatorship. Demand-revealing tax schemes may in some cases provide acceptable SCF's, although the assumption of separability for utility functions violates our universality requirement. Nor does the theorem say that no decision will be reached in a given situation. Some alternative is always chosen.

But the Gibbard-Satterthwaite Theorem does indicate that no acceptable SCF is always immune to strategic behavior. No acceptable

SCF makes truth telling a dominant strategy. No acceptable SCF is dominant-strategy implementable.

Therefore the theorem raises questions about how people will behave in making social decisions. For example, what sort of strategies might they adopt if some or all are acting dishonestly? It raises questions about the dynamics of decision making: How might people react to one another's preferences? How might 1's misstatements of preferences affect 2's?

It also raises questions about the optimality of any social decision. The rules we generally use are manipulable, so we cannot rely on Proposition 1 above to guarantee Pareto optimality. There is no a priori reason to believe that non-Pareto optimal choices won't be made. What choices are ultimately made depends on assumptions about strategic behavior, and optimality of the final equilibrium depends on these assumptions as well, but inefficiencies are to be expected. There is nothing like a "first welfare theorem" for strategic equilibrium concepts in game theory.

The Gibbard-Satterthwaite Theorem raises questions about political legitimacy: In a world in which many, or all voters are misrepresenting their preferences, it is difficult to say that an outcome is "right" or "correct" or "proper" or "legitimate."

In the economic sphere, the Gibbard-Satterthwaite Theorem raises questions about the provision of public goods like bridges, parks, schools, roads, national defense, and so on. How narrow are the circumstances under which one can truly say: this project should be carried through; this bridge should be built?

A final observation: Like the Arrow Impossibility Theorem, the Gibbard-Satterthwaite Theorem comes to a broad *negative* conclusion. The theorem demands that truth telling be a dominant strategy, and it finds no acceptable SCF. But requiring that truth telling be a dominant strategy, that is, requiring that no individual have an incentive to lie, no matter *what* the other individuals are doing, may be requiring too much. Suppose we require less; suppose we require that no individual have an incentive to lie, *if* the others aren't lying? That is, suppose we require that truth telling be a Nash equilibrium rather than a dominant-strategy equilibrium? This will be the topic of the next chapter.

5. Exercises

1. For the bridge example of this chapter, suppose T_i is defined as follows:

$$T_i = \begin{cases} \frac{1}{n-1}\sum_{j\neq i} v_j + [C - \sum_{j\neq i} v_j], \text{ if } \sum_{j\neq i} v_j < C \leq \sum_{i=1}^{n} v_i \\ \\ \frac{1}{n-1}\sum_{j\neq i} v_j, \text{ otherwise} \end{cases}$$

It is assumed here that the

$$\frac{1}{n-1}\sum_{j\neq i} v_j$$

is always collected, whether or not the bridge is built!

Show that no individual can profitably misrepresent his preferences, given this tax rule. Show that the Bridge Board will have enough revenue if it does build the bridge.

2. Recall the demand-revealing tax scheme from Chapter 8, where person i's tax is

$$T_i = \hat{x} - \sum_{j\neq i} v_j(\hat{x}) + \max_x \sum_{j\neq i} \left[v_j(x) - \frac{x}{n} \right].$$

In the bridge example of this chapter, the variable x can take on only 2 values, 0 and C, corresponding to "don't build" and " build" respectively. Similarly, $v_i(x)$ is either 0, or v_i.

(a) Show that in the dichotomous bridge model, demand-revealing taxes would be given by the following 2 equations, the first corresponding to "don't build," and the second to "build":

(i) $$T_i = \max \left[0, \sum_{j\neq i} v_j - \frac{n-1}{n} C \right]$$

(ii) $$T_i = C - \sum_{j\neq i} v_j + \max \left[0, \sum_{j\neq i} v_j - \frac{n-1}{n} C \right].$$

(b) It turns out that the demand-revealing tax scheme is the only scheme that (1) never provides any incentive for an individual

to misrepresent his preference, and (2) always ensures sufficient revenue if the bridge is build. Show with an example that the demand-revealing tax scheme will sometimes require that some individuals be taxed even when the bridge is *not* built.

3. Plurality rule is an SCF that works as follows. There are n candidates. Each person ranks the candidates in his order of preference (assuming no one is indifferent between two candidates). The Election Board collects the rankings, and, for each candidate, counts the number of times he is ranked first. Barring ties, the candidate with the highest number of firsts is the winner.

   Construct an example to show how plurality rule might be manipulated.

4. The Reverend C.L. Dodgeson (better known as Lewis Carroll, and author of *Alice's Adventures in Wonderland*) wrote several pamphlets on voting rules. (These pamphlets are reproduce in part in Duncan Black's *Theory of Committee and Elections.*) One of Dodgeson's rules is called the "method of marks," and it works as follows. Each voter gets a certain fixed number of points, or "marks," which he may distribute among the various candidates as he chooses. For instance, if a voter gets 10 points, he might assign 3.5 points to candidate A, 4.5 points to candidate B, 2 points to candidate C, and zero to all the rest.

   Construct an example with three candidates and three voters, where each voter gets 10 points to distribute, to show that Dodgeson's method of marks is liable to strategic voting.

5. A multistage procedure called exhaustive voting works as follows. There are n candidates for a position. In the first stage of the voting, each voter assigns one vote to each of the $n-1$ candidate he likes best — and assigns no vote to the candidate he likes worst. The votes are counted, and the candidate with the smallest total is dropped from the field. In the second stage, each voter assigns one vote to each of the $n-2$ candidates he likes best — and assigns no vote to the candidate he likes worst. Again, the votes are counted, and the candidate with the smallest total in this stage is dropped from the field. The process continues through states $3, 4, \ldots$, until only one candidate remains. He is the winner. Consider the following preferences:

Type 1	Type 2	Type 3
A	B	C
B	C	A
C	A	B
(3 people)	(2 people)	(2 people).

Carefully show how the exhaustive voting procedure can be manipulated, if voters have the preferences indicated above.

6. In the proof of the Gibbard-Satterthwaite Theorem, the nondegeneracy and nonmanipulability assumptions were used to fill the six cells of row one of Table 14.3. Continue the filling in process in the remainder of the thirty-six cells.

6. Appendix

We now turn to the proof of Proposition 1. We will, in fact, present a more general result than the one given above. We suppose that there are n people and many alternatives. The notation here is also slightly more complicated than what was used above. We let the function F represent an SCF, and $(P_1, P_2, \ldots, P_n)$ a preference profile for the n people. Alternatives will be denoted with X's and Y's. XP_iY then means person i prefers X to Y. F transforms preference profiles into winning alternatives, we can write, for example, $F(P_1, P_2, \ldots, P_n) = X$. If this is the case for some preference profile, alternative X is said to be in the range of F.

Proposition. Suppose the SCF F is nonmanipulable, and X is in the range of F. If XP_iY for all i, then, $F(P_1, P_2, \ldots, P_n) \neq Y$.

Proof. Define P_i' from P_i by moving $\{X, Y\}$ to the top of i's list, preserving the $\{X, Y\}$ ordering ($XP_i'Y$ for all i), and preserving the ordering among all elements other than X and Y.

First, we claim that $F(P_1', P_2', \ldots, P_n') = X$. Suppose to the contrary that $F(P_1', P_2', \ldots, P_n') \neq X$, and let $(P_1'', P_2'', \ldots, P_n'')$ be a preference profile which does give rise to the choice of X.

Define

$$
\begin{aligned}
X_0 &= F(P_1'', P_2'', \ldots, P_n'') \quad (= X) \\
X_1 &= F(P_1', P_2'', \ldots, P_n'') \\
X_2 &= F(P_1', P_2', \ldots, P_n'') \\
&\vdots \\
X_n &= F(P_1', P_2', \ldots, P_n') \quad (\neq X).
\end{aligned}
$$

Let j be the smallest number for which $X_j \neq X$. Then

$$F(P'_1, \ldots, P'_{j-1}, P''_j, \ldots, P''_n) = X$$

but

$$F(P'_1, \ldots, P'_{j-1}, P'_j, P''_{j+1}, \ldots, P''_n) = X_j \neq X.$$

By the construction of P'_j, XP'_jX_j. This implies F is manipulable by j at $(P'_{1j}, \ldots, P'_{j-1}, P'_j, P''_{j+1}, \ldots, P''_n)$, a contradiction. Therefore, $F(P'_1, \ldots, P'_n) = X$, as claimed.

Next, suppose that $F(P_1, P_2, \ldots, P_n) = Y$. Define

$$\begin{aligned}
Y_0 &= F(P'_1, P'_2, \ldots, P'_n) \quad (= X) \\
Y_1 &= F(P_1, P'_2, \ldots, P'_n) \\
Y_2 &= F(P_1, P_2, P'_3, \ldots, P'_n) \\
&\vdots \\
Y_n &= F(P_1, P_2, \ldots, P_n) \quad (= Y).
\end{aligned}$$

Let k be the largest number for which $Y_k \neq Y$. Then

$$F(P_1, \ldots, P_k, P'_{k+1}, \ldots, P'_n) = Y_k \neq Y$$

and

$$F(P_1, \ldots, P_k, P_{k+1}, \ldots, P'_n) = Y.$$

There are two cases to consider. (i) If $Y_k = X$, then $XP_{k+1}Y$ by assumption, and F is manipulable by $k+1$ at $(P_1, \ldots, P_k, P_{k+1}, \ldots, P'_n)$, a contradiction. (ii) If $Y_k \neq X$, then $YP'_{k+1}Y_k$ by the construction of P'_{k+1}, and F is manipulable by $k+1$ at $(P_1, \ldots, P_k, P'_{k+1}, \ldots, P'_n)$, again a contradiction. In either case the supposition that $F(P_1, \ldots, P_n) = Y$ is untenable, which completes the proof of the proposition. Q.E.D.

7. Selected References

(Items marked with an asterisk (*) are mathematically difficult.)

1. S. Barberá, "An Introduction to Strategy-Proof Social Choice Functions," *Social Choice and Welfare*, V. 18, 2001, pp. 619-653.

   Salvador Barberá, one of the main contributors to the literature on strategy proofness, published this very useful survey. Impossibility results, like the Gibbard-Satterthwaite theorem, and positive results, ways to get around it, are covered.

2. D. Black, *The Theory of Committee and Elections*, Cambridge University Press, London, 1958.

   See the Chapter 12 References section.

3. R. Farquharson, *Theory of Voting*, Yale University Press, New Haven, 1969.

   This remarkable short book was written a decade before it was published, and so it ranks as one of the earliest serious treatments of strategic voting. Farquharson writes in his preface that he studied the works of Kenneth Arrow and Duncan Black, but "found, however, that they both assumed that voters used neither strategy nor skill, while Von Neumann and Morgenstern, to whom their footnotes referred me, had not applied the theory of games to actual voting procedures." So Farquharson proceeded to do just that.

   The interesting and readable monograph develops several of the notions later used by Gibbard and Satterthwaite: particularly sincere voting, and straightforward (dominant) strategies.

*4. P. Gärdenfors, "A Concise Proof of a Theorem on Manipulation of Social Choice Functions." *Public Choice*, 1977.

   Gärdenfors provides a relatively clear version of the proof of the Gibbard-Satterthwaite Theorem in the general, many person, many alternative case.

5. P. Gärdenfors, "On Definitions of Manipulation of Social Choice Functions," in J.J. Laffont, ed., *Aggregation and Revelation of Preferences*, North-Holland Publishing Co., 1979.

   In this clearly written paper, Gärdenfors surveys possible definitions of manipulation for social choice rules that produce multiple winners, or multiple best alternatives. (One example is an election procedure that allows ties. Another is a rule that generates a whole set of good alternatives, like the rule that defines the set of Pareto optima.) When there are multivalued choice sets, manipulation has many possible interpretations, since there are many senses in which an individual might prefer one set of alternatives to another set.

*6. A. Gibbard, "Manipulation of Voting Schemes: A General Result," *Econometrica*, V. 41, 1973, pp. 587-601.

   This paper has the first published version of the Gibbard-Satterthwaite Theorem. Actually, Gibbard proves something more general. A game form is a rule that transforms strategies of all the players into outcomes, so an SCF is a special type of game form, in which the only

playable strategies are preference relations. And a game form is straightforward if each player always has a dominant strategy, that is, a strategy that will produce a best outcome for him no matter what the other players do.

Gibbard proves that no nontrivial game form is straightforward. The result on what we have called SCF's and what Gibbard calls voting schemes follows immediately.

The first half of the paper is nonmathematical and very clear.

*7. A. Gibbard, "Manipulation of Schemes that Mix Voting with Chance,' *Econometrica*, V. 45, 1977, pp. 665-681.

Some SCF's (or voting schemes, in Gibbard's terms) mix voting and lotteries. For instance, a two-way tie for first place might be broken by the flip of a coin. Or, voting might work this way. Everyone might write down the name of his favorite candidate. Let p_A = the proportion of ballots with A's name, p_B = the proportion of ballots with B's name, and so on. All the ballots might be mixed together, and the winner might be drawn at random from the mixed ballots. Then A's probability of winning would be p_A, etc.

Gibbard characterizes nonmanipulable mixed schemes in this paper.

8. T. Groves and J. Ledyard, "Some Limitations of Demand Revealing Processes," *Public Choice*, V. XXIX-2, 1977, pp. 107-124.

See the Chapter 8 References section.

*9. E. Maskin, "Nash Equilibrium and Welfare Optimality," Mimeo, MIT, 1977.

To the best of our knowledge, the term "implementation" is first used in this classic paper. See reference 4 in Chapter 15 for more information.

*10. M.A. Satterthwaite, "Strategy-Proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions," *Journal of Economic Theory*, V. 10, 1975, pp. 187-217.

This is Satterthwaite's original version of the Gibbard-Satterthwaite Theorem. Unlike Gibbard's, Satterthwaite's version of the proof does not rely on the Arrow Impossibility Theorem. In fact, Satterthwaite shows that his theorem can be used to provide a new proof of Arrow's Theorem.

11. R. Serrano "The Theory of Implementation of Social Choice Rules," *SIAM Review* V. 46, 2004, pp. 377-414.

This is a survey of implementation theory. A general proof of the Gibbard-Satterthwaite theorem is provided. See also the references to alternative recent proofs in this survey.

12. T. Nicolaus Tideman, ed., *Public Choice*, V. XXIX-2, Special Supplement to Spring, 1977.

    See the Chapter 8 References section.

Chapter 15

NASH IMPLEMENTATION

1. Introduction

In this chapter we continue to explore issues of manipulation of SCF's. In the last one we stumbled with an important impossibility result, the Gibbard-Satterthwaite theorem. Recall that this theorem establishes that, when there are at least three alternatives, there is no SCF that is nontrivial, universal, nonmanipulable and nondictatorial.

Given this negative result, one is forced to relax some of its assumptions in the hope of finding more encouraging news. For example, if society is choosing between only two alternatives, one can easily find SCF's that satisfy all these properties (such as majority voting). A second way out is based on the relaxation of universality: if the Central Authority has determined that certain kinds of preferences can be ruled out, possibility results arise. For example, if we assume that preferences are additively separable (as we did in Chapter 8), one can succeed in finding demand revelation tax schemes: these were nontrivial, nondictatorial and nonmanipulable SCF's.

The third way out of the impossibility theorem is the one that will be dealt with in this and the next chapter. The key will be a change in the way we check for manipulability. Recall that we use R_i to denote person i's preference relation, P_i to denote his strict preference relation, and I_i his indifference relation. Thus far, an SCF F was defined as manipulable if one could find an agent i and preferences $(R_1, \ldots, R_{i-1}, R_{i+1}, \ldots, R_n)$ for the others –we shall use the notation R_{-i} to denote the profile of preferences of all persons but i– such that, when the true preferences of agent i were R_i, he'd rather report R_i' instead, i.e., $F(R_i', R_{-i})P_iF(R_i, R_{-i})$. In other words, for nonmanipulability or strategy proofness, we were

requiring that, regardless of what the true R_{-i} are, agent i not have an incentive to misrepresent his preferences. This means that reporting the true preferences must be a dominant strategy in the mechanism in which each agent is asked to report his preferences.

Much weaker than requiring truth telling regardless of the other reports is to require truth telling when the others are also telling the truth, and this is what we will utilize in these two chapters. Formally, this is done by appealing to the game theoretic notion of an equilibrium, which was already used in Chapter 8 when the Groves-Ledyard tax scheme was covered. In the current chapter we shall assume complete information among the agents, and the corresponding equilibrium notion is that of Nash equilibrium. In the next chapter this assumption will be relaxed and we shall consider incomplete information environments, in which Bayesian equilibrium will be used.

2. An Example

We begin with an example that should be familiar. We go back to the model of a market economy of Chapter 3. For simplicity, we rule out production considerations. That is, our example will be based on a pure exchange economy with no externalities.

Consider the following two-agent two-goods exchange economy. Suppose the initial endowments of the commodities are $\omega_1 = (3, 9)$ for agent 1, and $\omega_2 = (9, 3)$ for agent 2. This is known by the Central Authority. Everyone knows, including the Central Authority, that agent 1's preferences can be represented by the utility function $u_1(x_{11}, x_{12}) = x_{11}x_{12}$. However, the Central Authority (but not agent 1!) is uncertain about agent 2's preferences. For simplicity, suppose that there are only two possibilities: either agent 2's preferences can be represented by the utility function $u_2(x_{21}, x_{22}) = x_{21}x_{22}$ (with a neutral position towards both goods), or they can be represented by $v_2(x_{21}, x_{22}) = x_{21}^2 x_{22}$ (showing a predisposition for good 1).

Suppose that the SCF that the Central Authority would like to implement is the Walrasian or competitive market equilibrium allocation. We next calculate it.

Suppose the economy is the one described by the utility functions u_1 and u_2. Then, letting p represent the competitive price of good 1, and assuming that the price of good 2 is normalized at 1, the competitive equilibrium is described by the following six equations. The first two give agent 1's optimal choice, the third and fourth agent 2's, and the last two take care of market clearing:

$$\begin{aligned}
\mathrm{MRS}_{u_1} &= \frac{x_{12}}{x_{11}} = p; \\
px_{11} + x_{12} &= 3p + 9; \\
\mathrm{MRS}_{u_2} &= \frac{x_{22}}{x_{21}} = p; \\
px_{21} + x_{22} &= 9p + 3; \\
x_{11} + x_{21} &= 12; \\
x_{12} + x_{22} &= 12.
\end{aligned}$$

The reader can check that the solution to this system of equations is $p = 1$, $(x_{11}, x_{12}) = (6, 6)$ and $(x_{21}, x_{22}) = (6, 6)$. Thus, at the competitive equilibrium when agent 2's utility function is u_2, goods are exchanged in a one-to-one ratio and the final allocation is the center of the Edgeworth box.

Let's now calculate the competitive equilibrium when agent 2's utility function is v_2. The corresponding conditions are:

$$\begin{aligned}
\mathrm{MRS}_{u_1} &= \frac{x_{12}}{x_{11}} = p; \\
px_{11} + x_{12} &= 3p + 9; \\
\mathrm{MRS}_{v_2} &= \frac{2x_{22}}{x_{21}} = p; \\
px_{21} + x_{22} &= 9p + 3; \\
x_{11} + x_{21} &= 12; \\
x_{12} + x_{22} &= 12.
\end{aligned}$$

The solution is $p = \frac{13}{9}$, $(x_{11}, x_{12}) = (\frac{60}{13}, \frac{20}{3})$, $(x_{21}, x_{22}) = (\frac{96}{13}, \frac{16}{3})$. That is, given agent 2's stronger preference for good 1, it become more valuable and now more units of good 2 have to be paid per unit of good 1 in the market. Agent 2 ends up consuming substantially more good 1 than before, while agent 1 ends up with more good 2.

If the Central Authority knew agent 2's preferences, it would want to implement the allocation $((6, 6), (6, 6))$ when his utility function is u_2, and $((\frac{60}{13}, \frac{20}{3}), (\frac{96}{13}, \frac{16}{3}))$ when it is v_2. Unfortunately, the Central Authority does not have this information. It could try to ask agent 2 what his preferences are, much in the spirit of what we were doing in the last chapter. But then, it would not receive a truthful report. The reason is that $u_2(96/13, 16/3) > u_2(6, 6)$. That is, when his true utility function is u_2, agent 2 has an incentive to report that it is v_2.

In the jargon of the last chapter, the competitive equilibrium SCF is not strategy proof. The question we ask now is whether the Central

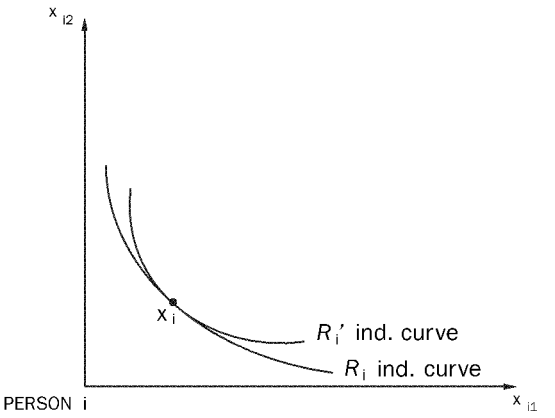


Figure 15.1.

Authority can exploit the fact that agent 1 knows agent 2's true preferences. Perhaps a more sophisticated mechanism can be designed where this bit is used in order to elicit the information truthfully.

As will be shown in the next section, the key to a successful answer to this question is a new requirement on SCF's that was first introduced in Maskin (1999). It is called Maskin monotonicity.

3. Maskin Monotonicity

We begin by defining the central condition in this chapter.

Maskin monotonicity. Suppose that the SCF assigns alternative x when the preference profile is $R = (R_1, R_2, \ldots, R_n)$, i.e., $F(R) = x$. If the preferences of each individual i change from R_i to R_i' in a monotonic way around x (that is, whenever xR_iy, one has that $xR_i'y$), then the alternative socially chosen should not change: $F(R') = x$.

Maskin monotonicity is related to the condition of NIM used in the last section of Chapter 13 (we will be specific about how both relate in the appendix to this chapter). The meaning of Maskin monotonicity is illustrated in Figure 15.1. In it, we are representing two indifference curves for an agent in an exchange economy with two goods. In such an example, an "alternative" is simply a "feasible allocation" of goods in the economy. Then, the change in preferences contemplated by the Maskin monotonicity requirement means that the lower contour set at x expands. (Recall from Chapter 1 that the lower contour set of R_i at x is the set $\{y|xR_iy\}$.) To reiterate, suppose that the alternative initially chosen is x. Further, suppose preferences change, but in such a way that for no individual it is true that x has fallen with respect to any other

alternative in his personal ranking. Then, Maskin monotonicity says that the social choice should remain x. Of course, if preferences change in any other way (so that at least for one agent the lower contour sets at x are not nested as in Figure 15.1), Maskin monotonicity does not restrict the social choice at all.

