THE PHYSICAL FOUNDATION OF ECONOMICS

An Analytical Thermodynamic Theory

Jing Chen
School of Business
University of Northern British Columbia
Prince George, BC
Canada V2N 4Z9
Phone: 1-250-960-6480
Fax: 1-250-960-5544
Email: chenj@unbc.ca
Web: http://web.unbc.ca/~chenj/

Comments and reviews are welcome. Published by World Scientific. More information can be obtained at http://www.worldscibooks.com/economics/5819.html

Preface

After the works of Schrodinger (1944), Wiener (1948) and others, there is a consensus that life processes in general and human activities in particular are thermodynamic processes. Economic activities may be characterized and studied as mechanical processes, but at a more fundamental level of nature they are thermodynamic processes. Thus one would expect that more accurate economic theories could be built on the foundation of thermodynamics instead of rational mechanics, from which neoclassical economic theory emerged. However, the theory of thermodynamics only has had a very limited impact on economic theory to date as no analytical thermodynamic theory of economics has been established. In *The Coal Question*, first published in 1865, Jevons systematically discussed many aspects of economic activities from an energy perspective. Georgeucu-Roegen (1971) and others subsequently provided a more updated economic theory based on thermodynamics. Their approaches, however, remain qualitative. The lack of an analytical paradigm prevented thermodynamic theory from offering a detailed and quantitative analysis applicable to concrete economic and business problems. In this book, we will develop an analytical thermodynamic theory of life systems, which include social system as a special case, and show how it provides much more realistic and intuitive understanding of economic and social phenomena than neoclassical economics.

The entropy law states that systems tend toward higher entropy states spontaneously. Living systems, as non-equilibrium systems, need to extract low entropy from the environment to compensate for continuous dissipation. This process is the most fundamental property of life. It can be represented mathematically by lognormal processes, which contain a growth term and a dissipation term. From the entropy law, the thermodynamic diffusion of an organic or economic system is spontaneous. The extraction of low entropy from the environment, however, depends on specific biological or institutional structures that incur fixed or maintenance costs. Higher fixed cost systems generally have lower variable costs. In Chapter 3 of this book, we derive the thermodynamic equation that variable cost of a production system should satisfy, determine the initial value and then solve the thermodynamic equation to derive an analytic formula that explicitly represents the relation among fixed costs, variable costs, uncertainty of the environment and the duration of a production system, which is the core concern in most economic decisions. This analytical representation of various factors in production processes enables us to directly compute and analyze the returns of different production systems under various kinds of environment in a simple and systematic way.

Human activities are predominantly economic activities, which are chiefly regulated by the exchange value of different economic commodities. "The problem of value must always hold the pivotal position, as the chief tool of analysis in any pure theory that works with a rational schema." (Schumpeter, 1954, p. 588) The mainstream value theory, which is systematically represented in Debreu's *Theory of Value*, does not provide a measurable quantity for value. Since all human activities represent extraction and transformation of low entropy from the environment, it is natural to relate economic value to low entropy. From the properties that the value of commodities should satisfy, we derive that the only mathematical formula to represent value, as a function of scarcity, is the entropy function. This is parallel to the case where the only mathematical formula to represent information, as a function of probability, is the entropy function (Shannon, 1948). The entropy theory of value offers a unified understanding of physical entropy, information and economic value. It provides a quantitative measure of value that is highly consistent with our intuitive understanding. Just like the entropy theory of information provided a clear understanding of the fundamental problems in communication theory, the entropy theory of value provides clear understanding of the fundamental problems in social activities. In Chapter 2 of this book, we will discuss some of the conceptual difficulties that prevent the development of the entropy theory of value in the past and how they are resolved.

Many people do not agree that theories of social sciences should be derived from physical laws. They argue that physical laws are fixed while the human mind is free. However, the human mind is shaped by natural selection and sexual selection (Pinker, 1997). Living organisms need to extract low entropy from the environment, to defend their low entropy sources and to reduce the diffusion of low entropy. The struggle to stay in low entropy non-equilibrium states is called natural selection. Sexual selection is the struggle between the individuals of one sex, generally the males, to communicate their attractiveness to the other sex in order to form a partnership for reproduction. Human beings, as well as other sexually reproducing species, are the successful descendants of the earliest sexually reproducing species about a billion years ago (Margulis, 1998). For the system of communication to be successful in different kinds of environments over such a long time, the mode of communication has to be simple, stable and universal. Since the entropy law, which states that closed systems tend towards states of higher entropy, is the most universal law of the nature, it is natural that the display of low entropy levels evolves as the universal signal of attractiveness in the process of sexual selection. As both natural and sexual selection favor low entropy state, the pursuit of low entropy becomes the main motive in human mind and animal mind. In Chapter 1, we will show that some psychological patterns reflect the constraints of thermodynamic laws. Others are evolutionary adaptations to enable efficient processing of information, which is the reduction of entropy. Still others are mental attitudes that help us survive the constant dissipation of energy endured by all non-equilibrium systems. Therefore, entropy theory offers a unified understanding of the human mind. In this way, the understanding of matter and mind is unified on the foundation of physical laws. Just like the movement of particles is governed by physical laws, thinking is governed by physical laws, albeit in more complex ways.

In the rest of the book, we will apply the theory to understand other fundamental problems in social sciences. All the problems are analyzed with the unified methodology. This is in sharp contrast to analysis based on neoclassical economic theory, where many different models are developed for different problems. Even patterns easily understood by most people, such as increasing returns, have to be described by extremely ingenious and arcane mathematical models. Very often reality cannot fit into these models very well. The discrepancy between theory and reality is often attributed to imperfection of reality, such as "imperfect market", "imperfect information", "imperfect contract", "imperfect competition", "inefficient property right", "market failure", "government failure", "externality". A brief review of the concept of imperfection in old astronomy will help us gain more understanding about it. Ancient people had long observed that stars moved in perfect harmony in the sky. Several planets, however, moved in irregular trajectories. It was thought that this was caused by the imperfectness of the planets. There were many elaborate theories why the planets were imperfect. Kepler, however, derived that all planets moved in perfect elliptic orbits around the sun. This story tells us that "imperfection of the world" often reflects imperfection of the theory that is used to understand the world instead of the world itself. In the book, we will show how this analytical thermodynamic theory offers a unified understanding of various "imperfection" or "externality".

Since a thermodynamic equation is of first order in temporal dimension, social and biological systems as thermodynamic systems are intrinsically evolutionary and dynamic, which is very different from the static view of general equilibrium theory. Given the dynamic nature of our economic activities, one may wonder why an equilibrium theory becomes the dominant paradigm in economics. Schumpeter apparently anticipated such a question when he discussed the concept of equilibrium:

> Now, an observer fresh from Mars might excusably think that the human mind, inspired by experience, would start analysis with the relatively concrete and then, as more subtle relations reveal themselves, proceed to the relatively abstract, that is to say, to start from dynamic relations and then proceed to working out the static ones. But this has not been so in any field of scientific endeavor whatsoever: always static theory has historically preceded dynamic theory and the reasons for this seem to be as obvious as they are sound --- static theory is much simpler to work out; its propositions are easier to prove; and it seems closer to (logical) essentials. The history of economic analysis is no exception. (Schumpeter, 1954, p. 964)

Neoclassical economics was founded around 1870 by Jevons, Walras and others, who believed that economics should be built on a sound physical foundation. When neoclassical economics germinated in the 1870s, work on statistical mechanics by Boltzmann happened to appear in the same decade. However it was more than thirty years later that Boltzmann's theory was generally accepted by the physicist community. Since the dominant platform of physics in Jevons and Walras' time was Newtonian mechanics, it was natural for them to adopt this platform. However, Jevons clearly pointed out that "I believe that dynamical branches of the Science of Economy may remain to be developed, on the consideration of which I have not at all entered." (Jevons, 1957, p. vii)

The analytical thermodynamic theory, as a dynamic theory of economics, is a natural continuation of their pursuits. This was very much like statistical mechanics was a natural extension to Newtonian mechanics, although it took many years for people to realize that (Isihara, 1971). Most people today agree that statistical mechanics is a sounder foundation to describe living systems than Newtonian mechanics, which is the physical foundation of general equilibrium theory in economics. As it is often the case, an analytical framework that is built on sounder foundation of physical science delivers more intuitive and simpler results. Most of the contents in the book have been taught at undergraduate classes and the students embrace the ideas enthusiastically. Since all results in this book are simple analytical formulas, they can be applied easily by researchers and students. Many examples on different applications are provided in the book. Although the book is written as a research monograph, it can be used as a textbook or reference book for many courses.

This analytical thermodynamic theory transforms the study of human societies into an integral part of physical and biological sciences. A unified framework enables us to utilize knowledge gained from biology and various branches of economics and social sciences systematically to understand particular problems. The mainstream economic theories often emphasize the uniqueness of human beings that somehow enables us to escape the physical constraints that is binding for all other living organisms. Of course, human beings are unique. All species are. But the progress of science is often marked by further recognition that human beings are not “higher” than other living organisms.

I have been contemplating an analytical thermodynamic theory of life systems since I was a college student more than twenty years ago. My neglect of other things often put me into difficult situations. I am very grateful to Nam Sang Cheng, Jin-Chuan Duan, Yonggeng Gu, Hemantha Herath, Ling Hsiao, Joel Smoller, Changyun Wang, Michael Wong, Lixin Wu, Shing-Tung Yau and many others for their kind help in my difficult times.

Many people have provided helpful comments or encouragement during various stages of the development of the theory. They include Rishi Basak, Avi Bick, David Carter, Liang Chen, Sungchul Choi, Steven Clark, Mick Common, Alan Conway, Ted Creese, Balbinder Deo, Dale Domian, Han Donker, Carmen Durney, Tony Fang, David Faulkner, Barry Feldman, Flobert Feng, Michael Ghiselin, Peter Hwang, Michael Jongmans, Kevin Keen, Xiaoying Kong, Jens Joerg Lauschke, Yan Ma, Donald Manson, Bill McGill, Alex Ng, Brendon Ogmundson, Peter Ostrowski, Raluca Iorgulescu Polimeni, Eric Rapaport, Frank Schmid, Myron Scholes, Wei Shao, Jeff Smith, Dietrich Stauffer, Joseph Vogel, Richard West, Bill Williams, Morgen Witzel, Yi-Cheng Zhang, Ralf Zurbrugg and many others. Ted Creese, Kevin Keen, Peter Ostrowski and Richard West read part of the manuscripts and offered very helpful comments. My papers, most of which are about various aspects of this theory, listed on SSRN have generated more than six thousand downloads. I would like to thank readers for taking an interest on it.

I met Yubing Zhai, Commissioning Editor of World Scientific, in ASSA meeting in San Diego and told her about my plan to write the book. She browsed my website and encouraged me to submit a book proposal. This was how the book was originated. I am very grateful to Yubing for taking an active interest to a new theory from an unknown author. Cheong Chean Chian and Huang Wei guide me through the details of publication matters.

A person is a product of the environment. The unconditional and complete love from my parents gives me the peace of mind to work at this theory for more than twenty years without any visible progress until recent years. My sister has been a constant source of support and encouragement.

My single minded pursuit of this theory over the years was unfortunately not conducive to steady employment. But this circumstance was mitigated by the unwavering support of my wife, Coral for whom I reserve the utmost gratitude and thanks. Even in our most difficult time, Coral allowed me to pursue my own interest. She shouldered all the burden and anxiety for the family while I worked on the theory, which would never have been developed without her love and devotion.

Jing Chen

Contents

Chapter 1

The Entropy Theory of Human Mind

1.1. Introduction

People generally think that physical laws only have limited utility in understanding human behavior because our mind is free. However, human mind is shaped by natural selection and sexual selection (Pinker, 1997). Living organisms need to extract low entropy from the environment, to defend their low entropy sources and to reduce the diffusion of the low entropy. The struggle to stay in low entropy states is called natural selection. In human societies, agriculture is the main low entropy source. Part of health care systems aim at defending our own low entropy sources to be accessed by viruses and bacteria. The military forces are established to extract low entropy from others and to defend own low entropy sources. Clothing and housing reduces the diffusion of low entropy.

Sexual selection is the struggle between the individuals of one sex, generally the males, to communicate their attractiveness to the other sex in order to form a partnership for reproduction. Human beings, as well as other sexually reproducing species, are the successful descendants of the earliest sexually reproducing species about a billion years ago (Margulis, 1998). For the system of communication to be successful in different kinds of environments over such a long time, the mode of communication has to be simple, stable and universal. Since the entropy law, which states that closed systems tend towards states of higher entropy, is the most universal law of the nature, it is natural that the display of low entropy levels evolves as the universal signal of attractiveness in the process of sexual selection.

As both natural selection and sexual selection favor low entropy state, the pursuit of low entropy becomes the main motive of human mind and animal mind. Indeed the low entropy state is the main way of advertisement for most sexually reproducing species. Large body size, colorful and highly complex feather patterns with large amount of information content and exotic structures are all different representations of low entropy states. Since a low probability event corresponds to a state of low entropy, a novel feature is often attractive in the competition for reproduction. It has been generally recognized that sexual selection is the main drive of diversity (Miller, 2000).

Besides communication with members of the opposite sex, social animals need to communicate their attractiveness and power in order to influence the behavior of others (Wilson, 1975). For the same reason as in sexual selection, the most general signal is display of low entropy. Among all social species, human beings have developed the most complex social structure. The creation of distinct art works, the demonstration of athletic prowess, the accumulation of wealth, and conspicuous consumption - all of which represent different forms of low entropy - are the major methods of advertising one's attractiveness.

As the social groups become larger and the division of labor becomes finer, people become less familiar with each other in their daily interactions, which make it more difficult for people to judge the ability of others. The need for people to advertise their attractiveness through external accumulation of low entropy also becomes stronger. People usually signal their capability by buying more expensive houses, cars, clothes, going to more expensive restaurants and attending more exclusive schools. The great efforts human beings put into non-food activities reflect the high cost of communication in a large and complex society. Historical evidences show that the transaction costs have been increasing over time (Wallis and North, 1986).

The main function of mind is information processing. The concept of information has been intimately related to entropy for over a century. In a thought experiment, Maxwell (1871) reasoned, if information is costless, the entropy of a system can be decreased. But this would violate the second law of thermodynamics. Maxwell went on to conclude that the physical cost of obtaining information must be at least as much as the value of information. Many years later Shannon (1948) identified information as entropy formally, at least at the mathematical level. (Shannon, 1956)

The remainder of the chapter is structured as follows. Section 1.2 introduces the generalized entropy theory of information. Information theory provides natural measures of the cost of obtaining information and of information asymmetry. Section 1.3 shows that entropy theory offers a unified understanding of the patterns of human psychology. Section 1.4 concludes.

1.2. What is Information?

The value of information is a function of probability and must satisfy the following properties:

(a) The information value of two events is higher than the value of each of them.
(b) If two events are independent, the information value of the two events will be the sum of the two.
(c) The information value of any event is non-negative.

The only mathematical functions that satisfy all the above properties are of the form

$$H(P) = -\log_b P \tag{1.1}$$

where H is the value of information, P is the probability associated with a given event and b is a positive constant (Applebaum, 1996). Formula (1.1) represents the level of uncertainty. When a signal is received, there is a reduction of uncertainty, which is information.

Suppose a random event, X, has n discrete states, $x_1, x_2, \ldots, x_n$, each with probability $p_1, p_2, \ldots, p_n$. The information value of X is the average of information value of each state, that is

$$H(X) = -\sum_{j=1}^{n} p_j \log(p_j) \qquad (1.2)$$

The right hand side of (1.2), which is the entropy function first introduced by Boltzmann in the 1870s, is also the general formula for information (Shannon, 1948).

After the entropy theory of information was developed in 1948, its technique has been applied to many different problems in economic and finance. (Theil, 1967; Maasoumi and Racine, 2002 and many others) However, the standard economic theory of information, represented by Grossman and Stiglitz (1980) was not built on the foundation of entropy theory. An entropy theory based economic theory of information can be simply stated as:

> Information is the reduction of entropy, not only in a mathematical sense, as in Shannon's theory, but also in a physical sense. The rules of information transmission developed in Shannon's theory, as mathematical rules, apply not only to communication systems, but also to all living organisms.

In the following, we will discuss some distinct properties of this new information theory. First, information that is more valuable is in general more expensive to obtain. From the second law of thermodynamics, Maxwell concluded that information of higher value is of higher physical cost. Since economic cost is highly correlated to physical cost, (Georgescu-Roegen, 1971) more valuable information is in general more expensive to obtain. The relation among entropy, information and economic value will be discussed in greater detail in Chapter 2.

Second, the amount of information one can receive depends on the person's background knowledge about that particular information. The most important result from Shannon's entropy theory of information is the following formula

$$R = H(x) - H_y(x) \qquad (1.3)$$

where R is the amount of information one can receive, H is the amount of information a source sent and $H_y(x)$, the conditional entropy, is called equivocation. Formula (1.3) shows that the amount of information one can receive would be equal to the amount of information sent minus the average rate of conditional entropy. Before Shannon's theory, it was impossible to accurately assess how much information one can receive from an information source. In communication theory, this formula is used to discuss how noises affect the efficiency of information transmission. But it can be understood from more general perspective. The level of conditional entropy $H_y(x)$ is determined by the correlation between senders and receivers. When x and y are independent, $H_y(x) = H(x)$ and $R = 0$. No information can be transmitted between two objects that are independent of each other. When the correlation of x and y is equal to one, $H_y(x) = 0$. No information loss occurs in transmission. In general, the amount of information one can receive from the source depends on the correlation between the two. The higher the correlation between the source and receiver, the more information can be transmitted.

The above discussion does not depend on the specific characteristics of senders and receivers of information. So it applies to human beings as well as technical communication equipments, which are the original focus in information theory in science and engineering. However, the laws that govern human activities, including mental activities, are the same physical laws that govern non-living systems.

$H_y(x)$ in Formula (1.3) offers the quantitative measure of information asymmetry (Akerlof, 1970). Since different people have different background knowledge about the same information, heterogeneity of opinion occurs naturally. To understand the value of a new product or new production system may take the investment public several years. To fully appreciate the scope of some technology change may take several decades. For example, the economic and social impacts of cars as personal transportation instruments and computers as personal communication instruments were only gradually realized over the path of several decades. This is why individual stocks and whole stock markets often exhibit cycles of return of different lengths. This property is very different from Grossman-Stiglitz information theory, where economic agents can recognize the value of information instantly and pay according to its value.

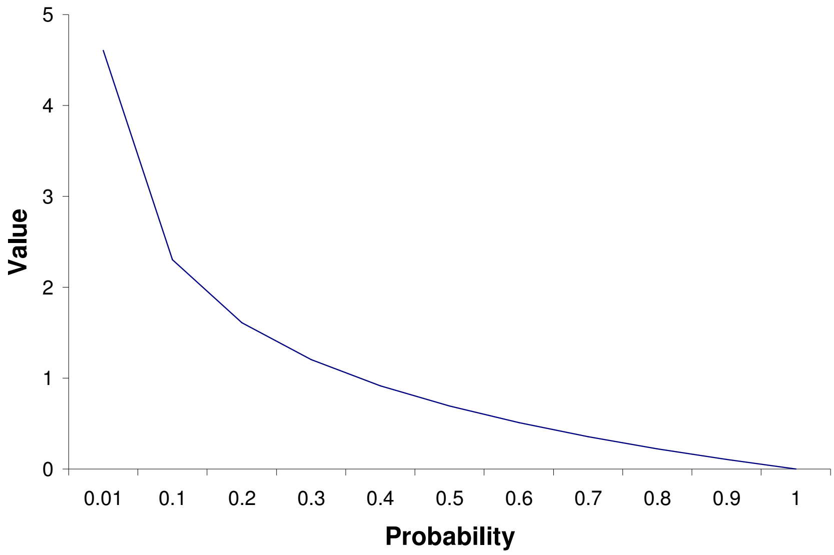

Figure 1.1 Information value and probability

Third, the same information, when known to more people, becomes less valuable. Figure 1.1 is a graph of (1.1), where H is a function of P, the probability of any given event. From Figure 1.1, value is a decreasing function of probability. In the standard information theory, P represents the probability that some event will occur. In this theory, P is generalized to represent the percentage of people or money that is controlled by informed investors. When $P = 1$, $-\log P = 0$. Thus the value of information that is already known to everyone is zero. When P approaches zero, $-\log P$ approaches infinity. Therefore, the value of information that is known to few is very high. The following example will illustrate this point. Figure 1.2 shows overnight rate of return and trading volume of shares of WestJet, a Canadian airline, surrounding the announcement of the bankruptcy of Jetsgo, the main competitor of WestJet. Jetsgo announced bankruptcy at the evening of March 10, 2005. If one bought stock at March 10, he would have made a return of 40% overnight. Judging from the trading volume of March 9, some people did buy WestJet stock before information was released to the public. After the announcement made the information public, trading volume was very high and the rate of return is near zero. Figure 1.2 neatly illustrates the relation between value of some information and the number of people who know the information.

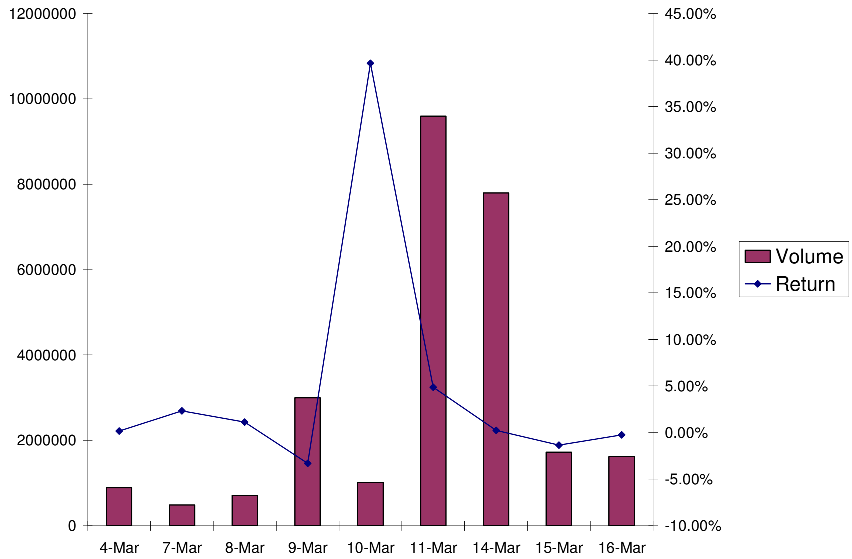

Figure 1.2: Overnight rate of return and trading volume of WestJet stock surrounding the date when Jetsgo announced bankruptcy

It is often said that the cost of information has dropped sharply over the years. But at the same time, the value of the same type of information has dropped sharply as well. Information of high value is usually carefully guarded and difficult to detect. For example, Warren Buffett, who has a very successful record for gaining and using insightful market information, would not announce to the public which stock(s) he is going to buy or sell. Animals have discovered this long ago. "In those cases where animal signals really are of mutual benefit, they will tend to sink to the level of a conspiratorial whisper: indeed this may often have happened, the resulting signals being to inconspicuous for us to have noticed them. If signal strength increases over the generations this suggests, on the other hand, that there has been increasing resistance on the side of the receiver." (Dawkins, 1999, p. 59)

Unlike Grossman-Stiglitz information theory, this information theory is a non-equilibrium theory. It does not assume a company possesses some intrinsic value waiting to be discovered by the investment public. Instead, the process of understanding the value of a company by the investment public is accompanied by the process of understanding the technology and market potential by its competitors, which generally reduce the value of that particular company. Empirical evidences that we will present in Chapter 6 support this statement.