Let's now go back to the kind of exchange economy that we were dealing with in our example. Specifically, consider a class of exchange economies with n agents. Assume there are m infinitely divisible commodities; each agent holds positive amounts of each good in his initial endowment and final consumption bundles are represented by m-dimensional vectors with all entries being nonnegative. Suppose preferences are represented by monotonic utility functions whose indifference curves are convex. Suppose also that we have made enough assumptions to guarantee that there is only one competitive equilibrium allocation and that this prescribes positive consumption of all goods for each agent (we will have more to say about this assumption at the end of the chapter).

We now claim that, over the considered class of exchange economies, the competitive equilibrium SCF satisfies Maskin monotonicity. To see this, consider an economy in which $\{1, \ldots, n\}$ is the set of agents, agents' preferences are $R = (R_1, \ldots, R_n)$ and their initial endowments are $(\omega_1, \ldots, \omega_n)$. Denote the competitive equilibrium SCF by F, and let $F(R) = x$, i.e., $x = (x_1, \ldots, x_n)$ is the unique competitive equilibrium allocation of this economy, where x_i is the final consumption bundle assigned to agent i in equilibrium. This implies that there exist competitive equilibrium prices p such that the bundle x_i is the optimal choice for each agent over the budget set determined by prices p and endowment ω_i. Furthermore, $\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} \omega_i$.

Next consider an economy consisting of the same agents with the same initial endowments, but in which preferences for agent i have gone through a monotonic change around x. That is, for any bundle $y_i P_i' x_i$, it was already true that $y_i P_i x_i$. Now, because x was a competitive equilibrium allocation in the economy with preferences R, any such bundle lies outside of the budget set for agent i and prices p. Therefore, for each agent $i = 1, \ldots, n$ x_i is also his optimal choice over the budget set determined by p and ω_i, even when his preferences are R_i'. Since $\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} \omega_i$, i.e., market clearing still holds, x is a competitive equilibrium allocation in the economy with preferences R': $F(R') = x$. But this means that the competitive equilibrium SCF F satisfies Maskin monotonicity, as we claimed.

4. Maskin's Theorem

The relevance of the requirement of Maskin monotonicity comes from Maskin's theorem, which we will present in this section. But first, we introduce some important notions.

To begin with, let's set up the implementation problem. There is a set of agents $N = \{1, \ldots, n\}$, a set of social alternatives A, and each agent has a preference relation R_i over the set A. Suppose the Central Authority wishes to implement a given SCF F for this society. Suppose the Central Authority knows what is feasible, i.e., it knows the set A. However, it does not know the preferences $R = (R_1, \ldots, R_n)$ of the individuals over the feasible alternatives. Thus, if the true preferences are R, it would like to implement alternative $F(R)$. But the question is: how can this be done without knowing R? Of course, there are certain SCF's for which this is an easy problem. If the SCF is constant, so that $F(R) = x$ for a fixed x no matter what R is, the Central Authority can simply enforce x. But SCF's that we are interested in for societies are not like this: we want our SCF to be sensitive to the preferences of individuals in society.

Next, we need to talk about how the agents will communicate with the Central Authority. In the analysis of the previous chapter, the implicit communication was one in which each agent reported his preferences to the Central Authority. Thus, if agent i's true preferences were R_i, he chose to report $\hat{R}_i$, where this could be a truthful or nontruthful report. With the information collected, i.e., $\hat{R} = (\hat{R}_1, \ldots, \hat{R}_n)$, the Central Authority implemented $F(\hat{R})$. This is a particular kind of communication, but there is no reason to restrict attention to such schemes: in an auction or market context, agents are asked to put a bid on the table instead of reporting their entire preferences, or in voting, agents may be asked simply for their top-ranked alternatives instead of having to report their entire rankings. This leads to the general notion of a mechanism.

A *mechanism* or *game form* G is a pair $G = ((M_i)_{i \in N}, g)$, in which M_i is a set of messages $\hat{m}_i$ that agent i can send to the Central Authority, and g is the outcome function. The function g collects the profile of messages $\hat{m} = (\hat{m}_1, \ldots, \hat{m}_n)$ sent by the agents, and delivers a feasible outcome $g(\hat{m}) \in A$.

A mechanism is also called a *game form*, not a game, because its specification is completely independent of agents's preferences or payoffs. Hence, a game form maps profiles of messages into feasible outcomes. In contrast, a *game* assigns a profile of payoffs or utilities to each profile of messages. Recall that the Central Authority knows the set A, but not the agents' preferences over A. Thus, while the Central Authority

can design a game form, it cannot design a game (as the latter requires information that it does not have).

In this chapter, we assume that there is *complete information* among the agents of the set N. That is, each agent i knows his own preferences R_i, and also knows the preferences R_{-i} of the others. Further, this is common knowledge among them: everyone knows everything, everyone knows that everyone knows everything, everyone knows that everyone knows that everyone knows everything, and so on. The complete information assumption is not often met in reality, although it is plausible if one is considering societies with a small number of individuals, who know each other well, much better than an outsider (the Central Authority) may know them. (E.g., think of a small community of neighbors upon whom Town Hall is trying to levy taxes. Or of a union of countries like the European Union from whom the United Nations may want to elicit payments to fund programs in the developing world. Or consider a father who wants to bring about the optimal allocation of toy time sharing among his two daughters, who have a much better idea about their true preferences for toys.) In general, outside of cases like these, we view the complete information assumption as a way to approximate situations in which informational asymmetries among the agents are small, compared to informational asymmetries between the agents on the one hand and the Central Authority on the other hand.

What is interesting now is that the mechanism or game form designed by the Central Authority to govern the communication with the agents is taken by these as a true game. That is, each of them knows his own preferences or payoffs, as well as those of the others. Thus, in choosing the message $\hat{m}_i$ out of the set M_i, strategic considerations in this game of complete information matter. In fact, because of this, we shall use the word *strategy* to talk about a message sent by agent i.

How does agent i choose his strategy $\hat{m}_i$ in the mechanism? Well, of course the answer will depend on his preferences over outcomes, as well as on the specification of the outcome function. Thus, a more appropriate question is: how does he choose his strategy in the game induced by the mechanism G when the preferences are R?

In such a game, he could choose a strategy whenever it is dominant. Strategy $\hat{m}_i$ is dominant whenever $g(\hat{m}_i, m_{-i})R_i g(m_i, m_{-i})$ for all messages $m_{-i} = (m_1, \ldots, m_{i-1}, m_{i+1}, \ldots, m_n)$ sent by the others and for all $m_i \in M_i$ (with at least one strong preference for each m_i). That is, agent i can never go wrong by choosing $\hat{m}_i$ if it is dominant, because, regardless of what the others choose to do, $\hat{m}_i$ always yields an outcome that is at least as good (and sometimes strictly better) than the outcome produced by any other message sent by agent i. This is a very strong

property: often dominant strategies do not exist. In most games, agents will not have it so easy, and their optimal strategy will not be independent of what the others do. Moreover, insisting on dominant strategies in the implementation problem leads us to the Gibbard-Satterthwaite impossibility theorem.

Since the concept of dominant strategies is too demanding, we shall weaken our game theoretic solution concept to that of an equilibrium.

A *Nash equilibrium* of the game induced by the mechanism G and preferences R is a profile of messages $m^* = (m_1^*, \ldots, m_n^*)$ such that for every $i \in N$ $g(m_i^*, m_{-i}^*) R_i g(m_i, m_{-i}^*)$ for every $m_i \in M_i$. That is, at a Nash equilibrium, each agent is choosing a strategy that is optimal given what the others are choosing. This is far weaker than requiring that agent i's strategy be optimal for *any* strategies chosen by the other agents. See our discussion of the concept in the section concerning the Groves-Ledyard tax scheme in Chapter 8.

We shall say that the SCF F is *Nash implementable* whenever one can design a mechanism $G = ((M_i)_{i \in N}, g)$ such that, for every possible preference profile R and for every Nash equilibrium m^* of the game induced by the mechanism G when the preferences are R, $g(m^*) = F(R)$.

Note how Nash implementability does not require that the equilibrium of the game be unique. Indeed, multiple equilibria –strategy profiles– are possible, but all of them must have the same outcome, which should be the socially desirable one. In this chapter our requirement of foolproofness of an SCF amounts to Nash implementability. That is, given that the use of dominant strategies is limited by the Gibbard-Satterthwaite theorem, can one design a mechanism such that, regardless of what the true preferences R are, all Nash equilibria of the corresponding induced game coincide with the one the Central Authority would like to implement if it knew R?

Another way to put the question is this: Can one describe the requirements on SCF's that are equivalent to Nash implementability? The answer is "yes," and the solution was provided by Eric Maskin. We present it in the next two results. The first result will deal with necessity and the second with sufficiency.

Maskin's Theorem 1; Necessity: If the SCF F is Nash implementable, it satisfies Maskin monotonicity.

Proof: Since the SCF F is Nash implementable, there exists a mechanism $G = ((M_i)_{i \in N}, g)$ that, when the true preferences are R, has a Nash equilibrium $m^*(R)$ whose outcome is the alternative specified by F: $g(m^*(R)) = F(R) = x$.

Now suppose there is a monotonic change of preferences around x. That is, the new preference profile is R' such that for every $i \in N$, if $yP_i'x$, it was already true that yP_ix. Since M^* was a Nash equilibrium of G under preferences R, any unilateral deviation from m^*, such as (m_i, m^*_{-i}) was producing an outcome $g(m_i, m^*_{-i}) = z$ such that xR_iz. Therefore, since the mechanism does not vary with a change in preferences, the same outcome z results following the message profile (m_i, m^*_{-i}) when preferences are R'. Because preferences have changed in a monotonic way around x, we know that $xR_i'z$. And since this holds for any unilateral deviation from m^*, this shows that m^* is also a Nash equilibrium of G when preferences are R'. Thus, if preferences are R', we have that x is a Nash equilibrium outcome of the mechanism. But since F is Nash implementable, it must be the case that $F(R') = x$, and then F satisfies Maskin monotonicity. Q.E.D.

Next we state and prove the other direction, almost a converse of the first result. As will become clear in the proof, the mechanism proposed makes use of the Japanese proverb "the nail that sticks up gets hammered down."

Maskin's Theorem 2; Sufficiency: Suppose there are at least three agents and the environment includes a private good. Then, if the SCF F satisfies Maskin monotonicity, it is Nash implementable.

Proof: The proof is based on the construction of a canonical mechanism that will work for any SCF F satisfying Maskin monotonicity, regardless of the implementation problem (exchange economy, production economy, allocation of public goods, voting, etc.)

Consider the following mechanism $G = ((M_i)_{i \in N}, g)$, in which each message $m_i \in M_i$ allowed to agent i consists of an alternative, a preference profile and a nonnegative integer. Thus, a typical message sent by agent i is denoted $m_i = (a^i, R^i, z^i)$. To be clear, $a^i \in A$ is an alternative, $R^i = (R^i_1, \ldots, R^i_i, \ldots, R^i_n)$ is the preference profile (preferences of all agents) reported by i, while $z^i = 0, 1, 2, \ldots$ is a number that i chooses. The outcome function g of the mechanism is defined with the following three rules, where $m = (m_1, \ldots, m_n)$:

(i) If all agents announce the same thing, $m_i = (a^i, R^i, z^i) = (a, R, 0)$ for all $i \in N$, and $F(R) = a$, then $g(m) = a$.

(ii) If there is an almost unanimous announcement as in part (i), i.e., if $n-1$ agents announce $m_i = (a, R, 0)$ with $F(R) = a$, but agent j announces $m_j = (a^j, R^j, z^j) \neq (a, R, 0)$, then we can have two

cases:

-If aR_ja^j, then $g(m) = a^j$.
-If a^jP_ja, then $g(m) = a$.

(iii) In all other cases, an integer game is played: identify the agent who announces the highest integer (if there is a tie at the top, pick the one with lowest index among them). This person is declared the winner of the integer game and the alternative implemented is the one that he picks.

We now have to prove two things: (1) $F(R)$ is a Nash equilibrium outcome of this mechanism when the true preferences are R, and (2) there is no other outcome supported by Nash equilibria when the true preferences are R. Therefore, fix an arbitrary preference profile R and let's analyze the game induced by the mechanism G and these preferences.

First, note that the unanimous announcement $m_i = (F(R), R, 0)$ for all $i \in N$ is a Nash equilibrium. These announcements are unanimous in reporting the agents' preferences truthfully, the alternative that is socially desirable under F for these preferences and the integer 0. If these are the announcements, the outcome is decided by rule (i) and it is $F(R)$. Note that unilateral deviations from this announcement cannot induce rule (iii), but only rule (ii). So suppose agent j considers deviating from the unanimously announced message, and instead announces $(a^j, R^j, z^j) \neq (F(R), R, 0)$. The outcome would then be determined by rule (ii). But then, notice that under rule (ii) the outcome would only change to be a^j, the one that j has proposed in his deviation, if aR_ja^j according to preferences R_j. Therefore, since R_j are j's true preferences, agent j will not benefit from such a deviation. It is important to observe how the fact that there are at least three agents is used in this last step, in order to determine what is an "almost unanimous report". This is what allows the mechanism to spot the liar and use the preference R_j for player j, which is being announced by $n-1$ individuals. Note how this would not be possible if one has only two agents: rule (ii) would not be well defined. Can you see why?

Therefore, we have established that the proposed strategy profile is a Nash equilibrium, whose outcome is the "right one," i.e., when the true preferences are R, the outcome is $F(R)$. This proves our goal (1) stated above. The rest of the proof will show that there is no other Nash equilibrium outcome of this mechanism when the true preferences are R, i.e., our goal (2).

To see this, observe first that one cannot have a Nash equilibrium under either rule (ii) or rule (iii). Fix an arbitrary candidate strategy profile that falls under either of these two rules. To prove our claim is simple: if the outcome is determined by either of these two rules, at least $n-1$ individuals are not receiving their top-ranked alternative. The presence of a private good guarantees that the top-ranked alternative for each agent i, call it x^i, is different. But then, any of these individuals can profit from a unilateral deviation: let agent i in this set announce $(x^i, \cdot, z^i)$, where z^i is a large enough integer that is larger than the integers announced by all other agents. With this deviation, agent i will be declared the winner of the integer game and x^i will be implemented, contradicting that the candidate profile was a Nash equilibrium.

Therefore, if there exists another Nash equilibrium of G when the true preferences are R, its outcome must be determined under rule (i). That is, there is a unanimous announcement $m_i = (F(R'), R', 0)$ for all $i \in N$, where $R' \neq R$. That is, the agents are unanimously reporting a false preference profile, but the deception is sophisticated in that the alternative they all mention is the one proposed by F for the reported preferences, and they are also all announcing integer 0. Given these reports, rule (i) would be applied and the outcome would be $F(R')$.

Well, if $F(R') = F(R)$, this is a collective lie that does not bother the Central Authority in the least. After all, the desired alternative is still implemented, so that's fine.

However, suppose that $F(R') \neq F(R)$. Then, this is a collective lie that interferes with the social goals, so this is to be taken seriously. But recall that the SCF F satisfies Maskin monotonicity. Since the alternative chosen by F has changed when preferences are R or R', i.e., $F(R') \neq F(R)$, this implies that alternative $F(R')$ must have fallen in the preference ranking of at least one individual with respect to some other alternative, in going from R' to R. That is, there exists agent j and alternative y such that $F(R')R'_j y$ and $yP_j F(R')$.

Recall that the candidate Nash equilibrium was $m_i = (F(R'), R', 0)$ for all $i \in N$. However, consider the following unilateral deviation announced by agent j: $(y, \text{anything}, \text{anything})$. Note then that the resulting outcome is determined by rule (ii). Moreover, the outcome implemented is y, because $F(R')R'_j y$. But this is great for agent j, whose true preferences are R_j, because $yP_j F(R')$, thereby contradicting the supposition that the candidate profile was a Nash equilibrium.

In conclusion, the only Nash equilibrium outcome of the mechanism G when the true preferences are R is $F(R)$. Since R was any arbitrary preference profile, this shows that the SCF F is Nash implementable and the proof is complete. Q.E.D.

5. Comments on Maskin's Theorem

Maskin's theorem provides an almost complete characterization of the SCF's that are Nash implementable. As such, it is a fundamental result that gives the solution to the implementation problem when there is complete information among the agents and they are assumed to play optimally given what they expect the others to do. That is, agents are assumed to play according to a Nash equilibrium, in which actions and expectations confirm each other. The expectations held by each agent are correct given the equilibrium actions, and given those expectations, the action taken by each agent is optimal.

It follows from the results proved in the previous section that, for implementation problems involving at least three agents and in which there is a private good, Nash implementability of an SCF is equivalent to Maskin monotonicity. This result allows implementation theory to move beyond the impossibility theorem of Gibbard and Satterthwaite. For example, as already argued earlier in this chapter, there are interesting classes of economies in which the competitive SCF will be Maskin monotonic, and therefore, in those domains it is Nash implementable.

As pointed out in the proof of the sufficiency part of the theorem, the requirement of there being at least three agents is used in the canonical mechanism. Indeed, the case of Nash implementation for two agents requires an additional condition. The reason is simple: it is easier to catch a liar in a community of at least three agents (by pointing fingers at the liar) than in one of only two ("my word against yours").

In cases where there is no private good, it is not possible to bribe people, for example, by offering more money. Then, to preserve Maskin's Theorem 2, the requirement of *weak no veto* on the SCF must be added. This requirement says that if at least $n-1$ individuals agree that an alternative is top-ranked, the SCF should choose it. It turns out that, while weak no veto is not necessary for Nash implementability, it is sufficient, together with Maskin monotonicity, for problems where there is no private good.

Finally, we are presenting this material for single-valued SCF's, for which a single alternative is picked out for each preference profile. The theory can be extended to multi-valued SCF's. A multi-valued SCF is called a *social choice correspondence*. Maskin's theorem continues to apply. For it, the condition of Maskin monotonicity must be adapted as

follows: if x is one of the alternatives chosen by the SCF F for preferences R and there is a monotonic change of preferences around x from R to R', then x must continue to be one of the alternatives chosen by F at profile R'. Also, the definition of Nash implementability now means that one can find a mechanism such that, for each preference profile, the set of its Nash equilibrium outcomes coincides with the set selected by the SCF. With essentially the same proof as in the previous section, one can show that for problems involving at least three agents and a private good, a multi-valued SCF is Nash implementable if and only if it satisfies Maskin monotonicity. It turns out that many correspondences of interest satisfy Maskin monotonicity, including the *Pareto optimality correspondence*, the *core* and a restriction of the Walrasian equilibrium correspondence called the *constrained Walrasian correspondence* (see the exercises section).

To illustrate the use of Maskin's theorem, consider the exchange economy example that we saw earlier in this chapter, but now let's add a third agent. Recall that agent 1's utility function is $u_1(x_{11}, x_{12}) = x_{11}x_{12}$ and his endowment is $\omega_1 = (3, 9)$. Agent 2's endowment is $\omega_2 = (9, 3)$, but his utility function could be either $u_2(x_{21}, x_{22}) = x_{21}x_{22}$ or $v_2(x_{21}, x_{22}) = x_{21}^2 x_{22}$. The new agent, agent 3, has a utility function $u_3(x_{31}, x_{32}) = \min\{x_{31}, x_{32}\}$ and his endowment is $\omega_3 = (8, 8)$.

You can check as an exercise that the only Walrasian equilibrium allocation in each of the two possible economies is the same as before: as a function of agent 2's utility function, the equilibrium bundles for agents 1 and 2 are the ones given before, while agent 3 receives his endowment bundle. You can also check that the only Nash equilibrium of the canonical mechanism in each of the two possible economies has each agent reporting the true utility function of agent 2's, the Walrasian equilibrium allocation for the true economy, and the integer 0. This implements the Walrasian bundle for each agent in each of the two economies.

Therefore, the Central Authority, by making the three agents send messages using this mechanism, will be able to bring about the Walrasian equilibrium allocation in each economy.

6. Limitations of Maskin Monotonicity and Approximate Implementation

We have argued that some interesting SCF's are Maskin monotonic, but it is true that some are not. Therefore, given Maskin's Theorem 1, there will be limits to the success of Nash implementability. For instance, consider the following example, taken from the First Book of Kings in the Bible.

The wise King Solomon was presented with the following implementation problem (1st Kings, chapter 3, verses 16-28). Two women, whom we shall call A and B, each claim to be the true mother of one baby. King Solomon, the Central Authority, wants to implement the SCF that allocates the baby to its true mother. However, he does not know who the true mother is.

Let's model the problem as follows. Following the Bible story, suppose that there are three possible alternatives: a (the baby is allocated to woman A), b (the baby is allocated to B) and c (the baby is divided with a sword, with half given to A and the other half to B).

Of course, the true mother does not want to see her baby cut in half. The true mother views c as the worst outcome, but the false mother does not. We assume that the preference profile corresponding to A being the true mother is $R = (R_A, R_B)$, given by the following table (all preferences are strict):

P_A	P_B
a	b
b	c
c	a.

Similarly, preference profile $R' = (R'_A, R'_B)$ corresponds to B being the true mother, given by the table below (again all preferences are strict):

P'_A	P'_B
a	b
c	a
b	c.

In our terminology, the SCF F that King Solomon wants to implement is the following: $F(R) = a$ and $F(R') = b$. But now we claim that F violates Maskin monotonicity. Indeed, since $a = F(R) \neq F(R') = b$, for F to satisfy Maskin monotonicity, it would be necessary that in the preference change from R to R' there is an agent i and an alternative that has risen in R'_i with respect to $a = F(R)$. But this agent is nowhere to be found: for agent A, a is top ranked in both preference profiles, and for B alternative a actually rises with respect to c in going from R_B to R'_B. In other words, the preference change from R to R' is a monotonic change around a, and therefore, Maskin monotonicity would require the social choice to stay put at a, but this does not happen.

It follows from the necessity part of Maskin's theorem that the Solomonic SCF F is not Nash implementable. That is, there does not exist any

mechanism that, making use of the complete information existing between the two women, has as its only Nash equilibrium outcome the one recommended by the SCF F. In short, implementing the Solomonic SCF is not trivial. It is impossible to do in Nash equilibrium.

Fortunately for Solomon, however, the false mother acted foolishly. She announced her true preferences, with c in the middle of her ranking instead of at the bottom: "But the other [woman] said: 'Let it be neither mine nor thine, but divide it.' " So, she gave herself away as the false claimant, and then Solomon gave the baby to the true mother.

If the false mother had put c at the bottom of her ranking, like the true mother, Solomon would have failed as Central Authority. If both women were fully strategic in playing the mechanism, Solomon could not have accomplished his goal of allocating the baby to its true mother since this SCF fails Maskin monotonicity.