When this theory was first proposed, it was treated as a natural extension of Shannon's entropy theory of information (Chen, 2002a). However, many have pointed out that Shannon (1956) would have a different view:

> Workers in other fields should realize that that the basic results of the subject are aimed at a very specific direction, a direction that is not necessarily relevant to such fields as psychology, economics, and other social sciences. Indeed, the hard core of information theory is essentially, a branch of mathematics, a strictly deductive system. (Shannon, 1956)

Recent authority reaffirmed this orthodox view:

> The efforts of physicists to link information theory more closely to statistical physics were less successful. It is true that there are mathematical similarities, and it is true that cross pollination has occurred over the years. However, the problem areas being modeled by these theories are very different, so it is likely that the coupling remains limited.
>
> In the early years after 1948, many people, particularly those in the softer sciences, were entranced by the hope of using information theory to bring some mathematical structure into their own fields. In many cases, these people did not realize the extent to which the definition of information was designed to help the communication engineer send messages rather than to help people understand the meaning of messages. In some cases, extreme claims were made about the applicability of information theory, thus embarrassing serious workers in the field. (Gallager, 2001, p. 2694)

Since this new information theory can be applied to much broader fields than Shannon's theory, it may be called the Generalized Entropy Theory of Information. In the next section and the rest of the book, we will show that establishing information theory on the foundation of statistical physics will yield great understanding of biological and social activities and turn "softer sciences" into a hard science.

1.3. The Entropy Theory of Human Psychology

Ideas occur at a blink of eye, which gives us impression that thinking is effortless. This seeming effortless, however, is the result of highly active brain system that consumes large amount of energy. Metabolically the brain is a very expensive organ. Representing only 2 percent of body mass, the brain uses about 20 percent of energy in humans (Pinker, 1997). Since mental activities are so costly, it is of great evolutionary advantage to have efficient information processing system to reduce the cost of thinking.

From the entropy law, we know that it is far easier for a system to disintegrate than to maintain its structure (Morowitz, 1992; Margulis, 1998). So there is a strong selective pressure for important knowledge to become genetically coded into heuristic principles to reduce the cost of learning (Tversky and Kahneman, 1974). Natural selection determines that human minds are born with many data and preferences stored (Pinker, 1997). In this section we will show that entropy theory offers a unified understanding of some frequently cited patterns of human psychology.

1. Conservatism

Conservatism in human beings may be characterized as behavior by individuals who possess a reluctance to update their beliefs in the face of new information. This property is a natural result from information theory. From Formula (1.3), the information one can receive is information sent minus equivocation, which is reduced gradually as the receiver's background knowledge about the source increases. Hence conservatism reflects the gradual reduction of equivocation by the receiver of any given information.

2. Herding behavior

From the second law of thermodynamics, a random action generally costs more than it gains. To concentrate actions into profitable ones, we, like wild animals, often learn from the experience of successful individuals and copy their behavior. It is generally very costly and impossible to repeat all of the experiences and mistakes that are possible. Therefore, we accept certain modes of behavior demonstrated by others without completely investigating the reasons behind them. Copying the actions of others directly is much easier, i.e., more efficient. Herding mentality developed because it is a cost-effective way of learning most of the time.

Human beings are social animals. Herding, or following the crowd, is good for survival. If you have walked alone in the wildness, you must have acute sense of vulnerability and powerlessness of human beings as individuals. Moose, bear and other animals can overrun human beings easily. It is only in crowds that we become powerful. So herding is essential for survival. A person who goes his own way usually cannot survive long. Herding mentality is evolved in this way.

It should be emphasized that all learning, especially institutionalized learning, are herding behavior. There are infinitely many things to explore. But we only have finite time and energy. The choice of subject to learn is from past experience. For example, when IT professionals earn high income, many people choose to get degrees in IT area.

3. Overconfidence and irrationality

"Extensive evidence shows that people are overconfident in their judgments" (Barberis and Thaler, 2003). From entropy law, any biological system, as a non-equilibrium system, faces constant dissipation of energy. Endless efforts are required to maintain a non-equilibrium system. Entropy law has been intuitively understood since ancient times. "The gods had condemned Sisyphus to ceaselessly rolling a rock to the top of a mountain, whence the stone would fall back of its own weight. They had thought with some reason that there is no more dreadful punishment than futile and hopeless labour. … If this myth is tragic, that is because its hero is conscious. … The workman of today works every day in his life at the same tasks and this fate is no less absurd. But it is tragic only at the rare moments when it becomes conscious." (Camus, 1955, p. 109) In the long course of evolution of our solar system, all life on earth will eventually go extinct in the far distant future (Lovelock, 1988). From a purely rational perspective, life is meaningless. Since human beings are self-conscious, the very question of why life is worth living lingers in many people's minds. "There is but one truly serious philosophical problem and that is suicide. Judging whether life is or is not worth living amounts to answering the fundamental question of philosophy" (Camus, 1955, p. 11). Overconfidence and irrationality are adaptive psychological traits that help us survive in this world.

The prevalence of irrationality is reflected in the prevalence of religious beliefs in various forms. A fundamental characteristic of various religions is that they are built on some miracles that are not consistent with physical or biological laws, such as virgin birth, sustainable growth or infinite human creativity (Daly, 1991). Marx (1844) once noted:

> Religion is the sigh of the exhausted creature, the heart of a heartless world and the soul of the soulless conditions. It is the opium of the people.
>
> The abolition of religion as the illusory happiness of the people is a demand for their true happiness. The call to abandon illusions about their condition is the call to abandon a condition that requires illusions.

Because of the inexorable increase of entropy in the universe, the condition that requires illusion will never leave us.

4. Loss aversion in winning and risk seeking in losses

Human beings often exhibit loss aversion in winning and risk seeking in losses. Kahneman and Tversky (1979) collected some responses to hypothetical choice problems. In one problem, the subjects were presented with two choices.

Choice A: There is an 80% probability of winning 4000 pounds and a 20% probability of winning nothing.

Choice B: There is a certainty of winning 3000 pounds.

The expected end wealth of choice A is 3200 and of choice B is 3000. Most respondents chose B, exhibiting loss aversion in winning. When the signs of the outcomes are reversed, the problems become the following:

Choice C: There is an 80% probability of losing 4000 pounds and 20% probability of losing nothing.

Choice D: There is a certainty of losing 3000 pounds.

The expected end wealth of choice C is -3200 and of choice D is -3000. Most respondents chose C, exhibiting risk seeking in losses. As money is a new invention in human evolutionary history, the preference for money must be derived from something else. Since food is the most important resource of our evolutionary past, our preference for wealth is probably derived from our preferences for food.

In the most part of the history of human evolution, we had not been able to store large amounts of extra food. If one goes without food for several days, he will starve. We translate the monetary numbers from the above four questions into days of food to obtain the following. In the case of gain, we can think of the choices of two possible strategies. In the first strategy, there is an 80% probability of getting food for 40 days and 20% chance of getting nothing. In the second strategy, there is a certainty of getting food for 30 days. It is easy to see why most people will prefer 30 days of food in certainty over a strategy that contains a 20% risk of getting nothing. In the case of loss, we can think of the choices of two possible strategies. In the first strategy, there is an 80% probability losing food for 40 days and a 20% chance of losing nothing for 40 days. In the second strategy, there is a certainty of getting no food for 30 days. Since without food for 30 days will represent sure death, people will naturally choose a 20% chance of survival. So people consistently avoid risk in both positive gain and negative loss. “Risk seeking” in loss is an unfortunate terminology borrowed from utility theory.

An important human institutional invention is the limited liability corporations, which alleviate people’s concern about unlimited loss. This greatly increases people’s willingness to explore new ideas and contributes to rapid economic growth. Human psychology has long been applied to the design of institutional structures.

5. Framing, representativeness and biases

We often frame, or sort problems into categories and assign different associated values based on the perceived relative levels of importance of each problem. Why do we do this? The following result from statistical physics helps answer this question.

If $\{p_1, \ldots p_n\}$ and $\{q_1, \ldots q_n\}$ are two sets of probabilities, then

$$-\sum_{j=1}^{n} p_j \log(p_j) \leq -\sum_{j=1}^{n} p_j \log(q_j) \qquad (1.4)$$

with equality achieved if and only if each

$$q_j = p_j, \qquad 1 \leq j \leq n$$

This result is called Gibbs inequality (Isihara, 1971). In Gibbs inequality, p_j can be understood as the probability of event j in nature and q_j is the subjective probability of our assessment of that event. The left hand side of Formula (1.4) is the average uncertainty of events and the right hand side is the uncertainty of our subjective assessment of those events. In general, the difference between the left hand side and right hand side of Formula (1.4) is smaller when q_j is closer to p_j. This means that information processing is more efficient when the subjective probabilities are closer to the objective probabilities. In particular, a mind with stored data about the natural environment is in general more efficient than a completely unbiased mind, where all subjective probabilities are to be learned from scratch. Natural selection determines that the human mind will evolve so that, "in general, instances of large classes are recalled better and faster than instances of less frequent classes; that likely occurrences are easier to imagine than unlikely ones; and that the associative connections between events are strengthened when the events frequently co-occur" (Tversky and Kahneman, 1974, p.1128).

Thus, "people rely on a limited number of heuristic principles which reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations. In general, these heuristics are quite useful, but sometimes they lead to severe and systematic errors"

(Tversky and Kahneman, 1974, p.1124). What causes these severe and systematic errors? Human minds are the result of natural selection, which "operates over thousands of generations. For ninety-nine percent of human existence, people lived as foragers in small nomadic bands. Our brains are adapted to that long-vanished way of life, not to brand-new agriculture and industrial civilizations" (Pinker, 1997, p. 42). This is why we observe systematic errors in judgment by human beings, i.e., the typical framework for processing information today was developed over millennia when environmental conditions were very different. For example, most of us still have a great fear of snakes, although they rarely pose a threat to urban dwellers today. On the other hand, fear of electricity has to be instilled into children's minds with great difficulty (Pinker, 1997).

From Gibbs inequality, the level of uncertainty in understanding a type of events is

$$-\sum_{j=1}^{n} p_j \log(q_j)$$

where p_i and q_i are objective and subjective probabilities respectively. Suppose this type of events has two possible outcomes, state 1 and state 2. The probability of state 1 is 90% and the probability of state 2 is 10%. An expert on this type of events may correctly estimate these probabilities and for her the uncertainty in prediction is

$$-0.9\ln 0.9 - 0.1\ln 0.1 = 0.33$$

A novice, who has no priori knowledge on these events, may assign 50% probability to each outcome. For her the uncertainty in prediction is

$$-0.9\ln 0.5 - 0.1\ln 0.5 = 0.69$$

It is clearly that the expert, who has accumulated knowledge through long time experience, has better estimation than novice in a stable environment.

Now assume the environment experiences some fundamental change and the new probabilities of state 1 and state 2 become 10% and 90%

respectively. This time, the uncertainty of the prediction by the expert, who still uses the old probability, is

$$-0.9\ln 0.1 - 0.1\ln 0.9 = 2.08$$

while the uncertainty of prediction by a novice is

$$-0.9\ln 0.5 - 0.1\ln 0.5 = 0.69$$

This shows that when environment changes suddenly, novice actually perform better than experts, whose priori knowledge often cause severe biases in prediction. This is one reason why scientific revolutions are often initiated by and new industries are often pioneered by newcomers or outsiders (Kuhn, 1996; Stearns and Allan, 1996).

1.4. Concluding Remarks

From the discussion in the last section, we find that some psychological patterns, such as conservatism, reflect the constraints of thermodynamic laws. Others, such as framing and herding, are evolutionary adaptations to enable efficient processing of information, which is the reduction of entropy. Still others, such as overconfidence, irrationality and loss aversion, are mental attitudes that help us survive the constant dissipation of energy endured by all non-equilibrium systems. Therefore, entropy theory offers a unified understanding of human mind. This theory shows that mind, like matter, is governed by the same physical laws.

Chapter 2

The Entropy Theory of Value

2.1. Introduction

Value theory occupies a peculiar position in the development of economic theory. Most of the time, it is a little treaded area to mainstream economists for it is generally thought to be completely resolved. But a major shift in economic thinking often begins with the emergence of new understanding about value. For example, Mill (1871) asserted that he had left nothing in the laws of value for any future economist to clear up, shortly before Jevons and Walras, in the 1870's, developed new theories of value that became the core of neoclassical economics. After neoclassical economics firmly established its dominance, research in the area of value theory became essentially dormant again in the last several decades.

Since all human activities represent extraction and transformation of low entropy from the environment, it is natural to relate economic value to low entropy (Schrodinger, 1944; Prigogine, 1980). Indeed "there have been sporadic suggestions that all economic values can be reduced to a common denominator of low entropy" (Georgescu-Roegen, 1971, p. 283). However, some conceptual difficulties prevented the development of an entropy theory of value.

Georgescu-Roegen thought that linking economic value to low entropy would not be of much help to economists because "he would only be saddled with a new and wholly idle task --- to explain why these coefficients differ from the corresponding price ratios" (Georgescu-Roegen, 1971, p. 283). To this argument we may compare the works of Shannon and Wiener on information theory. Both defined information as the reduction of entropy (Shannon, 1948; Wiener, 1948). However, Shannon further applied the mathematical definition of entropy to obtain some results that are of fundamental importance in information theory, while recognizing that the mathematical definition of information is not identical to the meaning of information common in our daily use. As a result, Shannon established information theory as a science. In this chapter, we show that while economic value is not identical to physical entropy, the entropy theory of value, an analytical theory based on fundamental physical laws, enables us to obtain some results that greatly clarify the meaning of economic value and are highly consistent with our intuitive understanding. Among other things, it offers a clear understanding how institutional structures affect economic value of commodities. Roughly speaking, economic value is the low entropy value of a commodity whose property rights are enforced by governments or other institutions. Since the costs and willingness to enforce property rights on different kinds of commodities are different, the levels of enforcement are different, which, among other factors, causes commodities of similar physical entropy level to be priced very differently.

How is the entropy theory of value related to the existing economic theories of value? Neoclassical economics, the current mainstream economic theory, was developed around 1870 by Jevons, Walras and others. Walras (1954), the chief architect of neoclassical economics, argued that value is a function of scarcity. From the properties that the value of commodities should satisfy, it can be derived that the only mathematical formula to represent value, as a function of scarcity, is the entropy function. This is parallel to the idea that the only mathematical formula to represent information, as a function of probability, is the entropy function (Shannon, 1948). Thus, the entropy theory of value is the analytical formalization of Walras' vision of value as a function of scarcity.

Why both information and economic value are the reduction of entropy? From the entropy law, the most universal law of the nature, the increase of entropy of a system is spontaneous. The reduction of entropy in a system, however, takes effort, which is the base for both information and economic value.

In today's mathematical environment, it is easy to envision an entropy theory of value as the formalization of Walras' theory of value. Historically, however, marginal utility theory of value, which was influenced by Jevons, was easier to define mathematically. Gradually, it becomes the standard economic theory (Fisher, 1925; Debreu, 1959). While marginal utility is easy to define mathematically, it is difficult to measure empirically (Mirowski, 1989). Indeed the current theory of value does not attempt to measure value empirically. This is reflected in the mathematical tools adopted in the theory: "In the area under discussion it has been essentially a change from calculus to convexity and topological properties, a transformation which has resulted notable gains in the generality and in the simplicity of the theory" (Debreu, 1959, p. x). At the same time, the convexity and topological methods leave no room for a quantitative measure of value. By contrast, the entropy theory of value is established on a measurable mathematical function with clear physical meaning.

Since information is the reduction of entropy, an entropy theory of value is inevitably an information theory of value. The success of Shannon's entropy theory of information stimulated many research efforts in economics (Theil, 1967). However, the information theory of value, or the entropy theory of value, was not developed. Very often, the direction of scientific research is shaped by the thinking of an authority. In an often cited passage, Arrow wrote, "the well-known Shannon measure which has been so useful in communications engineering is not in general appropriate for economic analysis because it gives no weight to the value of the information. If beforehand a large manufacturer regards it as equally likely whether the price of his product will go up or down, then learning which is true conveys no more information, in the Shannon sense, than observing the toss of a fair coin" (Arrow, 1983 (1973), p. 138). The Shannon measure actually carries weight of information. For example, N symbols with identical Shannon measure carry N times more information than a single symbol (Shannon, 1948). Similarly, the value of the information about the future price is higher to a large manufacturer than to a small manufacturer, other things being equal. Later in this chapter, we show that information as an economic commodity shares most of the important properties with physical commodities.

The rest of the chapter is structured as follows. In Section 2.2, we formally develop the mathematical theory of value as entropy. This part extends Shannon's (1948) classic work on information theory. The entropy theory of value provides a quantitative framework to understand how different factors affect the value of a commodity. The influence on value by factors such as scarcity, the number of producers, substitutability and market size of a commodity can be naturally understood from the entropy formula of value. Since scarcity of resources, including human resources, is often regulated by institutional measures such as immigration laws and patent laws, the values of economic commodities are in great part a reflection of institutional structures. In Section 2.3, we utilize the results from information theory, statistical physics and the theory of evolution to discuss the relation between physical entropy value and the economic value. We discuss how this entropy theory of value offers a unifying understanding of the objective and subjective theories of value. In Section 2.4, we discuss how informational and physical commodities share common properties in the light of this entropy theory of value. By resolving the conceptual difficulties that have confounded us for many years, we offer a unified understanding of physical entropy, information and economic value. In Section 2.5, we discuss the relation between economic value and social welfare. Section 2.6 concludes.

2.2. The Properties of the Entropy Theory of Value

Value is a function of scarcity. Scarcity can be defined as a probability measure P in a certain probability space. It is generally agreed that the value of any product satisfies the following properties:

(a) The value of two products should be higher than the value of each of them.

(b) If two products are independent, that is, if the two products are not substitutes or partial substitutes of each other, then the total value of the two products will be the sum of two products.

(c) The value of any product is non-negative.

The only mathematical functions that satisfy all of the above properties are of the form

$$V(P) = -\log_b P \tag{2.1}$$

where b is a positive constant (Applebaum, 1996).

In general, if the scarcity of a service or product, *X*, can be estimated by the probability measure $\{p_1, p_2, \dots p_n\}$, the expected value of this product is the average of the value of each possibility, that is

$$V(X) = \sum_{i=1}^{n} p_i(-\log_b p_i) \qquad (2.2)$$

Therefore, value, just as information, in its general form can be defined as entropy, given that they are the same mathematically. In information theory, the base of the logarithm function is usually chosen to be two because there are two choices of code in information transmission, namely, 0 and 1 (Shannon, 1948). In economics, the base *b* can be understood as the number of producers. In the following we will discuss the properties of value as entropy.

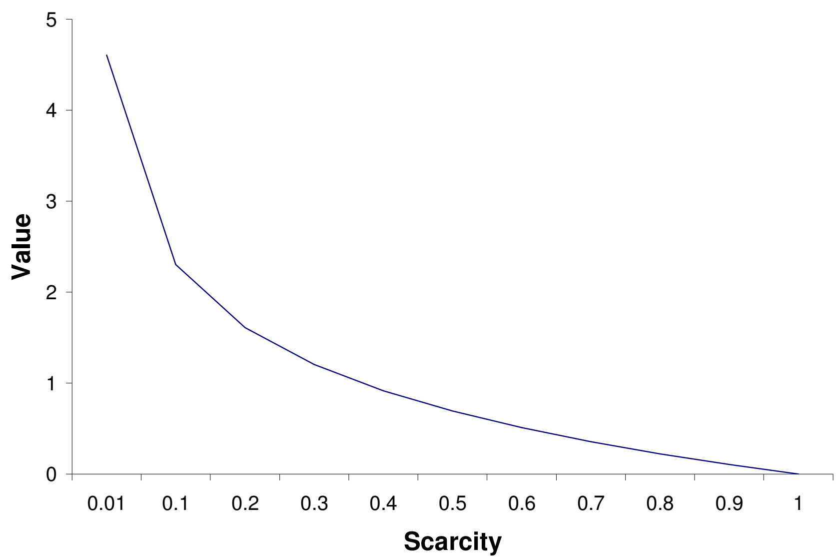

Figure 2.1 Value and scarcity

1. Scarcity and value

Figure 2.1 is a graph of (2.1), which shows that value is an increasing function of scarcity. That is why diamonds are worth more than water. In extreme abundance, i.e., when *P*=1, -log *P* =0, the value of a given commodity is equal to zero, even if that commodity is very useful. For example, food is essential for survival. Most countries subsidize food production in various ways to guarantee the abundance of food, which causes its low economic value. This shows that economic value and social value can have divergent valuations.

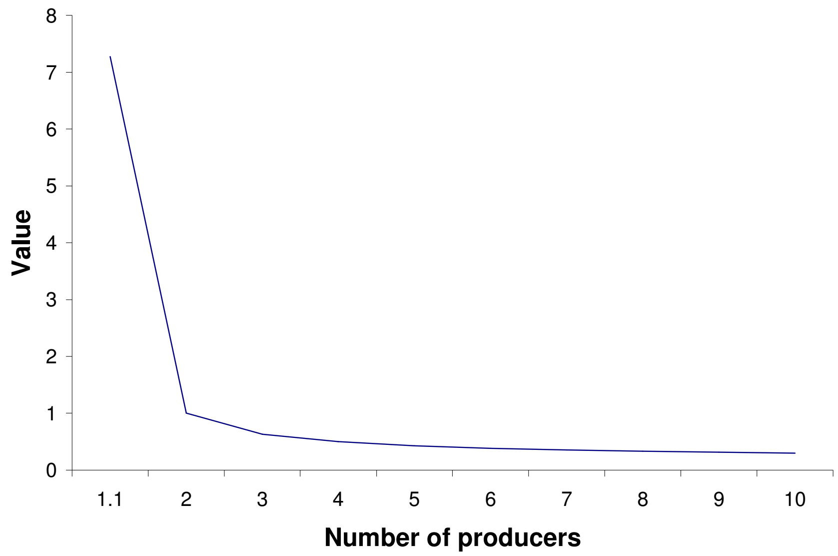

Figure 2.2 Value and the number of producers

2. Value and the number of producers

From (2.1), value is inversely related to the number of producers of a given product. Figure 2.2 displays the relationship between value and the number of producers. When the number of producers is small, the value of a product is high. That's why the products of monopolies and oligopolies are valued highly. If the base becomes one, i.e., absolute monopoly without substitution, value approaches infinity. This happens at some religious cults where only the spiritual leaders hold the key to heaven. In these types of organizations, the leaders often enjoy infinite power over their followers. The number of providers of most economic goods depends on many factors, some of which will be discussed in Chapter 3. In the following, we only give a brief discussion about the institutional structures that affect market entry and the number of suppliers for a given product.

Anti-trust regulations aim to prevent price fixing by existing providers of a service or product. They also intent to lower barriers to potential entry. Both measures, by increasing the number of choices, reduce the value of products, and hence the cost to consumers. For this reason, the value of a product will in general be lower in a more competitive market.

Patent rights and commercial secrets legislation, on the other hand, grant monopoly power and discourage the diffusion of knowledge. Patent rights and monopoly power allow the holders to maintain high product prices. Such a market virtually assures a firm's success. The social value embodied in such legislation encourages innovation, but discourages competition. The balance between fostering competition and protecting innovation so as not to stifle either is always a delicate one (Arrow, 1999).

It is often difficult to determine the exact number of providers of a service empirically. At the time of writing, Jetsgo, a Canadian airline, declared bankruptcy. There are three major operators in the air travel industry in Canada. They are Air Canada, WestJet and Jetsgo. There are some other regional carriers and international airlines competing for many routes. Together, we can assume four providers for the air travel service before Jetsgo declared bankruptcy. From (2.1), the value of each airline can be represented as

$$-\log_4 P \qquad \text{and} \qquad -\log_3 P$$

before and after Jetsgo declare bankruptcy. The change of value is therefore

$$(-\log_3 P)/(-\log_4 P) - 1 = \log_3 4 - 1 = 0.262$$

Jetsgo declared bankruptcy at the evening of March 10, 2005. The closing prices of stocks of WestJet and Air Canada at March 10 and 11 are 11.17, 15.6 and 32.19, 37 respectively. The price changes are

$$15.6/11.17-1=0.397 \qquad \text{for WestJet}$$

and

$$37/32.19-1=0.149 \qquad \text{for Air Canada}$$

respectively. The average change of price is

$$(0.397+0.149)/2=0.273$$

which is very close to the theoretical prediction of 0. 262.

Some theoretical and empirical results can be further refined. For example, this theory does not distinguish the sizes of different providers of a service. The refinement of the theory is left to the future research.

3. Substitutability and value

Many products and services are not identical but can substitute each other to a certain degree. The value of a single product can be defined as its entropy (2.2). The total value of two products, *X* and *Y*, can be defined as their joint entropy

$$V(X,Y)=-\sum_{j=1}^{n}\sum_{k=1}^{m}p_{jk}\log(p_{jk}) \qquad (2.3)$$

while the individual values of *X* and *Y* can be defined as

$$V(X)=-\sum_{j=1}^{n}p_j\log p_j$$

$$V(Y)=-\sum_{k=1}^{m}q_k\log q_k$$

It can also be proved that (Shannon, 1948)

$$V(X,Y)\le V(X)+V(Y) \qquad (2.4)$$

The equality holds only when *X* and *Y* are independent, i.e., *X* and *Y* are not substitutes or partial substitutes for each other. This means that substitutability reduces the value of a product, which is a very intuitive conclusion that is also verifiable via common sense observation in the marketplace. The purpose of brand name management and advertisement of a product is to make a product special and to reduce the perception and/or reality of substitutability for it, which increases the value of that product.

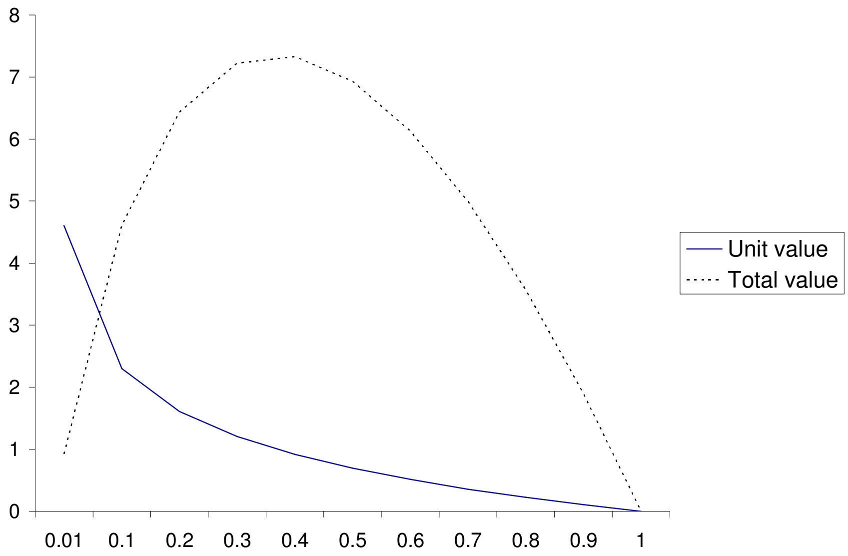

Figure 2.3 The unit value and total value of a product with respect to scarcity

4. Market size, product life cycle and product value

Suppose the potential market size of a product is *M*. The percentage of people who already have the product is *P*. Then the unit value of the product is

$$-\log P \qquad (2.5)$$

Since the number of people who have bought the product is *MP*, The total value of the product is

$$MP(-\log P) \tag{2.6}$$

From (2.6), the value of a product is higher with a larger market size. Figure 2.3 is the graph of unit value and total value of a product with respect to its abundance. From Figure 2.3, we can explore the relation between the value of a product and product life cycle. When a product is new and scarce, the unit value is high. Its total value is low. As the production increases, the total value will increase as the unit value decreases. When the production quantity is over a certain level, however, the total value of a product will start to decrease as well. Intuitively, this is easy to understand. The market values of manufacturers of mature products are generally low, although the production processes are very efficient. This observation shows that efficiency is not equivalent to value.

The above discussion shows that the implications of identifying value with the reduction of entropy are highly consistent with our intuitive understanding of economic value. It should be noted that in economic processes, a final product embodies many different kind of scarcities: labor, raw materials and capital. A detailed analysis of the value of a particular product will be much more involved. For example, black and white television sets are less common than color television sets and yet they have less economic value. This is because the process of making color TV takes more scarce resources such as labor. The value of a final product is the sum of total scarcity.

2.3. Physical Entropy, Information and Economic Value

The discussion about the relation between information and physical entropy began with the paradox of Maxwell's demon (Maxwell, 1871). In 1870s, Boltzmann defined the mathematical function of entropy, which Shannon (1948) identified as information many years later. From the discussion in Chapter 1, information is the reduction of entropy, not only in a mathematical sense, as in Shannon's theory, but also in a physical sense. How then can economic value as entropy be linked to physical entropy?

From Chapter 1, both natural selection and sexual selection indicate that human beings favor low entropy sources. This observation offers a connection between the entropy theory of value and the subjective utility theory of value. "Mind is an organ of computation engineered by natural selection" (Pinker, 1997, p. 429). It calculates the entropy level and sends out signals of pleasure for accumulating and displaying low entropy and signals of pain for dissipation of low entropy. Jevons "attempted to treat economy as a calculus of pleasure and pain" (Jevons, 1957, p. vi). Pleasure is generally associated with the accumulation or display of low entropy level, such as the accumulation of wealth, and conspicuous consumption. Pain is associated with dissipation of low entropy, such as work and the loss of money. So value in subjective theory, as a measure for pleasure and pain, is intrinsically linked to the level of entropy.

The current economic theory states that the value of a commodity is a function of supply and demand. From the theory of natural and sexual selection, the demand of an economic commodity is driven by its level of entropy. The supply of an economic commodity is constrained by its scarcity, with entropy as its unique measure (Applebaum, 1996). Therefore the level of entropy offers a natural measure of economic value.

It is easy to understand the objective theory of value from the entropy theory of value. Since the entropy level of a system increases spontaneously, the reduction of entropy in a system represents the effort that has been made. Entropy level may be the closest to an invariant measure of value of labor and other commodities. While economic values of commodities are highly correlated with the level of physical entropy, they are not identical for several reasons. In the following, we will discuss two: One from the perspective of information theory and another from the institutional structures that regulate scarcity.

First, the entropy level we perceive of a commodity is different from its objective entropy level. From information theory, the amount of information one can receive, R, is equal to the amount of information sent minus the average rate of conditional entropy.

$$R = H(x) - H_y(x) \tag{2.7}$$

The conditional entropy $H_y(x)$ is called the equivocation, which measures the average ambiguity of the received signal (Shannon, 1948). From our discussion in Chapter 1, equivocation arises because receivers don't have the complete background knowledge of signals. For example, gold, a scarce commodity, is highly valuable. Another commodity could be as scarce as gold, but unlike shiny and stable gold, it could be very difficult to identify. Most people will not invest much effort to gain knowledge needed to identify this commodity because the cost outweighs the potential benefit. Thus, it registers less attention and is valued less by human beings.

Second, scarcity of a commodity is regulated by the institutional structures that enforce property rights. For example, the value of an invention is influenced by how long and how broad patent protection is granted. The value of a patent is higher in a system where patents are valid for twenty years than one for ten years. From (2.4), substitutability reduces the value of a product. If patent protection is defined broader, the value of an invention is higher. Economic value, as a function of scarcity, is to a great extent regulated by institutional structures. Among all the institutional measures that regulate scarcity, the most important regulation is the immigration laws that regulate the scarcity of labor forces, which makes persistent wage differential across regions possible.

2.4. The Entropy Theory of Value and Information

Because of the equivalence of entropy and information, an entropy theory of value is inevitably an information theory of value. Information is often regarded as a rather unusual commodity. In this section, we will show that informational and physical commodities share most of the fundamental properties from the perspective of entropy theory. Since Arrow (1999) offers a clear description about the special characteristics of information as an economic commodity, our discussion is based on his writing.

> The algebra of information is different from that of ordinary goods. … Repeating a given piece of information adds nothing. On

> the other hand, the same piece of information can be used over and over again, by the same or different producer(s). (Arrow, 1999, p.21)

From (2.7), the amount of information received is the information of source minus equivocation. Repeating a signal of information helps reduce equivocation. That is why the same commercials are repeated many times on TV. A more detailed analysis of commercials by a company, say Coca Cola, will illustrate the concept more clearly. Most commercials of Coca Cola spread the same information: Drink Coca Cola. The purpose of the commercials is to reduce the equivocation in information transmission between the sender, Coca Cola company and the receivers, the potential consumers. Usually same commercial will be repeated many times and different commercials are designed to relate the viewers to Coca Cola in different ways. However, the efforts of Coca Cola will not automatically reduce the equivocation between the sender and the receivers. Other soft drink companies and other matters in life compete for attention. As a result, the equivocation between Coca Cola and the general public may increase, despite the efforts from Coca Cola. From the thermodynamic theory that all low entropy sources have a tendency to diffuse, repeating the same piece of information is essential to keep it valuable. The essence of a living organism is to repeat and spread the information encoded in its genes.

It is often thought that the use of information does not involve rivalry, since "the same piece of information can be used over and over again, by the same or different producer(s)". This property is not confined to information. The same hammer "can be used over and over again, by the same or different producer(s)". However, the value of the same information will be different for different users or at different time. For example, if an unexpected surge of corporate profit is known by very few people, i.e., when P is very small and $-\log P$ is very high, this information would be highly valuable. Huge profit could be made by trading the underlying stocks. But when it is known to many people, the value of such information is very low. In general, when some knowledge is mastered by many people, its market value is very low.

> The peculiar algebra of information has another important implication for the functioning of the economic system. Information, once obtained, can be used by others, even the original owner has not lost it. Once created, information is not scarce in the economic sense.

> This fact makes it difficult to make information into property. It is usually much cheaper to reproduce information than to produce it in the first place. In the crudest form, we find piracy of technical information, as in the reproduction of books in violation of copyright. Two social innovations, patents and copyrights, are designed to create artificial scarcities where none exists naturally, although the duration of the property is limited. The scarcities are needed to create incentives for undertaking the production of information in the first place. (Arrow, 1999, p. 21)

Information is a type of low entropy source. Utilization of low entropy source from others is a universal phenomenon of living systems. "Once again animals discover the trick first. … butterflies, did not evolve their colors to impress the females. Some species evolved to be poisonous or distasteful, and warned their predators with gaudy colors. Other poisonous kinds copied the colors, taking advantage of the fear already sown. But then some nonpoisonous butterflies copied the colors, too, enjoying the protection while avoiding the expense of making themselves distasteful. When the mimics become too plentiful, the colors no longer conveyed information and no longer deterred the predators. The distasteful butterflies evolved new colors, which were then mimicked by the palatable ones, and so on." (Pinker, 1997, p. 501)

So the perceived uniqueness of copying information products in human societies is actually quite universal within living systems. Once we look at the living world from the entropy perspective, it can hardly be otherwise. In human societies, the attempt to copy and reproduce valuable assets, whether informational or physical assets, is also universal.

The fashion industry offers an example that illustrates the dynamics of innovation and copying clearly. When a new fashion style is created, it is scarce and hence valuable. This valuable information will then be copied by others. As more people copy the style, P increases, $-\log P$ decreases and the value of the fashion decreases. To satisfy the demands for high value fashions, new fashion styles "are designed to create artificial scarcities where none exists naturally".

Protection of an organism's source of low entropy to prevent access by others is also a universal phenomenon of living systems. Animals develop immune systems to protect their low entropy source from being accessed by microbes. Plants make themselves poisonous to prevent their low entropy from being accessed by animals. When space is a limiting factor in survival or reproduction, animals defend their territory vigorously (Colinvaux, 1978). Whether to enforce the property rights depends on the cost of enforcement and the value of the low entropy source. When information products become an important class of assets, the property rights of physical assets are naturally extended to informational assets.

2.5. Economic Wealth and Social Welfare

From the above discussion, it is clear that economic wealth and social welfare are two distinct concepts. Economic activities provide low entropy sources for the survival and comfort of human beings. From the second law of thermodynamics, the reduction of entropy locally is always accompanied by the increase of high entropy waste globally. So "externality" is not a form of "market failure" but a direct consequence of fundamental natural laws. Since low entropy product is more concentrated while high entropy waste is more diffuse, the economic value of a product is easier to measure than the harmful effects of the wastes. Usually, a product is developed to satisfy certain market demand. Its utility is easily appreciated by the customers, who are willing to pay for the product. This is the source of economic value. The harmful effects of the wastes, being more diffuse, affect more people but usually at very low level. These effects often take very long time to get noticed. When the human population density and consumption level is low, most of the high entropy wastes that humans generate are absorbed by microbes and other natural forces with little human effort (Margulis, 1998). This vital recycling business is accorded no economic value. As the population density and the level of consumption increases, however, direct human intervention is needed to move the high entropy waste away from where people reside. The economic value of the waste management business is a function of the level of effort invested by the public in the recycling of wastes. This value is not equivalent to environmental quality of human habitats. Since high entropy waste is more diffusive, the market of recycling businesses is generally created by legal and regulatory methods to prevent the degeneration of the environment.

Theories in institutional economics often suggest that externalities can be internalized with institutional measures. These measures, however, are achieved with costs. For example, companies often pursue internal stability in their working environment. However, more training and better pay packages are often required to maintain internal stability. In living systems, warm blooded animals are able to maintain stable core temperature near the biochemical optimum. However, they achieve stable internal environment with much higher energy costs of metabolism than cold blooded animals, whose internal environment fluctuates with external environment (Smil, 1999, p. 62).

While economic wealth is not equivalent to social welfare, economic value, as a reflection of human efforts, is generally geared toward human welfare over the short term. This is why economic prosperity is often consistent with the improvement of social welfare in a particular point of time. However, wealth, as low entropy of human society, is ultimately supported by low entropy from nature. In the last several hundred of years, world wide consumption of energy has been increasing steadily with the economic progress (Smil, 2003). Since our current civilization is based on fossil fuel, the eventual depletion of fossil fuel will shake the foundation of today's lifestyle.

In general, wealth represents the total dependence of each other in a society. The increase of one's wealth means the increase of the dependence of others on him and hence the increase of his power. While it is natural for an individual or a company to pursue strategies that maximize wealth, national and international policies often concern more about long term sustainability of ecological and social systems.

2.6. Concluding Remarks

Theories built on a sound physical foundation often provide simple and intuitive results on practical problems. Shortly after Shannon's work of 1948 that identified information as entropy, Weaver commented, "Thus when one meets the concept of entropy in communication theory, he has a right to be rather excited --- a right to suspect that one has hold of something that may turn out to be basic and important" (Shannon and Weaver, 1949, p. 13). The development of information theory in the last half century has proved his foresight. This entropy theory of value, which establishes an explicit link between economic value and physical entropy, offers an analytical theory that is highly consistent with our intuitive understanding of economic value.

Chapter 3

Production and Competition: An Analytical Thermodynamic Theory

Economic and biological systems need to extract low entropy from the environment to compensate for continuous dissipation (Schrodinger, 1944; Georgescu-Roegen, 1971; Prigogine, 1980). This process can be represented by lognormal processes, which contain an extraction term and a dissipation term. The lognormal processes in turn can be mapped into a thermodynamic equation. From the entropy law, the thermodynamic diffusion of an organic or economic system is spontaneous. The extraction of low entropy from the environment, however, depends on specific biological or institutional structures that incur fixed or maintenance costs. The fixed costs help reduce variable costs in extracting low entropy from the environment. In this chapter, we solve the thermodynamic equation to derive an analytic formula that explicitly represents the relation among fixed costs, variable costs, uncertainty of the environment and the duration of a project, which is the core concern in most economic decisions. This analytical framework directly based on thermodynamic foundation greatly simplifies the understanding of economic activities and their relation with environmental changes.

3.1. Some Historical Background

Neoclassical economics was founded around 1870 by Jevons, Walras and others, who believed that economics should be built on a sound physical foundation. Since the dominant platform of physics in Jevons and Walras' time was rational mechanics, it was natural for them to adopt this platform.

About the same time when Jevons and Walras tried to establish economic theory on rational mechanics, physics had experienced a revolution whose impact is just to be felt in social sciences recently. Around 1870, Maxwell discussed the famous intelligent demon problem and Boltzmann established the theory of statistical mechanics. Boltzmann's theory met strong opposition because it seemed to contradict Newtonian mechanics. In despair, he wrote: "I am conscious of being only one individual struggling weakly against the current of time. But it still remains in my power to make a contribution in such a way that, when the theory of gases is again revived, not too much will have to be rediscovered." (Quotes from Isihara, 1971, p. 18) Boltzmann's growing pessimism led him to commit suicide in 1906, shortly before his theories were accepted.

Rational mechanics studies the movement of single particles. Statistical mechanics studies the dynamic properties of the systems of many particles. It is a generalization from rational mechanics. However, in its early days, people thought it was inconsistent with the rational mechanics. In rational mechanics, time is reversible. In statistical mechanics, time is irreversible and systems are evolutionary.

Gradually, the theory of statistical mechanics gained more influences. The first chapter of Wiener's (1948) *Cybernetics* was titled Newtonian Time and Bergosonian Time, in which he stated that most systems we encounter can be more precisely described by statistical mechanics instead of Newtonian mechanics. In Schrodinger's (1944) *What is Life*, he stated that the most fundamental property of life is their ability to extract negative entropy from the environment to compensate continuous dissipation. Since then some analytical theories related to thermodynamics have been developed, such as Lorenz' chaos theory and Prigorgine's far from equilibrium thermodynamic theory. Lorenz, a meteorologist, simplified weather equations, which are thermodynamic equations, into ordinary differential equations. He found chaos properties from these equations. Prigorigine developed the theory from some chemical reactions. These theories greatly influenced the thinking in biology and social sciences. However, they do not model life process or social activities directly. Chaos theory and Prigorgine's theory, while providing good insights to the research in social sciences, are mostly analogies.

As it is often the case, something from a totally different area turns out to be the key to the new development. In 1973, Black and Scholes developed an analytical theory of option pricing on financial assets. Initially it was a very technical subject. But gradually, it has been applied to many different areas. We will have a look at the basic properties of Black-Scholes theory to understand why it can be applied to so many different areas.

In Black-Scholes option theory, the price movement of financial assets is modeled with lognormal processes

$$\frac{dS}{S} = rdt + \sigma dz .$$

where r is the rate of expected return and σ is the rate of uncertainty. Option prices, as functions of prices of its underlying assets, satisfy the following Black-Scholes equation.

$$\frac{\partial C}{\partial t} + rS\frac{\partial C}{\partial S} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 C}{\partial S^2} = rC$$

I had been thinking about an analytical thermodynamic theory of life systems for many years when I learned about the Black-Scholes option theory. The most fundamental property of life is their ability to extract low entropy from the environment to compensate continuous dissipation. Soon I realized this property can be represented by lognormal processes, where r is the rate of extraction of low entropy and σ is the rate of diffusion. Every stochastic process can be mapped into a deterministic thermodynamic equation, which is often easier to handle and yields more results. So I hope Black-Scholes equation and option theory may offer some insight for an analytical thermodynamic theory of life systems. After several years, I first developed such a theory based an analogy between option theory and living systems (Chen, 2000). Later I was able to derive the theory directly without depending on its analogy with option theory (Chen, 2002b). While my original interest was in life sciences, I now make a living teaching financial economics. That is why currently, this theory is mostly applied to social sciences. In the next section, we will describe this theory.

3.2. Basic Theory

All biological systems, human or non-human, need to extract low entropy from the environment to compensate for continuous dissipation. In human societies, most human activities are measured by economic value. Suppose *S* is the amount of low entropy of a biological system, *r*, the rate of extracting low entropy from the environment and σ, the rate of diffusion of the low entropy into the environment. Similarly in an economic system, *S* represents economic value of a commodity, *r*, the rate of change of the value of this commodity and σ, the rate of uncertainty. Then the process of *S* can be represented by the lognormal process

$$\frac{dS}{S} = rdt + \sigma dz \qquad (3.1)$$

A production system is parallel to a biological system. A firm, which has blueprint to produce a product, such as cars, is similar to a biological entity, which has genes to produce offspring. The production of a good involves fixed cost, *K*, and variable cost, *C*, which are functions of the *S*, the value of the product. If the discount rate of a firm is *q*, from Feynman-Kac formula, (Øksendal, 1998) the variable cost, as a function of *S*, satisfies the following equation

$$\frac{\partial C}{\partial t} = rS\frac{\partial C}{\partial S} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 C}{\partial S^2} - qC \qquad (3.2)$$

To solve for variable cost from this equation, we need to determine the initial condition that the variable cost has to satisfy at time zero. We perform a thought experiment about a project with a duration that is infinitesimally small. When the duration of fixed cost is infinitesimal small, the project has only enough time to produce one piece of product. If the fixed cost is lower than the value of the product, the variable cost should be the difference between the value of the product and the fixed cost to avoid arbitrage opportunity. If the fixed cost is higher than the value of the product, there should be no extra variable cost needed for this product. Mathematically, the initial condition for variable cost is the following

$$C(S,0) = \max(S - K,0) \qquad (3.3)$$

where S is the value of the product and K is the fixed cost. Suppose the duration of a project is T. Solving the equation (3.2) with the initial condition (3.3) yields the following solution

$$C = Se^{(r-q)T} N(d_1) - Ke^{-qT} N(d_2) \qquad (3.4)$$

where

$$d_1 = \frac{\ln(S/K) + (r + \sigma^2/2)T}{\sigma\sqrt{T}}$$

$$d_2 = \frac{\ln(S/K) + (r - \sigma^2/2)T}{\sigma\sqrt{T}} = d_1 - \sigma\sqrt{T}$$

The function $N(x)$ is the cumulative probability distribution function for a standardized normal random variable. When q, the discount rate, is equal to r, the rate of change, Formula (3.4) becomes

$$C = SN(d_1) - Ke^{-rT} N(d_2) \qquad (3.5)$$

It takes the same form as the well-known Black-Scholes (1973) formula for European call options. A more technical discussion of this theory can be found in Chen (2002c).

This theory, for the first time in economic literature, provides an analytic theory that explicitly represents the relation among fixed costs, variable costs, uncertainty of the environment and the duration of a project, which is the core concern in most economic decisions. "The progress of science is marked by the transformation of the qualitative into the quantitative. In this way not only do notions become turned into theories and lay themselves open to precise investigation, but the logical development of the notion becomes, in a sense, automated. Once a notion has been assembled mathematically, then its implications can be teased out in a rational, systematic way." (Atkins, 1994, p. 29)

A new theory is ultimately justified by its implications. We will look at the properties and implications of this theory. For simplicity, we will only explore the special case when the discount rate is equal to the rate of change, that is, Formula (3.5). Several properties can be derived from (3.5). First, when the fixed cost investment, K, is higher, the variable cost, C, is lower. Second, for the same amount of fixed investment, when the duration, T, is longer, the variable cost is higher. Third, when uncertainty, σ, is higher, the variable cost increases. Fourth, when the fixed cost approaches zero, the variable cost will approach to the value of the product. Fifth, when the value of the product approaches zero, the variable cost will approach zero as well. All these properties are consistent with our intuitive understanding with production processes.