One way out of the limitations imposed by Maskin monotonicity is the approach of *approximate implementation.* To talk about approximate implementation, we shall introduce lotteries over alternatives. That is, if the set of alternatives is $A = (a_1, \ldots, a_k)$, we let $(q_1, \ldots, q_k)$ be a lottery over alternatives. For $j = 1, \ldots, k$, alternative a_j is implemented with probability q_j. Of course, $q_j \geq 0$ for $j = 1, \ldots, k$ and $\sum_{j=1}^{k} q_j = 1$. The interpretation is that the Central Authority may now use a random device by which each alternative is implemented with some probability.

Instead of *exact implementation*, as we had so far, in which for each preference profile R alternative $F(R) \in A$ was implemented with probability 1, we shall now speak of *approximate implementation*: for any arbitrarily small $\epsilon > 0$ and for any preference profile R, alternative $F(R)$ is implemented with probability $1 - \epsilon$.

To evaluate lotteries, we shall assume that agents have preferences that can be represented by *expected utility* functions. The reader is referred to the relevant section of Chapter 1. In the present context, for each agent i there exist numbers $u_i(a_j)$ for each $j = 1, \ldots, k$ (agent i's utilities associated with each pure alternative), such that the utility that agent i derives from a lottery $l = (q_1, \ldots, q_k)$ is its expected utility, i.e., $u_i(l) = \sum_{j=1}^{k} q_j u_i(a_j)$. Now recall that the indifference curves corresponding to expected utility are parallel straight lines. As an illustration, Figure 15.2 depicts the probability simplex for the case of three alternatives (a_1, a_2, a_3) and corresponding utilities $u_i(a_1) = 2$, $u_i(a_2) = 1$ and $u_i(a_3) = 0$ according to preferences R_i. For these preferences, the indifference curve of level $\bar{u}$ is the locus of points in the simplex whose equation is $2q_1 + q_2 = \bar{u}$. Not surprisingly, the top ranked point in the simplex is the degenerate lottery that puts all the weight on a_1, while the worst lottery is the degenerate one with all the weight on a_3.

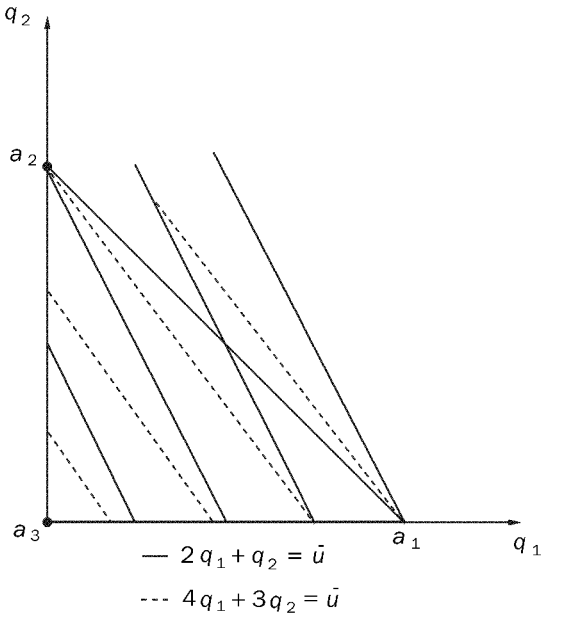


Figure 15.2.

Figure 15.2 also shows an indifference map with different expected utility preferences over lotteries. In it, $u_i'(a_1) = 4$, $u_i'(a_2) = 3$ and $u_i'(a_3) = 0$, and we call these preferences R_i'. For them, the indifference curve of level $\bar{u}$ has the equation $4q_1 + 3q_2 = \bar{u}$.

Recall that, despite the fact that u_i' is a monotone transformation of u_i, both utility functions do not represent the same preferences over lotteries.

Now consider SCF's whose range is the interior of the probability simplex, i.e., SCF's that assign to each preference profile R a lottery $F(R)$ that puts positive probability on each pure alternative. Therefore, the point $F(R)$ cannot be on the sides of the probability simplex, but must be contained in its interior. Now it is easy to see that any such SCF satisfies Maskin monotonicity. To see this, it suffices to observe that no preference change is a monotonic change around such $F(R)$: when preferences over lotteries change from R to R', it is never the case that the lower contour set of lotteries at an interior point $F(R)$ when preferences are R is contained in the lower contour set of the same $F(R)$ when preferences are R'. Therefore, Maskin monotonicity does not impose any restriction on what F should be for preferences R'.

But then, using Maskin's theorem, we have established that if there are at least three agents and a private good, every SCF whose range is in the interior of the probability simplex is (exactly) Nash implementable. This implies that every SCF (even those whose range includes the sides of the simplex) is *approximately* Nash implementable. This is a very insightful result, first discovered by Dilip Abreu, Arunava Sen and Hitoshi Matsushima in two papers written independently in the late 1980's. If we allow approximate implementation, we obtain a universal possibility result.

7. Exercises

1. In exchange economies where the initial endowment is $(\omega_i)_{i\in N}$, consider the following SCF F. Let x^* be a feasible allocation. Then, let $F(R) = x^*$ if x^* is Pareto optimal when the preferences are R, and $F(R) = (\omega_1, \ldots, \omega_n)$ otherwise. Is this SCF Maskin monotonic? Provide a proof for your answer.

2. Recall the Groves-Ledyard mechanism of Chapter 8. This was proposed in order to implement the allocation that satisfies the necessary Samuelson condition of efficiency, and it was a self-funded mechanism. Show that the SCF consisting of the described allocation rule satisfies Maskin monotonicity.

3. Consider King Solomon's problem. One alternative elaboration of the story is the following (we say it is an elaboration because the Bible does not describe the outcome that Solomon would have implemented for every possible contingency that could have happened). Suppose that Solomon thought first of using mechanism $G_1 = ((M_i)_{i=A,B}, g_1)$, where $M_i = \{\hat{A}, \hat{B}\}$ is simply a declaration of who is the true mother, and the outcome function g_1 was:

$$g_1(\hat{A}, \hat{A}) = a;$$
$$g_1(\hat{B}, \hat{B}) = b;$$
$$g_1(\hat{A}, \hat{B}) = g_1(\hat{B}, \hat{A}) = c.$$

   a Illustrate why this mechanism does not work to Nash implement the Solomonic SCF. Specifically, find the Nash equilibrium outcomes of the game induced by the mechanism when A is the true mother and when B is the true mother.

   b However, when he asked the women and the reports turned out to be $(\hat{A}, \hat{B})$, he was ready to implement outcome c by turning to his sword. Then, the true mother (say, A) changed her report

to $\hat{B}$. Now, with both reports being $\hat{B}$, Solomon did not use the outcome function g_1. Instead, he implemented $g_2(\hat{B}, \hat{B}) = a$. The Bible does not tell us the rest of the outcome function g_2, but perhaps woman B could appeal his wise decision: "eh! your Wise Majesty is not using g_1!" If this fact had been known before the women sent their messages, a different strategic analysis would have probably led to a different report profile. The problem of course is that we are not given the complete description of the mechanism that King Solomon was using.

So, complete the mechanism based on outcome function g_2 as follows:

$$g_2(\hat{A}, \hat{A}) = g_2(\hat{B}, \hat{B}) = a;$$
$$g(\hat{A}, \hat{B}) = g_2(\hat{B}, \hat{A}) = c.$$

Evaluate this mechanism from the point of view of implementation theory if one wishes to use it in order to implement the Solomonic SCF.

4. Show that the weak Pareto correspondence of any implementation problem satisfies Maskin monotonicity. (The weak Pareto correspondence when preferences are R prescribes the set of all its weak Pareto optimal alternatives, i.e., the set of feasible alternatives x such that there does not exist another feasible alternative y that every agent *strictly* prefers to x).

5. Show that the Pareto correspondence of any implementation problem need not satisfy Maskin monotonicity. (The Pareto correspondence when preferences are R prescribes the set of all its Pareto optimal alternatives as defined in Chapter 2, i.e., the set of feasible alternatives x such that there does not exist another feasible alternative y that every agent *weakly* prefers, and at least one agent *strictly* prefers, to x). Show, however, that in exchange economies with continuous and monotone preferences the Pareto correspondence satisfies Maskin monotonicity.

6. Recall the core, a solution concept introduced in Chapter 2, as the set of coalitionally stable allocations of an exchange economy. Define its weak version based on strict blocking (as we have just done in exercise 4 for the Pareto correspondence). Show that the weak core correspondence satisfies Maskin monotonicity.

7. Show that the correspondence that assigns to each exchange economy the set of all its Walrasian or competitive equilibrium allocations

may violate Maskin monotonicity. On the other hand, define the constrained Walrasian equilibrium correspondence as follows. Let A_i denote the set of bundles x_i for agent i such that for each good j, $0 \leq x_{ij} \leq \sum_{i=1}^{n} \omega_{ij}$. Given prices p and agent i's endowment ω_i, his budget set $B_i(p)$ is the set of bundles $x_i \geq 0$ such that $p \cdot x_i \leq p \cdot \omega_i$. An allocation x is a constrained Walrasian equilibrium allocation if $\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} \omega_i$ and for each agent i x_i maximizes i's utility over the set $B_i(p) \cap A_i$. Show that the constrained Walrasian equilibrium correspondence satisfies Maskin monotonicity.

8. Appendix

This appendix establishes connections between the properties of independence-monotonicity and Maskin monotonicity. NIM (for neutrality-independence-monotonicity) was used in Chapter 13 to prove a general version of Arrow's impossibility theorem.

Notational preliminaries: $\bar{A}$ is the set of all alternatives. Alternatives are x, y, z, etc. $\mathcal{A}$ is the set of subsets of $\bar{A}$. A, B, C, etc. are subsets of the set of all alternatives. R_i is person i's preference relation; P_i indicates strict preference for i. $R = (R_1, R_2, \ldots, R_n)$ is a preference profile. R_s is a social preference relation; P_s indicates strict preference for society. We assume that all the R_i's are complete and transitive. A social preference relation R_s may or may not be complete and transitive.

Definitions: An *Arrow social welfare function* is a mapping from the set of preference profiles into the set of social preference relations.

An Arrow social welfare function satisfies $I-M$ (independence-monotonicity) if the following holds:

For any pair of alternatives $\{x, y\}$ and any pair of profiles R and R', if $xP_iy \Rightarrow xP'_iy$ for all i and $yP'_ix \Rightarrow yP_ix$ for all i, then $xP_sy \Rightarrow xP'_sy$.

A *social choice function* is a mapping from the set of preference profiles into the set of all alternatives $\bar{A}$. A *generalized social choice function* is a mapping from the set of preference profiles $\times\mathcal{A}$ into $\bar{A}$; in particular, it takes a preference profile R and a subset A, and produces one element of A.

We write $F(R)$ for a social choice function, and $F(R, A)$ for a generalized social choice function.

We say a social choice function $F(R)$ satisfies *Maskin monotonicity* if for any pair of alternatives $\{x, y\}$ and any pair of profiles R and R', if $xR_iy \Rightarrow xR'_iy$ for all i, then $F(R) = x \Rightarrow F(R') = x$. Given that individual preference relations are assumed complete, $xR_iy \Rightarrow xR'_iy$ is equivalent to $yP'_ix \Rightarrow yP_ix$. Therefore $F(R)$ satisfies Maskin monotonicity if for any pair of alternatives $\{x, y\}$ and any pair of profiles R and R', if $yP'_ix \Rightarrow yP_ix$ for all i, then $F(R) = x \Rightarrow F(R') = x$.

We say that a generalized social choice function $F(R, A)$ satisfies *Maskin monotonicity* if for any pair of alternatives $\{x, y\} \subset A$, and any pair of profiles R and R', if $yP_i'x \Rightarrow yP_ix$ for all i, then $F(R, A) = x \Rightarrow F(R', A) = x$.

A generalized social choice function $F(R, A)$ can be used to define a strict social preference relation P_s as follows: Say xP_sy whenever $F(R, \{x, y\}) = x$. We'll call P_s the social preference relation *induced* by $F(R, A)$. Such a social preference relation can be generated for any preference profile R. The rule that transforms preference profiles into social preference relations in this fashion will be called the Arrow social welfare function *induced* by $F(R, A)$.

Proposition 1. Suppose the generalized social choice function $F(R, A)$ satisfies Maskin monotonicity. Then the Arrow social welfare function induced by F satisfies $I - M$.

Proof: Suppose for any pair of alternatives $\{x, y\}$, and any pair of profiles R and R', the following holds:

(a) $xP_iy \Rightarrow xP_i'y$ for all i and
(b) $yP_i'x \Rightarrow yP_ix$ for all$i$.

We need to show $xP_sy \Rightarrow xP_s'y$ for the induced strict social preference relations P_s and P_s'. Let xP_sy, i.e., $F(R, \{x, y\}) = x$. We want to show that $xP_s'y$, i.e., $F(R', \{x, y\}) = x$.

At the preference profile R' restricted over the pair $\{x, y\}$, for any individual i, either $xR_i'y$ or $yP_i'x$. But in the latter case, yP_ix, by (b) above. Thus, restricted over the pair $\{x, y\}$, the change in preferences from R to R' has been a monotonic change around x: alternative y has not become strictly better than x for any agent i at the profile R'. Thus, since the generalized social choice function satisfies Maskin monotonicity, $F(R', \{x, y\}) = x$. Therefore, the Arrow social welfare function induced by F satisfies I-M. Q.E.D.

We now turn to a near converse proposition.

We will now restrict our attention to Arrow social welfare functions that map into the set of *complete*, *transitive* and *strict* social preference relations. Let P_s be the social preference relation produced by such an Arrow social welfare function. Then, $\bar{A}$ has a unique top-ranked alternative under P_s, that is, there is an x such that xP_sy for all $y \neq x$ in $\bar{A}$.

We define the social choice function $F(R)$ *induced* by the Arrow social welfare function by $F(R) = x$, where x is the top-ranked alternative under P_s.

In the proposition that follows, we also assume $\bar{A}$ is the set of allocations in an exchange economy, and we make the assumptions that we had made in Chapter 3 regarding individuals' preferences. That is, we assume that all the individuals have continuous, monotonic, and self-interested preferences.

Proposition 2. Let $\bar{A}$ be the set of allocations in an exchange economy. Suppose an Arrow social welfare function maps into the set of complete, transitive and strict social preference relations. Let $F(R)$ be the induced social choice function. Assume the Arrow social welfare function satisfies I-M.

Then, $F(R)$ satisfies Maskin monotonicity.

Proof. Suppose that $F(R) = x$, which means that xP_sy for all $y \neq x$. Now suppose that there is a monotonic change of preferences around x. That is, we consider a preference profile R' such that for all agents i and all y, $yP'_ix \Rightarrow yP_ix$.

We argue by contradiction. Suppose that F does not satisfy Maskin monotonicity. This means that $F(R') = z \neq x$. By the definition of F, this implies that zP'_sx. And of course, it was the case that xP_sz.

Because the Arrow social welfare function satisfies I-M, one of the two premises in the definition of I-M (i.e., $xP_iz \Rightarrow xP'_iz$ for all i, and $zP'_ix \Rightarrow zP_ix$ for all i) must not hold. Therefore, either

[a] xP_iz and zR'_ix for some individual i, or

[b] zP'_ix and xR_iz for some i.

However, case [b] is impossible because the preference change from R to R' has been a monotonic change around x. So the only possibility is case [a]. But in exchange economies with continuous, monotonic and self-interested preferences, this case is also impossible, because it would imply that we can find another alternative $\bar{z}$ arbitrarily close to z such that $xP_i\bar{z}$ and $\bar{z}P'_ix$, which would also contradict our assumption that the preference change has been monotonic.

Thus, both cases are impossible and the proof is complete. Q.E.D.

9. Selected References

(Items marked with an asterisk (*) are mathematically difficult.)

*1. D. Abreu and A. Sen, "Virtual Implementation in Nash Equilibrium," *Econometrica* V. 59, 1991, pp. 997-1021.

This is one of the two original articles that introduce the idea of approximate or virtual implementation (see also reference 5).

*2. L. Hurwicz, E. Maskin and A. Postlewaite, "Feasible Nash Implementation of Social Choice Rules when the Designer Does not Know Endowments or Production Sets," in Ledyard, J. O. (ed.) *The Economics of Informational Decentralization: Complexity, Efficiency and Stability*, Kluwer Academic Publishers, Amsterdam, 1995.

This paper studies implementation in contexts in which the Central Authority does not necessarily know the set A of feasible alternatives. The paper contains many examples and the most relevant results concerning the Walrasian and constrained Walrasian correspondences.

*3. M. Jackson, "A Crash Course in Implementation Theory," *Social Choice and Welfare* V. 18, 2001, pp. 655-708.

This is a clear survey on implementation theory.

*4. E. Maskin, "Nash Equilibrium and Welfare Optimality," *Review of Economic Studies* V. 66, 1999, pp. 23-38.

This is Maskin's classic paper, which contains his theorem. The first version of this important article circulated as an MIT working paper in 1977. Evidently, Maskin had some difficulties implementing its publication. Publishing in economics sometimes is associated with interesting funny stories, like this.

*5. H. Matsushima, "A New Approach to the Implementation Problem," *Journal of Economic Theory* V. 45, 1988, pp. 128-144.

This is the other original paper on approximate implementation (see also reference 1).

*6. J. Moore, "Implementation, Contracts and Renegotiation in Environments with Complete Information," in J. J. Laffont (ed.) *Advances in Economic Theory, VI World Congress of the Econometric Society* (vol. I), Cambridge University Press, 1992.

This is an excellent survey on implementation theory under complete information. It is divided into two parts, and the first is less technical. Its coverage of the King Solomon problem is delightful.

*7. R. Repullo, "A Simple Proof of Maskin Theorem on Nash Implementation," *Social Choice and Welfare* V. 4, 1987, pp. 39-41.

This paper contains the proof of the sufficiency part of Maskin's theorem that we have presented.

*8. R. Serrano, "The Theory of Implementation of Social Choice Rules," *SIAM Review* V. 46, 2004, pp. 377-414.

In this survey, one can find a specific mechanism that approximately implements the Solomonic SCF in Nash equilibrium. This can be done, even though the problem involves only two agents.

Chapter 16

BAYESIAN IMPLEMENTATION

1. Introduction

This will be our last chapter on the theory of implementation. We continue to study the issue of manipulation of SCF's. The game theoretic solution concept that we shall employ, as we did in Chapter 15, is that of equilibrium. However, we shall not assume that agents have complete information about one another. We'll consider, instead, *incomplete information* environments. In such environments, the game theoretic equilibrium concept is called *Bayesian equilibrium.*

Many "real world" situations fit well the assumption of incomplete information. For example, a person may know his own preferences over which people to date, but this is private information, unknown by his roommates. Often people are not completely informed about their own preferences. Consider, for example, a bidder in an auction for an oil drilling permit. There is uncertainty about the true amount of oil under ground. If the bidder knew it exactly, he would know his willingness to pay for the permit. But he doesn't know; all he can do is to rely on the opinions of expert geologists he hires. These experts perform studies that inform him about how much oil each expert believes is recoverable. In the terminology of Chapter 6, this bidding agent receives a *signal* (the report of the expert that he hires) about the true *state of the world* (the true amount of oil). Then, given his signal, he is left with a "noisy" willingness to pay. (With some probability, he believes there is much oil and the drilling right is very valuable, while with some probability he wouldn't pay much for it.)

The uncertainty need not be about the agents' preferences; it might be intrinsic uncertainty about the state of the world. Suppose for exam-

ple that there is only one good –money– in each state, and suppose that the utility function of each agent in each state is the same. Agents differ, however, in their information. (Some may know the true state with certainty, while others might assign different probabilities to each state.) In this context, differences in agents' information result in different interim preferences across agents in society. The Central Authority may then want to implement different alternatives in different states of the world (perhaps optimal insurance contracts, which are called for even if ex-post preferences are identical across agents and states).

To unify the treatment of all the previous examples, rather than the Central Authority not knowing the agents' preferences (as we had in Chapters 14 and 15), it is more general to say that it does not know the agents' information, whatever that may involve (preferences, beliefs, or both). This means that the Central Authority does not know the state of the world. And in general, the SCF it wants to implement is a mapping that prescribes a feasible alternative in each state.

In this chapter we return to the ex-ante, interim, ex-post model of Chapter 6. We focus on the *interim stage*, the stage at which there is asymmetric information among the agents. Each agent has received his private signal about the true state, and this information is not commonly known. Instead of talking about signals, we will use the notion of types, following Harsanyi's (1967, 1968) formalization of games of incomplete information.

A *type* of an agent is simply the description of all his private information. This is largely a matter of jargon, but as we shall see, it will prove very useful. Instead of defining the interim stage as that stage in which "each agent has received his signal," we shall define it as the stage in which "each agent has learned his type." Therefore, agents, like the Central Authority, may not know what state will arrive ex-post. However, each agent knows his own type, which is more than the Central Authority knows. The novelty of this chapter is that asymmetric information is also present among the agents of society. We turn to a description of the general model of types that we shall employ.

2. Asymmetric Information in the Model of Types

Our task in this section is to explain the model of types, and how it connects with the model based on signals that we followed in Chapter 6.

Consider Example 1, as follows. There are three equally likely states of the world $\{1, 2, 3\}$ and two agents, agent 1 and agent 2. Suppose at the interim stage agent 1 becomes fully informed about the state that

will prevail ex-post, while agent 2 remains completely uninformed. In the jargon of Chapter 6, agent 1 can receive one of three possible signals: σ_1 if the state is 1, σ_1' if the state is 2, and σ_1'' if it is 3, while agent 2's unique signal, σ_2, is completely uninformative. Now, we shall say that there are three possible types of agent 1: type t_{11} is the one that knows that the state is 1, type t_{12} is the one that knows that the state is 2, and type t_{13} the one that knows that the state is 3. And there is only one type of agent 2, who is uninformed. To continue with the oil drilling story, suppose there are three possible states, consisting of the amount of recoverable oil being low, medium or high. Agent 1 is lucky to hire a competent geologist, who learns the state and informs him. Agent 2 cannot afford to hire any expert, and cannot update his ex-ante beliefs with any new information. Each agent knows about each other's access to new information. Agent 1 knows that agent 2 has no new information. Agent 2 knows that the geologist hired by agent 1 is excellent and will find out the true state for sure, whatever that is.

We shall now identify a state of the world with a profile of types, one for each agent. That is, state 1 is the profile (t_{11}, t_{21}), state 2 is (t_{12}, t_{21}) and state 3 is (t_{13}, t_{21}). This identification implies that at each ex-post state no uncertainty remains. That is, at the ex-post stage, by observing each other's types, agents learn everything there is to be learned. If this is not the case, we would be considering underlying residual uncertainty beyond the information held by anyone in the model. In this case, agents would still use expected utility to evaluate outcomes at the ex-post stage. Nothing essential is different, and hence, we assume this case away. Thus, we will denote a typical state of the world by t and the set of all states by T.