Unlike a conceptual framework, this analytical theory enables us to make precise calculation of returns of different projects under different kinds of environments. First, we examine the relation between fixed cost and variable cost at different levels of uncertainty. For example, a product can be manufactured with two different technologies. One needs ten million dollars of fixed cost and the other needs one hundred million fixed cost. Assume the other parameters are unit value of the product, to be one million, discount rate, to be 10% and duration of the project, to be twenty-five years. When uncertainty of the environment is 30% per year, variable cost for the low fixed cost project is 0.59 million and variable cost for the high fixed cost project is 0.14 million, calculated from (3.5). When uncertainty of the environment is 90% per year, variable cost for the low fixed cost project is 0.98 million and variable cost for the high fixed cost project is 0.94 million. In general, as fixed costs are increased, variable costs decrease rapidly in a low uncertainty environment and decreases slowly in a high uncertainty environment. This is illustrated in Figure 3.1. In the extreme environment when the uncertainty reaches infinity, the variable cost is equal to the value of the product, regardless of the level of the fixed cost investment. In this environment, the value of any fixed asset is zero. For example, the elaborate institutional structures of the Roman Empire, which were once of great value, became worthless during the empire's chaotic collapse. In general, the value of any physical capital, institutional capital and human capital depends highly on the environment.

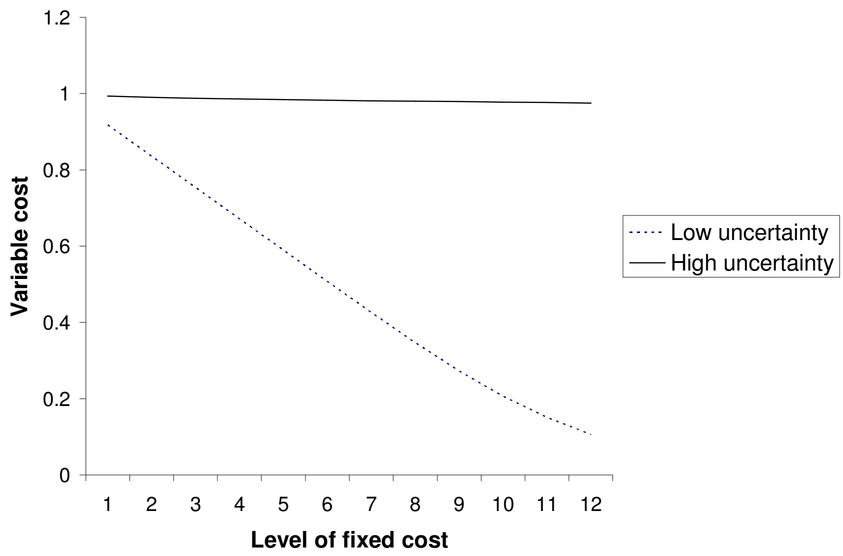

Figure 3.1 Uncertainty and variable cost

Next we discuss the returns of investment on different projects with respect to the volume of output. *K* is the fixed cost of production and *C* is the variable cost. Suppose the volume of output during the project life is *Q*, which is bound by production capacity or market size. We assume the present value of the product to be *S* and variable cost to be *C* during the project life. Then the total present value of the product and the total cost of production are

$$SQ \text{ and } CQ + K \tag{3.6}$$

respectively. The return of this project can be represented by

$$\ln\left(\frac{SQ}{CQ+K}\right) \tag{3.7}$$

Continuing the example on two technologies with different fixed costs, we now discuss how the expected market sizes affect rates of return. Suppose the level of uncertainty is 30% per year and other parameters are the same. If the market size is 100, the return of the low fixed cost project, calculated from (3.7), is 37% and the return of the high fixed cost project is -12%. When the market size is 400, the return of the low fixed cost project is 48% and the return of the high fixed cost project is 97%. Figure 3.2 is the graphic representation of (3.7) for different levels of fixed costs. In general, higher fixed cost projects need higher output volume to breakeven. At the same time, higher fixed cost projects, which have lower variable costs in production, earn higher rates of return in large markets.

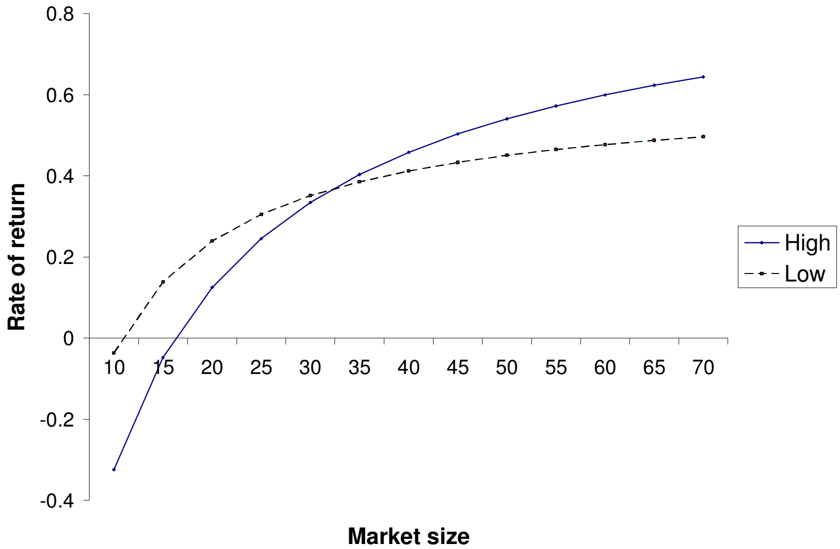

Figure 3.2 Market size and return with different levels of fixed costs

From the above discussion the level of fixed investment in a production system depends on the expectation of the level of uncertainty of production technology and the size of the market. When the outlook is stable and market size is large, production systems with high fixed investment earn higher rates of return. When the outlook is uncertain or market size is small, production systems with low fixed cost breakeven easier.

In the following, we will use the example of industry life cycles to illustrate the evolution of production systems. When an industry is new, there is a lot of uncertainty about future development. This environment offers opportunities for small companies with low fixed costs as they are more flexible. When uncertainty is high, increases in fixed costs will not reduce variable costs very much. (Figure 3.1) Economy of scale is not significant under high risk condition, which permits many low fixed cost companies to compete. When an industry becomes mature and uncertainty decreases, increases in fixed costs, *K*, (capital investments and accumulated human capital), drive down variable costs rapidly, which permits leading companies to lower product prices and drive out small high variable cost companies. So, in a mature industry only very few big companies can stay in business. In fact, without anti-trust legislation, many industries with high capital intensity, or high fixed assets would probably end with monopolies or regional monopolies (Acs and Audretsch, 1988; Mazzucato, 2000).

While large firms with high fixed cost are highly competitive in their own fields, they are less effective in entering new markets with high uncertainty. First, from Figure 3.2, high fixed cost systems need a large market size to break even. So they tend to hesitate in moving into new markets, which are typically of small size in their early stages. Second, the high concentration of wealth in large companies often attracts litigation and other attempts to extract wealth from them. So large companies often incur high legal costs and are cautious in pursuing new opportunities. Third, internal coordination in large companies is much more complex and difficult than in small companies. While it is relatively easy to incorporate innovative ideas by adjusting firm structures in simple and small companies, innovation is often very disruptive to a highly coordinated and efficient complex structure. Large companies often develop highly optimized structures to reduce uncertainty and bring down variable costs in producing particular products. This however often stifles the innovative spirit inside the companies and makes it difficult for them to adjust in a changing environment.

In general, large companies, which have invested a great amount on existing technologies, may be unwilling or unable in the short term to switch to new and potentially better technologies. This opens opportunities for new and typically small companies when new industries emerge. For example, the champions of the IT revolution, such as Microsoft, Intel, CISCO, Oracle, AOL, are all relatively new companies that reaped tremendous profits by being able to quickly respond to and take advantage of newly emerging markets. Microsoft is now a mature company in a rapidly evolving industry. Its success and the resulting high concentration of wealth within Microsoft has attracted litigation and other attempts to extract from it wealth and market share.

3.3. A Comparison with Neoclassical Economic Theory

Since its birth, the foundation or "assumptions" of neoclassical economic theory has been criticized for its lack of relevance to reality. To this, Friedman replied:

> In so far as a theory can be said to have "assumptions" at all, and in so far as their "realism" can be judged independently of the validity of predictions, the relation between the significance of a theory and the "realism" of its "assumptions" is almost the opposite of that suggested by the view under criticism. Truly important and significant hypotheses will be found to have "assumptions" that are wildly inaccurate descriptive representations of reality, and in general, the more significant the theory, the more unrealistic the assumptions (in this sense). The reason is simple. A hypothesis is important if it "explains" much by little, that is, if it abstracts the common and crucial elements from the mass of complex and detailed circumstances surrounding the phenomena to be explained and permits valid predictions on the basis of them alone. To be important, therefore, a hypothesis must be descriptively false in its assumptions; it takes account of, and accounts for, none of the many other attendant circumstances, since its very success shows them to be irrelevant for the phenomena to be explained. (Friedman, 1953, p. 16)

He further challenged:

> As we have seen, criticism of this type is largely beside the point unless supplemented by evidence that a hypothesis in one or another of these respects from the theory being criticized yield better predictions for as wide a range of phenomena. (Friedman, 1953, p. 31)

In the next several chapters, we will offer more detailed discussion on how the new theory, which is based on more realistic assumptions, does "yield better predictions for as wide a range of phenomena". In the following, we will compare the new theory briefly with neoclassical economic theory.

Consistency with physical and biological theories

Neoclassical economics was founded around 1870 by Jevons, Walras and others, who believed that economics should be built on a sound physical foundation. Since the dominant platform of physics in Jevons and Walras' time was Newtonian mechanics, it was natural for them to adopt this platform. However, theories derived from rational mechanics often do not offer good explanation to economic behaviors. Gradually, explicit identification with physics disappears while analogies between physics and economics are frequently mentioned. The following quote from Samuelson's Nobel lecture is quite representative:

> There is really nothing more pathetic than to have an economist or a retired engineer try to force analogies between the concepts of economics. How many dreary papers have I had to referee in which the author is looking for something that corresponds to entropy or to one or another form of energy.

In the very next paragraph, however, Samuelson found some analogy himself.

> However, if you look upon the monopolistic firm hiring ninety-nine inputs as an example of a maximum system, you can connect up its structural relations with those that prevail for an entropy-

> maximizing thermodynamic system. Pressure and volume, and for that matter absolute temperature and entropy, have to each other the same conjugate or dualistic relation that the wage rate has to labor or the land rent has to acres of land.

Mirowski observed, “The key to the comprehension of Samulson’s meteoric rise in the economics profession was his knack for evoking all the outward trapping and ornament of science without ever coming to grips with the actual content or implications of physical theory for his neoclassical economics” (Mirowski, 1989, p. 383).

Life systems are non-equilibrium thermodynamic systems. The current dominant economic theory is general equilibrium theorem. Social system is a special case of living systems. When a theory about a special case is inconsistent with general foundation, either the general foundation or the special theory is wrong. So far, economists have not challenged the validity of the non-equilibrium thermodynamic theory of life systems. This theory shows that an analytical theory of economics can be directly derived from basic physical and biological laws. By this, it establishes social sciences as an integral part of physical and biological sciences.

A comparison with production functions

Production functions, such as Cobb-Douglas production function, form the fundamental blocks in general equilibrium production theory. Cobb-Douglas function takes the form

$$Y = AL^{\alpha} K^{\beta}$$

where Y, L and K denote output, labor (variable cost) and capital (fixed cost) respectively. Solow had made following comment about the production function:

> I have never thought of the macroeconomic production function as a rigorously justifiable concept. In my mind it is either an illuminating parable, or else a mere device for handling data, to be used so long as it gives good empirical results, and to be discarded as soon as it doesn't, or as something better comes along. (Solow, 1966, p. 1259)

By contrast, the analytical production theory developed here is derived rigorously from the fundamental property of life systems. It gives simple and clear results of returns to investment under different market conditions. The form and parameters of Cobb-Douglas function are given without rigorous justification. *A*, the coefficient in Cobb-Douglas function, "has been called, among other things, 'technical change', 'total factor productivity', 'the residual' and 'the measure of our ignorance'" (Blaug, 1980, p. 465).

Since production functions are widely used in economic literature in constructing economic models, even a small improvement on this topic should have a big impact in understanding economics.

Optimality vs. tradeoff

Optimization theory holds the central position in neoclassical economics. Paul Samuelson's Nobel Lecture is titled *Maximum Principles in Analytical Economics*. Alchian (1950) and Friedman (1953) tried to reconcile the maximization principle with evolutionary theory. Friedman stated:

> Confidence in the maximization-of-return hypothesis is justified by evidence of a very different character. … unless the behavior of businessmen in some way or other approximated behavior consistent with the maximization of returns, it seems unlikely that they would remain in business for long. Let the apparent immediate determinant of business behavior be anything at all --- habitual reaction, random chance, or whatnot. Whenever this determinant happens to lead to behavior consistent with rational and maximization of returns, the business will prosper and acquire resources with which to expand;

> whenever it does not, the business tend to lose resources and can be kept in existence only with addition of resources from outside. The process of "natural selection" thus helps to validate the hypothesis --- or rather, given natural selection, acceptance of the hypothesis can be based largely on the judgment that it summarized appropriately the conditions for survival. (Friedman, 1953, p. 22)

We will use an example of project investment to illustrate the problem of Friedman's argument. Assume the relevant parameters are unit value of the product to be one million, discount rate to be 4%, diffusion to be 40%, duration of the project, to be thirty years and market size to be 150 over the project life. It can be calculated that a project with a fixed cost of 25 million dollar will be optimal. However, if any parameter changes, the optimal value of fixed cost investment will change as well. For example, if diffusion increases to 60%, the optimal value of fixed investment will become 11 million. Since fixed cost is spent or committed at the beginning of the project while other parameters may change over the course of project life, it is impossible to determine optimality in advance. Furthermore, higher fixed cost systems, which are often the winners of earlier market competition, suffer more from the increase of uncertainty. This means that long term survival is not necessarily consistent with short term optimization.

Earlier, we have shown that systems with higher fixed costs earn higher rates of return in large markets and stable environments than those with lower fixed costs. These systems may appear superior. However, the performance of high fixed cost systems deteriorates in high volatile environments. In Chapter 4, we will show that the main theme of economic and biological evolution is the tradeoff between competitiveness of high fixed cost systems in a stable environment and flexibility of low fixed cost systems in a volatile environment. Biologists haven't found a universally applicable measure of fitness (Stearns, 1992, p. 33). Our theory shows that there does not exist such a measure. For the same reason, there will not exist a universally applicable measure of optimality.

On imperfection

In neoclassical economics, many phenomena that are not consistent with theories are considered "imperfect". For example, from Modigliani and Miller (1958) theory, in a perfect market, capital structure is irrelevant. Since capital structure is relevant in reality, the real market is imperfect. There are many similar terms, such as "imperfect information", "imperfect contract", "imperfect competition", "inefficient property right", "market failure", "government failure", "externality". Before discussing these imperfections, we briefly review the idea of imperfection in old astronomy.

Ancient people had long observed that stars moved in perfect harmony in the sky. Several planets, however, moved in irregular trajectories. It was thought that this was caused by the imperfectness of the planets. There were many elaborate theories why the planets were imperfect. Kepler, however, derived that all planets moved in perfect elliptic orbits. This story tells us that "imperfection of the world" often reflects imperfection of the theory that is used to understand the world. In later chapters, we will explain how this analytical thermodynamic theory offers a unified understanding of various "imperfection" or "externality". In the following we will briefly discuss "imperfect competition". Again, we quote Friedman as a reference point.

> The theory of monopolistic and imperfect competition is one example of the neglect in economic theory of these propositions. The development of this analysis was explicitly motivated, and its wide acceptance and approval largely explained, by the belief that the assumptions of "perfect competition" or "perfect monopoly" said to underlie neoclassical economic theory are a false image of reality. And this belief was itself based almost entirely on the directly perceived descriptive inaccuracy of the assumptions rather than on any recognized contradiction of predictions derived from neoclassical economic theory. The lengthy discussion on marginal analysis in the American Economic Review some years ago is an even clearer, though much less important, example. The articles on both sides of the controversy largely neglect what seems to me clearly the main issue --- the conformity to experience of the implications of the marginal analysis --- and concentrate on the largely irrelevant question whether businessmen do or do not in fact reach their decisions by consulting schedules, or curves, or multivariable functions showing marginal cost and marginal revenue. Perhaps these

> two examples, and the many others they readily suggest, will serve to justify a more extensive discussion of the methodological principles involved than might otherwise seem appropriate. (Friedman, 1953, p. 16)

From value theory developed in Chapter 2, the unit value of a product is

$$-\log_b P$$

where b is the number of producers and P is the abundance of the product. If the market is highly competitive, that is, b and P are very large, then the value of the product is very low. This shows that goods produced in nearly perfect competitive markets are of low economic values and models of perfect competition are of little use in describing most important and dynamic economic sectors. For example, companies of high market valuation such as Microsoft and Intel all have dominant positions in their own industries. So imperfectness in competition is the very source of economic value. To appreciate the magnitude of valuation, we calculate the values of two products, one in highly competitive market and other in less competitive market. Suppose the first product is produced by 10 firms and with a market saturation of 90% and the second product is produced by 2 firms with a market saturation of 20%. The values of two products are

$$-\log_{10} 0.9 = 0.046 \quad \text{and} \quad -\log_2 0.2 = 2.322$$

respectively. This means highly competitive markets are of extremely low economic value.

Strictly speaking, perfect competition can not even be properly defined. Form (3.5), variable cost is a function of fixed cost, uncertainty and other factors. Suppose one product is manufactured by several companies. If from today's perspective, the technologies or organizational structures of all competing companies are not "perfect", does it always induce some new firms or existing firms to adopt better technologies or organizational structures? This is not necessarily the case because a new technology always involves fixed cost. Empirical evidences and real option theory suggest that required rates of return can be substantial before new entry can occur. Therefore, perceived "imperfection" can persist indefinitely.

Is marginal cost equal to marginal revenue?

Traditional economic theory suggests that companies will keep increasing the output until the marginal cost of the product is equal to its marginal revenue (Friedman, 1953, p. 16). Empirical evidences show that companies generally charge a substantial price mark up on their products. This analytical theory offers a simple and clear understanding about price markup. For example, if a software is targeted to sophisticated users, its interface can be simple, which reduce development cost and its sales effort can be small, which reduce variable cost. If the software developer considers increasing the market size by targeting general users, the interface of the software needs to be very intuitive with many help facilities, which increase development cost and its sales effort and after sales service can be substantial for less sophisticated users, which increase variable cost. Since the increase of market size often involve both the increase of variable cost and fixed cost, most projects are designed that the marginal cost to be much lower than the product value to maximize potential profit.

To keep increasing the output until its marginal cost equal to marginal revenue often means that the company may have to enter difficult areas, which will have repercussion on its earlier units. For example, when employees in a WalMart store in Quebec decided to unionize, WalMart closed down that store although that store would remain profitable. To keep a unionized store open will affect the margin of other stores, whose staff will attempt to unionize as well. From (3.7), the rate of return not only depends on the market size, but also depends on other factors. If the increase of market size will increase the diffusion rate as well, companies have to consider the total effect on long term profitability.

From (2.5) and (2.6), the increase of production will decrease unit value and ultimately total revenue. Therefore, if possible, companies will not keep increasing the level of output until its marginal cost is equal to marginal revenue.

Market and regulation

In neoclassical economics, regulation is justified when there is a "market failure". From this theory, it can be shown that regulation is largely driven by industries themselves to keep high rate of returns. We will use an example to illustrate how the level of regulation is influenced by sizes of the markets.

From the value theory, the unit value of a product is

$$-\log_b P$$

where b is the number of producers and P is the abundance of the product. Suppose the market size is M. MP is the number of customers. Then on average, each producer's revenue is

$$\frac{MP}{b}(-\log_b P)$$

To produce the product, it takes fixed cost and variable cost. The value of a product will increase if the number of producers can be reduced. Since monopoly often brings a lot of legal actions, we assume companies will aim to achieve duopoly by raising fixed costs with regulatory or other means.

Suppose the cost structure of each company is the same. Then the total cost of each firm is

$$K + C\frac{MP}{b}(-\log_b P)$$

where K is the fixed cost and C is the variable cost, which is defined by the percentage of the product value. The rate of return for each company is

$$\ln\left(\frac{\frac{MP}{b}(-\log_b P)}{K + C\frac{MP}{b}(-\log_b P)}\right)$$

Table (3.1) displays the relation among fixed costs, market sizes and returns of companies where there are three or two companies producing the same product. Variable cost is computed from (3.5), assuming S =1, r= 0.03, T = 25, σ = 0.55. Here variable cost means the percentage of the value of the product. *P* is assumed to be equal to 0.5.

If possible, firms will raise fixed costs so that three competing firms will lose money on average. This will allow only two firms in the market. Then we calculate the average returns for these two firms. As can be shown from Table 3.1, as market sizes increase, the level of fixed costs needed to maintain duopoly increases as well, so are the rates of return. This indicates that leading companies in industries with large market size often have strong incentives to help introduce costly regulatory requirements for their products to reduce the number of competitors to gain higher rate of return. This is why industries of vital importance to most families, such as education, health care and pharmaceutical industry, are heavily regulated. In fact, primary and secondary educations in most countries have achieved total monopoly. In a later chapter, we will discuss why trade is mostly free, while migration, which of much larger economic impact, is highly regulated (Hamilton and Whalley, 1984; Moses and Letnes, 2004).

Market size	200	400	800
Fixed cost	5.55	16.8	44.75
Variable cost	0.737	0.601	0.4676
Revenue to each firm when there are three firms	21.03	42.06	84.124
Rate of return	0.000	0.000	0.000
Revenue to each firm when there are two firms	50	100	200
Rate of return	0.165	0.263	0.3692

Table 3.1: Required levels of fixed cost to achieve duopoly with different market size

The above calculation also explains why biological and chemical weapons are banned by international treaties while nuclear weapons, which can cause much more destruction than chemical weapons, are not. Biological and chemical weapons, which are sometimes called poor men's nuclear weapons, are cheaper to make. If these weapons are not banned, many people can make them, which will reduce the value of weapons of mass destruction. To maintain the high value of such weapons, international treaties, which are generally initiated by leading political powers, banned weapons of mass destruction that are cheap to make.

3.4. What is Next?

North (1981) pointed out some missing parts in the current economic theory. They are (p. 68)

1. Theory of demographic change.
2. Theory of the development of military technology: Military technology and changes in military technology were crucial to the structure and size of the state in history.
3. Deficiency in the model of state, especially on modern pluralist state.
4. Neat supply function of new institutional arrangements.
5. A positive theory of the sociology of knowledge.

In the next two chapters, we will apply the theory developed in the first three chapters to major economic and social problems. In the process, five problems listed by North will also be resolved.

Chapter 4

Natural Resources, Technology and Institutions: A Historical Perspective

What are the driving forces of human history? What is the future of human society? Mainstream economists believe that “the performance of an economy is dependent upon its organizational structure” (North, 1981, p. 90). The gradual development of technology and increase of human population in the past ten thousand years is often interpreted as the gradual improvement of institutional structures. From this understanding, human societies will or at least would become more prosperous in the future as human institutions keep on improving. From this perspective, natural resources are not essential in economics.

Mainstream economists argued that the improvement of technology will reduce the overall consumption of resources and hence technology can substitute resources. Recently, Arrow et al. (2004) did an empirical work to conclude that “genuine investment was positive in all the rich nations of the world and in many of the poorer nations as well” (p. 160). This means that the current living standard at rich countries is sustainable, although the authors conceded that natural capital may be under priced.

Jevons, however, pointed out long ago in *The Coal Question* that, “It is wholly a confusion of ideas to suppose that the economic use of fuel is equivalent to a diminished consumption. The very contrary is the truth.” (Jevons, 1965(1865), p. 140) In the last several hundred years, the consumption of energy has increased tremendously while energy conversion has become more efficient overtime (Smil, 2003). If the consumption level of natural resources is higher than what the nature can regenerate, mostly by energy from the sun, then this consumption level is unsustainable. Historically, all the high consumption societies collapsed because of the depletion of natural resources. When natural resources are depleted, the accumulated technology and human capital, which are really means to utilize natural resources, lose their value as well.