An SCF F is a mapping from the set T of states of the world (or type profiles) to the set A of feasible alternatives. Thus, $F(t_{11}, t_{21})$ denotes the alternative chosen by F in state (t_{11}, t_{21}), $F(t_{12}, t_{21})$ the alternative chosen by F in state (t_{12}, t_{21}), and so on. Implicit in this definition is that the Central Authority, like the agents, knows the model (including the set of possible types of each agent and his ex-ante beliefs). Also, if the Central Authority knew the type of an agent, it would know all his private information (his interim preferences and beliefs). However, it does not know what the true types are.

Let's suppose that each agent has an ex-ante belief over the set of states. We shall denote by $q_i(t)$ the ex-ante probability that agent i assigns to state t.

The ex-ante stage is the one that happens before agents learn their types. Continuing with the example, only agent 1 changes his beliefs from the ex-ante stage to the interim stage, since agent 2 does not learn

anything by learning his type. How do agents evaluate an SCF F at the ex-ante stage? They simply calculate ex-ante expected utility. That is, for $i = 1, 2$, the ex-ante utility of F is:

$$u_i(F) = \sum_{t \in T} q_i(t) u_i(F(t), t).$$

In the example, each $q_i(t) = 1/3$ for $i = 1, 2$ and for all $t \in T$, because all three states are equally likely. Note also how the ex-post utility function may change with the state, which explains its second argument. Thus, the ex-ante utility of F is simply the expected utility that will be derived from F in each state, and the probabilities attached to each state are the ex-ante beliefs.

As noted above, the interim stage, the one we shall be concerned with here for the most part, happens when each agent learns his type. Learning one's type results in the corresponding updating of one's beliefs, using Bayes' rule. Each agent i updates his beliefs upon learning his type t_i. That is, letting $t_{-i} = (t_1, \ldots, t_{i-1}, t_{i+1}, \ldots, t_n)$ represent an arbitrary profile of types for agents other than i, we write type t_i's interim beliefs as follows:

$$q_i(t|t_i) = \begin{cases} \frac{q_i(t)}{\sum_{t_{-i}} q_i(t_i, t_{-i})} & \text{for states compatible with } t_i, \\ 0 & \text{otherwise} \end{cases}$$

Therefore, the *interim/conditional expected utility* of F, for agent i of type t_i is given by the expression:

$$u_i(F|t_i) = \sum_{t \in T} q_i(t|t_i) u_i(F(t), t).$$

In the example, the interim beliefs for type t_{11} of agent 1 are $q_1((t_{11}, t_{21})|t_{11}) = 1$, $q_1((t_{12}, t_{21})|t_{11}) = q_1((t_{13}, t_{21})|t_{11}) = 0$, which means that he is fully informed about the fact that the state is (t_{11}, t_{21}). The reader can write down the interim beliefs for the other types of agent 1. Of course, for agent 2, his interim beliefs coincide with his ex-ante beliefs, i.e., $q_2(t|t_{21}) = q_2(t) = 1/3$, i.e., he is completely uninformed. Correspondingly, each type of agent 1 cares only about his ex-post evaluation of the alternative chosen by F in the state he knows is true, while agent 2 still calculates ex-ante expected utility over the three states.

Finally, the ex-post stage is now defined as that stage in which each agent learns the entire profile of types t. Again, the appropriate updating of beliefs will happen, leading to the ex-post utility evaluation of F, i.e., if state s has happened:

$$u_i(F|s) = \sum_{t \in T} q_i(t|s) u_i(F(t), t) = u_i(F(s), s).$$

In the example, agent 1's ex-post utility was already his interim utility, given his information at that point. Agent 2's ex-post utility, however, is different from his interim utility, because now he can identify the true state, and expected utility is no longer necessary.

To fix ideas better, let's turn to Example 2. Suppose there are three equally likely states at the ex-ante stage, and there are two agents. But now each agent can receive one of two possible signals. Signal σ_1 is fully informative, it tells agent 1 that the state is 1; we shall call the type of agent 1 that receives this signal type t_{11}. On the other hand, signal σ_1' lets him know that the state is not 1, but nothing else beyond that; we shall call type t_{12} the type of agent 1 that receives this signal. As for agent 2, signal σ_2 is not fully informative, but it tells agent 2 that the state is not 3; let t_{21} be the type of agent 2 that receives signal σ_2. Finally, signal σ_2' is fully informative about the state being 3, and we'll call the type of agent 2 who receives it t_{22}.

So, there are four states in principle: (t_{11}, t_{21}), (t_{11}, t_{22}), (t_{12}, t_{21}) and (t_{12}, t_{22}). However, the ex-ante probability that both agents assign to state (t_{11}, t_{22}) is 0. It is not possible that simultaneously agent 1 receives signal σ_1 and agent 2 receives σ_2', as they comprise incompatible or contradictory information. Each of the other three states receives a $1/3$ probability at the ex-ante stage.

Therefore, when agents evaluate F ex-ante, they view as irrelevant what F assigns in state (t_{11}, t_{22}) because it has zero probability. On the other hand, they weight equally each of the three possible utility levels derived from the alternatives assigned by F in the other states.

The interim stage, however, will lead to the following updating. For type t_{11}, $q_1((t_{11}, t_{21})|t_{11}) = 1$, while he believes that the other states have zero probability. Agent 1 knows the state, and knows that agent 2 knows only that the state is not 3. However, for type t_{12}, his interim beliefs are $q_1((t_{12}, t_{21})|t_{12}) = q_1((t_{12}, t_{22})|t_{12}) = 1/2$, while he believes that the other state is impossible. He knows for sure that the state is not 1. It could be 2 or 3 with the same probability. Now agent 1 is interacting with the two types of agent 2. If agent 2 is type t_{22}, agent 2 knows the state is 3. In the eyes of type t_{12}, this occurs with probability $1/2$. If agent 2 is type t_{21}, agent 2 knows the state is not 3 (according to t_{12}, this type of agent 2 also occurs with probability $1/2$). In this case, (t_{12}, t_{21}), if the agents could credibly communicate with each other, they would be able to learn the state. That is, since agent 1 knows the state is not 1 and agent 2 knows that it is not 3, they would infer that the state is 2. However, for our purposes, we shall continue to study mechanisms in which each agent sends a message to the Central Authority only, instead of communicating among themselves, which would amount to

coordinated or coalitional choice of messages. The reader can write as an exercise the interim beliefs of agent 2 in this example.

3. Incentive Compatibility

At this point it is important to recall the elements of the implementation problem. The Central Authority wishes to implement an SCF F, which is a mapping from the set T of states of the world to the set A of feasible alternatives. This is taking place in the presence of asymmetric information among the n agents of the set N, at the interim stage. We now introduce the condition of *incentive compatibility* of an SCF, which will be necessary for its implementation in incomplete information environments. To define it, it will be convenient to resort to notation employed earlier. So, as already defined above, let $t_{-i} = (t_1, \ldots, t_{i-1}, t_{i+1}, \ldots, t_n)$. Then, the condition can be stated as follows:

Incentive compatibility. For every agent i and for every type t_i,

$$\sum_{t \in T} q_i(t|t_i) u_i(F(t), t) \geq \sum_{t \in T} q_i(t|t_i) u_i(F(t_i', t_{-i}), t)$$

for every t_i'.

Before explaining incentive compatibility, we have to take a slight detour. It is instructive to view any SCF F as a simple kind of *mechanism*. In it, the set of messages available to agent i is the set of his types, say T_i, and the outcome function is simply F. Thus, if each agent i of type t_i chooses to report $\hat{t}_i$, which could be a true or a false report, the outcome implemented by this mechanism is $F(\hat{t}_1, \ldots, \hat{t}_n)$. These mechanisms whose sets of messages are the sets of types are called *direct mechanisms*, or *direct revelation mechanisms*. The direct mechanism whose outcome function is F is called the direct mechanism associated with the SCF F.

In general, a mechanism under incomplete information is the same object as we defined in Chapter 15. That is, it is a pair $G = ((M_i)_{i \in N}, g)$, where M_i is agent i's set of messages and g is the outcome function, mapping profiles of messages into A. The difference now is that different types of an agent may be sending different messages. Thus, let $m_i(t_i)$ be the message sent in the mechanism by type t_i of agent i.

The notion of equilibrium used in mechanisms with incomplete information is called Bayesian equilibrium, and it is the generalization of Nash equilibrium to these environments. We define it next.

A *Bayesian equilibrium* of the mechanism G is a profile of messages $m^* = (m_i^*(t_i))_{t_i \in T_i, i \in N}$ such that for every $i \in N$ and for every $t_i \in T_i$,

$$\sum_{t \in T} q_i(t|t_i) u_i(g(m^*(t)), t) \geq \sum_{t \in T} q_i(t|t_i) u_i(g(m_i, m_{-i}^*(t_{-i})), t)$$

for every $m_i \in M_i$.

That is, at a Bayesian equilibrium m^* of the mechanism G, each type t_i of each agent is best-responding to the others' messages by sending his equilibrium message $m_i^*(t_i)$. No type has a unilateral incentive to send a different message if all the others behave according to the equilibrium. And he should expect the others to do so, because it is consistent with their rationality if they also expect m^* to be played. This justification should sound familiar to the reader. Indeed, a Bayesian equilibrium is simply a Nash equilibrium in the game played by the types of each agent.

Now we can explain incentive compatibility of an SCF F in terms of Bayesian equilibria of direct mechanisms. That is, if all the other agents tell the truth and report their true types to the Central Authority when it is using the direct mechanism associated with F, no type of any agent has an incentive to manipulate the SCF by pretending that he is of a different type. In other words, telling the truth about one's type is a Bayesian equilibrium of the direct mechanism associated with F. Next, let's understand the relevance of the condition.

Incentive compatibility has sometimes been referred to as *informational feasibility*. This is the terminology used by Roger Myerson, one of the first proponents of the condition in the 1970's. Here's the idea: First, we have to take into account the restrictions imposed by *classical feasibility*, i.e., that total consumption not exceed the endowment (if we are considering an exchange economy), or that the outcome belong to the feasible set A (in abstract social choice settings). But now, additionally, in the presence of asymmetric information the truly feasible SCF's must satisfy incentive compatibility. For there is no hope that an SCF F that violates incentive compatibility will ever be implemented. That is, if individuals have incentives to misrepresent their types to the Central Authority, the Authority will act on the basis of incorrect information, and the alternatives prescribed by the SCF F will not be implemented.

We shall say that an SCF F is *truthfully implementable* if, in the direct mechanism associated with F, truth telling is a Bayesian equilibrium. Therefore, truthful implementability of an SCF is equivalent to its incentive compatibility. Note that this is weaker than the requirement of *full implementation* for an SCF, as we used it in Chapter 15. Full implementability requires that *all* the equilibrium outcomes of a mechanism coincide with the SCF. That is, even if F is incentive compatible, its

direct mechanism may have an additional (nontruthful) undesired equilibrium. We shall come back to this important point soon, when we address full implementation.

Recall that in this chapter agents send messages to the Central Authority in the mechanism at the interim stage (types in direct mechanisms, and, as we will see, arbitrary messages in more general mechanisms). The following examples are meant to illustrate the convenience or not of imposing incentive compatibility.

First, if the implementation of the SCF takes place also at the interim stage, it is clear that incentive compatibility is a necessary condition for its implementability. This is simply because it is necessary and sufficient for truthful implementability. For instance, think of an insurer, who plays the role of the Central Authority here, offering health insurance contracts to consumers. There are several kinds of contracts, which differ in premiums, percentage of coverage in hospitalizations, etc. The goal is that the likelihood of getting sick will make consumers self-select. The SCF here is the menu of contracts offered by the company. Each consumer, who knows his type (the general state of his health), will self-select and choose the kind of contract that he prefers. Incentive compatibility would imply, for example, that a contract with a lower premium should have a sufficiently lower level of coverage, for otherwise people who are very likely to get sick would choose it, and the company would go out of business.

Second, a different context where incentive compatibility is also necessary is that in which, while the messages are sent at the interim stage, the Central Authority implements outcomes ex-post but it cannot verify the true state of the world. In some cases, this verification may be impossible. For example, think of the shareholders of a firm (who are the Central Authority here) designing different contracts in order to hire potential managers. Again, the SCF is the menu of contracts. While all of them coincide in laying out basic fundamental features, they differ in the specifics of compensation packages and the amount of extra responsibilities that the manager is expected to fulfill. The idea is that there are two types of potential managers out there. The first type really cares about the firm and is truly concerned with the interests of its shareholders. The second is more interested in his own career. Even though the complete execution of the contract –including the payment– may well happen at the ex-post stage, after the shareholders have learned the type of the manager they are dealing with, it is very difficult to prove to an outsider, like the courts enforcing contracts, that the true type of the manager is one or the other. Therefore, it would be a terrible idea to offer a menu of contracts that fails to be incentive compatible. This

implies, for instance, that the contract that pays better should include enough additional tasks, so that the "career concerned" manager prefers not to take it.

On the other hand, there are circumstances in which incentive constraints are not necessary. Recall that the ex-post stage is that in which information becomes common knowledge among the agents, but this may or may not be true also for the Central Authority. If implementation takes place ex-post and the state is verifiable by the Central Authority, incentive compatibility ceases to be a constraint. The Central Authority can impose a draconian punishment on those who send a false report. For example, suppose the Central Authority is a Criminal Court imposing a prison sentence on a defendant, with the sentence suspended if he is not subsequently arrested on a new charge. The Court does not know whether the defendant is the type who will recommit the crime. The Court's sentence, the SCF, is contingent on an ex-post event, the commission of the new crime, which is easily observable, and a severe punishment is imposed on a defendant who reported at the interim stage that he would not reoffend, but is caught lying at the ex-post stage. There is no need to impose the incentive compatibility constraints on this SCF, because the Central Authority can discover ex-post if an agent has lied at the interim stage and make it costly for him to have done so. In the remaining sections of this chapter, though, we will rule out easy cases like these, and so, incentive compatibility will be necessary.

4. From Truthful to Full Implementability

Although truthful implementability, equivalent to incentive compatibility, is a necessary condition for implementation, it is often far from sufficient to satisfy the goals of the Central Authority. The reason is that the mechanism used may have additional unwanted equilibria, whose outcomes differ greatly from the social goal. Example 3 will suffice to make this point.

Consider an environment involving three agents and two equally likely states. Agent 1 is uninformed about the state, and agents 2 and 3 are fully informed. In our notation, the ex-ante beliefs are $q_i(t_{11}, t_{21}, t_{31}) = q_i(t_{11}, t_{22}, t_{32}) = 1/2$ for $i = 1, 2, 3$. At the interim stage, agent 1 retains this belief, while agents 2 and 3 update it to become fully informed, i.e., for $i = 2, 3$,

$$q_i((t_{11}, t_{21}, t_{31})|t_{i1}) = 1, \qquad q_i((t_{11}, t_{22}, t_{32})|t_{i1}) = 0,$$

and

$$q_i((t_{11}, t_{21}, t_{31})|t_{i2}) = 0, \qquad q_i((t_{11}, t_{22}, t_{32})|t_{i2}) = 1.$$

Let $t = (t_{11}, t_{21}, t_{31})$ and $t' = (t_{11}, t_{22}, t_{32})$. Also, let us suppose there are three alternatives: a, b and c. Assume the following preferences for the three agents.

In state t:

Agent 1	Agent 2	Agent 3
b	a	a
c	b	b
a	c	c.

In state t':

Agent 1	Agent 2	Agent 3
a	b	b
c	a	a
b	c	c.

Agents 2 and 3 have identical preferences. For them, outcome c is always the worst. But their preferences between a and b change with the state. As for agent 1's preferences, since he will be calculating expected utility, we need extra assumptions. We shall assume that for agent 1 a top ranked alternative in either state is a "moderately good" thing, a middle ranked alternative in either state is a "somewhat good" thing, and a bottom ranked alternative in either state is a "very bad" thing. It then follows that c implemented in both states is better than a in both states, and also better than b in both states.

Suppose that the SCF F that the Central Authority wants to implement is the one that maximizes the welfare of agent 1, i.e., $F(t) = F(t_{11}, t_{21}, t_{31}) = b$ and $F(t') = F(t_{11}, t_{22}, t_{32}) = a$. To do this, the Central Authority sets up a mechanism.

Suppose the Authority's first attempt is to employ the direct mechanism for F, i.e., it simply asks agents 2 and 3 for their private information. Note that agent 1 would be irrelevant in any direct mechanism since he has no information (trivially, he has only one type). Note also how we have to complete the direct mechanism associated with F by specifying what to implement if the reports from agents 2 and 3 are incompatible, i.e., (t_{21}, t_{32}) or (t_{22}, t_{31}). These reports are incompatible, because they correspond to states that have zero probability. In these cases, the Authority is getting little cooperation from agents 2 and 3, since at least one of them is lying. Then, let outcome c be implemented. (This makes good sense. Outcome c is bottom ranked by agents 2 and 3, so we want to make this uncooperative behavior costly to them.)

We represent the direct mechanism for F just described in the table below:

1's report	2's report	3's report	Outcome
t_{11}	t_{21}	t_{31}	b
t_{11}	t_{21}	t_{32}	c
t_{11}	t_{22}	t_{31}	c
t_{11}	t_{22}	t_{32}	a.

The reader can check that the SCF F is truthfully implementable, i.e., it is incentive compatible. To see this, simply convince yourself that having each type of agent 2 and 3 tell the truth, i.e., $m_2(t_{21}) = t_{21}$, $m_2(t_{22}) = t_{22}$, $m_3(t_{31}) = t_{31}$ and $m_3(t_{32}) = t_{32}$, constitutes a Bayesian equilibrium of the proposed direct mechanism associated with F. For example, take type t_{21}. In equilibrium, he is looking at the first row of the above table, and outcome b is implemented. He prefers this to outcome c, which would be implemented if he changed his report (third row of the table). The same is true for all other types. It is clear that this equilibrium indeed yields the best outcome for agent 1.

However, there is a glaring problem with this direct mechanism. There is an additional equilibrium, in which all the types of agents 2 and 3 lie to the Authority. Indeed, $m_2(t_{21}) = t_{22}$, $m_2(t_{22}) = t_{21}$, $m_3(t_{31}) = t_{32}$ and $m_3(t_{32}) = t_{31}$ is also a Bayesian equilibrium profile of messages. To see this, take for example type t_{21} again. In this equilibrium, he is placed at the fourth row of the table, leading to outcome a, and he prefers this to the outcome of the second row, which would result if he deviates and reports his true type. Again, the same can be checked for all other types, so that this utterly untruthful profile is also a Bayesian equilibrium. Moreover, both types of agents 2 and 3 strictly prefer it to the truthful equilibrium, since alternative a is implemented in state t, and alternative b in t'. Finally, to make things worse, this equilibrium produces the alternative that is bottom-ranked by agent 1 in each state.

The reader can check that there are in fact more equilibria of this mechanism. And these other equilibria also yield outcomes different from F.

It has sometimes been argued that the truthful equilibrium is more appealing, more intuitive, and that it has a "focal point" attraction that makes it more plausible than any other equilibrium. Examples like these, however, clearly show that this is a flawed argument.

Our solution to the problem will be to resort to full implementation. We will require that all the Bayesian equilibrium outcomes of the mechanism we design agree with F. Note that this is similar to what we did

in Chapter 15, where we required (for Nash implementability) that all the Nash equilibrium outcomes of the mechanism agree with F.

5. Bayesian Implementability: Example 3 Continued

As hinted above, the definition of (full) Bayesian implementability is the following:

We shall say that the SCF F is *Bayesian implementable* whenever one can design a mechanism $G = ((M_i)_{i \in N}, g)$ whose unique Bayesian equilibrium outcome is F.

Let's go back to the example of the previous section. As we see in the definition of Bayesian implementability, nothing forces us to consider only the direct mechanism for F. We can construct a mechanism of our choice, and if we are clever enough in its design, attain our goal. In fact, we show now that this is the case: the SCF F of the example is Bayesian implementable.

Consider the following mechanism. Agent 1's message is an integer z_1, chosen from the set of nonnegative integers: $M_1 = \{0, 1, 2, \ldots\}$. For agent $i = 2, 3$, the message consists of a report of his type $\hat{t}_i$ and also a nonnegative integer z_i.

The outcome function g is defined by the following rules:

(i) If agent 1 chooses $z_1 > 0$, the outcome is the one that maximizes the interim utility of the agent who announces the highest integer, with any tie-breaking rule in case of ties. Thus, if the winner of this integer game is agent 1, outcome c is implemented. If it is 2 or 3, either outcome a or b is implemented as a function of the type that the winner of the integer game reports.

(ii) If agent 1 chooses $z_1 = 0$, the outcome is determined by the direct mechanism employed in the previous section, i.e.,

(ii.1) $g(0, (t_{21}, z_2), (t_{31}, z_3)) = F(t) = b$,

(ii.2) $g(0, (t_{22}, z_2), (t_{32}, z_3)) = F(t') = a$,

(ii.3) $g(0, (t_{21}, z_2), (t_{32}, z_3)) = g(0, (t_{22}, z_2), (t_{31}, z_3)) = c$.

We will now show that this mechanism works to implement F. That is, the unique Bayesian equilibrium outcome of this mechanism is F. We prove this claim in several steps.

First, any profile $m_1 = 0$, $m_2(t_{21}) = (t_{21}, \text{anything})$, $m_2(t_{22}) = (t_{22}, \text{anything})$, $m_3(t_{31}) = (t_{31}, \text{anything})$ and $m_3(t_{32}) = (t_{32}, \text{anything})$ is a Bayesian equilibrium, and according to rules (ii.1) and (ii.2), outcome b is implemented in state t and outcome a in state t'. This is

exactly F. To see that this is a Bayesian equilibrium, note first that agent 1 obtains his top-ranked outcome, so he does not want to deviate. Given this and that the other informed agent is reporting the true type, each type of agent 2 and 3 is also best-responding by telling the truth.

Second, there is no equilibrium if $z_1 > 0$. The reason is that at least one agent will not be getting his top-ranked outcome at any of these message profiles. Therefore, a deviation by such an agent to name the highest integer is profitable. Thus, rule (i) never produces an equilibrium.

Third, $z_1 = 0$ and incompatible type reports are impossible in equilibrium. Outcome c would result, but then either agent 2 or 3 should change his type report to make it compatible with the other agent's type report. Thus, in each state, the type reports must result in a compatible (positive probability) state.

Fourth, $z_1 = 0$ and each type of agents 2 and 3 reporting their types falsely is not an equilibrium. This would implement outcome a in state t and outcome b in state t'. But then agent 1 could deviate, and announce the highest integer in order to implement outcome c.