North noted that neoclassical economics was unable to explain the dynamics of change of the human history:

> Neoclassical economics applied to economic development or economic history may account very well for the performance of an economy at a moment of time or, with comparative statistics, contrasts in the performance of an economy over time; but it does not and cannot explain the dynamics of change. (North, 1981, p. 57)

In this chapter we will show how the new theory provides a simple and coherent understanding of the co-evolution of natural resources, technology and human institutions. This theory also provides a simple description of the process of biological evolution, which includes the evolution of human societies as a special case. Finally, we will offer a detailed analysis of the process of industrialization in the last several hundred years to show this theory provides a very coherent understanding of the process. This analysis unequivocally indicates that the current high level of consumption is unsustainable.

4.1. Dynamics of Change in Human Civilization

We often associated many changes in human civilization as progress or setback. But they can be more precisely described as the changes of systems with different levels of fixed cost and variable costs. The beginning of civilization is often identified with the transformation from gathering/hunting to agriculture. We can compare agriculture and gathering activities as an example.

Gathering only needs working at harvesting time. Agriculture, however, needs working in the whole process from planting to harvesting. Hence, agriculture activity is of higher fixed costs than gathering of wild grains and fruits. Agriculture activities, as higher fixed cost systems than gathering activities, also generate higher density of resources, which support higher density of population and make agriculture societies more powerful than gathering societies. Agriculture societies, therefore, have the incentive and power to take more land, by force if necessary, and convert it into agricultural use. The major part of human civilization is the process of expansion of agricultural activity (Diamond, 1997).

While the high concentration of resources in an agriculture society enables itself to expand, it will also attract looting from outsiders. Only when military capability of a society is strong enough related to its neighbors, agriculture becomes possible.

Therefore agriculture activities are intrinsically linked to military activities and the use of violence. Among all human activities, the application of violence often offers the highest returns. So the military industry was developed very early and like most other industries without anti-trust constraints, achieved regional monopolies in most places. Each unit of the monopoly on violence is a state. A monopoly on violence provides a more stable social environment than a free market of violence (Diamond, 1997). More equal opportunity to conduct violence is conducive to the creation of anarchy and chaos. From Formula (3.5), systems with lower uncertainty have lower variable cost in production and transaction. So state as in institutional structure is very competitive and gradually replaces all other forms of social systems as the most important form of social institutions.

Since all economic activities are ultimately regulated by the states, the history of economic development is largely shaped by the evolution of the structures of the states, which in turn is largely shaped by the development of military organizations and military technologies. (North, 1981) Better military technologies are often more expensive and better military organizations often need longer time of training, which means that the development of military capabilities is a process toward higher fixed cost, which requires the support of higher level of resource abundance and larger market size.

Historically, the emergence of new military technologies is generally associated with the new technologies on basic materials or energy use. The progress of technology is often marked by the increase of energy demand. Therefore, the emergences of new military technologies, which require higher fixed cost, often cause financial difficulties for the states (North, 1981). Higher fixed cost systems require bigger market to sustain itself. So the emergence of new military technologies, by changing the cost structure of states, makes the existing political units unstable. At the same time, higher fixed cost systems earn higher rates of return in larger markets. When market size is large, it pays to maintain costly social structures with well defined property right that enable them to build up human and physical capital and to reduce uncertainty in exchanges. These states enjoy lower variable costs in various economic activities. They tend to expand to achieve even greater economies of scale and accumulate great wealth in the process. The stage of expansion of the empires is often associated with rising living standard and greater political freedom, at least for the citizens of the empires. This type of periods can easily be classified as progress. However, most of these wealthy states eventually collapsed. Since most theories about history and economics are intrinsically progressive, the explanations for the collapses of highly developed societies are often very colorful and imaginative. However, from a physical, or a natural resource based theory, understanding collapses becomes very simple and straightforward.

First, wealthy states are often the results of the development of new technologies with higher fixed costs, which require more resources or low entropy to sustain themselves. Economic development is largely determined by the amount of low entropy that human beings can extract from the natural environment. The agricultural revolution is a transformation of food production from collecting grains or fruits in the wild to actively managing the extraction of large quantities of low entropy from the sun through certain plants in highly concentrated manner. The industrial revolution is a transformation from collecting low entropy from living organisms to the development of technologies that systematically utilize the deposits of biological low entropy, like petroleum and coal. Biological low entropy that was deposited over hundreds of millions of years is transformed for use by human societies over the last several hundreds years. This development forms the foundation of the economic prosperity many the world over have enjoyed in recent times. Each step of progress is a step toward higher fixed cost system, which requires more resource to sustain itself. Many believe that ultimately natural resources are practically infinite. Although this could be true, there is no logical certainty that a scientific breakthrough will always occur in time to provide new ways to utilize resources when the prevailing ways of resource development and use no longer sustain the existing social structures and life styles. Historical evidence shows that most highly developed societies collapsed before they could find new ways to utilize resources.

Mainstream economists argue that today's capital investment will always benefit future for accumulated human capital can substitute natural capital (Samuelson and Nordhaus, 1998, p. 328). However, the accumulated capital, including knowledge capital, is geared toward the current level of energy consumption and other natural conditions. If these conditions change, the assets could become liability. For example, highways can facilitate rapid movements of goods in an era of cheap fuels. However, if cheap energy supplies exhaust in the future, highways, which take vast amount of land, cannot be reverted to agriculture land easily. So today's investment could harm future's sustainable economic environment. In fact, in most sites of early civilizations, once rich and productive areas have been turned into desolate regions through over exploitation (Tainter, 1988; Ponting, 1991).

While over exploitation of natural resources is often thought to be caused by some specific mistakes, it is an intrinsic consequence from competitive pressure. If a community can develop methods to extract more natural resources faster than its neighbors, it will become stronger and gain ability to take over its neighbors resources. So competitiveness is intrinsically incompatible with sustainability.

Some of societies that Diamond (2005) described as success stories did manage to maintain sustainable lifestyle by restricting energy consumption. It will be helpful to compare the paths of development of two island states, Japan and Britain. As noted by Diamond (2005), Japan maintained a sustainable lifestyle. It even abolished the use of firearms, which require tremendous amount of energy in their manufacturing and use. By contrast, Britain actively developed its industry, with the result of complete deforestation (Jevons, 1965). Because of the unsustainability of its ecological environment, Britain adopted emigration as a systematic way to reduce ecological pressure, which was made possible with its superior military power (Colinvaux, 1980). As a result, Britain colonized North America and many other places to become the largest empire in human history. On the other hand, the Tokugawa-era Japan, being low energy intensive, was quickly defeated by America, a descendent of British Empire, which utilized high energy intensive weapons. Soon after an American fleet under Commodore Perry forced Japan to open its ports, Japan swiftly abandoned its policy to maintain a sustainable lifestyle. It quickly modernize itself with high energy intensive industries and start conquering neighboring countries for land and other natural resources to satisfy the increasing demand for an expanding economy. This comparison explains why the policy of sustainability cannot be sustained under free competition. The subtitle of Diamond's new book is "How societies choose to fail or succeed". However, the reality is more subtle. A society that is succeeded in maintaining a sustainable lifestyle will eventually be failed by a society that is more aggressive and adventurous in developing high energy consuming technologies.

Second, economic development is usually achieved by more concentration of resources or wealth. For example, compared with grains in the wild, grains of cultivated plants are much more concentrated, which provides great incentive for looting. The concentration of wealth makes it necessary and profitable to create and enforce property rights of exclusion. At the same time, the concentration of wealth by only a small number of people or states increases the incentives for others to extract wealth from them. Externally, as the wealth of a state accumulates, it becomes more and more profitable for other states to extract wealth from it. Internally, as the level of wealth increases, the demand of redistribution increases as well. Most wealthy states evolve into a system to provide safety net for the poor to reduce this incentive. The decline of the Roman Empire offers a classical example. "The prospering economy faces ever-growing costs of either bribing the invaders or making increased military expenditures. … Not only were larger and larger payments in gold made to barbarian groups to bribe them not to invade, but the expenses of the legions rose.... At the same time, Rome was feeding 120,000 of its citizens free." (North, 1981, p. 115, 122) In a word, the cost of maintaining a wealthy system increases with the level of wealth.

Third, the accumulation of wealth increases the size of market. It encourages higher fixed cost investment, which is mainly achieved through division of labor and specialization of knowledge. While specialization increases the depth of knowledge of a person, it also reduces the breadth of knowledge and makes one less able to determine the value of others' work, which increases uncertainty in exchanges and hence increases transaction costs (Barzel, 1982; Wallis and North, 1986). Highly complex formal and informal constraints are developed to reduce uncertainty and transaction costs (North, 1990). But the constraints often restrict the path of future development of the system as well and make it less responsive to changes in the environment.

Human history demonstrates that social systems become more and more sensitive to uncertainties when the level of fixed costs and living standards increase. In the end, exhaustion of natural resources or changes in the environment sparked the inevitable decline of the old systems. Frequently, this dynamics was often accompanied by the rise of simpler structured and lower fixed costs systems on the periphery (Colinvaux, 1980; Tainter, 1988).

Along with a repeated cycle of rise and fall of states and civilizations, the higher fixed cost systems, with greater scale economies, gradually diffuse and replace the lower fixed cost systems. But will this trend last forever? Will human societies ever evolve to an optimal equilibrium state?

All living organisms are characterized by the struggle to extract low entropy from the environment. Since biological data contain much richer samples over a much longer time horizon than data on the human species, in the next section, we will investigate the general patterns in biological evolution in order to gain deeper insights into the evolution of human societies.

4.2. General Patterns in Biological Evolution

Biological species are sometimes classified, according to their relative level of fixed and variable costs, into two categories, the *r*-strategists and the *K*-strategists. Organisms with low fixed costs are termed *r*-strategists. They are usually of small size, produce abundant offspring and invest very little in each one. They are the species that prosper in a volatile environment for low fixed costs make them flexible. But they cannot compete well with other species in a stable environment for their marginal costs are high. In contrast, fixed costs are high for the *K*-strategists. They are usually large in size, produce fewer offspring but invest much more in each one. They are conservative species that are able to out-compete the *r*-strategists in stable environments, for their marginal costs are low. But they cannot adjust quickly when environment changes. Between the extreme *r*-strategists such as bacteria and the extreme *K*-strategists such as elephants, there lies the *r* and *K* continuum (MacArthur and Wilson, 1967).

Over geologic time, stable periods with relatively uniform environment were punctuated with sudden changes from time to time, such as a huge asteroid hitting the earth or the quick emergence of a dominant species, such as human beings. During short periods of volatile change, large size *K*-strategy species are more prone to extinction since it is more difficult for the high fixed cost systems to adjust to the changing environment. Specifically there are three reasons. First, large species need more resources to survive and are often more severely affected by changes in the environment (Withers, 1992; Gould, 1996). Second, large species often contain big concentration of low entropy sources and become the prime targets of other species. For example, improvement in the hunting skills of human beings quickly leads to the extinction of most large mammal species. Third, large species usually have more complex structures than small species, which make it more difficult for large species to develop variations that are well coordinated internally. So large species usually have much lower genetic diversity than small species. For example, two species of fruit flies may only have about 25 percent of their DNA sequences in common while humans and chimpanzees have over 98 percent in common, even though they belong to different genera (Stebbins, 1982). There are more than five hundred species of fruit flies in Hawaii (Stebbins, 1982). But there is only one species in the genus *Homo*. The ultimate in genetic diversity and reproductive speed can be found in micro-organisms, such as bacteria (Margulis, 1998). This lack of genetic diversity among large species makes them much more vulnerable to sudden environmental changes. The mass extinction of species, especially the dominant species, during periods of volatile environmental changes, clears the ground for a new round of evolutionary competition. During long periods of relative stability, the small size and less specialized *r*-strategy species tend to branch into new species. Among these new species, those larger or more specialized, which incur higher fixed costs but are more efficient with lower marginal costs, have a competitive advantage in a stable environment and gradually dominate the ecosystem (Brown and Maurer, 1986; Colinvaux, 1986; Gould and Eldredge, 1993).

Simpson (1944) was the first biologist to apply careful statistical methods for interpreting pattern of evolution (Stebbins, 1982). He summarized his findings as the following:

> Liability to extinction tends to be directly proportional to rate of evolution. Bradytelic (*slow evolving*) lines are almost immortal. The majority of tachytelic (*fast evolving*) lines quickly become extinct and those that survive cease to be tachytelic (*fast evolving*). ...When

> related phyla die out in the order of their rates of evolution or in the reverse order of their times of origin, it follows that this order is also usually that of degrees of specialization and that more specialized phyla tend to become extinct before less specialized. This phenomenon is also far from universal, but it is so common that it does deserve recognition as a rule or principle in evolutionary studies: the rule of the survival of the relatively unspecialized. (Simpson, 1944, p. 143)

The statistical results show that the fast evolving and highly specialized species, which are often more efficient and tend to dominate ecological systems, are more prone to extinction than the slow evolving and unspecialized species, which are less efficient, less competitive but more flexible.

Biologists have not found a universally applicable measure of fitness. (Stearns, 1992) From the analytical thermodynamic theory, such measure doesn't exist. Our analysis shows that the main theme of economic and biological evolution is the tradeoff between competitiveness of high fixed cost systems in a stable and resource rich environment and flexibility of low fixed cost systems in a volatile or resource poor environment. Since there is no dominant strategy in all environments, the beautiful and diverse ecological system does not reach an equilibrium state, even after four billion years of biological evolution. For the same reason, economic organizations and systems will not converge to an equilibrium state.

The tradeoff in performance of organic systems ultimately rests on the tradeoff in performance of inorganic systems, which Wiener observed more than fifty years ago:

> While the prediction apparatus which we had at first designed could be made to anticipate an extremely smooth curve to any desired degree of approximation, this refinement of behavior was always attained at the cost of an increasing sensitivity. The better the apparatus was for smooth waves, the more it would be set into oscillation by small departures from smoothness, and the longer it would be before such oscillations would die out. Thus the good prediction of a smooth wave seemed to require a more delicate and sensitive apparatus than the best possible prediction of a rough curve, and the choice of the particular apparatus to be used in a specific case

> was dependent on the statistical nature of the phenomenon to be predicted. (Wiener, 1948, p. 9)

Wiener believed that the problems “centered not around the technique of electrical engineering but around the much more fundamental notion of the message, whether this should be transmitted by electrical, mechanical, or nervous means” and thought they had something in common with Heisenberg’s Principle of Uncertainty, which itself is a tradeoff between two factors.

The following passage from Atkins (1995) may shed further light about some basic properties of biological and social evolution:

> The region of sulfur was also explored by nature --- in nature’s serendipitous, purposeless, but effective way --- in an early investigation of the opportunities for life. Nature discovered that in some respects hydrogen sulfide (H_2S), the analog of water (H_2O), can be used by organisms in much the same way as water is used in the process of photosynthesis --- as a source of hydrogen. The great difference to note is that when hydrogen is removed from a water molecule by a green plant, the excrement is gaseous oxygen, which then mingles with the globally distributed atmosphere. However, when hydrogen is removed from hydrogen sulfide in the interior of a bacterium, the excrement is sulfur. Sulfur, being a solid, does not waft away, so the colony of organisms has to develop a mode of survival based on a gradually accumulating mound of its own sewage. We still mine those ancient mound of sulfur excrement from beneath the Gulf of Mexico. Sulfur’s northern neighbor, oxygen, turned out to be a much more viable alternative to sulfur in nature’s blind efforts to generate the transmittal and accumulation of information, and sulfur is now used only by primitive species that occupy a minor niche of nature. (Atkins, 1995, p. 22)

Several basic properties in biological and social evolution may be observed. First, technology development is usually achieved with a higher fixed cost methodology. The chemical bond between hydrogen and oxygen is stronger than the chemical bond between hydrogen and sulfur. Therefore it is more difficult to obtain hydrogen from water than from hydrogen sulfide. Human evolution is often divided into the Stone, Bronze, and Iron Ages. Each age is characterized by the ability of human beings to utilize energy more intensively. The melting point of bronze is 232 ºC and the melting point of iron is 1535 ºC, which is much higher. The substitution of bronze by iron is a long-drawn process that required much higher intensity of energy use, or much higher fixed cost (Smil, 1999). Second, new technologies are often accompanied by wider diffusion of waste and can be very harmful to general community. "When hydrogen is removed from a water molecule by a green plant, the excrement is gaseous oxygen, which then mingles with the globally distributed atmosphere." The release of oxygen as waste into the atmosphere killed most of the living organisms at that time. Early industrial sites are often built by rivers or lakes so industrial wastes are diffused to wider communities while profits from products are retained by the producers. Today, the advance of telecommunication enables the production facilities with harmful side effects to be located in poor countries. Third, organisms, including human beings, often find extraordinary ways to adapt to the new environment. Initially, oxygen was deadly to most organisms. Gradually, some surviving organisms evolved mechanisms to resist the poisoning of oxygen. Eventually, oxygen, which permeates in air, becomes the vital source of energy that enables the evolution of highly mobile animals. Major evolutions in human societies, such as agriculture revolution and industrial revolution, are successful responses to resource depletions. The ability for organisms to adapt to new environment makes the timing of resource constraint highly unpredictable.

In the next section, we will apply the insights gained to understand the origin and future of our industrial civilization.

4.3. Industrialization: Origin and its Sustainability

4.3.1. The origin of the Industrial Revolution

The steady state of the Middle Ages was punctuated by the introduction of guns.

> Once better confined and directed, the force of exploding gunpowder begun propelling increasingly heavier projectiles at larger distances. Manufacture of such guns begun in China just before the year 1200, and the true guns were cast in Europe only a few decades later. (Smil, 1999, p. 129)

Since China was a unified country and Europe was comprised of many small states, the competition among European states was much more intense, which led to continuous improvement in the techniques of gun making in Europe.

> The combination of more maneuverable vessels with more accurate guns (made possible by advances in the smelting of copper and iron and by the invention of gunpowder) produced an energy converter of unprecedented speed, range, and destructive power that helped to usher a new era of world history. (Smil, 1999, p. 106)

Because of the unprecedented destructive power of guns, "casting of field guns became one of the first mass-production industries of the early modern world" (Smil, 1999, p. 128). Manufacture and use of guns, however, is very energy intensive and financially expensive.

> Survival now required not only a larger army, but a trained, disciplined fighting force supported by costly equipment in the forms of cannon and muskets, …Warfare on land and at the sea (where the size and armament of naval ships increases dramatically) had dramatically altered the size of the financial and resources necessary for survival. (North, 1981, p. 138)

The high cost of new military technology led to the growth of the size of viable states, which was often achieved by wars. It also forced rulers to make political concession in exchange for financial contribution from broader spectrum of society (North, 1981).

While North clearly recognized the direction of causality from technology to institutions in many individual historical events, he reversed the direction when he tried to interpret general patterns of social changes, such as the Industrial Revolution:

> While gunpowder, the compass, better shipdesigns, printing and paper all played a part in the expansion of western Europe, the results were widely divergent. The technological changes associated with the Industrial Revolution required the prior development of a set of property rights, which raised the private rate of return on invention and innovation. (North, 1981, p. 147)

We shall briefly review the process of the Industrial Revolution, which will show that a resource based explanation offers a much simpler understanding. A resource based understanding explains why the economic performances of different states, especially neighboring states, can diverge significantly, although states can learn from each other's institutional and technical innovations.

Iron was the essential material in modern warfare. Since the making of iron needs tremendous amount of energy, by the middle of eighteenth century, whole England was deforested as wood was the main source of energy in iron making. Because of the depletion of wood, the substitution of coal for charcoal became a necessity (Jevons, 1965). The first successful experiment to process iron with coal was around 1750, the time that is generally considered as the starting point of the Industrial Revolution. Hence the Industrial Revolution is a successful technology response to a resource crisis by vastly increasing the use of coal in iron making and steam engine. Of course, many institutional measures were adopted to respond the resource and technology conditions of the time. But the fundamental reason of the Industrial Revolution is the inventions that enable the use of the vast amount of coal as a reliable energy source.

	1860	1875	1889	1903
Great Britain	80,043	133,306	176,917	230,334
United States	14,334	46,686	126,098	319,068
Germany	12,000	47,756	83,614	159,846
France	7,900	16,686	23,915	34,345

Table 4.1: The annual production of coal in thousands of tons in some coal-producing countries. Adapted from Jevons (1965).

Coal had been the main source of energy before the widespread use of petroleum. Table 4.1 records the annual production of coals in the chief coal-producing countries from 1860 to 1903. Several observations can be made from the table. First, in 1860, by far the largest producer of coal was Great Britain, which was also the most powerful country. By 1903, United States became the largest coal producer, which soon demonstrated its unparallel economic and military might. Second, among European countries, the share of coal production from Germany rose sharply from 1860 to 1903, which gradually but decisively changed the balance of economic and military power in Europe. Third, France entered nineteenth century as the strongest continental power. However, as Europe became more industrialized, the paucity of coal mine on its territory, relative to Britain and Germany, greatly inhibits its economic development. France tried its best to increase its coal supply. For example, after World War I, France demanded repatriation of coal from German mines. But the fact that the mines were on German territory made it difficult to enforce the treaty for long times. The lack of energy resources explains why France, facing German emergence, was always on the defensive. The above discussion shows that ultimately, it is the amount of energy and other natural resources a country can control that determines the economic wealth and military power of a nation.

In the following, we will further discuss the institutional view of historical development expressed by North:

> It was the reduction of transaction costs due to the establishment of private property rights of and competition in trade and commerce that allowed England to escape the Malthusian check that both France and Spain suffered during the seventeenth century. (North, 1981, p. 157)

Table 4.2 displays the statistics of population of England and Wales from 1570 to 1901. From it, we can find that English population was declining in the early eighteenth century, when the resource depletion caused by iron making industry lowered the carrying capacity of the country. This means that Britain, like France and Spain, suffered from Malthusian check at that time. British population started to grow steadily after 1750, which coincides with the large scale production of iron with coal and Watt's improvement on steam engines. It is the technology breakthrough which allowed coal to be used in producing iron and other commodities, which can be used to trade for food and other necessities that allowed British population to grow steadily. In Jevons' own words, "The nation, as a whole, has rapidly grown more numerous from the time when the steam-engine and other inventions involving the consumption of coal came into use" (Jevons, 1965, p. 9). However, France was a much smaller coal producer than Britain and Spain's coal out was negligible even compared with France (Jevons, 1965). Therefore, it is the differences in resources, which cannot be changed easily, and not the differences in institutional structures, which can be changed, that determined the long run social performances.

Year	Population	Numerical increase for ten years	Percentage increase for ten years
1570	4,160,321		
1600	4,811,718	217,132	5
1630	5,600,718	262,933	5
1670	5,773,646	43,282	1
1700	6,045,008	90,454	2
1701	6,121,525		
1711	6,252,105	130,580	2
1721	6,252,750	645	0
1731	6,182,972	-69,778	-1
1741	6,153,227	-29,745	0
1751	6,335,840	182,613	3
1761	6,720,547	384,707	6
1771	7,153,494	432,947	6
1781	7,573,787	420,293	6
1791	8,255,617	681,830	9
1801	8,892,536	636,919	11
1811	10,164,256	1,271,720	14
1821	12,000,236	1,835,980	18
1831	13,896,797	1,896,561	16
1841	15,914,148	2,007,351	14
1851	17,914,148	2,007,351	14
1861	20,066,224	2,138,615	12
1871	22,712,266	2,646,042	13
1881	25,974,439	3,262,173	14
1891	29,002,525	3,028,086	12
1901	32,527,843	3,525,318	12

Table 4.2 Population of England and Wales from 1570 to 1901. Reproduced from Jevons (1965).