Fifth, $z_1 = 0$ and all types of agents 2 and 3 reporting the types corresponding to the unprimed state is not an equilibrium either. The same deviation by agent 1 described in the previous paragraph would be profitable. And the same goes for $z_1 = 0$ and all types of agents 2 and 3 reporting the types of the primed state.

Therefore, the only equilibrium outcomes correspond to $z_1 = 0$ and truth telling on the part of the informed agents. This means that F is Bayesian implementable.

6. Bayesian Implementability: Theory

Example 3 is very instructive. It shows that, while incentive compatibility of F is necessary for Bayesian implementability, it is not sufficient. Establishing in general the first of these two claims is important, and we do it next.

The Revelation Principle. If the SCF F is Bayesian implementable, it satisfies incentive compatibility.

Proof: Suppose that F is Bayesian implementable. Therefore, there exists a mechanism G and a Bayesian equilibrium m^* whose outcome is F. That is, $g(m^*(t)) = F(t)$ for each state t. Because m^* is a Bayesian equilibrium, we have, for every $i \in N$ and for every $t_i \in T_i$,

$$\sum_{t \in T} q_i(t|t_i) u_i(g(m^*(t)), t) \geq \sum_{t \in T} q_i(t|t_i) u_i(g(m_i, m^*_{-i}(t_{-i})), t)$$

for every $m_i \in M_i$.

We now argue by contradiction. Suppose now that F violates incentive compatibility. This would mean that we can find a type of agent i, say t_i, that, when everyone else is telling the truth, strictly prefers to report $t_i' \neq t_i$. That is,

$$\sum_{t \in T} q_i(t|t_i) u_i(F(t), t) < \sum_{t \in T} q_i(t|t_i) u_i(F(t_i', t_{-i}), t).$$

But recall that $g(m^*) = F$ because the equilibrium m^* yields the outcome prescribed by F in each state. But then, the last inequality, which says that type t_i would prefer to pretend that he is of type t_i', implies that type t_i strictly prefers to send the message $m_i^*(t_i')$ instead of the one he is supposed to, which is $m_i^*(t_i)$. That is,

$$\sum_{t \in T} q_i(t|t_i) u_i(g(m^*(t)), t) < \sum_{t \in T} q_i(t|t_i) u_i(g(m_i^*(t_i'), m_{-i}^*(t_{-i})), t).$$

This contradicts m^* being a Bayesian equilibrium of G, since at least type t_i of agent i has found a profitable deviation. Q.E.D.

This result is called the revelation principle because it says that *any* Bayesian equilibrium of *any* mechanism must be outcome equivalent to the truth-telling equilibrium of a direct revelation mechanism. Therefore, a necessary condition for the Bayesian implementability of an SCF is its truthful implementability, i.e., that F satisfy incentive compatibility. It is a result of fundamental importance. It tells us that, in order to investigate the possible SCF's that one could implement with any mechanism, one must restrict attention to the set of incentive compatible SCF's.

However, as shown by Example 3 above, incentive compatibility is not sufficient for Bayesian implementability. In fact, there is another condition that is also necessary and that is needed for sufficiency as well. It is called *Bayesian monotonicity*. Bayesian monotonicity is a generalization of the condition of Maskin monotonicity that we used in the complete information model of Chapter 15. The precise statement of Bayesian monotonicity is involved, and we will not give it here. Roughly speaking, it says that the SCF must prescribe different alternatives in different states, when certain interim preferences of the types of agents present in those states change. The role of this condition is to take care of the undesired equilibria that are sometimes present in direct mechanisms. The main result in this area is due to Matthew Jackson, and we state a simple version of it next without providing a proof. We shall say that

an environment is *economic* if different agents have different top-ranked alternatives. Here is the theorem:

Jackson's Theorem. Suppose we have an economic environment and there are at least three agents. Then, an SCF F is Bayesian implementable if and only if it satisfies incentive compatibility and Bayesian monotonicity.

7. Exercises

1. Find all the Bayesian equilibria of the direct mechanism associated with the SCF of Example 3.

2. Consider an alternative mechanism designed to implement the SCF F that maximizes the welfare of agent 1 in the same problem. Specifically, consider the mechanism $H = ((M_i)_{i \in N}, h)$, in which $M_1 = \{0, 1\}$, while for $i = 2, 3$, M_i is the set of i's types. Let the outcome function h be the following, where we denote by $m = (m_1, m_2, m_3)$ a typical message profile:

   (i) If $m_1 = 1$, $h(m) = c$;

   (ii) Otherwise, i.e., if $m_1 = 0$, $h(m) = F(m_2, m_3)$.

   Show whether or not this mechanism works in order to fully implement F in Bayesian equilibrium.

3. Consider the demand revealing mechanism of Chapter 8 in order to implement the efficient level of public good. Can you propose a modification of this mechanism by which this level of public good is truthfully implementable in Bayesian equilibrium, but where the budget surplus is eliminated? Hint: See D'Aspremont and Gerard-Varet (1979).

8. Selected References

(Items marked with an asterisk (*) are mathematically difficult.)

*1. C. D'Aspremont and L.A. Gerard-Varet "Incentives and Incomplete Information," *Journal of Public Economics* V. 11, 1979, pp. 25-46.

This article analyzes interesting variants of demand revealing tax schemes. The emphasis is the use of incentive compatibility (truthful implementability).

*2. J. Harsanyi, "Games with Incomplete Information Played by Bayesian Players," Parts I, II and III, *Management Science* V. 14, 1967-1968, pp. 159-182, 320-334 and 486-502.

This path-breaking series of articles introduces the study of games with incomplete information.

*3. B. Holmstrom and R. Myerson, "Efficient and Durable Decision Rules with Incomplete Information." *Econometrica* V. 51, 1983, pp. 1799-1819.

This article introduces incentive efficiency notions, similar to the concepts of ex-ante efficiency, interim efficiency and ex-post efficiency from Chapter 6. Their definitions differed only in the timing at which the Central Authority contemplates the improvement, i.e., whether it takes place at the ex-ante stage, the interim stage or the ex-post stage, respectively.

To save having to write all these definitions again here, we shall mention the change in the definitions and refer the reader to Chapter 6 for details. The only difference now is that incentive compatibility is imposed both over the allocation rule (or SCF) and over the improvements. The justification is simple: in a context where information is asymmetric, both the status quo (the allocation rule or SCF) and the improvements considered by the Central Authority must be incentive compatible. This leads to the following definitions.

An SCF F is said to be *ex-ante/interim/ex-post incentive efficient* if it is incentive compatible and if there does not exist an incentive compatible SCF F' that is an ex-ante/interim/ex-post improvement over F, respectively.

It follows from the definitions that if an allocation is incentive compatible and satisfies ex-ante/interim/ex-post efficiency, it also satisfies ex-ante/interim/ex-post incentive efficiency, respectively. Also, there are versions of the no-trade theorem for these incentive efficiency concepts.

*4. M. Jackson, "Bayesian Implementation," *Econometrica* V. 59, 1991, pp. 461-477.

This article contains Jackson's theorem. The main result is his Theorem 2. We stated here a version of the simpler result, his Theorem 1.

5. R. Myerson, "Mechanism Design," in J. Eatwell, M. Milgate and P. Newman (eds.) *The New Palgrave: Allocation, Information, and Markets*, Norton, New York, 1989.

This is an accessible article that introduces incentive compatibility and some of its uses.

*6. R. Myerson, *Game Theory: an Analysis of Conflict*, Harvard University Press, 1991.

This is an excellent advanced graduate text on game theory. Its treatment of Bayesian games is superb.

*7. T. Palfrey and S. Srivastava "On Bayesian Implementable Allocations," *Review of Economic Studies* V. 54, 1987, pp. 193-208.

This article contains many examples to illustrate the strength of the different conditions that are necessary for Bayesian implementability. The main focus is Bayesian monotonicity.

*8. A. Postlewaite and D. Schmeidler "Implementation in Differential Information Economies," *Journal of Economic Theory* V. 39, 1986, pp. 14-33.

This is the first article that poses the full implementation question in contexts of incomplete information. Our Example 3 is an elaboration of one example described in their introduction.

*9. R. Serrano, "The Theory of Implementation of Social Choice Rules," *SIAM Review* V. 46, 2004, pp. 377-414.

in this survey, there is a relatively simple proof of the version of Jackson's theorem provided here.

*10. R. Serrano and R. Vohra "A Characterization of Virtual Bayesian Implementation," *Games and Economic Behavior* V. 50, 2005, pp. 312-331.

One can find well-behaved examples in which Bayesian monotonicity is shown to be a very strong condition, so that most interesting SCF's violate it. It then follows from Jackson's theorem that exact implementation is very limited in those settings. For cases like these, the approach of approximate implementation has been suggested, as was done in the case of complete information. In most environments, approximate implementation turns out to be very powerful. Specifically, this article proves a theorem that says that an SCF is approximately implementable in Bayesian equilibrium if and only if it satisfies incentive compatibility and another condition called virtual monotonicity. The paper also shows that, in almost every environment, virtual monotonicity is trivially satisfied by all SCF's. It is in this sense that approximate implementation is as powerful as it can be. Its only limitation is the incentive compatibility condition.

Chapter 17

EPILOGUE

Welfare economics and social choice theory are about the "goodness" or "badness" of economic arrangements in particular, and social choices in general. They offer partial answers to questions like these: Does a competitive economic system produce good results? Is there a necessary connection between market competition and optimality? What are the implications of externalities, public goods or informational asymmetries in the allocation of economic resources? Are there criteria for judging among distributions of goods, or wealth, other then market criteria? For example, is fairness an appropriate criterion for deciding when one distribution of goods is superior to another? Are traditional nonmarket decision methods, like voting, able to show us what social choices are good and what are bad? Can a Central Authority get around not having enough information and bring about socially desirable outcomes? Are there intrinsic logical limitations to all methods, market and nonmarket, for deciding what social alternatives are good and what are bad?

In this book we were guided by the principal that any judgment about social welfare, or about what is good or bad for society, has to be directly based on the welfare of the individuals involved. And the best judge of a particular individual's welfare is that individual himself. This means that we took individuals' preferences as given, as the foundation upon which all the analysis rests. What is good for society depends fundamentally on what individuals consider to be good for themselves.

Since individual preferences were viewed as fundamental here, Chapter 1 began with an analysis of individual preferences. An individual was characterized by a preference relation defined over an abstract set of alternatives. We saw that if a preference relation is complete and transitive, or even only complete and acyclic, it is possible to use it to

find best alternatives. We also saw that if the preference relation is complete and transitive, it can be used to derive a utility function. These preference relations, or equivalently, these utility functions, provide the basic givens throughout all the subsequent analysis.

With utilities or preferences in hand, in Chapters 2, 3 and 4 we analyzed barter exchange, and market (or price governed) exchange. Chapter 2 developed elementary analytical tools, such as the Edgeworth box diagram, and also developed modern game-theoretic exchange notions, in particular, blocking and the core. Roughly speaking, a group of traders can block a proposed allocation if it can achieve with its own resources a distribution of goods that its members prefer to the proposed one. The core is the set of unblocked allocations. This is the set of allocations of goods that would evolve in a world with costless, and frictionless bargaining, and voluntary trading. It is the set of solutions in a perfect exchange economy, and it clearly has significance for welfare economists. An allocation in the core is good, is voluntary, in the sense that no group can oppose it on the grounds that it can do better by itself. Conversely, an allocation outside of the core is involuntary in the sense that some group can oppose it on these grounds. Any allocation in the core must be Pareto optimal; the concept of the core is a refinement of the concept of Pareto optimality.

Chapter 3 introduced prices into the barter model. It developed a market model of exchange. Traders chose what to buy and sell according to market-clearing prices. A competitive equilibrium was defined as a situation in which every person chooses the bundle of goods that maximizes his utility, subject to his budget constraint, and in which supply equals demand for every good. In this chapter the two fundamental theorems of welfare economics were introduced, discussed, and in one case proved. The First Fundamental Theorem traces its roots to Adam Smith. It says, roughly, that the free market will bring a society to a good economic organization. Competitive buying and selling among self-interested and possibly greedy traders leads to a social optimum. In Chapter 3, the theorem took this strong form: Any competitive equilibrium allocation of goods must be a core allocation, and must be Pareto optimal.

The Second Fundamental Theorem is almost a converse of the first. It says, roughly, that any Pareto optimal allocation must be achievable through the competitive mechanism, given an appropriate cash transfer system. The theorem implies that any economy, even a socialist economy, can turn to the market mechanism to achieve its ends. If a planner in a socialist state, for example, wants to bring about a particular egalitarian and Pareto optimal distribution of goods, he need not rely on a

gigantic bureaucracy to distribute all goods to all people. He can instead use a relatively simple cash transfer program, and let the market work by itself.

We interpret Chapter 3 like this: in an exchange economy, where trade is voluntary, the market mechanism leads to distributions of goods that are Pareto optimal, and in this sense good for society, and the market mechanism, suitably modified, can be used to get to any particular Pareto optimal distribution of goods. Similar results may hold in economies where trade is not voluntary. This is the subject of Chapter 4, with its analysis of "jungle" exchange. The major theoretical difference between market exchange and "jungle" exchange, however, is that market exchange, which is voluntary, leads to the core, whereas "jungle" exchange, which is involuntary, does not.

In Chapter 5 we developed a model of production. This model starts with production sets, defines optimality for production plans, and then relates production to the market mechanism. The production versions of the two fundamental theorems of welfare economics were discussed and proved.

The main thrust of Chapters 3 and 5 was that the competitive mechanism produces results that are good for society. However, there are several basic shortcomings in all this. First, information asymmetries may compromise the very notion of a "market." For example, if sellers in a market know precisely what they are selling, but buyers do not (a "pig in a poke," a miracle cure for cancer, a labor contract, a "lemon" in a used car market), market prices may not properly reflect the characteristics of goods being sold, and optimality may suffer.

Second, there might be external or spillover effects. One person's price-governed consumption, or one firm's price-governed production, might directly affect another person's utility level, or another firm's production set. These effects might not be reflected properly in market prices either, and consequently a competitive equilibrium might not be optimal after all.

Third, there might be public goods, which ought to be provided and paid for by a government; that is, there might be goods whose use is nonexclusive, and which would, if provided privately, break the link between the competitive market mechanism and optimality.

Chapter 6 was devoted to exchange economies under uncertainty. Its main finding was two-fold. If uncertainty is symmetric, the two fundamental theorems of welfare economics survive. However, if there is asymmetric information, the theorems fail. Prices no longer serve as signals to allocate resources efficiently.

Chapter 7 provided a brief treatment of externalities, and, with examples, sketched how the competitive mechanism might be modified to correct spillover effects. The basic principal there was the idea of a Pigouvian tax or subsidy. A person or firm which is imposing external costs (or benefits) on others ought to be taxed (or subsidized) at a rate equal to marginal external cost (or marginal external benefit).

Chapter 8 provided a lengthy treatment of public goods. The crucial question there was: What system of public decision making and taxation will bring about the optimal output of a public good? The Wicksell-Lindahl tax scheme satisfies Samuelson's optimality condition, and it links a person's tax to his marginal utility from the public good. However, its disadvantage is its extreme susceptibility to abuse by free riders, people who claim a low marginal utility and consequently pay a low tax, even though they get substantial benefits from the public good. The fixed tax shares majority voting scheme is not so vulnerable to free rides, but it will generally lead to levels of output for the public good that do not satisfy the Samuelson optimality condition. The demand-revealing tax scheme solves the free rider problem since it provides no incentives for people to misrepresent their preferences. And it leads to an optimal level of output for the public good. However, it might collect more taxes than necessary to produce the public good, and, because of the special logic of this tax system, it might be necessary to waste this budget surplus. Consequently, the demand-revealing tax scheme might lead to a situation that is overall not Pareto optimal, even though it involves an optimal level of output for the public good. The Groves-Ledyard tax scheme also leads to an optimal level of output for the public good, and, unlike the demand revealing tax scheme, it produces no wasteful budget surplus. However, it does not solve the free rider problem so much as it sidesteps it. In short, the public goods problem is complicated and not yet completely solved, but there are decentralized ways to bring about optimality in the provision of these goods. The market mechanism should not be abandoned because of the existence of public goods.

The fourth shortcoming of the competitive mechanism is its neglect of equality: even if it leads to a Pareto optimal allocation, that allocation might be quite unequal, or quite unfair. The Second Fundamental Theorem does say that any equitable Pareto optimal allocation can be reached via a modified competitive mechanism. But how do we judge among allocations to find the one that is not only Pareto optimal, but also best for society? What are the criteria for judging among Pareto optima? How do we choose the best from among the optimal? The market provides no clue.

Chapter 9 and the subsequent chapters addressed the fourth shortcoming of the competitive mechanism: How do we judge among allocations to find one that is not only Pareto optimal, but also socially best? Chapter 9 reviewed the compensation criteria or social improvement criteria traditionally used in applied fields of microeconomics. These criteria were first developed by Nicholas Kaldor, John Hicks, Tibor Scitovsky, and Paul Samuelson, in order to refine the economist's ability to judge among social alternatives; in order to allow the economist to say "a is better for society than b," even when either a or b might come about through the market mechanism. The underlying theme here is something like this: a is better than b if the gainers in a move from b to a gain so much that they could, in theory, compensate the losers for their losses in the move from b to a. We saw that this criterion, and the more sophisticated ones derived from it, all had logical drawbacks. For instance, neither the Kaldor, the Scitovsky, nor the Samuelson criterion can successfully distinguish between alternative Pareto optimal allocations in an edgeworth box diagram. We also saw that the measures of consumers' surplus, used extensively in applied work, are generally inconsistent, unless utility functions are assumed to have a special quasilinear form.

Chapter 10 jumped straight away from the traditional compensation criteria, and away from the usual welfare-competition analysis. It examined the idea of fairness in the non-envy sense, and the Rawls criterion, as ways to judge when one distribution of goods is better for society than another. But again there are problems. Fairness is not in fact a criterion that one can join to the market mechanism, or to the Pareto criterion, in order to find out what is really best for society. For it turns out that the criterion of fairness might conflict with the other usual economic criteria, and that fairness might not survive trade. Fairness and either barter or market exchange, do not mix well. The Rawls criterion, the idea of maximizing the welfare of the worst off person, has the serious disadvantage of being overly biased toward the poor. It implies a social preference for a state in which everyone is poor over a state in which one person is very poor but thousands of others are very rich. In short, if we want to narrow down the range of good social alternatives, if we want to decide among Pareto optima, we ought to look beyond these two egalitarian criteria.

Chapter 11 looked at the welfare implications of life and death choices. We examined the standard economic model of value of life and found it produces paradoxical results. We considered various utilitarian methods to evaluate population changes, including what appears to be a reasonable extension of the Pareto criterion, and we found that the utilitarian

measures generally had implications that are odd, or shocking, or repugnant.

Chapter 12 and the subsequent chapters looked at other (mostly) nonmarket choice mechanisms. The market mechanism generally cannot tell us when alternative a is better for society than alternative b. What about voting rules? Can they serve this purpose? Can the "will of the people" be discovered in election procedures, rather than in the market? Can voting rules answer the questions about what is socially best that the market mechanisms leave unanswered? Chapter 12 examined the most common voting procedure, majority rule. The voting paradox of Condorcet was presented, and the paradox was "solved" for the special case of single-peaked preferences, first analyzed by Duncan Black. The significance of Black's Theorem is that for particular types of individual preferences, majority voting provides a logically ideal way to decide when alternative a is socially preferable to alternative b, and, in fact, to decide what alternatives are socially best. Unfortunately, single-peakedness is a rather stringent condition, and it cannot be easily generalized to multidimensional alternative sets. Since the distribution of goods or income is inherently a multidimensional problem, majority voting fails as a logically satisfactory way to decide when a distribution a is socially preferable to another distribution b.

The search for means to judge among social alternatives continued in Chapter 13. There we presented the general question raised by Kenneth Arrow: Does there exist a method to judge between social alternatives that has no serious flaws? The Pareto criterion is flawed by its inability to judge among the many Pareto noncomparable states. The majority voting rule is flawed by its possible inconsistencies, its paradoxes. The criterion of fairness is flawed by its possible incompatibility with voluntary, utility-increasing trades. Is there any foolproof way to discover which alternatives are socially better than which? Chapter 13 presented a model in which the question is answered with Arrow's Impossibility Theorem. This theorem says that there is no completely satisfactory way to aggregate individual preference relations. There is no completely satisfactory collective choice rule or Arrow social ranking. Every mechanism for deciding whether a is socially "better" than b has, inevitably, serious shortcomings.

Chapters 14, 15 and 16 turned to the theory of implementation. Implementation theory is about the possibility of achieving socially desirable outcomes in contexts in which the Central Authority (the Central Planning Board, the Public Good Board) does not have all the relevant necessary information. It then needs to elicit it from the agents. To do so, it designs an institution in which the agents will act. This institu-

tion is called a mechanism. The Central Authority's goal is defined by a social choice function (SCF), a rule that produces unique socially best alternatives, rather than an Arrow social welfare function as in Chapter 13.

The negative theme that came from Arrow's theorem was carried further in Chapter 14. There, we picked up a criterion for a foolproof rule that first appeared in Chapter 8, on public goods: We required that a social choice function be completely immune to strategic misrepresentation of preferences, in order to be considered foolproof. If person i is asked about his preferences, reporting his true preferences must be a dominant strategy for him. The idea is that if people misrepresent their true feelings, and the outcome of (say) a voting process is alternative a, it is implausible to argue that a is really socially "best"! Does there then exist a foolproof SCF, an SCF that is not liable to manipulation in this sense, and that is also nondictatorial and nontrivial? The negative answer is given in the Gibbard-Satterthwaite Impossibility Theorem. This theorem says that every procedure will in some circumstances induce people to misrepresent their preferences. Consequently, the decisions of any social choice procedure, any voting procedure, any rule to decide what is socially best, are suspect.

Evidently, requiring truth-telling as a *dominant strategy* was asking too much. Requiring that they tell the truth *if* they expect others to do so also would be a lower hurdle. This led us to replace the search for a dominant strategy equilibrium with the search for an alternative game theoretic solution concept.

In Chapter 15 we studied environments with complete information and used the notion of Nash equilibrium; in Chapter 16 we analyzed environments with incomplete information and used the notion of Bayesian equilibrium. In both chapters, the news for implementation theory was more encouraging than in Chapter 14. We identified the conditions that imply implementability in these environments, namely Maskin monotonicity, incentive compatibility, and Bayesian monotonicity. These conditions generally allow nondictatorial SCF's to be implemented. Furthermore, we showed that if the Central Authority uses approximate implementation instead of exact implementation, there are strong positive implementability results.

Solutions to Exercises

Chapter 1

1. R_i is transitive: $xR_iy,\ yR_iz$ then xR_iz.

   (i) To show: xP_iy and yI_iz implies xP_iz.

   Proof. xP_iy means xR_iy and not yR_ix; yI_iz means yR_iz and zR_iy. Since xR_iy and yR_iz, xR_iz follows by transitivity of R_i. If zR_ix were also true, then, $yR_iz,\ zR_ix$ implies yR_ix which would be a contradiction. □

   (ii) To show: xI_iy and yI_iz implies xI_iz.

   Proof. xI_iy means xR_iy and yR_ix; yI_iz means yR_iz and zR_iy. By transitivity, $xR_iy,\ yR_iz$ gives xR_iz; also $zR_iy,\ yR_ix$ gives zR_ix. Thus we have xI_iz. □

2. a. The preference order is complete because for any pair of teams, we can either say one is better than the other or they are equal.
   Not transitive because then A better-than B, B better-than C, should imply A better-than C. Not quasi-transitive either for the same reason as above. Acyclic because we have A better-than B, B better-than C and A ties with C, so C is not better-than A.

   b. No.

   c. Let $u(A) = 10 > u(B) = 5$, A defeats B.
   $u(B) = 5 > u(C) = 3$, B defeats C.
   So $u(A) > u(C)$, but A doesn't defeat C.

   d. With $v(A) = 10,\ v(B) = 5,\ v(C) = 3$, we can have a pseudo quality function with the property that if x defeats y then $v(x) > v(y)$.

3. If preferences satisfy independence, then for all lotteries l_1, l_2 and l_3, one has that $l_1 I_i l_2$, if and only if $(\alpha l_1 + (1-\alpha) l_3) I_i (\alpha l_2 + (1-\alpha) l_3)$ for every $\alpha \in [0,1]$.

   Proof. If $l_1 I_i l_2$, then $(\alpha l_1 + (1-\alpha) l_3) I_i (\alpha l_2 + (1-\alpha) l_3)$ for every $\alpha \in [0,1]$.
   $l_1 I_i l_2$ means $l_1 R_i l_2$ and $l_2 R_i l_1$.
   By independence,
   $l_1 R_i l_2 \Leftrightarrow (\alpha l_1 + (1-\alpha) l_3) R_i (\alpha l_2 + (1-\alpha) l_3)$, for every $\alpha \in [0,1]$.
   Also, $l_2 R_i l_1 \Leftrightarrow (\alpha l_2 + (1-\alpha) l_3) R_i (\alpha l_1 + (1-\alpha) l_3)$, for every $\alpha \in [0,1]$.
   Combining the above, $l_1 I_i l_2 \Rightarrow (\alpha l_1 + (1-\alpha) l_3) I_i (\alpha l_2 + (1-\alpha) l_3)$ for every $\alpha \in [0,1]$.

   To show if $(\alpha l_1 + (1-\alpha) l_3) I_i (\alpha l_2 + (1-\alpha) l_3) \Rightarrow l_1 I_i l_2$, put $\alpha = 1$ and we have $l_1 I_i l_2$. □

4. To show: if preferences over lotteries are representable by an expected utility function, they must satisfy completeness, transitivity, continuity and independence.

   Proof. By the expected utility theorem, for any given lottery we have a real number $u(l) = \sum_{j=1}^{k} u_j q_j$ that represents the utility from that lottery.
   Completeness is satisfied because for any given pair of lotteries, l and l', using the order property of real numbers, we can either have
   $u(l) > u(l') \Leftrightarrow l P_i l'$ or
   $u(l) < u(l') \Leftrightarrow l' P_i l$ or
   $u(l) = u(l') \Leftrightarrow l I_i l'$. Hence complete.
   Transitivity requires: if $l R_i l'$, $l' R_i l''$, then we should have $l R_i l''$. By expected utility theory, $l R_i l' \Leftrightarrow u(l) \geq u(l')$ and $l' R_i l'' \Leftrightarrow u(l') \geq u(l'')$. So we have $u(l) \geq u(l'')$,implying $l R_i l''$.
   Continuity says that if $l_1 P_i l_2 P_i l_3$, then there exists $\alpha \in [0,1]$ such that $\alpha l_1 + (1-\alpha) l_3) I_i l_2$. By expected utility theorem, $l_1 P_i l_2 P_i l_3$ implies $u(l_1) > u(l_2) > u(l_3)$. Since these are three real numbers, so we can always solve for an α such that the weighted sum of expected utilities from lotteries l_1 and l_3(with weights being α and $(1-\alpha)$, respectively) should equal the expected utility from lottery l_2. That would then give us a lottery which would be indifferent to l_2.
   Independence states that $l_1 R_i l_2 \Leftrightarrow (\alpha l_1 + (1-\alpha) l_3) R_i (\alpha l_2 + (1-\alpha) l_3)$, for every $\alpha \in [0,1]$. By expected utility theorem,

$l_1 R_i l_2 \Leftrightarrow u(l_1) \geq u(l_2)$. Multiply both sides by α and add the term $(1-\alpha)u(l_3)$ to both sides. Note that the above operations do not affect the sign of the inequality. Then, on left we have expected utility from the lottery $\alpha l_1 + (1-\alpha)l_3$ and on right we have expected utility from the lottery $\alpha l_2 + (1-\alpha)l_3$. Thus $(\alpha l_1 + (1-\alpha)l_3) R_i (\alpha l_2 + (1-\alpha)l_3)$.
To prove the converse, put $\alpha = 1$ in the expected utility expressions for the compound lotteries and we get $l_1 R_i l_2$. □

5. a. Suppose the alternatives are x, y, z, t and u. And the committee members are numbered 1 to 5.
1 2 3 4 5
$x\ y\ z\ t\ u$
$y\ z\ t\ u\ x$
$z\ t\ u\ x\ y$
$t\ u\ x\ y\ z$
$u\ x\ y\ z\ t$
Consider x and y: Persons 1,3,4 and 5 vote for x. So xPy.
Consider y and z: Persons 1,2,4 and 5 vote for y. So yPz.
Consider z and t: Persons 1,2,3 and 5 vote for z. So zPt.
Consider t and u: Persons 1,2,3 and 4 vote for t. So tPu.
Consider u and x: Persons 2,3,4 and 5 vote for u. So uPx.
So we have a cycle : xPy , yPz , zPt , tPu , uPx.

b. Now suppose we have just four alternatives x, y, z and t and 5 committee members numbered 1 to 5 as before.
1 2 3 4 5
$x\ y\ z\ t\ x$
$y\ z\ t\ x\ y$
$z\ t\ x\ y\ t$
$t\ x\ y\ z\ z$
Consider x and y: Persons 1,3,4 and 5 vote for x. So xPy.
Consider y and z: Persons 1,2,4 and 5 vote for y. So yPz.
Consider z and t: Persons 1,2 and 3 vote for z. So we cannot have zPt.
Consider t and x: Persons 2,3 and 4 vote for t. So we cannot have tPx.
Thus, there cannot be a voting cycle.

Chapter 2

1. a. We leave the diagram to the student.
b. For Pareto optimality we need $MRS_1 = MRS_2$, implies $\frac{3}{2} = \frac{x_{22}}{x_{21}}$. The set of Pareto optimal allocations is given by

the locus of points $3x_{21} = 2x_{22}$ plus a part of x_{11} axis from person 1's origin to $(\frac{10}{3}, 0)$. The core is given by points on the Pareto set above the indifference curves passing through the endowment point for the two persons. Let us consider person 1's indifference curve through the endowment point and find the point where it intersects the Pareto locus. To do this, we solve $3x_{11} + 2x_{12} = 3 \times 10 + 2 \times 9$ and $3(10 - x_{11}) = 2(10 - x_{22})$. This gives $x_{11} = 9$ and $x_{12} = 8.5$. So the core is given by all points on the line $3x_{21} = 2x_{22}$ from person 2's origin to the point $(9\ ,\ 8.5)$.

2. a. No two people can block ω.
   b. No single person can block ω.
   c. ω is not Pareto optimal because we can have the following allocation: $(1,0,0),(0,0,1),(0,1,0)$. Now each individual gets 3 utils as compared to the 2 utils he was getting under ω.
3. a. $MRS_1 = \frac{x_{12}}{x_{11}} = 2.$
   $MRS_2 = 2.$
   $MRS_3 = \frac{x_{32}}{x_{31}+1} = 2.$
   b. $u_1(x) = 2 > u_1(\omega_1) = 0.$
   $u_2(x) = 20 > u_2(\omega_2) = 6.$
   $u_3(x) = 8 > u_3(\omega_3) = 0.$
   c. To show that x is not in the core, consider a coalition of 1 and 3. The allocation $(4,4),(4,4)$ is feasible and $u_1(4,4) = 16 > u_1(x) = 2$ and $u_3(4,4) = 20 > u_3(x) = 8$.

Chapter 3

1. $MRS_1 = 3 = \frac{p_1}{p_2} = MRS_2 = \frac{x_{22}}{x_{21}}$.
   Assume $p_2 = 1$. Then $p_1 = 3$. Now from 2's budget constraint, $3x_{21} + x_{22} = 3 \times 1 + 2$, and using the above condition, we get the competitive allocation as $(\frac{13}{6}, \frac{1}{2})\ (\frac{5}{6}, \frac{5}{2})$.
2. a. 1's offer curve :
   For $p_1 > 1,\ x_{11} = 0,\ x_{12} = 100p_1$.
   For $p_1 < 1,\ x_{11} = 100,\ x_{12} = 0$.
   For $p_1 = 1$, any (x_{11}, x_{12}) such that $x_{11} + x_{12} = 100$.
   b. We leave it to the student to check that the indifference curve for person 2 is indeed symmetrical around the line $x_{21} = x_{22}$.
   c. $MRS_1 = 1 = p_1 = MRS_2 = \frac{x_{22}}{x_{21}}$. Using 2's budget constraint $p_1 x_{21} + x_{22} = p_1 \times 0 + 50$, and using the above condition, we get the competitive allocation as $(75, 25), (25, 25)$.
3. a. Note that for any price ratio $\neq 1$, one individual's maximization exercise would require him to consume at a point on the

outside of the box. So the only possible competitive price ratio is 1, with the following allocation: $(1,0),(0,1)$.

b. $MRS_1 = 2 \neq MRS_2 = \frac{1}{2}$.

c. The competitive allocation is Pareto optimal although the MRS are not equal because to make one person better off we need to worsen the other.

4. Note that person 2 consumes both the goods, which means that the price ratio should coincide with the slope of his indifference curve, which is $\frac{1}{2}$. So $p = \frac{1}{2}$. Now using the given allocation in person 1's budget constraint,
$\frac{1}{2} \times \frac{3}{4} + 0 = \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} + T_1$, which gives $T_1 = -\frac{3}{8}$.
Similarly for person 2:
$\frac{1}{2} \times \frac{1}{4} + 1 = \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} + T_2$, which gives $T_2 = \frac{3}{8}$.

5. a. Set of Pareto optimal allocations is given by all points which satisfy:
$MRS_1 = \frac{x_{12}}{x_{11}} = MRS_2 = 2$, plus a part of x_{21} axis from 2's origin to $x_{21} = 1/2$ and $x_{22} = 0$.
The core is the set of all points such that $x_{12} = 2x_{11}$.

b. We leave it to the student.

c. $MRS_1 = \frac{x_{12}}{x_{11}} = \frac{p_1}{p_2} = MRS_2 = 2$. Assume $p_2 = 1$. Then $p_1 = 2$. Now from 1's budget constraint, $2x_{11} + x_{12} = 2 \times 1 + 0$ and using the above condition we get the competitive allocation as $(\frac{1}{2}, 1), (\frac{1}{2}, 0)$.

d. Note that person 2 consumes both goods, which means that the price ratio should coincide with the slope of his indifference curve which is 2. So $\frac{p_1}{p_2} = 2$. Now using the given allocation in person 1's budget constraint, $2 \times \frac{1}{4} + \frac{1}{2} = 2 \times 1 + T_1$, which gives $T_1 = -1$.
Similarly for person 2:
$2 \times \frac{3}{4} + \frac{1}{2} = 2 \times 0 + 1 + T_2$, which gives $T_2 = 1$.

Chapter 4

1. The allocation (h_2, h_4, h_1, h_3) is not in the core because person 4 prefers h_4, which he originally has, to h_3 which he would get. So the above allocation is blocked by person 4.
The allocation (h_3, h_4, h_1, h_2) is not in the core because person 1 prefers h_1, which he originally has, to h_3 which he would get. So the above allocation is blocked by person 1.
The allocation (h_2, h_4, h_3, h_1) is not in the core because person 4

prefers h_4, which he originally has, to h_1 which he would get. So the above allocation is blocked by person 4.

2. If the power relation is 1, 2, 3, 4 then person 1 gets to choose first and he chooses house h_2 because he likes it the best. Next person 2 gets to choose and he takes h_4 because that is his favorite. Person 3 then has to choose between h_1 and h_3 and he takes h_1. Finally person 4 has to take h_3. Thus we have the allocation (h_2, h_4, h_1, h_3).
   If the power relation is 1, 3, 2, 4 then person 1 takes h_2, person 3 then takes h_1, person 2 takes h_4 and person 4 is left with h_3. So we have the same allocation as we had under the power relation 1, 2, 3, 4.
   If the power relation is 4, 3, 2, 1 then person 4 gets to choose first and he chooses h_2, then person 3 chooses h_1, followed by person 2 who gets h_4 and finally person 1 gets h_3. So the allocation is (h_3, h_4, h_1, h_2).

3. The power relations that lead to each of the Pareto optimal allocations:
   (h_2, h_3, h_1, h_4) is supported by 3, 1, 4, 2.
   (h_2, h_4, h_1, h_3) is supported by 1, 2, 3, 4.
   (h_3, h_4, h_1, h_2) is supported by 2, 3, 4, 1.
   (h_2, h_4, h_3, h_1) is supported by 1, 2, 4, 3.
   (h_1, h_4, h_3, h_2) is supported by 2, 4, 1, 3.

Chapter 5

1. a. One possible set of production vectors is $(1/4, -1/27, -27/8)$, $(1, -1, -1)$, $(0, 2, -4)$.
   b. Net input-output vector is $(5/4, 26/27, -67/8)$. One possible production plan for the economy with zero net input of capital is $(1, 0, -3)$ with the set of production vectors being $(0, -0, -1)$, $(1, -1, -1)$, $(0, 1, -1)$.
   c. Consider the following adjustment in the plan: firm 1 uses 1 unit less of labor and firm 2 uses 1 unit more. Then the set of production vectors is given by- $(0.1974, -1/27, -19/8)$, $(5/3, -1, -2)$, $(0, 2, -4)$. Check that net output of good 1 is strictly higher as compared to the initial plan, while the quantities of capital and labor remain unchanged.

2. For Firm 1:
$\frac{MP_{12}}{MP_{13}} = \frac{p_2}{p_3}$ which gives $y_{13} = y_{12}$. Thus any production plan that uses equal amounts of the two inputs and produces the same amount of output is a profit maximizing vector.
For Firm 2:
$\frac{MP_{22}}{MP_{23}} = \frac{1}{2} = \frac{p_2}{p_3}$. So any point is a profit maximizing vector.
For Firm 3:
$p_2 MP_{33} = p_3$, which gives $y_{33} = -\frac{1}{16}$, $y_{32} = \frac{1}{4}$. Profits at the optimum equal $\frac{1}{8}$.

Chapter 6

1. To show: set of ex-ante efficient allocations is contained in the set of interim efficient allocations.

Proof. We argue by contradiction. Suppose that allocation x is ex-ante efficient but not interim efficient. Then there exists an allocation y such that

$$\sum_{s=1}^{k} q_i(s|\sigma_i)u_{is}(y_{is}) \geq \sum_{s=1}^{k} q_i(s|\sigma_i)u_{is}(x_{is})$$
$$\text{for all } i = 1, 2, \ldots, n \text{ and for all } \sigma_i$$
$$\sum_{s=1}^{k} q_i(s|\sigma_i)u_{is}(y_{is}) > \sum_{s=1}^{k} q_i(s|\sigma_i)u_{is}(x_{is}) \text{ for at least one pair } i, \sigma_i.$$

Let $\pi(\sigma_i)$ denote the probability that agent i receives signal σ_i.Then

$$\sum_{s=1}^{k}(q_{is}u_{is}(y_{is})) = \sum_{\sigma_i}(\sum_{s=1}^{k} q_i(s|\sigma_i)u_{is}(y_i(s))) \cdot \pi(\sigma_i).$$

Using this and the fact that $\pi(\sigma_i)$ is positive on the above system of inequations, we get the following system of inequalities:

$$\sum_{s=1}^{k} q_{is}u_{is}(y_{is}) \geq \sum_{s=1}^{k} q_{is}u_{is}(x_{is}) \text{ for all } i = 1, 2, \ldots, n$$
$$\sum_{s=1}^{k} q_{is}u_{is}(y_{is}) > \sum_{s=1}^{k} q_{is}u_{is}(x_{is}) \text{ for at least one i .}$$

This contradicts the fact that x is an ex-ante efficient allocation. The intuition behind the result is that if agents have decided on

an optimal allocation prior to obtaining their private signals, then they cannot do any better once they actually receive their private information. □

To show: set of interim efficient allocations is contained in the set of ex-post efficient allocations.

Proof. We argue by contradiction. Suppose that allocation x is interim efficient but not ex-post efficient. Since it is not ex-post efficient, there exists a state s' and an alternative allocation y that makes everyone in that state s' at least as good as x, and at least one individual strictly better. Now define the following allocation z:

$$z = \begin{cases} y_s \text{ if } s = s', \\ x_s \text{ otherwise} \end{cases}$$

Since $q_i(s|\sigma_i)$ is non-negative for all s and σ_i, so we have the following system of inequalities:

$$\sum_{s=1}^{k} q_i(s|\sigma_i)u_{is}(z_{is}) \geq \sum_{s=1}^{k} q_i(s|\sigma_i)u_{is}(x_{is}) \text{ for all } i = 1, 2, \ldots, n$$

$$\sum_{s=1}^{k} q_i(s|\sigma_i)u_{is}(z_{is}) > \sum_{s=1}^{k} q_i(s|\sigma_i)u_{is}(x_{is}) \text{ for at least one } i.$$

But this means that x is not interim efficient either, which is a contradiction. Here again the intuition is that if agents have contracted on an optimal allocation depending on the private signals that they have received then, there cannot be an ex-post allocation that makes everyone at least as well off as before and someone strictly better off. □

2. The set of ex-ante efficient allocations is given by all the points on the line joining the origins of persons 1 and 2, because along this line, $MRS_1 = MRS_2$, satisfying the Pareto requirement of equality of MRS across agents.
   The set of ex-post efficient allocations is given by the entire Edgeworth box. This is because, ex-post, it is a one-good world and so it is not possible to make any one person better off without worsening the other.
3. Trader 1 solves max $q_{11}\sqrt{x_{11}} + q_{12}\sqrt{x_{12}}$ subject to $p_1 \cdot x_{11} + p_2 \cdot x_{12} \leq p_1 \cdot \omega_{11} + p_2 \cdot \omega_{12}$
   The first order conditions when re-arranged give: $\frac{x_{12}}{x_{11}} = (\frac{p_1 \cdot q_{12}}{p_2 \cdot q_{11}})^2$.

Normalizing the price of good 2 to 1, and using the above condition in person 1's budget constraint, $p_1 \cdot x_{11} + (\frac{p_1 \cdot q_{12}}{q_{11}})^2 \cdot x_{11} = p_1 \cdot 1 + 1 \cdot 0$,
gives $x_{11} = \frac{1}{1+(q_{12}/q_{11})^2 \cdot p_1}$
and $x_{12} = \frac{p_1}{1+(q_{11}/q_{12})^2 \cdot (\frac{1}{p_1})}$.

Trader 2 solves max $q_{21}\sqrt{x_{21}} + q_{22}\sqrt{x_{22}}$ subject to $p_1 \cdot x_{21} + p_2 \cdot x_{22} \leq p_1 \cdot \omega_{21} + p_2 \cdot \omega_{22}$
The first order conditions when re-arranged give : $\frac{x_{22}}{x_{21}} = (\frac{p_1 \cdot q_{22}}{p_2 \cdot q_{21}})^2$.
Normalizing the price of good 2 to 1, and using the above condition in person 2's budget constraint, $p_1 \cdot x_{21} + (\frac{p_1 \cdot q_{22}}{q_{21}})^2 \cdot x_{21} = p_1 \cdot 0 + 1 \cdot 1$,
gives $x_{21} = \frac{1}{p_1(1+p_1(q_{22}/q_{21})^2)}$
and $x_{22} = \frac{p_1}{p_1+(q_{21}/q_{22})^2}$.
Market clearing condition: $x_{11} + x_{21} = 1$, gives the following equation:

$$p_1^3 a^2 b^2 + p_1^2 a^2 - p_1 a^2 - 1 = 0$$

where $a = (q_{12}/q_{11})$ and $b = (q_{22}/q_{21})$. This is an equation in third degree which would yield complicated roots. But even without explicitly solving for p_1 we can see how it would be affected with changes in agents' beliefs about the occurrence of the two states. For example, when both think that the states are equally likely, then $a = b = 1$ and we get $p_1 = 1$ as the solution. And the allocations are $(1/2, 1/2)$ and $(1/2, 1/2)$. This can be explained by the symmetry in utility functions and endowments. Also, when person 1 doesn't expect state 2 to happen at all, i.e a equals 0, then p_1 is undefined and $x_{11} = 1$, meaning person 1 consumes whole of his endowment in the first state. Similarly, if b equals 0, then $x_{22} = 1$, meaning person 2 consumes whole of his endowment in state 2.

4. To show: that there is a unique REE in this economy, which is the one given in the text.
Claim: Any REE in this economy cannot be a non-revealing one.

Proof. Suppose we have a non-revealing REE. Then $p^*(1) = p^*(2) = p^*$, say. Since person 2 cannot distinguish between the two states, so he maximizes the expected utility: $(1/2)[\ln(x_{21}(1)) + \ln(x_{22}(1))] + (1/2)[\ln(x_{21}(2)) + 2 \cdot \ln(x_{22}(2))]$ subject to his budget constraints in each state namely, $p^* \cdot x_{21(1)} + x_{22(1)} \leq p^* \cdot 2 + 1$ and $p^* \cdot x_{21(2)} + x_{22(2)} \leq p^* \cdot 2 + 1$. From

the first order conditions we can see that his consumption bundles in the two states are independent of each other. In other words, it is as if he is maximizing two different utility functions for the two different states subject to the same constraint, namely $p^* \cdot x_{21} + x_{22} \leq p^* \cdot 2 + 1$. This does not give identical consumption bundles for the two states. And so the REE cannot be non-revealing. □

So now we just need to look at revealing REE , that is equilibria with prices $p^*(1)$ and $p^*(2)$ such that $p^*(1) \neq p^*(2)$.