It is often claimed that in today's high tech society, natural resource plays an insignificant role for a nation's economic wealth. Japan, which doesn't possess many natural resources, is often cited as an example. While Japan does achieve very high living standard, during this process, its fertility rate has dropped to 1.38, which is among the lowest in the world and far below the replacement level. This means that high living standard without the support of natural resources is unsustainable. Next chapter will provide more detailed discussion about natural resource, living standard and fertility.

Ultimately, it is the resource endowment, which cannot be copied from country to country, that determines nations' power. However, when resources are discovered in poor or weak countries, it often take long time before people in these countries can take effective control over resources on their land. For example, oil rich developing countries are often the targets of political and military actions from major powers and political reforms in resource rich countries, such as Iran, were often suppressed. While resource endowment determines long term economic performance, internal and external institutional structures determine short term fortunes, which leaves the impression that institutional structures, instead of natural resources are the decisive factor in economic performance.

We have shown that the resource based theory of human history offers a simpler and more consistent understanding than the institutional theory. In the following subsection, we will consider the future of our fossil fuel based civilization from the resource theory.

4.3.2. The future of our fossil fuel based civilization

While it is difficult to predict the pace of technology advance and its consequence to resource exploitation in the near future, the collapses of all past highly developed civilizations due to resource depletion offer a clear understanding of long term patterns, which should make us prepare for a soft landing when resources become very scarce. Instead, our human society keeps on increasing consumption whenever it is possible. Why this is the case? We will analyze from individual and institutional levels.

At individual level, as we have explained in Chapter 1, the only credible signal of attractiveness is the possession and consumption of low entropy resources. All other signals of attractiveness, which require less effort, can be copied easily and hence lose credibility. Therefore, in a consumer society, maximizing consumption becomes the standard way in social communication. For example, legislations often require high fuel efficiency for cars to reduce oil consumption. However, SUVs, which consume much more energy than ordinary cars, are not classified as cars, "although they are clearly used as passenger cars"(Smil, 2003, p. 96). SUVs become very popular for the exact reason that they are costly to buy and operate, which signals the wealth and status of SUV owners.

At the institutional level, institutions, such as firms, have fixed and variable costs. From Figure 3.2, higher output generally means higher rate of return. Therefore higher level of consumptions is encouraged at institutional levels. This is why saving is often called private virtue but public vice. Mainstream economic theory, which is a religion that justifies the modern lifestyle, often advocates the high consumption strategy (Nelson, 2001). Samuelson and Nordhaus argued:

> The substitutability of natural capital and other kinds of capital is shown by the production indifference curve … That output can be produced … with a conservationist policy that emphasizes reducing energy use today, leaving much oil and gas and relatively little human capital for the future. Or it might be produced with a low-energy-price and high-education strategy … Either of these is feasible, and the more desirable one would be the one that has a higher consumption both now and in the future. (Samuelson and Nordhaus, 1998, p. 328)

What will be the consequence when the fossil energy is no more available at low cost? Will it only affect our lifestyle, such as extensive traveling by cars and airplanes? In fact, not only industry depends utterly on fossil energy, but also agriculture, our very food source, depends very much on the fossil energy.

> The next cause of cheapness in food came from applying the new cheap energy to agriculture. Tractors, harvesting and planting machines and, above all, chemical fertilizers lowered the costs of growing food even as they increased the total supply. …There then came yet one further push to cheap food. This was the development of crops such as hybrid corn, a new agriculture that goes by the name of the "green revolution" in the contemporary press. This agriculture is completely and inextricably dependent on a large flux of cheap energy. … We have taken over many of the functions that a wild plant had to do for itself, and have done it for the plant ourselves, in factories. We do not let plant hunt out scarce minerals with its roots, we give it super abundant supplies of fertilizer so that it does not have to work for its nutrients. We take away plant's ability to protect itself against disease and pests, because the plants used to spend part of the energy reserves of its grain to do the job itself. Instead we protect the plant with chemicals. In other words we keep alive, with fertilizers

> and chemicals, a plant that would have no chance of hacking it alone, and the energy that its ancestor would have spent in fighting its own battles is then freed for the plant to make more grain. This extra grain, therefore, is entirely dependent on cheap fuels supplied to our chemical industries; indeed, in a real sense the energy of this extra grain is some of the energy from the chemical industry. (Colinvaux, 1980, p. 335)

Smil (1999) put it in a more quantitative scale.

> Between 1950 and 1995, global synthesis of fertilizer ammonia rose from less than five million to about eighty million tones, …perhaps already two out of every five people, gets the protein in his or her diet from synthetic nitrogenous fertilizers. (Smil, 1999, p. 171)

Therefore our very survival depends on the continuing flow of fossil energy. The discussion by Colinvaux and Smil also shows that the main function of knowledge is to utilize energy and other natural resources. This is distinctly different from the mainstream theory that knowledge is used to substitute natural resources (Johnson, 2000). Since knowledge is used to utilize natural resources instead of substitute them, it is easy to understand in most sites of highly developed early civilizations, once rich and productive areas have been turned into desolate regions through over exploitation while in primitive human societies, environmental quality does not deteriorate as much for people there had less knowledge to extract as much resources from nature.

Will the progress of science resolve the resource problem in the future? Jovons made the following observation.

> It is credibly stated, too, that a manufacturer often spends no more in fuel where it is dear than where it is cheap. But persons will commit a great oversight here if they overlook the cost of improved and complicated engine, is higher than that of a simple one. The question is one of capital against current expenditure. … It is wholly a confusion of ideas to suppose that the economic use of fuel is equivalent to the diminished consumption. The very contrary is the truth. As a rule, new modes of economy will lead to an increase of consumption according to a principle recognized in many parallel instances. (Jevons, 1965, p. xxxv and p. 140)

Put it in another way, the improvement of technology is to achieve lower variable cost at the expense of higher fixed cost. Since it takes larger output for higher fixed cost systems to breakeven, to earn a positive return for higher fixed cost systems, the total use of energy has to be higher than before. That is, technology advancement in energy efficiency will increase the total energy consumption. Jevons' statement has stood the test of time. Indeed, the total consumption of energy has kept growing, almost uninterrupted decades after decades, in the last several centuries, along with the continuous efficiency gain of the energy conversion (Smil, 2003).

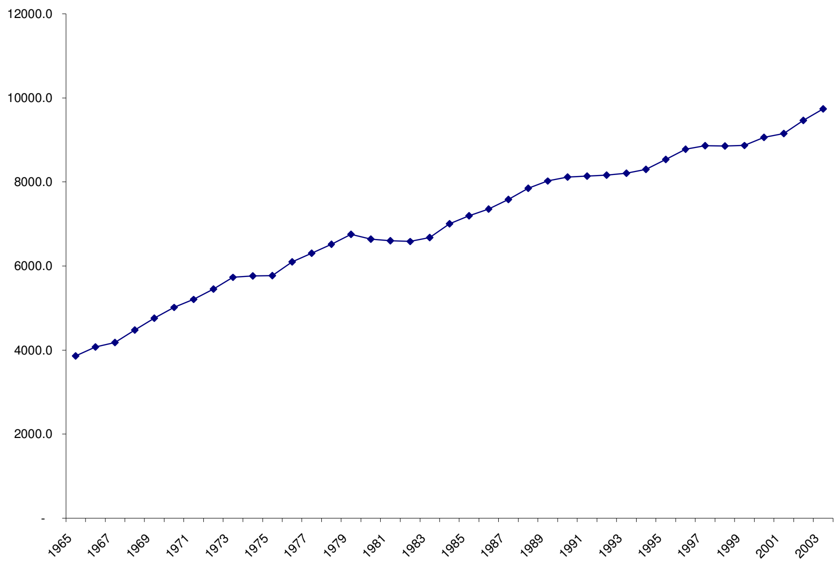

Figure 4.1: Global energy consumption (Million ton oil equivalent) from 1965 to 2003. Source: BP.

Figure 4.1 displays the total primary energy consumption world wide from 1965 to 2003, a period of rapid technology progress. During this period, energy consumption grew steadily, with only two brief interruptions. The first is from 1979 to 1982, a period of Iranian Revolution. Oil price jumped from 13.60 US dollars per barrel in 1978 to 35.69 in 1980. The second is from 1997 to 1998, a period of Asian financial crisis. Oil price dropped from 19.09 US dollars per barrel in 1997 to 12.72 in 1998 for lack of demand. Figure 4.2 displayed the oil price change from 1965 to 2003. The drops of energy consumption in both periods were not related to technology progress.

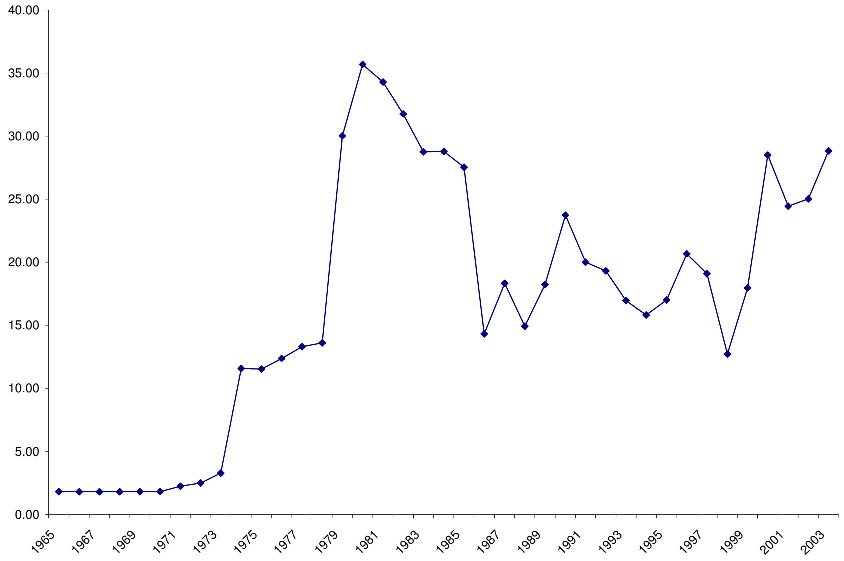

Figure 4.2. Oil price (US dollars per barrel) from 1965 to 2003. Source: BP.

At the crest of every civilization, it is difficult to imagine its ultimate demise. But are there clear signs of unsustainability of our civilization at this moment? Yes. The fertility rates in most developed countries are already far below replacement level, a clear indication that the current living standard in most developed countries are not adequately supported by the available resources. In the next chapter, we will make a detailed discussion of fertility problem.

4.4. Concluding Remarks

The world history is an interaction between natural resources and technology, which is the ability to utilize natural resources. Institutional structures are evolutionary adaptations to the level of technology and resources. Because of the existence of fixed assets accumulated over time, the adaptations are gradual and are constrained by existing structures. The progresses we observe are really the increase of fixed costs that bring down the variable costs. Since higher fixed costs systems need large amount of resources to maintain themselves, the grand systems of ancient civilizations all collapsed spectacularly when resources were depleted.

Chapter 5

Migration, Trade, Education and Fertility: A Spatial Perspective

From the entropy law, there is a tendency for systems to become homogeneous. Biological systems require selectively permeable barriers to keep them separate from the environment (Morowitz, 1992; Margulis, 1998). If one social group, for one reason or another, achieves higher level of status than others, entropy law determines that this superior status cannot last long unless an institutional barrier is established to prevent the free migration of members of different social groups. In general, social systems with more natural or institutional resources are more exclusive and enforce property rights more vigorously. Today, this exclusiveness is most profoundly represented at the state level. A state is a common property right of its citizens. To protect the high value from dilution, countries with more natural or institutional resources restrict access through strict immigration laws. In this chapter, we show that how the patterns of trade, education and fertility in wealthy and poor countries can be easily understood when these problems are considered together with migration restriction.

5.1. Migration and Trade

While the standard economic theory claims equivalence of trade and migration, the countries that adopt the most restrictive migration laws are often the strongest promoters of free trade. In the last two decades, tremendous numbers of papers have been written about the new trade theory. The work by Avinash Dixit and Joseph Stiglitz (1977) is often credited as the crucial innovation that made the new trade theory possible

(Fujita, Krugman, Venerables, 1999). In Dixit and Stiglitz (1977) and later works, models of production costs involve fixed costs and variable costs, which make economy of scale easy to describe. These models can finally explain the obvious fact that free trade benefits some countries at the cost of other countries (Krugman and Venables, 1995; Puga, 1999). In these models, fixed costs and variable costs are independent constants. Although it is intuitively clear that a production system with higher fixed cost generally has lower variable costs, a lack of analytical theories make it difficult to implement this intuition. The analytical theory developed in this book enables us to make more precise analysis of gains or losses from trade between different nations with different fixed costs at different stages of economic development.

Adam Smith noted that the division of labor is determined by the size of the market. A system of finer division of labor has higher fixed cost because the production involves more people and more coordination. This system is also of lower variable cost because of the efficiency gain. From Figure 3.2, a finer division of labor, which has higher fixed cost, is more profitable when the market size is large. As division of labor becomes finer, final products are the results of corporation among many more people or firms. Since physical closeness facilitate communication and transportation, firms often agglomerate into clusters.

The clustering of firms and distribution systems cause the clustering of human population. The level of concentration of human population is largely determined by the fixed cost, or living standard of the society. When the living standard is higher, population becomes more concentrated. For example, if education is not required, families may live very far apart. When primary education is required, population densities have to be high enough to support the cost of a primary school. When education in secondary school is required, even higher population density is required for the cost of a secondary school is higher than a primary school. The establishment of high fixed cost facilities requires a large population. This is why as living stand increases, an increasing percentage of people live in metropolitan areas. This pattern of population concentration with social development was also evident in ancient societies such as the Roman Empire (Colinvaux, 1980).

Free movement ensures that real wages are roughly equal across regions. Globally, the efficient processing of goods and information in developed countries creates huge demand in the labor market. From the value theory, value, including the value of labor, is a function of scarcity.

To maintain high real wages, developed countries adopt strict migration control to limit the supply of labor. This restriction of labor mobility imposes a tremendous global economic cost that is much higher than the gain from trade (Hamilton and Whalley, 1984; Moses and Letnes, 2004). To a country as a whole, labor cost can be better understood as fixed cost. Developed countries, with high fixed costs and low variable costs, generally promote free trade to increase the size of the market of their products. Less developed countries or "latecomers", being of lower fixed assets and higher variable costs, are less competitive. At the same time, from Figure 3.2, lower fixed cost systems do not need large market size to sustain themselves. So they prefer protective trade policies.

While mainstream economics claims free trade benefit all nations, a look at the historical development of the trade policies of England and America, arguably the two most successful countries in the last two hundred years, will illustrate how countries at different stages of development adopt different trade policies.

In the early nineteenth century, England was the dominant force in manufacturing. David Ricardo, an English economist, wrote the following:

> Under a system of perfectly free commerce, each country naturally devotes its capital and labor to such employments as are most beneficial to each. This pursuit of individual advantage is admirably connected with the universal good of the whole. By stimulating industry, by rewarding ingenuity, and by using most efficaciously the peculiar powers bestowed by nature, it distributes labor most effectively and most economically: while, by increasing the general mass of production, it diffuses general benefit, and binds together, by one common tie of interest and intercourse, the universal society of nations throughout the civilized world. It is in this principle which determines that wine shall be made in France and Portugal, that corn shall be grown in America and Poland, and that hardware and other goods shall be manufactured in England. (Ricardo, 1817)

If the United States adopted this comparative advantage theory, today it would still be mainly a corn growing country. While Ricardo saw it as optimal that corn should be grown in America and manufacturing should be in England, Alexander Hamilton, an American lived at roughly the same time as Ricardo, apparently had different opinions. Since his argument is as relevant today as was two centuries ago, we will quote him in length.

> Experience teaches, that men are often so much governed by what they are accustomed to see and practice, that the simplest and most obvious improvements, in the most ordinary occupations, are adopted with hesitation, reluctance, and by slow graduations. The spontaneous transition to new pursuits, in a community long habituated to different ones, may be expected to be attended with proportionally greater difficulty. When former occupations ceased to yield a profit adequate to the subsistence of their followers, or when there was an absolute deficiency of employment in them, owing to the superabundance of hands, changes would ensure; but these changes would be likely to be more tardy than might consist with the interest either of individuals or of the Society. In many cases they would not happen, while a bare support could be insured by an adherence to ancient courses; though a resort to a more profitable employment might be practicable. To produce the desirable changes as early as may be expedient, may therefore require the incitement and patronage of government.
>
> The apprehension of failing in new attempts is perhaps a more serious impediment. There are dispositions apt to be attracted by the mere novelty of an undertaking – but these are not always the best calculated to give it success. To this, it is of importance that the confidence of cautious sagacious capitalists, both citizens and foreigners, should be excited. And to inspire this description of persons with confidence, it is essential, that they should be made to see in any project, which is new, and for that reason alone, if, for no other, precarious, the prospect of such a degree of countenance and support from government, as may be capable of overcoming the obstacles, inseparable from first experiments.
>
> The superiority antecedently enjoyed by nations, who have preoccupied and perfected a branch of industry, continues a more formidable obstacle, than either of those, which have been mentioned, to the introduction of the same branch into a country in which it did not before exist. To maintain between the recent establishment of one country and the long matured establishments of another country, a competition upon equal terms, both as to quality and price, is in most cases impracticable. The disparity, in the one or in the other, or in

> both, must necessarily be so considerable as to forbid a successful rivalship, without the extraordinary aid and protection of government.

Because of the protectionist policy of the US government, US manufacturing industries were able to become established. After US industries became highly competitive, government policy gradually shifted to free trade to access larger market for domestic products.

Why does the establishment of a new industry often need the support from the government? Often substantial fixed cost investment for a new industry is required before it becomes viable. From Figure 3.1 and 3.2, high fixed cost investment is highly sensitive to market uncertainty and require large market size to breakeven. Government support, by lowering uncertainty and guaranteeing market share, increases the chance of success for a new industry.

The theory developed in this book offers a unified understanding of life processes. This universality helps relate trade policies to problems in other disciplines. For example, a common argument against infant industry policy is that once government provides support to a new industry, it is difficult to decide when this industry should become independent. It is also difficult for parents to decide when their children should become independent. But should we expose infants to "market competition" immediately after they are born because of this difficulty? The fact that we can not time the "optimal" moment to let children become independent does not mean infants should not be supported at all. For the same reason, the difficulties involved in supporting new industries should not discredit infant industry policies completely.

All superpowers throughout time advocate free trade, while poor countries prefer protectionist policies. So it is easy to link free trade with prosperity, and protection with poverty. But historically, all latecomers that turned into industrial powers, including England, the USA, Japan, and Germany, adopted protectionist policies in the period of takeoff (Bairoch, 1993).

Why is domestic competition generally encouraged while international competition needs subtler policy? The fundamental difference is labor mobility. Labor mobility is often considered the most important factor in economic efficiency. There are always winners and losers in market competition. Firms which lose in competition often have to close down. Workers who work in these firms often develop specialized skills that are highly valued in particular contexts. When labors are free to move, which is the case in domestic markets, these workers will move to more successful companies or regions with more job opportunities. This mobility tends to equalize the real wages across regions. In international labor markets where migration is strictly controlled, these skilled workers often don't have the opportunity to utilize their skills and become low income unskilled labors. As a result, some countries with more competitive industries gradually deindustrialize other countries that do not, or cannot adopt protective trade policies (Krugman and Venerables 1995). Historical data show that the gap in living standard between third world countries and developed countries has been increasing over time (Moses and Letnes, forthcoming).

Trading is often explained by comparative advantages, which are often classified as labor intensive and capital intensive. Some developed countries are densely populated. In general, all metropolitan areas in developed countries are densely populated. Yet no one calls these areas labor intensive. So labor intensive is really a euphemism of low wage. Since capital is highly mobile, there is no real comparative advantage being "capital intensive". Capital intensive really means high wage. While low wage is an advantage to employers, it is never an advantage, comparative or absolute, to labors, who are always the majority in any country. In general, the countries that unconditionally adopt free trade policies all sink into poverty over time (Bairoch 1993). Mexico under NAFTA offers a new example. "There has been a steady erosion in the purchasing power of both minimum and average wages in the 1990s. … The dramatic fall in real wages explains why labor income as a percentage of GDP fell from levels over 40% in the early 1980s to 30.9% in 1994, and a mere 18.7% in 2000." (Ramirez 2003, 881)

The fundamental problem with the theory of comparative advantage is that it treats labor income as static. In reality, labor income in a country is greatly affected by the number and type of firms that are located in that country. So comparative advantages change over time and with policy.

Ricardo used an example of trade between two countries to illustrate the idea of comparative advantage. Most of the later literature essentially follows the same argument. In the following, we will use a similar example that incorporates a more precise understanding of product value and production process.

In the real world, a small number of countries specialize in high tech industries, which are of high R&D costs and low marginal costs. Most other countries specialize in low tech industries, which are mature industries with low R&D costs and high marginal costs. To model this pattern, we assume there are twelve countries of equal size in the world. There are no intrinsic differences among them. Two countries specialize in high fixed cost, low variable cost industry while the other ten countries specialize in low fixed cost, high variable cost industry. Assume, for the first industry, the fixed cost and variable cost is 5 and 50% of the product value respectively and for the second industry the fixed cost and variable cost is 0.5 and 80% of the product value respectively. One can think of the first product as the high tech product, while the second one as low tech product. Assume the market size is 1000 and market saturation for both products is 80%. From the value theory, the unit value of the first product is

$$-\log_2 0.8 = 0.32$$

while the unit value of the second one is

$$-\log_{10} 0.8 = 0.09$$

Suppose the outputs of each product are equally shared by producing countries, then the output value of the first product for each of the producing countries is

$$\frac{1}{2}(1000 * 0.8 * (-\log_2 0.8)) = 128.8$$

The return on the first product to each of the two producing countries is therefore

$$\ln(\frac{128.8}{5 + 0.5 * 128.8}) = 0.61$$

With similar calculations, the return on the second product to each of the ten producing countries is 15%, which is much lower than 61% of the first product.

A highly profitable industry from one country inevitably attracts competition from many other countries. Since the production of the first product requires a high fixed cost, it is difficult for countries specializing in low tech products to gain a foothold in high tech industries. However, if one country manages to do so, with the help of government policy or for other reasons, the following is the new rate of return for producing two products.

Since the first product now has three producers, the unit value of the product is now

$$-\log_3 0.8 = 0.20$$

while the unit value of the second product, which now has nine producers, becomes

$$-\log_9 0.8 = 0.10$$

Using the same methodology, it can be calculated that the return on the first product decreases to 52% while the return on the second product increases to 16%.

These results are easy to understand intuitively. The calculations show that the country that is able to move into the production of high tech industry earns a higher rate of return. Overall, the rate of return on the high tech industry is lowered because new entry increase competition and reduce product value, while the rate of return on the low tech industry was increased because of the reduction of competition.

The above analysis shows that later comers, when possible, will move to industries with high economic value and will not stick to their comparative "advantage" of low wage.

How would early comers respond when an industry becomes less competitive than in other countries? Standard trade theory suggests that the gain from trade is always bigger than the loss. So there is no need for the country to adopt an active trade policy. To think about this problem more clearly, we perform a thought experiment. Suppose Detroit is a country specialized in the auto industry. Suppose, over time, Japanese auto makers make cars more cheaply than auto makers in Detroit. If Detroit's government did not intervene in the car industry, its auto companies, with little time to adjust, would collapse. All the physical assets and human assets built over many decades would be lost permanently. Most people associated with the car industry, who are highly paid employees, will become unskilled people. In fact, most of them probably could not even be employed as unskilled workers because they have been used to work only in "decent" environments. While cars in Detroit may be cheaper, consumers, as unemployed or unskilled workers, cannot afford these cheaper cars. So Detroit as a country becomes worse off.