When person 1 receives signal σ_1, he knows that it is going to be state 1 and so he maximizes $u_{11} = x_{11}(1) + \ln(x_{12}(1))$ subject to $p^*(1) \cdot x_{11}(1) + x_{12}(1) \leq p^* \cdot (1)1 + 1$. Since the utility function is monotonic, the budget constraint will hold with equality at the optimum. Also an equilibrium in this case cannot be at a boundary point. (Why?)
The first order conditions give $x_{12}(1) = p^*(1)$ and using the budget constraint, we get $x_{11}(1) = (1/p^*(1))$.
Person 2 observes $p^*(1)$, knows it is going to be state 1 and maximizes $u_{21} = \ln(x_{21}(1)) + \ln(x_{22}(1))$ subject to $p^*(1) \cdot x_{21}(1) + x_{22}(1) \leq p^*(1) \cdot 2 + 1$. Again, person 2 exhausts his budget at the optimum because the utility function is monotonic.
The first order conditions give $\frac{x_{22}(1)}{x_{21}(1)} = p^*(1)$ and using the budget constraint, we get $x_{21}(1) = 1 + (1/2p^*(1))$ and $x_{22}(1) = p^*(1) + (1/2)$.
From the market clearing condition for good 1, $x_{11}(1) + x_{21}(1) = 3$, we solve for the price as $p^*(1) = (3/4)$. So the allocation is (4/3, 3/4), (5/3, 5/4).

When person 1 receives signal σ_1', he knows that it is going to be state 2, he maximizes $u_{12} = x_{11}(2) + 2\ln(x_{12}(2))$ subject to $p^*(2) \cdot x_{11}(2) + x_{12}(2) \leq p^*(2) \cdot 1 + 1$. Budget constraint holds with equality at the optimum because utility function is monotonic and again we need to look at just the interior points for an equilibrium.
The first order conditions give $x_{12}(2) = 2p^*(2)$ and using the budget constraint, we get $x_{11}(2) = (1/p^*(2)) - 1$.
Person 2 observes $p^*(2)$, knows it is going to be state 2 and maximizes $u_{22} = \ln(x_{21}(2)) + 2\ln(x_{22}(2))$ subject to $p^*(2) \cdot x_{21}(2) + x_{22}(2) \leq p^*(2) \cdot 2 + 1$. Again, since the utility function is monotonic the budget is exhausted.

The first order conditions give $\frac{x_{22}(2)}{2x_{21}(2)} = p^*(2)$ and using the budget constraint, we get $x_{21}(2) = (2/3) + (1/3p^*(2))$ and $x_{22}(2) = (2(2p^*(2)+1)/3)$.
From the market clearing condition for good 1, $x_{11}(2) + x_{21}(2) = 3$, we solve for the price as $p^*(2) = (2/5)$. So the allocation is (3/2, 4/5), (3/2,6/5).
These are exactly the ex-post competitive allocations as given in the text. Since there is only one value of price that clears the market in any state, we see that the equilibrium is unique.

Chapter 7

1. a. We are going to guess that the competitive equilibrium is an interior point. Let us look at person 2's maximization: he maximizes the selfish part of his utility function. The slope of his indifference curve is 1. So the price ratio is 1. Person 1 equates his MRS to this price ratio and we get $\frac{x_{12}}{x_{11}} = 1$. Using 1's budget constraint:
   $1x_{11} + x_{12} = 1 \times 1 + 0$, we get $x_{11} = \frac{1}{2}$, $x_{12} = \frac{1}{2}$. And so, $x_{21} = \frac{1}{2}$, $x_{22} = \frac{1}{2}$. Thus our guess was correct.
   b. To find the Pareto optimal points, we-rewrite the utility function of person 2 as: $u_2 = 2x_{21} + 2x_{22} - (1 - x_{21})$. Then, $MRS_1 = \frac{x_{12}}{x_{11}} = MRS_2 = \frac{3}{2}$,
   gives the set of interior Pareto optimal points.
   c. Impose a per unit tax t on person 1's consumption of good 1. Then person 1's optimality condition becomes $MRS_1 = \frac{p_1+t}{p_2}$. Person 2's optimality condition however, remains the same: $MRS_2 = \frac{p_1}{p_2}$. At the given allocation, $MRS_2 = 1$. So let $p_1 = p_2 = 1$. Since $MRS_1 = \frac{3}{2}$, so $t = \frac{1}{2}$.
   Now we look at the budget constraint for 2:
   $1 \times \frac{2}{3} + 1 \times \frac{1}{2} = 1 \times 1 + T_2$, which gives $T_2 = \frac{2}{3}$.
   And the budget constraint for person 1:
   $\frac{3}{2} \times \frac{1}{3} + 1 \times \frac{1}{2} = 1 \times 1 + T_1$, which gives $T_1 = 0$.
2. The competitive equilibrium is derived by equating the selfish MRS for 1 to the MRS for 2, that is, $MRS_1 = \frac{x_{12}}{x_{11}} = \frac{p_1}{p_2} = MRS_2 = \frac{1}{2}\frac{x_{22}}{x_{21}}$.
   Using the budget constraints for 1 and 2 and the above condition, we get the competitive allocation as $(\frac{1}{2}, \frac{1}{3})$, $(\frac{1}{2}, \frac{2}{3})$.
   To find the Pareto optimality condition, we re-write person 1's utility function as $u_1 = x_{11}x_{12} - (1 - x_{11})(1 - x_{12})$ and equate the MRS for 1 and 2:

$MRS_1 = \frac{(x_{12}+(1-x_{12}))}{(x_{11}+(1-x_{11}))} = 1.$
$MRS_2 = \frac{1}{2}\frac{x_{22}}{x_{21}}.$
Equating them, we get $x_{22} = 2x_{21}$. It can be easily seen that this condition is not satisfied at the competitive equilibrium that we calculated above.
Impose tax t_1 on person 2's consumption of good 1, and let T_1 and T_2 be the lump sum taxes or transfers. We want the equilibrium to be Pareto optimal; so we find a point that lies at the intersection of person 2's offer curve and person 1's "real"indifference curves. The latter are straight lines with slope 1. That point is given by $(\frac{2}{3}, \frac{1}{3}), (\frac{1}{3}, \frac{2}{3})$. For this point to be chosen, we need $MRS_1 = \frac{x_{12}}{x_{11}} = \frac{p_1}{p_2} = \frac{1}{2}$. Let $p_2 = 1$. Then $p_1 = \frac{1}{2}$.
Using person 1's budget constraint: $\frac{1}{2} \times \frac{2}{3} + \frac{1}{3} = \frac{1}{2} \times 1 + T_1$, which gives $T_1 = \frac{1}{6}$.
For person 2, we need: $MRS_2 = \frac{1}{2} \times \frac{x_{22}}{x_{21}} = \frac{p_1+t_1}{p_2}$, which gives $\frac{p_1+t_1}{p_2} = 1$. Since $p_1 = \frac{1}{2}$ and $p_2 = 1$, so $t_1 = \frac{1}{2}$. Using person 2's budget constraint:
$1 \times \frac{1}{3} + \frac{2}{3} = 1 + T_2$, which gives $T_2 = 0$.

3. a. Firm 1: For optimality, $p_2 MP_{11} = p_1$, which gives $y_{11} = -\frac{1}{4}$.
   Firm 2: For optimality, $p_2 MP_{21} = p_1$, which gives $y_{21} = -\frac{1}{4}$.
   Putting in the respective production functions we get, $y_{12} = \frac{2}{7}$, $y_{22} = \frac{2}{7}$. Aggregating for the economy, the net input-output vector is $(-\frac{1}{2}, \frac{4}{7})$.

   b. Close firm 2. Set $y_{11} = -\frac{1}{2}$. Then $y_{12} = \sqrt{\frac{1}{2}} = 0.707 > \frac{4}{7} = 0.571$.

Chapter 8

1. a. Aggregate net benefit from the public good is given by $\sum_{i=1}^{n} v_i(x) - x$.

   b. Maximize $\sum_{i=1}^{n} v_i(x) - x$ with respect to x. The first order conditions give $\sum_{i=1}^{n} v_i'(x) = 1$ which is the Samuelson optimality condition.

   c. Person 1 chooses x to maximize his utility $u_1 = v_1(x) + y_1$ subject to his budget constraint $1 \cdot x + 1 \cdot y_1 = 1 \cdot \omega_1$. The solution is given by a point that satisfies $v_1'(x) = 1$.
   Let $\hat{x}_1$ denote the quantity of the public good that 1 would

choose to purchase privately. So in the equilibrium, we will have $v_1'(\hat{x}_1) = 1$. Since we have assumed $v_i'(x) > 0$, so we have $\sum_{i=1}^{n} v_i'(\hat{x}_1) > 1$ which violates the optimality condition mentioned in (b).

2. a. The Samuelson optimality condition is:
$v_1'(x) + v_2'(x) + v_3'(x) = 1$
So we have
$\frac{a}{x} + \frac{b}{x} + \frac{c}{x} = 1$
which gives $\hat{x} = a + b + c$.

   b. Lindahl-Wicksell tax shares are calculated as $t_i = v_i'(\hat{x})$ which gives
$t_1 = \frac{a}{a+b+c}$.
$t_2 = \frac{b}{a+b+c}$.
$t_3 = \frac{c}{a+b+c}$.

   c. The virtues of the Lindahl-Wicksell scheme are: it ensures optimal provision of the public good and it links tax payments to benefits for each individual.The shortcomings of the scheme are that it does not ensure truthful demand revelation and may lead to inefficient outcome if people misrepresent their preferences.

3. a. For optimality we need,
$v_1'(x) + v_2'(x) = 1$
which gives $\frac{a}{x} + \frac{b}{x} = 1$,
so the optimal $\hat{x} = a + b$.

   b. Lindahl-Wicksell tax shares are calculated as $t_i = v_i'(\hat{x})$. So we have
$t_1 = \frac{a}{a+b}$
$t_2 = \frac{b}{a+b}$

   c. 1's after tax utility is given by-
$v_1(\hat{x}) - t_1(\hat{x})\hat{x} = a\ln(a+b) - \frac{a}{a+b} \times (a+b) = a\ln(a+b) - a$.
In case he claims his valuation to be 0, the amount of public good that will be provided is equal to b. Then person 1's after tax utility is given by $a\ln(0+b) - 0.b = a\ln b$. So misreporting his valuation to be 0 is better if $a\ln b > (a\ln(a+b) - a)$, that is, $\frac{a+b}{b} < e$.

   d. Demand revealing taxes can be formulated as:
$T_1 = \hat{x} - v_2(\hat{x}) = (a+b) - b\ln(a+b)$ and
$T_2 = \hat{x} - v_1(\hat{x}) = (a+b) - a\ln(a+b)$.

4. a. Given tax shares, person i solves:
$\max v_i(x_i) - \frac{1}{3}x_i$

The first order condition is:
$v_i'(x_i) = t_i$, which gives
for $i = 1$, $\frac{1}{2}ax_1^{-\frac{1}{2}} = \frac{1}{3}$, or, $\hat{x}_1 = (\frac{3}{2}a)^2$;
for $i = 2$, $\frac{1}{2}bx_2^{-\frac{1}{2}} = \frac{1}{3}$, or, $\hat{x}_2 = (\frac{3}{2}b)^2$;
for $i = 3$, $\frac{1}{2}cx_3^{-\frac{1}{2}} = \frac{1}{3}$, or, $\hat{x}_3 = (\frac{3}{2}c)^2$.
Since $a < b < c$, so $\hat{x}_1 < \hat{x}_2 < \hat{x}_3$. The majority voting equilibrium output for the public good is $\hat{x}_2$.

b. For optimality we need
$v_1'(x) + v_2'(x) + v_3'(x) = 1$
which gives $a\frac{1}{2}x^{-\frac{1}{2}} + b\frac{1}{2}x^{-\frac{1}{2}} + c\frac{1}{2}x^{-\frac{1}{2}} = 1$,
so the optimal $\hat{x} = (\frac{a+b+c}{2})^2$.

c. We need $(\frac{a+b+c}{2})^2 = (\frac{3}{2}b)^2$, that is $b = \frac{a+c}{2}$ for the answers in (a) and (b) to coincide.

Chapter 9

1. To show xCy implies xKy.

Proof. Consider $z_i = y_i + \frac{(\sum_{i=1}^n x_i - \sum_{i=1}^n y_i)}{n}$, for all $i = 1, \ldots, n$. Then $z \in S(x)$ and $z_i P_i y_i$, for all $i = 1, \ldots, n$. So zPy. Thus there exists $z \in S(x)$ such that zPy, implies xKy. □

To show xKy implies xCy.

Proof. xKy implies there exists $z \in S(x)$ such that zPy. $zPy \Leftrightarrow z_i R_i y_i$, for all $i = 1, \ldots, n$, and $z_i P_i y_i$, for at least some i. With monotonic preferences over a single good, $z_i R_i y_i \Leftrightarrow z_i \geq y_i$, for all $i = 1, \ldots, n$, and $z_i P_i y_i \Leftrightarrow z_i > y_i$. Summing over i, $\sum_{i=1}^n z_i > \sum_{i=1}^n y_i$. But $z \in S(x)$. Therefore, $\sum_{i=1}^n z_i = \sum_{i=1}^n x_i > \sum_{i=1}^n y_i$. Hence xCy. □

To show xKy implies xSy.

Proof. Suppose there exists $z \in S(y)$ such that $z_i R_i x_i$ for all $i = 1, \ldots, n$ and $z_i P_i x_i$, for some i. Then $z_i \geq x_i$, $\forall\ i = 1, \ldots, n$ and $z_i > x_i$, for some i. Summing we have, $\sum_{i=1}^n z_i > \sum_{i=1}^n x_i$. But $z \in S(y)$, so $\sum_{i=1}^n z_i = \sum_{i=1}^n y_i < \sum_{i=1}^n x_i$. Contradiction. □

To show xSy implies xNy.

Proof. That is to show that for any $z \in S(y)$, xKz. Suppose not. Then there exists $z \in S(y)$ such that zKx. Since $z \in S(y)$,

so $\sum_{i=1}^{n} z_i = \sum_{i=1}^{n} y_i$. Now from (ii) we know that zKx implies zCx, that is $\sum_{i=1}^{n} z_i > \sum_{i=1}^{n} x_i$. Contradiction. □

To show xNy implies xCy.

Proof. To show that $\sum_{i=1}^{n} x_i > \sum_{i=1}^{n} y_i$, given that for any z in $S(y)$, xKz. For any z in $S(y)$, there exists a $w \in S(x)$, such that wPz. That is, $w_i R_i z_i$, for all $i = 1, \ldots, n$ and $w_i P_i z_i$, for at least some i. Since there is only one good and we have monotonic preferences, we should have $w_i \geq z_i$ for all $i = 1, \ldots, n$ and $w_i > z_i$, for some i. Summing, $\sum_{i=1}^{n} w_i > \sum_{i=1}^{n} z_i$. But $w \in S(x)$, so $\sum_{i=1}^{n} w_i = \sum_{i=1}^{n} x_i$; and $z \in S(y)$, so $\sum_{i=1}^{n} z_i = \sum_{i=1}^{n} y_i$. So $\sum_{i=1}^{n} x_i > \sum_{i=1}^{n} y_i$. □

Chapter 10

1. a. By proposition 1, we know that if we begin from an equal allocation then the resulting competitive equilibrium is one that is Pareto optimal and fair. Let us calculate the competitive equilibrium when $\omega = (\frac{1}{2}, \frac{1}{2}), (\frac{1}{2}, \frac{1}{2})$.
$MRS_1 = \frac{x_{12}}{x_{11}} = \frac{p_1}{p_2} = MRS_2 = 2$. Let $p_2 = 1$.Then $p_1 = 2$. Using person 1's budget constraint:
$2x_{11} + x_{12} = 2 \times \frac{1}{2} + \frac{1}{2}$, and using the above condition we get the competitive equilibrium allocation as $(\frac{3}{8}, \frac{3}{4}), (\frac{5}{8}, \frac{1}{4})$. That it is Pareto optimal follows from the first welfare theorem. To check that it is fair,
$u_1(x_1) = \frac{9}{32} > u_1(x_2) = \frac{5}{32}$.
$u_2(x_2) = \frac{3}{2} = u_2(x_1) = \frac{3}{2}$.
Neither envies other's bundle. To make the economy attain this point, we impose lump sum taxes (or transfers) T_1 and T_2 . From person 1's budget constraint,
$2 \times \frac{3}{8} + \frac{3}{4} = 2 \times 1 + T_1$, which gives $T_1 = -\frac{1}{2}$.
Similarly from person 2's budget constraint, $2 \times \frac{5}{8} + \frac{1}{4} = 2 \times 0 + 1 + T_2$, which gives $T_2 = \frac{1}{2}$.

   b. Begin from the allocation $(\frac{1}{4}, 1), (\frac{3}{4}, 0)$. We can check that it is fair. Now consider the allocation given by $(\frac{5}{8}, \frac{1}{4}), (\frac{3}{8}, \frac{3}{4})$. This is Pareto superior to the original allocation because it makes person 1 strictly better-off while person 2 remains indifferent. But this unfair because person 1 envies person 2's bundle.

2. $u_1(x_1) = \frac{3}{2}$, $u_1(x_2) = \frac{3}{2}$. $u_1(y_1) = 2$, $u_1(y_2) = 1$.
$u_2(x_1) = \frac{3}{2}$, $u_2(x_2) = \frac{3}{2}$. $u_2(y_1) = 2$, $u_2(y_2) = 1$.

$Min(u_1(x_1), u_1(x_2) = \frac{3}{2} > Min(u_1(y_1), u_1(y_2) = 1$. Person 1 prefers x to y from behind a veil of ignorance.

$Min(u_2(x_1), u_2(x_2) = \frac{3}{2} > Min(u_2(y_1), u_2(y_2) = 1$. Person 2 prefers x to y from behind a veil of ignorance.

So x is Rawls superior to y.

Without the veil of ignorance, person 1 prefers y to x. And same for person 2. So allocation y is Pareto superior to x. Thus the Rawls criterion contradicts Pareto criterion.

Chapter 11

1. Expected Utility is given by: $u(x, y) = q(y) \cdot x^{\alpha} + (1 - q(y)) \cdot (x^{\alpha} - K) = x^{\alpha} - (1 - q(y))K$.
   Differentiate $u(x, y)$ with respect to x and y to get:
   $$\partial u(x, y)/\partial x = \alpha x^{(\alpha-1)}.$$
   $$\partial u(x, y)/\partial y = q'(y)K.$$
   Equating them we get , $q'(y) = \frac{\alpha}{K} \cdot x^{(\alpha-1)}$. The value of life is $1/q'(y)$, so VOL $= \frac{K}{\alpha} \cdot x^{(1-\alpha)}$.
2. Expected utility is given by: $u(x, y) = q(y) \cdot \ln(x) + (1 - q(y))(\ln(x) - K) = \ln(x) - (1 - q(y))K$.
   Differentiate $u(x, y)$ with respect to x and y to get:
   $$\partial u(x, y)/\partial x = \frac{1}{x}.$$
   $$\partial u(x, y)/\partial y = q'(y)K.$$
   Equating them we get , $q'(y) = \frac{1}{x \cdot K}$. Therefore, VOL $= x \cdot K$.
3. Compute TU$(a) = \sum_{i \epsilon N_a} u_i(a) = 0.01 + 80 + 80 = 160.01$.
   Compute TU$(b) = \sum_{i \epsilon N_b} u_i(b) = 80 + 20 = 100$.
   As per total principle, a is better because TU$(a) >$ TU(b).
   As per average principle, a is better because $\frac{1}{3} \cdot$TU$(a) = 53.3667 > \frac{1}{2} \cdot$ TU$(b) = 50$.
   As per critical principle, a is better because TU$(a) - 3 \cdot 20 = 100.01 >$ TU$(b) - 2 \cdot 20 = 60$.
   Let $\overline{u}$ be the critical level such that a and b are equally good. Then we should have TU$(a) - 3 \cdot \overline{u} =$ TU$(b) - 2 \cdot \overline{u}$, which gives us $\overline{u} = 60.01$.

4. The critical level principle weighs $\sum_{i \in N_a}(u_i(a) - \bar{u})$ and $\sum_{i \in N_b}(u_i(b) - \bar{u})$ when considering the alternatives a and b. If we have a person in the timeless population under a and b who enjoys the same utility under both the alternatives,say u_x, then the difference $u_x - \bar{u}$ affects the two terms under consideration equally, irrespective of the value that u_x takes. So utility independence is guaranteed.

Chapter 12

1. Black's theorem states that if single-peakedness is satisfied, the majority voting relation M is quasi-transitive.

   Proof. With 3 distinct alternatives x, y, z we need to show that if x beats y , y beats z then x beats z. If alternative a beats alternative b, then the number of people who vote for a or abstain in the a vs b contest, is greater than $\frac{n}{2}$.
   Case(1): The single-peakedness ordering is x, y, z. If i votes for x over y, then $u_i(x) > u_i(y)$ and $u_i(x) > u_i(z)$ for single-peakedness to hold. If i abstains from voting over x and y, then $u_i(x) = u_i(y)$. Again $u_i(x) > u_i(z)$ must hold. Since x beats y, the number of people voting for x or abstaining in the x vs y contest is greater than $\frac{n}{2}$. And all these people must vote for x over z. So x beats z.
   Case(2): The single-peakedness ordering is x, z, y. If i votes for y over z, then $u_i(y) > u_i(z)$ and $u_i(y) > u_i(x)$ for single peakedness to hold. If i abstains from voting over y and z, then $u_i(y) = u_i(z)$. Again $u_i(y) > u_i(x)$ must hold. Since y beats z, the number of people voting for y or abstaining in the y vs z contest is greater than $\frac{n}{2}$. And all these people must vote for y over x. So y beats x. Contradiction.
   Case(3): The single-peakedness ordering is y, x, z. Suppose z beats x. If i votes for z over x, then $u_i(z) > u_i(x)$ and $u_i(z) > u_i(y)$ for single- peakedness to hold. If i abstains from voting over z and x, then $u_i(z) = u_i(x)$. Again $u_i(z) > u_i(y)$ must hold. Since z beats x, the number of people voting for z or abstaining in the z vs x contest is greater than $\frac{n}{2}$. And all these people must vote for z over y.But this contradicts y beats z. So the supposition was wrong and x beats z.
   Case(4),(5) and (6) The single-peakedness orderings are y, z, x; z, x, y and z, y, x respectively. These cases are mirror images of cases 2, 3 and 1, respectively and the corresponding arguments apply. □

2. Corollary(2) states that if the set of alternatives is finite, the number of people n is odd, and the single-peakedness property is satisfied, then (1) there exists a unique alternative x that beats any other alternative y. (2) Moreover, there exists a social utility function U that exactly reflects the social preferences defined by majority voting. That is, for all alternatives x and y, xMy if and only if, $U(x) \geq U(y)$.