In reality, Detroit is part of the US. Many people moved away from Detroit to places with more opportunities. The US government imposed an import quota on Japanese cars to reduce the impact. Even so, anyone who visited Detroit in early 1990's would agree that the import of Japanese cars had a negative impact on the overall welfare of Detroit.

This understanding is of course common sense. If the sole bread winner in a family loses a highly paid job, which requires long and specialized training, because imported goods are cheaper, the family will be worse off even though some particular goods can be purchased at lower prices. A country is a bigger version of a family. While common sense tells that a country may or may not benefit from a particular international trade, common sense is not highly valued in mainstream economics, as Samuelson made clear in the following statement:

> Good sense economics is not all obvious. The common sense you bring with you from your home to college will not let you understand why a rich country and a poor country can both gain great benefit from free international trade at the same time. (p. xxvii, Samuelson and Nordhaus, 1998)

Empirical evidences and our theory show that common sense is right after all.

Whether a country will benefit from a particular trade will depend on many factors. If an industry is of low fixed cost and low marginal profit, government in general intervenes less because of the low market value. If an industry is of high fixed cost, low marginal cost and of large market size, government will often actively intervene because huge value is at stake. So, while high fixed cost developed countries support free trade in principle, they will actively intervene to protect key industries. They will also not leave commodities of fundamental importance to the mercy of the market.

> Polayni is certainly correct in emphasizing that Athens could not afford a free international market in grain when its survival depended upon the uninterrupted flow of that import. The strategic importance of grain imports to Athens (and later Rome) is paralleled in the late twentieth century by the strategic importance of oil to the importing countries; the parallel in political economy are quite striking (but no one would argue that the motivation in securing oil imports was noneconomic). (North, 1981, p. 106)

In general, because of the low transaction cost of markets, most products are allocated with free trade. Terms of trade of economic products of vital importance, however, are often resolved at institutional levels by diplomatic or military means, as the high cost of diplomatic or military actions are compensated with even higher benefits in influencing the terms of trade. For example, the high cost of military actions in the Middle East is compensated by the low oil price to major oil consuming countries.

In the following, we will discuss some specific trade policies, such as quota systems and tariffs. Extensive literature surrounds the issue of cost and benefit of import quotas and tariffs (Romer, 1994). From the value theory developed in Chapter 2, the value of a product is inversely related to the number of its producers. If a product is mainly made in one country, the value of the product, or the cost to consumers, is high. Quota systems force the producers that are subject to quota constraints to move production facilities to other countries. The transfer of production technology from leading export countries to other countries will increase the number of producers and intensify competition. From Formula (2.1), the increase of the base will reduce the value of the imported goods. This will result in a net welfare gain to the importing countries over the long term, instead of the welfare loss suggested in most literature.

However, the quota system, by restricting the scale of production and reducing the value of the products does increase the cost and reduce the profit to the original leading exporters. By limiting the size of the market, it will also discourage exporters from adopting higher fixed cost production systems and expensive innovation.

Next we will discuss an example from the natural resource industry to illustrate the function of tariff policy. In natural resource industries, production output is often regulated to maintain sustainability. From the value theory, product value is a function of scarcity. Tariff policy can often significantly influence output quantity and hence product value, especially when a certain commodity has one big producer and one big consumer. For example, Canada is a big producer of softwood lumber while the USA is a big consumer. From Formula (2.6), the value of the lumber market is represented by –VP*ln*P, where *P* is the proportion of lumber that is on the market and *V* is the total volume of all lumber. For a consumer country, it will benefit from a trade policy that increases the production of lumber since it will reduce the value of imported lumber.

Suppose the cost structure of the lumber industry is the following. The total fixed cost of lumber production in country *C* is 100. The marginal cost is 60% of product value. So the total value of the lumber products is –VP*ln*P and the total cost of production is 100+0.6*V*P*(-lnP). Suppose every year, 1.5% of the all lumber is harvested and *V* is 10000. The profit on lumber production is

$$\begin{aligned}
&-VP\ln P-(100+0.6*(-VP\ln P))\\
&=-10000*0.015*\ln(0.015)\\
&\quad-(100+0.6*(-10000*0.015*\ln(0.015)))\\
&=152
\end{aligned}$$

In 2001, the USA imposes a 27% import duty on softwood lumber from Canada. If the volume of production remains at the same level, the profit for lumber production would be

$$\begin{aligned}
&-VP\ln P*(1-0.27)-(100+0.6*(-VP\ln P))\\
&=-10000*0.015*\ln(0.015)*(1-0.27)\\
&\quad-(100+0.6*(-10000*0.015*\ln(0.015)))\\
&=-18
\end{aligned}$$

which means that the lumber industry will lose money. Production of lumber has to be increased to avoid loss. If the production level is increased to P = 2%, the profit for the lumber industry will becomes

$$
\begin{aligned}
&-VP\ln P*(1-0.27)-(100+0.6*(-VP\ln P))\\
&=-10000*0.02*\ln(0.02)*(1-0.27)\\
&\quad-(100+0.6*(-10000*0.02*\ln(0.02)))\\
&=1.71
\end{aligned}
$$

As the production is increased from 1.5% of the total reserve to 2%, the unit value of lumber is decreased from –ln(0.015)= 4.2 to –ln(0.02) = 3.9. So the USA collects 27% tariff on lumber import and enjoy lower price on lumber. Table 5.1 is a summary statistics of softwood lumber futures price, annual production from Canada, revenues and profits from Canfor, Canada's largest softwood producer, in 2000 and 2002, one year before and after USA imposed the 27% tariff on softwood lumber import from Canada.

The empirical data confirm the theoretical predictions that after the tariff, production increased, prices dropped, and corporate profits from lumber producers tumbled. This shows that a tariff is an effective way to shift wealth from producing countries to consuming countries and contradicts the standard theory that tariffs hurt importing countries with higher prices for consumers.

	2000	2002
Softwood lumber futures price (January closing)	346.6	268.7
Production (thousands of cubic meters)	68557	71989
Canfor revenue (millions of dollars)	2265.9	2112.3
Canfor profit (millions of dollars)	125.6	11.5

Table 5.1: summary statistics of softwood lumber futures price, annual production from Canada, revenues and profits from Canfor. Sources of data: CME, indexmundi, Canfor annual reports

From the above discussion, we find that trade policies, such as quota and tariff, are effective measures to improve terms of trade. That these policies are not universally implemented is due to the threat of retaliation. The dynamics of trade policies reflects the differences and changes in economic and political power of the trading parties. If free trade really was an optimal policy, then competition over the last several thousands years would have ensured that the whole world would evolve into a big free trade zone. However, this has not happened.

Trade leads to specialization. If some specialization in a region turns out to be very successful, such as the automobile industry in Detroit and the IT industry in Silicon Valley, the high land and labor cost in that region will make it difficult for other industries to survive. Therefore, these successful regions will become highly specialized. Can these successful stories be sustained over a long time? The history of Detroit and Pittsburgh suggests that the answer is no. Cope (1896) summarized the pattern of biological evolution more than a hundred years ago.

> The 'Doctrine of the Unspecialized' ... describes the fact that the highly developed, or specialized types of one geological period have not been the parents of the types of succeeding periods, but that the descent has been derived from the less specialized of preceding ages. ... The validity of this law is due to the fact that the specialized types of all periods have been generally incapable of adaptation to the changed conditions which characterized the advent of new periods. ... Such changes have been often especially severe in their effects on species of large size, which required food in great quantities. ... Animals of omnivorous food-habits would survive where those which required special foods would die. Species of small size would survive a scarcity of food, while large ones would perish. … An extreme specialization … has been, like an overperfection of structure, unfavorable to survival. In general, then, it has been the 'golden mean' of character which has presented the most favorable condition of survival, in the long run. (Cope, 1896, p. 173-174)

Economists tend to link specialization with progress and efficiency. However, biologists have come a long way since the time of Darwin:

> If one of two groups "is in any way more primitive than the other, then its primitiveness must in itself be an adaptation to some less specialized mode of life which it can pursue successfully; it cannot be a sign of inefficiency" (Cain, 1964, p. 57, quoted from Dawkins, 1999, p. 31).

5.2. Education and Fertility

5.2.1. Theory of education and fertility

Because of migration control, technological advantage in developed countries is translated into high labor income. At the same time, high labor costs make it difficult for developed countries to compete with low wage countries in mature industries. Because of knowledge diffusion, developed countries have to keep innovating to maintain their lead, which requires long education for most people at tremendous cost. Education in most countries, especially in developed countries, is heavily subsidized. Why is it that in a market economy, the most important activity is essentially sponsored by governments? In mainstream economics, education is said to generate "positive externality" or "endogenous growth". In fact, anything government subsidizes is supposed to generate "positive externality". This is to justify a policy, not to understand it. Our analytical theory provides a much simpler and consistent approach.

Education is largely a fixed cost. To the majority of people, it is also their largest investment in time and money. From Figure 3.1, high fixed cost investments are very sensitive to uncertainty. The government subsidy, by reducing the level of uncertainty in funding, enhances the effect of education.

While state subsidized education benefits many economic sectors, it affects different sectors differently. In particular, education offers systematic subsidization for new or infant industries by training for free a large quantity of potential workers and users of new technologies.

Figure 3.2 shows the rate of return as a function of output for systems with different levels of fixed costs. For a higher fixed cost system, it takes a higher output before the return becomes positive. The return curve for a higher fixed cost system is also steeper. This means that a high level of investment in education, which is largely a fixed cost, is riskier and when the investment is successful, the return is higher. To ensure the investment to be successful, more resources are put on each child. That is why the regions where a high level of education is required for good jobs usually have lower fertility rates. The long time for education and training forces many potential parents to delay reproduction and reduce their number of children they have. As competition intensifies, the fertility rates in most developed countries have dropped far below the replacement rate. This raises big concerns about the sustainability of the social system in developed countries.

Peripheral regions, on the other hand, can't compete with the developed countries except in mature industries. The opening up of trade creates a consistent labor surplus in developing countries, which depress wages. Because of the difficulties in learning the latest knowledge for most people in developing countries, the return on education is low. As a result, most people adopt a low fixed cost strategy and choose not to spend lengthy years in school. From Figure 3.2, it takes less effort to make a positive return in a low fixed cost environment. Empirical investigation showed that it is of great benefit for a poor family to have more children because each child offers new potential for the future of the family (Bledsoe 1994). Although low income families are often advised to reduce their fertility rates, "neither evidence nor analysis has yet disproved the notion that the poor in poor countries know, at least in a rough manner, what is in their best interest" (Dasgupta 1995, p. 1899). The increased population further worsens the labor surplus problem and depresses labor income, this makes the idea that "growth is the best contraceptive" impossible to realize. Poverty and high population growth feeds on each other to form a positive feedback loop. Like population in wealthy countries, population in poor countries will eventually be checked by the constraint of natural resources. However, since the resource requirement of a poor person is much lower than of a rich person, the check comes later. This is the same in ecological systems. Colinvaux wrote an ecological book, which was aptly titled *Why Big Fierce Animals Are Rare*. That the fertility rates in resource intensive developed countries have dropped far below replacement rate is a clear sign that human beings, like any other organisms, are constrained by Malthusian forces.

The fertility pattern of rich and poor is not a new one, as we can find from Darwin's words:

> A most important obstacle in civilized countries to an increase in the number of men of a superior class has been strongly urged by Mr. Greg and Mr. Galton, namely, the fact that the very poor and reckless, who are often degraded by vice, almost invariably marry early, whilst the careful and frugal, who are generally otherwise virtuous, marry late in life, so that they may be able to support themselves and their

> children in comfort. Those who marry early produce with a given period not only a greater number of generations, but, as shewn by Dr. Duncan, they produce many more children. The children, moreover, that are born by mothers during the prime of life are heavier and larger, and therefore probably more vigorous, than those born at other periods. Thus the reckless, degraded, and often vicious members of society, tend to increase at a quicker rate than the provident and generally virtuous members. Or as Mr. Greg puts the case: "The careless, squalid, unaspiring Irishman multiplies like rabbits: the frugal, foreseeing, self-respecting, ambitious Scot, stern in his morality, spiritual in his faith, sagacious and disciplined in his intelligence, passed his best years in struggle and in celibacy, marries late, and leaves few behind him. Given a land originally peopled by a thousand Saxons and a thousand Celts --- and in a dozen generations five-sixths of the population would be Celts, but five-sixths of the property, of the power, of the intellect, would belong to the one-sixth of Saxons that remained. In the eternal 'struggle for existence,' it would be the inferior and less favoured race that had prevailed --- and prevailed by virtue not of its good qualities but of its faults." (Darwin, 1981(1871), p. 174)

Statistically, the tradeoff between intelligence and fertility or education and fertility is clear. It shows that, intelligence and education, like any other perceived good quality, has cost. This tradeoff is very clear not only between wealthy and poor countries, but also among different social groups in the same country. For example, white, black and Hispanics are the three largest ethnic groups in the United States. Whites have the highest level of education and lowest level of fertility while Hispanics have the lowest level of education and highest fertilities. Blacks, who are on average better educated than Hispanics, often found themselves less competitive than Hispanics in the job market. If we treat education always as a positive asset, this phenomenon is hard to understand. But if we regard education as a fixed cost, then it is easy to understand that the level of education should match the opportunity in the job market, not the other way around. This is why over education may not pay, just like a high fixed cost project will not pay if the market size if not big enough.

Human fertility rate is a reflection of efficiency of energy consumption in social life. Statistics show that low fixed cost cold blooded carnivores are more efficient in converting zoomass than high fixed cost warm blooded carnivores.

> Carnivorous ectotherms convert as much as 30-35 percent of ingested zoomass into their body tissues, but the rate for their endothermic counterparts is only 2-3 percent. Low efficiency of energy transfer is a price mammals and birds pay for their high rates of metabolism and for their highly mobile way of life. (Smil, 1999, p. 65)

The high energy cost of modern life is the fundamental reason why fertility rates in developed countries are low. Fertility rates in all developed countries, with the exception of United States, which, as the only superpower in the world, possesses unparallel natural and institutional resources, are far below the replacement rate. This indicates that the current living standard in developed countries is much higher than resource can support. However, it is politically unpopular to call for the reduction of consumption.

5.2.2. A comparison with the mainstream theory

The mainstream economists' view is represented by Becker's works. To explain the fertility reduction with the rising living standard, Becker wrote:

> A reduction in the number of children born to a couple can *increase* the representation of their children in the next generation if this enables the couple to invest sufficiently more in the education, training, and "attractiveness" of each child to increase markedly their probability of survival to reproductive ages and the reproduction of each survivor. (Becker, 1993, p. 137)

This analysis is highly consistent with the maximization principle of the general equilibrium theory. However, it is not consistent with the empirical regularity that those who get more investment in education, training, and "attractiveness" actually reduce the rates of reproduction. This regularity has been observed for more than a hundred years (Spengler, 1938, p. 156 – 174).

The main difference between the analysis of our theory and Becker is that in Becker's theory, resource allocation is chiefly decided by parents themselves while in our theory a large part of resource allocation is determined by social environment as the fixed cost. For example, after cars become abundant, residential areas and business areas are separated by zoning codes. This makes owning cars a necessity in life, whether or not you want to own a car. Mandatory education in developed countries, which put a high fixed cost on parents and society, is determined by social environment and not by parental choices.

In biological population theory, fitness is defined purely quantitatively. Current theories on human fertility rates introduce the concept of quality to explain away the low fertility rates in developed countries as a substitution of quality for quantity. In essence, these theories indicate that the mechanisms of human reproductive patterns differ from that of other living organisms. In our theory, fertility patterns of human beings are determined by the level of fixed cost and abundance of natural resources, just like any other plants or animals. Both fixed cost and abundance of natural resources, unlike quality, are directly measurable. This theory on human fertility is a natural extension from biological analysis of r and K reproduction strategies. (Colinvaux, 1986; Rushton, 1996) It offers a simple and unified model of empirical fertility patterns in human societies throughout history, which is viewed "as one of the most significant research challenges facing economists interested in growth and development (Galor and Weil, 2000, p. 806)".

5.3. Long Term Consequences of Migration Restriction

Many policies with short term benefit entail long term cost. For the developed countries, restriction of migration may have the following negative long term consequences.

First, with declining shares of population, the market power of developed countries will decline over time. Many attempts have been made to encourage higher fertility rates in wealthy countries and lower fertility rates in poor countries. But the results are not promising. This is, in fact, not a new problem.

> It suggested that selective marriages among the talented would be a possible way to 'eugenics' and the voluntary development of a caste of the naturally gifted, among whom intermarriage would be socially

> encouraged, and exogamy socially discouraged. Positive eugenics soon fell into doldrums, however. It was not that intermarriage among the gifted was unpopular. On the contrary, it was a flourishing practice, not least among the great literacy and scientific families of the last half of the nineteenth century, including the Darwins and the Keyneses. Rather, one problem was that these marriages had continuously falling fertility rates; and the other problem was that there was nothing much which could be done to increase intermarriage among the gifted or to raise the fertility of such unions. (Toye, 2000, p. 117)

Second, to maintain a high level of living standard, there is permanent pressure to innovate in developed countries. Innovation is usually termed as something positive in the current social environment. However, for biological species, "liability to extinction tends to be directly proportional to rate of evolution (Simpson, 1944, p. 143)". Intuitively, this is easy to understand. The constant need for innovation is due to the constant threat of loss of our niche in the world. In fact, the low fertility rate is a direct reflection of our difficult situation.

Third, a fertility that is below replacement level inevitably causes the aging of population. In a democratic system, the aging of population will increases the political power of senior citizens, which will help shift resources from young parents and children to senior citizens. The reduction of resources for children will further reduce fertility. This positive feedback poses a serious challenge to the viability of democratic systems.

In many developed countries, health care cost is the main government spending. However, politicians, with voters in mind, often proclaim that the spending on health care is not enough. The high cost of health care puts a heavy burden on society, making average families unable to support two children. For example, in the province of British Columbia, where I live, the expenditure on health care as the percentage of total government revenue has increased from 28% in 1989 to 40% in 2004. (Figure 5.1)

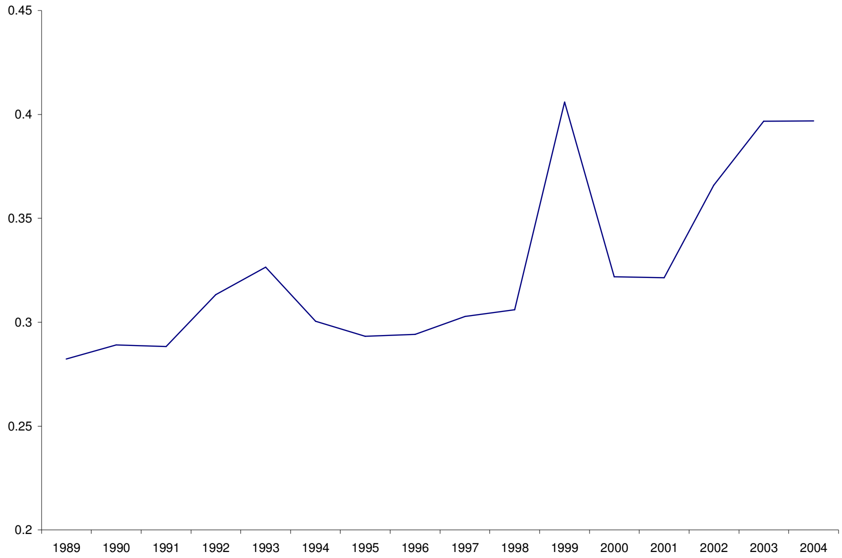

Figure 5.1 The expenditure on health care as the percentage of total government revenue in the province of British Columbia. Source: CANSIM

In the last several decades, GDP per capita actually increases as population ages. Why does the aging of the population not negatively affect economic performance?

> The answer to this puzzle is surprising simple: Workers support not only retirees but also children … So, although the proportion of retirees in the population has steadily grown over the past 40 years, the proportion of the population in their working years has also steadily risen because the number of children per working-age person fell more than twice as fast as the number of retirees was rising. While the elderly were becoming an increasing portion of the population, the boomers left childhood and the percentage of children fell, so in the aggregate, the demand from dependents of all categories for the goods and services workers produced actually fell, even as the number of retirees rose. (Arnott and Casscells, 2003, p. 23)

The above passage indicates that economic growth in the last several decades is a transition state. Eventually, as proportion of working population declines, the economic performance will decline as well. Countries with high living standard and low resource reserve, such as Japan, will experience earlier reversal. Indeed, the Nikkei 225 index, the major stock market index of Japan, a country with a low fertility rate and minimal immigration, has dropped from its peak near 40,000 more than a decade ago to less than 12,000 at the time of writing.

Fourth, because of the great difference in living standard, wealthy and poor countries adopt different and often conflicting strategies in domestic and international policies (West, 2005). High fixed cost wealthy countries require low uncertainty in the environment and large market size. Low fixed cost poor countries, on the other hand, thrive in a volatile environment and do not depend on large market size. In the early days when Western powers had unparallel military advantages, gunboats were routinely used to open foreign markets. As people in the rest of the world begin to master advanced technologies, military actions become both ways to achieve their respective objectives. However, high fixed cost wealthy countries are much more sensitive to the increase of uncertainty. For example, Afghanistan, one of the poorest countries in the world, has been subject to heavy bombing for decades. But it functions pretty much the same. The United States, the wealthiest country in the world, was hit only once by some hijacked airplanes. It was estimated that the cost for hijacking was only about half a million dollars and the loss due to that incident was about half a trillion dollars. What is worse, it permanently increases the level of uncertainty in many business activities in the United States. As long as the living standard of North and South diverge, the strategies they adopt will diverge as well.

While the current policy of migration restriction has some serious long term consequences, there is no easy solution. As the difference in living standard and difference of population size between North and South increase over time, it will be even more difficult to resolve the problem in the future.

5.4 Concluding Remarks

Migration policies, or more generally, policies about inclusion and exclusion, are determined by two conflicting factors. On the one hand, more exclusive policies can retain more resources for the dominant groups. On the other hand, more inclusive policies can gain support from larger population, which makes the system more stable. When resources are plentiful, societies in general are more inclusive. For example, liberty flourishes in conquering states, from Athens, to Rome, to the United States (Colinvaux, 1980). As resources become scarcer, political systems will turn more repressive.

The restriction on migration, complemented by free trade, affects wealthy and poor countries in different ways. To maintain high living standard, wealthy countries adopt policies to reduce variable costs in production by increasing fixed cost in social investment, which ultimately push the fertility rates below the replacement rate. Poor countries, because of their inability to compete in high fixed cost industries, remain in a low fixed cost economic structure. The low resource intensity way of life enables their fertility rates to stay higher than the replacement rate, for the moment.

The divergence of living standard and population size between rich and poor countries will make the world increasingly unstable. Gibbon once attributed the decline of Roman Empire to barbarians and the rise of Christianity. Barbarians and Christians are, according to Toynbee, external and internal proletariats. As the number of external and internal proletariats keeps increasing in today's world, what is our response?

Chapter 6

From Modern Astronomy to Modern Finance: A New Theory of Finance

6.1. What Finance and Astronomy Have in Common?

Some theories, originated from a very technical area, because of their deep insight and analytical power, become the foundation of much broader fields. Modern astronomy, founded by Copernicus, Kepler, Newton and others, was originally developed to understand the movements of several planets. Rational mechanics, the physical foundation of modern astronomy, later became the foundation of all modern science. Starting from 1870, Jevons, Walras and others applied the principle of rational mechanics to establish neoclassical economic theory. Today the philosophy and techniques originated from modern astronomy become the foundation of both natural science and social science.

In this chapter, we will examine the parallels between finance and astronomy to understand why the progress in financial research will revolutionize the whole foundation of social science. In the process, we will build up a new theory of finance.

We will discuss several common properties of finance and astronomy that make them the pioneer subjects that trigger much deeper changes in the foundation of sciences.