   Proof. (1) By Black's theorem 2, we know that if the single-peakedness property is satisfied then the majority voting relation is quasi-transitive. But quasi-transitivity implies acyclicity (Proposition 1, Chapter 1). Next, we look at the finite set of alternatives available and since M is complete and acyclic, then there must be at least one alternative that beats or ties any other alternative, that is xMy, for all y. (Proposition 3, Chapter 1).

   We want to show that this is unique. Suppose to the contrary that x and y are distinct, both beating or tying all other alternatives. Then x ties y. Since n is odd, there must be one person who is indifferent between x and y. And there must exist an alternative z lying between x and y such that he prefers z to both x and y for single-peakedness to hold. Now if we have a vote between x and z then, this person and everyone to the right, would vote for z over x. Moreover, this group of people would comprise a majority. Therefore z would beat x, a contradiction. Thus we cannot have a tie.

   (2) By Black's theorem 1, when the number of people is odd and the single peakedness property is satisfied then the majority voting relation M is transitive. Now if the set of alternatives is finite and M is complete and transitive, we can assign numerical values $U(x)$, $U(y)$ etc to the alternatives such that $U(x) \geq U(y)$, if and only if, xMy. (Proposition 5, Chapter 1). □

3. $m = \frac{n+1}{2}$. To show that when n is odd and single- peakedness ordering is $x_1, x_2, \ldots, x_n$, then x_m will win a majority over all other alternatives.

   Proof. Consider any x_i vs x_m. We can have two cases:
   Case(1): x_i is to the left of x_m. Then everyone to the left of x_i and person i will vote for x_i. And everyone to the right of x_m, including person m will vote for x_m. Since m is the median, we have more than half the votes for x_m. x_m beats x_i.

Case(2): x_i is to the right of x_m. Then everyone to the left of x_m, including person m will vote for x_m. Again since m is the median, we have more than half the votes for x_m. x_m beats x_i. □

Chapter 13

1. Suppose a person's first choice gets a weight of 7 points, the second choice gets a weight of 6 points, the third choice gets a weight of 4 points and the fourth choice gets a weight of 1 point. Then y gets 15 points, z gets 14 points, x gets 13 and w gets 12 points.
   Next suppose a person's first choice gets a weight of 8 points, the second choice gets a weight of 6 points, the third choice gets a weight of 2 points and the fourth gets a weight of 1 point. Then z gets 15 points, x gets 13 points, y gets 12 points and w gets 11 points.

2. a. This rule satisfies Pareto consistency because: consider xP_iy for all i, when $y \neq x_0$, then we have xPy because person 1 also prefers x to y. And when y is the same as x_0, then if xP_ix_0, for all i, we still have xPx_0 because 1 and 2 and one more person does prefer x to x_0.
   To check independence, note that since person 1 is the dictator for all alternatives except x_0, so if his feelings change about some irrelevant alternatives, but do not change about the pair in question, then the collective choice rule preserves the social ordering of the alternatives. For the special alternative: as long as the preferences of person 1, 2 and at least of one other person remains the same over x and x_0, with changes in preferences for irrelevant alternatives, the collective choice will remain unchanged.

   b. Consider the following preference ordering for four people:
   person 1: x to y to z to x_0.
   person 2: z to y to x to x_0.
   person 3: x to z to x_0 to y.
   person 4: x_0 to y to z to x.

   The collective choice rule says xPy, xPz, yPz, xPx_0, yRx_0, zPx_0.
   The rule is not quasi-transitive because then yPz, zPx_0, should imply yPx_0, but that is not the case.

   c. From the above ordering we can see that it must be acyclic.

   d. Collegium for the above choice rule is person 1.

e. No. It is not the case that when all members of the collegium prefer a certain alternative the social choice rule prefers that alternative. As we can see that though person 1 prefers y to x_0, the social choice rule says yRx_0.

Chapter 14

1. Person i reports v_i (which might be true or false). He is called *pivotal* if his report v_i satisfies:
$\sum_{j\neq i} v_j < C \leq \sum_{i=1}^{n} v_i$.
In this case, his report causes the bridge to be built. Note that in case he is pivotal, he has to pay a greater tax. Let us denote a false report by $\tilde{v_i}$ for person i. We will now examine the following cases:
Case(1): $\sum_{j\neq i} v_j < C \leq \sum_{j\neq i} v_j + v_i$, that is, person i's true report makes him pivotal. His payoff will be $v_i + \omega_i - T_i(v_i)$. Now suppose person i reports a false $\tilde{v_i}$ such that $\sum_{j\neq i} v_j + \tilde{v_i} < C$. Then the bridge will not be built. His payoff will be $0 + \omega_i - T_i(\tilde{v_i})$. His gain from misreporting is: $(\omega - T_i(\tilde{v_i})) - (v_i + \omega_i - T_i(v_i))$which gives $[\sum_{j\neq i} v_j + \tilde{v_i}] - C$ after a little simplification, and this is negative because $\sum_{j\neq i} v_j + \tilde{v_i} < C$. Thus he has no incentive to misreport.
Case(2): $\sum_{j\neq i} v_j < C \leq \sum_{j\neq i} v_j + v_i$, that is, person i's true report makes him pivotal. His payoff will be $v_i + \omega_i - T_i(v_i)$. Now suppose person i reports a false $\tilde{v_i}$ such that $\sum_{j\neq i} v_j + \tilde{v_i} \geq C$. Then the bridge will continue being built. His tax payment will not change, and so his payoff will also be the same. Thus again, no gain from misreporting.
Case(3): $\sum_{j\neq i} v_j + v_i < C$, that is, under person i's truthful report the bridge is not built. His payoff will be $\omega_i - T_i(v_i)$. Now suppose person i reports a false $\tilde{v_i}$ such that $\sum_{j\neq i} v_j + \tilde{v_i} \geq C$. Then the bridge will be built. His payoff will be $v_i + \omega_i - T_i(\tilde{v_i})$. His gain from misreporting is: $v_i + \omega_i - T_i(\tilde{v_i}) - [\omega_i - T_i(v_i)]$ which gives $\sum_{j\neq i} v_j + v_i - C$ after a little simplification, and this is negative because $\sum_{j\neq i} v_j + v_i < C$. Thus he has no reason to misreport.
Case(4): $\sum_{j\neq i} v_j + v_i < C$, that is, under person i's truthful report the bridge is not built. His payoff will be $\omega_i - T_i$. Now suppose person i reports a false $\tilde{v_i}$ such that $\sum_{j\neq i} v_j + \tilde{v_i} < C$. Then the bridge continues not being built. His tax payments will not change, and so his payoff will be the same. He has no incentive to misreport.

Cases(5), (6), (7) and (8) would deal with individual i not being pivotal in the building decision and reporting a false valuation which can either change the decision or which can leave it unaffected. We leave it to the student to show that in all of these cases there is no incentive to misreport.

To show: if $\sum_{i=1}^{n} \geq C$, then $\sum_{i=1}^{n} T_i \geq C$.

$T_i = \frac{1}{n-1} \sum_{j \neq i} v_j$, when i is not pivotal and
$T_i = \frac{1}{n-1} \sum_{j \neq i} v_j + C - \sum_{j \neq i} v_j$, when i is pivotal. So, T_i when i is pivotal is greater than T_i when i is not pivotal. Then $\sum_{i=1}^{n} T_i > \sum_{i=1}^{n} \frac{1}{n-1} \sum_{j \neq i} v_j$.
$= \frac{1}{n-1}(\sum_{j \neq 1}^{n} v_j + \sum_{j \neq 2}^{n} v_j + \sum_{j \neq 3}^{n} v_j + \ldots + \sum_{j \neq n}^{n} v_j)$.
$= \frac{1}{n-1}((n-1)(v_1 + v_2 + v_3 + \ldots + v_n)$.
$= \sum_{i=1}^{n} v_i$.
Since $\sum_{i=1}^{n} v_i \geq C$, we have $\sum_{i=1}^{n} T_i > C$.

2. a. Under the demand-revealing tax scheme, person i's tax is $T_i = \hat{x} - \sum_{j \neq i} v_j(\hat{x}) + \max_x \sum_{j \neq i}[v_j(x) - \frac{x}{n}]$.
If the bridge is not built, then $\hat{x}$ takes the value 0. Then the first two terms in the expression disappear. The third becomes:
$T_i = \max[\sum_{j \neq i}[v_j - \frac{C}{n}], 0]$ which can be rewritten as-
$T_i = \max[0, \sum_{j \neq i} v_j - \frac{n-1}{n} C]$.
If the bridge is built, then $\hat{x}$ takes value C. By an argument similar to one above, we have
$T_i = C - \sum_{j \neq i} v_j + \max[0, \sum_{j \neq i} v_j - \frac{n-1}{n} C]$.

b. Consider an example with 10 people, where persons 1 to 9 have a valuation of 11 dollars for the bridge if it is built and 0 otherwise and person 10 has a valuation of 0 dollars for the bridge. Let the cost of the bridge be 100 dollars. Since $\sum_{i=1}^{10} v_i = 99 < 100$, the bridge is not built.
We calculate the taxes for each person using the above formula: $T_i = \max[0, \sum_{j \neq i} v_j - (9/10) \cdot 100]$. So for persons 1 to 9, the tax is given by $T_i = \max[0, 8 \times 11 - 90] = \max[0, -2] = 0$.
But the tax for person 10 is given by $T_{10} = \max[0, 9 \times 11 - 90] = \max[0, 9] = 9$. So even though the bridge is not built, someone has to pay positive taxes.

3. Consider the following order of preferences:
   1 2 3 4
   x x y z
   y z z w
   z y x y
   w w w x

   If everyone ranks honestly, then the outcome is x. Person 4 may rank y to be first, because he prefers y to x. Then the social outcome would be y indifferent to x.

4. Consider the following order of preferences:
   1 2 3
   x z y
   y y z
   z x x

   If everyone gives 5 points to his most preferred alternative, 3 points to the second and 2 to the worst, y wins. But person 2 favors z to y. So he may put 10 points to z and 0 to the rest. Then the social rule would place z over y.

5. Result of exhaustive voting: in stage 1, A gets 5 votes, B gets 5 votes and C gets 4 votes. So C drops out. In Stage 2, A gets 5 votes and B gets 2 votes. So A wins. But people of type 2 like B over C over A. So they may manipulate in the first stage by giving one vote to C and one vote to A and none to B. That way, after the first stage voting, we have A getting 7 votes, B getting 3 votes and C getting 4 votes. So B drops out. In the next stage, A gets 3 votes and C gets 4 votes. So the outcome is C. Note that this is manipulation by a coalition of voters; the 2 type 2 people.

6. We leave it to the student.

Chapter 15

1. Consider the following hypothetical ranking of 4 feasible alternatives for two people.
   $R_1 : x_1 P_1 x_2 P_1 x^* P_1 \omega$
   $R_2 : x_1 P_2 x_2 P_2 \omega P_2 x^*$
   It can be easily verified that x^* is not Pareto optimal , as it is Pareto dominated by x_1 and x_2. So $F(R) = \omega$. Now consider a monotonic change of preferences around ω. One such profile can be
   $R_1' : x_1 R_1' x_2 R_1' x^* P_1' \omega$

R_2' : $x_1 R_2' x_2 R_2' \omega R_2' x^*$
Under this changed profile however, x^* is Pareto optimal. So $F(R') = x^*$. Thus $F(R)$ is not Maskin monotonic.

2. Note that the Groves-Ledyard tax scheme is Nash implementable. To see this, note that any Nash equilibrium of the Groves-Ledyard mechanism satisfies the Samuelson condition and is self-funded. (See Chapter 8).
Next, we show that for any $\hat{x}$ that satisfies the Samuelson condition, we can derive a set of $\triangle_j$'s and taxes such that they form an equilibrium set of announcements. Let $(\triangle_1, \triangle_2, \ldots, \triangle_N)$ solve the following N equations:

$$v_i'(\hat{x}) = \frac{1}{N} + \gamma \frac{N-1}{N}(\triangle_i - A_i), \text{ for all } i = 1, 2, \ldots, N$$

and define T_i as in the text. It is now straight-forward to show that $(\triangle_1, \triangle_2, \ldots, \triangle_N)$ is an equilibrium of the Groves-Ledyard mechanism.
So by Maskin's theorem, the Groves-Ledyard mechanism satisfies Maskin monotonicity.

3. a. Solomonic SDF says , $F(R) = a$ and $F(R') = b$. Mechanism g_1 is given by:
$g_1(\hat{A}, \hat{A}) = a$
$g_1(\hat{A}, \hat{B}) = g_1(\hat{B}, \hat{A}) = c$
$g_1(\hat{B}, \hat{B}) = b$
Case 1: True preference profile is R, i.e. A is the true mother. Then individual B will always announce $\hat{B}$. Given this the true mother will also announce $\hat{B}$ to prevent the baby being cut. Thus $(\hat{B}, \hat{B})$ is a Nash equilibrium . So $F(R) = b$, which doesn't satisfy the Solomonic SDF.
Case 2: True preference profile is R', i.e. B is the true mother. Then individual A will announce $\hat{A}$, and B fearing the baby being cut in half, will announce $\hat{A}$. Thus $(\hat{A}, \hat{A})$ is a Nash equilibrium. So, $F(R') = a$, which is different from the outcome of Solomonic SDF.

   b. The alternative mechanism does not Nash implement the Solomonic SDF because: if preference profile is R, i.e. A is the true mother, there is no pure strategy Nash equilibrium and so it is not Nash implementable. If preference profile is R', i.e. B is the true mother, then B will announce $\hat{B}$ and A will also announce $\hat{B}$ and get the baby, i.e. $(\hat{B}, \hat{B})$ is a Nash equilibrium. So $F(R') = a$ which, again does not implement the Solomonic SDF.

4. To show that weak Pareto correspondence of any implementation problem satisfies Maskin monotonicity.

   Proof. The weak Pareto correspondence prescribes the set $F^{WPO}(R) = \{x \in A \mid \text{there is no } y \in A \text{ such that } y_i P_i x_i \text{ for all } i\}$.
   To show that F^{WPO} is Maskin monotonic, we have to show that if $x \in F^{WPO}(R)$ then, $x \in F^{WPO}(R')$ where R' is a monotonic change of preferences around x.
   Suppose not. That is, when the preferences have undergone a monotonic change around x, and the new preference profile is given by R' , then $x \notin F^{WPO}(R')$. This implies, under R', there exists $y \in A$ such that $y_i P_i x_i$ for all i. But given that R' is derived from R by a monotonic change, we have $y_i P'_i x_i \Rightarrow y_i P_i x_i$ for all i. But that would mean x was Pareto dominated by y, even under R and hence could not have belonged to $F^{WPO}(R)$. A contradiction! □

5. To show that the Pareto correspondence of any implementation problem need not satisfy Maskin monotonicity.

   Proof. The Pareto correspondence prescribes the set $F^{PO}(R) = \{x \in A \mid \text{there is no } y \in A \text{ such that } y_i R_i x_i \text{ for all } i \text{ and } y_j P_j x_j \text{ for some } j\}$.
   To show that $F^{PO}(R)$ may not satisfy Maskin monotonicity, consider the following hypothetical ranking of 2 feasible alternatives for two people :
   $R_1 : x_1 P_1 x_2$
   $R_2 : x_2 P_2 x_1$
   Verify that the Pareto set consists of x_1 and x_2. So $F^{PO}(R) = \{x_1, x_2\}$. Now consider a monotonic change of preference around x_2 for the two persons. One such profile can be
   $R'_1 : x_1 P'_1 x_2$
   $R'_2 : x_2 I'_2 x_1$
   Under the changed profile, x_2 is no longer Pareto optimal. Thus, $x_2 \notin F^{PO}(R') = \{x_1\}$ which violates Maskin monotonicity. □

   To show: in exchange economies with continuous and monotone preferences,the $F^{PO}(R)$ will satisfy Maskin monotonicity.

   Proof. Suppose $x \in F^{PO}(R)$, but $x \notin F^{PO}(R')$ where R' is a monotonic change of preferences around x. This implies that

there exists y such that $y_i R_i' x_i$ for all i and $y_j P_j' x_j$ for some j. Consider the following allocation:

$$\hat{y} = \begin{cases} y_j - \epsilon_j \\ y_i + \frac{\epsilon_j}{N-1} \text{ for all } i \neq j \end{cases}$$

where $\epsilon = (\epsilon, \epsilon, \ldots \epsilon)$. Since preferences are continuous, there exists an ϵ such that $\hat{y}_j P_j' x_j$. Also, $\hat{y}$ is feasible and because of monotonicity of preferences, $\hat{y}_i P_i' x_i$ for all i. But because of Maskin monotonicity, we must have $\hat{y}_i P_i x_i$ for all i, which would be a contradiction to x being the optimal point under R. □

6. To show weak core correspondence satisfies Maskin monotonicity.

Proof. A feasible allocation $x \in A$ is improved upon by a coalition $S \subseteq N$ under R, if there exists a feasible S allocation $a_S \in A_S$ such that $x_i' P_i x_i$ for all $i \in S$. The weak core for R is the set of feasible allocations that cannot be improved upon by any coalition $S \subseteq N$. Let $C(R)$ denote the weak core correspondence under R.
To show : for all $x \in A$, if $x \in C(R)$ and if the preferences change from R to R' in a monotonic way around x, then $x \in C(R')$.
If $x \in C(R)$, then it has the property that it cannot be improved upon by a coalition $S \subseteq N$. Suppose now preferences change to R'. And suppose x is improved upon by a coalition $S \subseteq N$. That is, there exists, $x_S' \in A_S$ such that $x_i' P_i' x_i$ for all $i \in S$. But the preferences have changed in a monotonic way. So, we must have had $x_i' P_i x_i$ for all $i \in A_S$, which means x could have been improved upon by x_S and so x could not have belonged to the weak core. A contradiction! □

7. To show that the correspondence that assigns to each exchange economy the set of *all* its competitive equilibria may violate Maskin monotonicity. To see this, denote the competitive equilibrium SDF by F, and let $F(R) = x$ and x_i is the final consumption bundle assigned to agent i. Let x be a boundary point. This implies that there exist competitive equilibrium prices p such that the bundle x_i is the optimal choice for each agent over the budget set determined by prices p and endowment ω_i. Furthermore, $\sum_i x_i \leq \sum_i \omega_i$. Now consider a monotonic change of preferences around x_i for some or all i. The indifference curve passing through x_i now becomes steeper. If we allow $x_{ij} > \sum_i \omega_{ij}$,for any i, j,

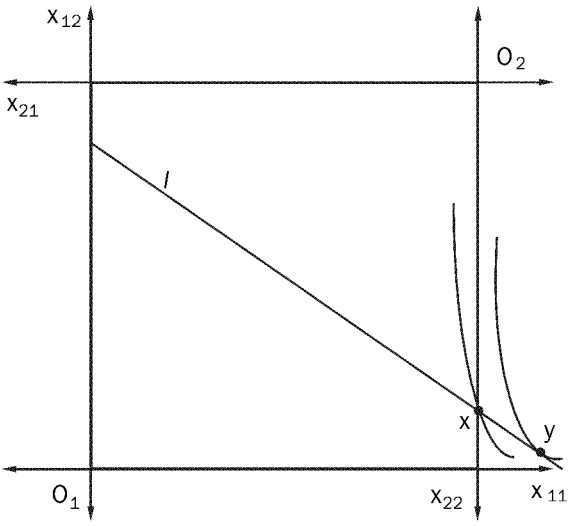


Figure 17.1.

then it might happen that x_i is no longer the optimal choice for i under prices p. Thus $F(R')$ is not x which violates Maskin monotonicity.The attached figure will help make this clear.

Here we have two agents both having straight line indifference maps initially. The competitive price ratio coincides with the slope of indifference curves, denoted by l, and so the competitive equilibrium is at some point x on the boundary where person 1 consumes whole of good 1 and some of good 2, while person 2 consumes only good 2. Now there is a monotonic change of preference around x for person 1. The new indifference curve has a steeper slope at x than p and so the indifference curves of the two individuals intersect at x, hence it cannot be a competitive equilibrium.

To show: The constrained Walrasian correspondence satisfies Maskin monotonicity.

Proof. For constrained Walrasian equilibrium, suppose x is a competitive equilibrium under R with prices p. Now suppose preferences change to R', monotonically around x. We can show that x continues to be the Walrasian equilibrium. To see this, note that under fixed p, the budget constraints remain the same for the two persons and since R' is a monotonic change from R, there cannot exist anything in the budget set that is preferred to x. □

Chapter 16

1. There are four Bayesian equilibria given the mechanism. They can be listed as:
   1. $m_2(t_2) = t_2,\ m_3(t_3) = t_3,\ m_2(t'_2) = t'_2,\ m_3(t'_3) = t'_3$. This is the truth telling equilibrium.
   2. $m_2(t_2) = t_2,\ m_3(t_3) = t_3,\ m_2(t'_2) = t_2,\ m_3(t'_3) = t_3$. In this equilibrium, truth telling holds for state t and not for state t'.
   3. $m_2(t_2) = t'_2,\ m_3(t_3) = t'_3,\ m_2(t'_2) = t_2,\ m_3(t'_3) = t'_3$. In this equilibrium, truth telling holds for state t' and not for state t.
   4. $m_2(t_2) = t'_2,\ m_3(t_3) = t'_3,\ m_2(t'_2) = t_2,\ m_3(t'_3) = t_3$. This is an equilibrium in which they all lie.

   That each of the above message profile is an equilibrium can be easily verified.
2. This mechanism does not fully implement F in Bayesian equilibrium. When the true state is t (or t') and agents 2 and 3 announce the truth, agent 1 knows that if he does not accuse, the outcome will be b (or a) which gives him a utility of 1. But if agents 2 and 3 have been lying then agent 1 will end up with a utility of -5. Since both states are equally likely, so his expected utility from not accusing is negative. On the other hand if he accuses, he ends up with a utility of 0. So agent 1 child chooses to accuse and the outcome is c always. Thus F is never implemented.
3. To do away with the budget surplus and maintain strategy-proofness of the mechanism, we can construct the transfers in the following way: $t_i(\hat{x}) = \sum_{j\neq i} v_j(\hat{x}) - \frac{1}{n-1}\sum_{j\neq i}(\sum_{k\neq j} v_k(\hat{x}))$. As can be readily checked that $\sum_i t_i(\cdot) = 0$ so this balances the budget. Also, since the transfer for any person does not explicitly depend on his announcement of his valuation, so the mechanism does not provide one with incentives to misreport.

Index