First, both the study of astronomy and finance are heavily data driven. Astronomy is the oldest precise science. Data of celestial observation have been accumulated over thousand of years. As observation became more accurate over time, it became easier to test alternative theories.

Financial data are the most frequently and abundantly recorded data set. For each stock, each transaction price, bid-ask price, the size of the trade, and many other information are recorded on computerized systems. The abundance of real time financial data makes financial theories much easier to test than economic theories, since economic data are less frequent, less reliable and often subject to different interpretation.

The second is the simplicity of astronomy and finance. Astronomy, which Wiener (1948) termed as "an ideally simple science", studied the orbits of isolated planets with little disturbance from other sources. Finance studies cash flows under uncertainty or price innovation of market securities, abstracted from all the intrigues of social and organizational complexities. Although many complicated forces are at work in the financial market, the low transaction costs determines that prices alone already reflect most of the interaction of these forces.

The simplicity of astronomy and finance makes the alternative theories easy to test. In astronomy, two alternative theories at the time were the earth centered universe and the sun centered universe. As the data became very accurate and sophisticated mathematical tools were developed, the alternative theories became easy to test. In finance, the default theories are the irrelevance of financial structure in corporate finance and efficient market theory in investment. Both theories are empirically testable. This is in sharp contrast to the general economic theory, where utility function can be defined in many different ways. It is difficult to test whether economic agents maximize "utility" because it cannot be precisely defined (Mirowski, 1989).

Third, it is usually in simple subjects where sophisticated mathematical theories are developed and applied with great effectiveness. Calculus was invented by Newton to calculate the planet orbits. Before the invention of calculus, the calculation of nonlinear curves was almost impossible. After calculus was invented, the same calculation can be performed by any competent person with proper training. In social sciences, the complex relationship often prevents the development of sophisticated mathematical tools that are relevant and effective. In finance, however, the simplicity of research subjects enables people to develop mathematical models with deep analytical power and far reaching consequences. While calculus was first invented to calculate nonlinear planet orbits, stochastic calculus was successfully applied to solve the nonlinear payoff problems in options by Black and Scholes (1973).

Fourth, both astronomy and finance are of immense practical value. A new theory doesn't grow from vacuum. Before a new theory germinates, there always exists an established paradigm in any area of research. Before Copernican theory, theology was the foundation of the cosmology. Currently, general equilibrium theory is the foundation of economics and finance. Given the dominant status of theology then and the neoclassical economics today, why Copernican theory in astronomy and new theories in finance, such as behavioral finance, got established. This is because both astronomy in Copernicus' day and finance today have tremendous practical values. People tend to neglect the ideological differences on issues of practical values. For example, in general economic theory, dissident opinions rarely surface in major economic journals, although the problems of neoclassical economics are apparent to many people. In finance, however, papers on alternative theories, such as behavioral finance, have already occupied top tier finance journals for a long time.

Fifth, both astronomy and finance were pure mathematical theories initially but gradually turned to physics. Astronomy was a part of mathematical science in the ancient time. But starting from Kepler, people looked for physical causes to offer a unified understanding of celestial movements. (Voelkel, 2001) Eventually, Newton established the universal gravitational laws to unify the understanding of not only celestial bodies but also objects on earth. Financial theories are predominantly mathematical today. However it has been shown recently that the analytical thermodynamic theory of social sciences can be developed based on the techniques similar to those of Black-Scholes option pricing theory. (Chen, 2000, 2002b)

The simplicity of research subjects, the objectivity of research methodologies and powerful mathematics enabled the researchers in astronomy to break into the dominant paradigms of the time: the earth centered universe. In the process, a physical theory, rational mechanics, was developed to understand much broader phenomena. Since then, rational mechanics has become the dominant paradigm in natural science and social science. These same qualities in financial research help us establish a new paradigm: the analytical thermodynamic theory in social science and life science.

In Section 6.2, we will describe the trajectory of thinking from modern astronomy to modern finance and how the development of modern finance eventually offers a fertile ground for a new paradigm in social sciences. In Section 6.3 and 6.4, we will develop new theories of corporate finance and investment.

6.2. From Modern Astronomy to Modern Finance: A Brief Review

As inhabitants of the earth, we naturally think that the earth is the center of the universe. Since ancient time, people noticed that several planets, Mercury, Venus, Mars, Jupiter and Saturn, move in irregular ways while all stars move in perfect circles. It was thought that these planets were imperfect. To describe the orbits of these planets, many epicycles were needed. What is worse, as the observation became increasingly accurate, the models that were used to describe the movements of the planets had to became increasingly complex as well. This theoretical difficulty made it possible for new theories to emerge (Kuhn, 1996). Copernicus proposed a new theory of astronomy, in which the sun was the center of the universe and the planets, including the earth, move around the sun. In a sun centered universe, the movements of planets became much simpler. But more importantly, the new theory made it possible to attribute the movements of planets from a universal physical force. At that time, people already knew that the sun was much larger than the planets. This sun centered theory inspired Kepler to conjecture that all planets were attracted by some force from the sun. This idea turned astronomy from a purely mathematical theory into a physical theory. Since all planets were driven by a unified force, their movements must follow some simple laws, Kepler conjectured. Eventually Kepler was able to show that the movements of all planets follow three simple laws: the Kepler's Laws. (Voelkel, 2001) Newton further showed that the movements of planets can be derived from gravitational law and his three Laws of Motion. So the "imperfectness" of planets comes from our understanding of planets but not from planets themselves.

Furthermore, "Newton presented his three Laws of Motion, which applied celestial and terrestrial behavior alike. Here was another revolution, for it sealed the doom of the long-held belief that those two realms were different." (Schorn, 1998, p. 34) The success of rational mechanics to offer a unified theory on celestial and terrestrial behavior encourages people to apply it to broader fields, including social sciences.

The pioneers of neoclassical economics, such as Jevons and Walras, attempted to transform economics into a rigorous science based on rational mechanics. From there, rational expectation theory and optimization theory was developed. In neoclassical economics, rational individuals try to maximize their utility. The interaction of all individuals ends up in a general equilibrium state. A major problem of this theory is that utility is usually difficult to measure or even define. (Mirowski, 1989)

Finance is a part of economics. There are two major branches in finance: corporate finance and investment theory. In corporate finance, the theoretical foundation is Modigliani and Miller theorem, which states that financial structure of a firm is irrelevant to its value in a perfect market. In investment theory, the main theory is the efficient market theory. Unlike utility maximization theory in economics, both capital structure irrelevance theory and efficient market theory are empirically testable and are the subjects of investigation in many research works.

Empirical evidences demonstrate that financial structures are relevant for corporations. This relevance is often attributed to market imperfection. If an imperfection is identified, this type of imperfection would be gradually reduced over time from competition or regulation. So we might expect capital structures of firms will be less and less relevant and financial decision making becomes simpler and simpler over time. Fifty years ago when Modigliani and Miller first published their paper, theories and practices in finance were relatively simple. Since then, problems in corporate finance, instead of getting simpler and simpler, have become more and more complicated. In the process, many complex financial instruments, such as convertible bond, which is a hybrid of debt and equity, have been created in the financial markets. Number of finance professionals also increase tremendously in the last fifty years. All these indicate that the relevance of financial structure is not a matter of market imperfection but of more fundamental reasons.

With few exceptions, tests on efficient market theory are joint tests on asset pricing models. (Chen, 1999, 2003) The standard asset pricing model is Capital Asset Pricing Model (CAPM). The beauty of CAPM is that it offers a conceptually simple and testable relation between risk and return. As empirical results accumulate, it becomes clear that the relation predicted by CAPM doesn't hold in the asset markets. Many new models, often containing three or four factors, are developed to fit data better. However, these increasingly complex models have lost the simple property of risk return tradeoff in the original CAPM theory. This is very much like astronomy in Copernicus' day. The original Ptolemy system was a very elegant model. The earth is near the center of the universe.

All celestial bodies move around the earth in circles. As the astronomical observation became more accurate, it became clear that planets didn't move around the earth in circles. Astronomers developed more complex models of circles over circles to fit the observed data better. While these new models fit the data very well, they lost the beauty and simplicity of the original Ptolemy model. What is worse, as the observation became progressively more accurate, the models had to became progressively more complicated to fit the data. This pattern seems to repeat itself in today's financial researches. As data accumulate, asset pricing models become progressively more and more "sophisticated".

As problems mount, we need to reexamine the very foundation of corporate finance theory and investment theory.

Miller states, "The M&M propositions are the finance equivalents of conservation laws" (Miller, 1991, p. 483). From the theory of conservation laws, entropy condition has to be introduced to obtain realistic solutions (Smoller, 1994). So the concept of entropy emerges naturally from conservation laws.

Efficient market theory states that market can process information instantly. From Maxwell (1871) and Shannon (1948), information is the reduction of entropy. So a study of market behavior needs a better study of information as entropy.

In both corporate finance theory and investment theory, the concept of entropy emerges naturally from discussion. While theories based rational mechanics do not offer clear understandings about problems in finance, we demonstrate in the next two sections that theories based on thermodynamics do offer simple and consistent understanding of financial problems.

6.3. A New Theory of Corporate Finance

Since Modigliani and Miller (1958) presented the result on capital structure irrelevance in a perfect market, researchers have searched for imperfections in the capital market and how these imperfections determine the capital structure in real life. Fama and French (2002) summarized the current state of research:

> The finance literature offers two competing models of financing decisions. In the tradeoff model, firms identify their optimal leverage by weighing the costs and benefits of an additional dollar of debt. The

> benefits of debt include, for example, the tax deductibility of interest and the reduction of free cash flow problems. The costs of debt include potential bankruptcy costs and agency conflicts between stockholders and bondholders. At the leverage optimum, the benefit of the last dollar of debt just offsets the cost. …Myers (1984) develops an alternative theory known as the pecking order model of financing decisions. The pecking order arises if the cost of issuing new securities overwhelm other costs and benefits of dividends and debt. The financing costs that producing pecking order behavior include the transaction costs associated with new issues and the costs that arise because of management's superior information about the firm's prospects and the value of its risky securities. Because of these costs, firms finance new investments first with retained earnings, then with safe debt, then with risky debts, and finally, under duress, with equity. As a result, variation in a firm's leverage is driven not by the tradeoff model's costs and benefits of debts, but rather by the firm's net cash flows (cash earnings minus investment outlays). (Fama and French, 2002, p. 2)

While some empirical patterns can be explained by tradeoff and pecking order theories, many questions remain unanswered. For example, Miller once commented:

> The standard finance explanation that we were seeing a conscious and well-considered tradeoff between the tax benefits of borrowing and the expected costs of bankruptcy was hard to take seriously. The relative numbers were just too far out of line. Firms in 1976 were throwing at the government $40 billion in corporate taxes, a sum that would cover the true deadweight costs of all bankruptcies for decades. There had to be something else at work. (Miller, 1998, p. 116)

Before we show how our theory can provide parsimonious understanding of the empirical results, we discuss the concept of "imperfection".

If the relevance of capital structure is caused by some kind of market imperfection, it would be reduced gradually. Over time, financial decisions will become less and less important. But finance theory and practice has become more and more important since M&M first propose their theory about fifty years ago. Some start to question whether "imperfections", such as agency costs can account for the observed patterns of capital structure:

> The possibility that managers might let their own interest override that of the shareholders was something that Franco Modigliani and I were certainly aware of back in 1958 and through all our subsequent revisions and extensions; and we knew that anecdotal evidences of non-value maximizing behavior by under-diversified managers would always easy to come by. But we doubted that such nonoptimizing behavior would lead to *systematic* departures from the model. We believed that the stock holders would learn to solve, or at least greatly to mitigate any excessive risk aversion of their managers by appropriate compensation and incentive mechanisms. The stockholders, after all, could always persuade the managers to act more like stockholders by giving the managers stock or stock appreciation rights or stock options of any of the a number of kinds. Given the defenses available to stockholders to recapture value they believe belongs to them --- including defenses such as large-shareholder influence and hostile takeover by outsiders --- it is hard to believe that a sum as large as $150 billion a year would be left lying on the table. (Miller, 1998, p. 118)

M&M theory becomes the standard theory in corporate finance because it is logically consistent. There could be, however, many different logically consistent theories. For example, both earth centered theory and sun centered theory about our universe are logically consistent. But sun centered theory gives a simpler description of the movements of the planets. In earth centered theory, the movements of planets are very imperfect. In the new theory of sun centered universe, the planets move perfectly according to Kepler's laws. Very often, "imperfection" of some empirical patterns leads to the discovery of new and unified theories.

A theory of corporate finance is built on a theory of the firm (Zingales, 2000). After analyzing the shortcomings of current theories of the firm, Zingales stated that a new theory of the firm should answer the following questions:

> How does an organization succeed in acquiring power that differs from ordinary market contracting? How is this power maintained,

> enhanced, or lost? How is this power used in a way that differs from ordinary market contracting? How is the surplus generated by the firm divided among its members? (Zingales, 2000, p. 1651)

We will show how the theories developed in this book will provide a new theory of firm that gives simple and coherent answers to these questions.

An organization succeeds in acquiring power that differs from ordinary market contracting because of some information advantage. From Formula (1.3),

$$R = H(x) - H_y(x)$$

the amount of information one can receive depends on the background knowledge of receiver. Those who understand information better can utilize information better. The production of a good needs fixed assets and variable costs, which is a function of fixed cost and rate of diffusion. A firm's competitive power comes from its internal structure that reduces diffusion in the production process. Formula 1.3 explains the function and stability of firms. In a firm, because of the common knowledge of many details of the business, $H_y(x)$, the equivocation is small and the information flow is more fluent, which reduces transaction cost (Coase, 1937). If a person's talent is of amount $H(x)$, the value that is appreciated by others is $H(x)-H_y(x)$. In another firm that people are less familiar with his knowledge, equivocation is high. A person is usually valued higher in his current position than in a potential new position in other firms. That explains the relative stability of employment and long-term stability of firms (Arrow, 1999). Since the common knowledge and custom of a firm is built over a long time, market cannot replicate its structure instantly. That is how a firm acquires power that different from ordinary market contracting.

The power of a firm can be maintained or enhanced by further reducing diffusion or by increasing fixed cost, both reducing variable costs. The detailed operation should depend on market conditions. For example, if market size increase, the firm should increase fixed costs. (Figure 3.2) Established firms often lose market power in a fast changing environment since the existing fixed assets make it less adaptive in a new environment than small and new firms. (Figure 3.1)

Whether a company decides to develop and produce something in house or contract out by market transactions depends on several factors. First, it depends on how existing facilities are compatible with the new products. If simple modification or extension of existing facilities can help produce the product, the company probably will produce in house. That is, if the existing fixed assets can help reduce variable costs in the production, it will produce in house. Otherwise, it will contract out. Second, it will depend on the value of the product. From the value theory developed in Chapter 2, if many companies can product the product, its value is low. Contracting out will not reduce the company's competitive power and will keep the company flexible in future developments.

A corporation is a collection of assets. Although shareholders are nominally the owners of these assets, they own the assets only to the extent they can effectively control the assets. While incentives may be designed to align management interest with shareholders' interest, the only way the management interest completely align with shareholders is for management to completely own the companies, that is, shareholders relinquish all their claims. Therefore a better alignment is more costly. Inside a company, management cannot monitor everything. They have to delegate responsibility to other employees, which creates information asymmetry. So every employee owns part of the company. So are some large customers and suppliers.

How the surplus generated by the firm divided among its members will depends on the power structure of the firm. If top manager are major shareholders who understand businesses very well, which is typically the case in small and new firms, major portion of surpluses will be retained by the shareholders. If the ownership is very diffuse and top managers have very little share ownership, which is typically the case for large, mature firms, managers and employees retain higher shares of surpluses. This is why pay and benefit in large companies are generally better than in small companies.

We then present the new theory of capital structure, which is a combination of information theory and production theory. Information theory alone can derive the pecking order model, that firms finance new investments first with retained earnings, then with safe debt, then with risky debts, and finally, with equity. From Figure 3.1 and 3.2, fixed cost, or operating leverage, matters to the performance of a company. For the same reason, capital structure, or financial leverage, matters to the performance of a company. Production theory indicates that firms should choose a proper combination of fixed cost and variable cost to achieve highest rate of profit. Since debts are fixed income instruments for investors, they are fixed costs for issuers. From the production theory, high fixed cost systems perform well in a low uncertainty environment and perform badly in a high uncertainty environment. So we expect small growth firms, which are of high uncertainty, to be heavily financed by equities instead of debts. This prediction differs from the pecking order theory and is confirmed by empirical results (Fama and French, 2002). In general, firms with high fixed costs in operation, such as heavy R&D spending, may choose lower fixed cost strategy in financing, to reduce the impact of uncertainty on the performance of the firms (Titman, 1984). Firms operating in low level of uncertainty, such as utility firms in regulated environment, will be financed with high level of debt to increase their fixed costs, or financial leverage.

In the original Modigliani and Miller (1958) paper, the value of a firm, in the sense of the total value of its securities, depends only on the earning power and risk of its operating assets and not at all on the debt/equity composition of the liabilities. Later works add financial risk into the picture since operating risk is related to financial risk. In this theory, financial risk and operating risk are integrated into a single analytical model so that operating leverage and financial leverage can be considered together. The following discussion shows that this unified approach greatly simplified the understanding of financing decisions.

In the tradeoff theory, the cost of debt is essentially the cost of bankruptcy. Miller (1998) had pointed that the cost of bankruptcy is too small to justify the low debt ratio. In the new theory, variable cost of operation is a function of fixed cost and uncertainty, which are affected by the debt level. So the level of debt, by affecting fixed cost and variable cost of operation, has much broader impact on firms than the cost of bankruptcy. This is why firms do not fully utilize the tax advantage of debt financing.

6.4 A New Theory of Investment

Currently, the default theory of investment is the efficient market theory, which states that stock market incorporates new information instantly and hence stock prices always reflect the true value of the underlying companies. The standard economic theory of information was developed by Grossman and Stiglitz (1980). This theory is based on rational expectation. It assumes that investors can accurately assess the value of some information and pay some fixed amount accordingly to obtain the information. Recently, various models relax the rational expectation assumption to explain major market patterns. Most of these models rely on some kind of human psychological biases and are generally grouped under the category of behavioral finance.

> To date, behavioral asset pricing models have been more ad hoc, mainly constructed to provide behaviorally based explanations of particular empirical phenomena, rather than to develop a general approach.
>
> The ad hoc approach that has characterized most behavioral asset pricing theories to date has a theory mining flavor, mainly building custom models to fit the empirical facts. These models have tended to combine one or two behaviorally realistic assumptions with other assumptions that are highly unrealistic. (Shefrin, 2005, p. 5)

More importantly, formal asset pricing models in behavioral theories are still based on general equilibrium theory. In Chapter 1, we showed that major psychological patterns frequently cited in behavioral finance literature can be naturally derived from the principles of statistical mechanics and natural selection. In this section, we will show that the generalized entropy theory of information developed in Chapter 1 will provide parsimonious understanding of major market patterns and trading behaviors of market participants.

First, since information with higher value is more costly, only large investors will spend time and resources to acquire valuable information. Hence, the return of small investors should be lower than the return of large investors and the general market index, which is the average of all investors.

The above result may be called the generalized efficient market hypothesis: investors without informational advantage, who compose the majority of investors, can not outperform the general market. Empirical results indicate that small investors' returns are harmed by active trading or active choices in investment vehicles (Barber and Odean, 2000; Cronqvist and Thaler, 2004).

This result has very strong social implications. Economists generally believe that individuals are best equipped to choose what is in their best interest and more freedom will increase social welfare. The new theory and the supporting empirical evidences, however, cast doubt over this strongly held belief.

It is not economical for small investors to spend a lot of effort to study individual stocks. Can they expect superior returns by investing in mutual funds with talented managers? The information theory predicts that the value generated by the talented managers will be mainly retained by the managers, who spend time to collect and analyze information, and not be distributed to mutual fund investors, who mostly have limited means to uncover valuable information. For example, as the average transaction costs decrease over the years, average expense ratios increase, making the sum of these two costs remaining relatively constant over the years (Wermers, 2000). Wermers also noted that during the 1977 to 1994 period, mutual funds returned an average of 13.3 percent per year to investors, which is the same as the Vanguard fund, the largest index fund. In essence, for small investors without special information advantage or people whose cost of collecting and analyzing information outweighs its benefit, which are the majority of the population, they cannot expect a higher rate of return on stocks than the general market. This statement is the same as what have been prescribed from the efficient market theory. For large investors or investors with special informational advantages, their investment strategies or corporate strategies can be very different

Second, from Formula (1.3), understanding information is easier when receivers have lower level of equivocation. This indicates that investors will earn higher rate of return if they choose to invest in securities that they are familiar with. This is supported by some recent empirical investigations. Professional managers' and individual investors' local investments outperform their remote investments (Coval and Moskwitz, 2001; Ivkovic and Weisbenner, 2005). Mutual funds with high industry concentration, where fund managers can focus on particular industries they are familiar with, are more successful in selecting securities than diversified funds (Kacpercyzk, Sialm and Zheng, 2005).

This information theory and empirical evidences challenge the current theory of wide diversification in security investment. From the theory, if one has informational advantage in certain area, she should relatively concentrate her investment in that particular area to earn higher rate of return. Otherwise, she should diversify to reduce her risk.

Third, the gradualness for the investment public to understand corporate investment and strategy will produce systematic patterns in return and trading volume. Many of these patterns cannot be explained by current behavioral literature (Lee and Swaminathan, 2000; Hvidkjaer, 2001). However, they can be explained easily by this information theory. The detailed exposition is developed in Chen (2004). Here we will give a brief description of the patterns.

To an investor, the choice of information gathering is a matter of cost. More valuable information is more costly to obtain in general. For large investors, it pays to spend a lot of effort and money to research the fundamentals. For small investors, it doesn't pay to dig into the fundamentals. They depend on processed and easy to understand information that is readily available at low cost, such as news from popular media and price movement of the shares. When an investor will be able to access a certain information also depends on her particular background, which determines her level of equivocation in receiving that information. In the following, we will illustrate the patterns of return and trading volumes of a stock of a typical company from the information processing cycle.

Suppose a company develop a new technology, which is expected to bring the company high profit in the future. From Formula (1.3), those who are familiar with the technology and company will have low equivocation in receiving the information. They understand the significance of the information first and buy the company shares. Since they are a small number of people, the trading volume is low.

As the technology goes through various stages from R&D to production, the potential becomes clearer to more people. This means that the level of equivocation gradually reduces to more people, which sustains buying interest and share prices increase gradually. As the technology becomes adopted in production and profit figures become public, the level of equivocation decrease further and the pool of investors increases further. Eventually, both the sustained increase of stock price and stable pattern of profit increase, which are very easy to understand by the general public, attract large amount of buyers, which results high trading volume and push the stock prices further up.

From Figure 1.1, the value of some information that is known to everyone is zero. As the good news reaches most investors, the security is probably already fully or over priced. Among the increased pool of investors, more and more investors understand very little of the fundamentals behind the technology and depend on easy to understand signals such as coverage from popular media and stock price movement to make trading decisions. For this group of investors, they will stop buying only when the opinion of public media changes and the trend of price increase reverses significantly. As stock price keeps increasing, momentum trading becomes highly profitable, which will eventually push the share prices higher than the fundamental value. Since large investors spend more resources in investment, they are generally better informed than small investors. As share prices become highly overvalued relative to fundamentals, large investors start to unload positions while small investors keep buying. As the selling pressure from large investors becomes greater than the buying pressure from small investors, the trend of price increase reverses to price decrease.

As the pattern of price drop becomes clear, more and more people joined the selling. After large investors and some of the small investors have finished unloading the positions, the volume of trading will decline, which is the period of low volume loser in momentum life cycle. This period is characterized by active selling of small investor (Hvidkjaer, 2001). Since small investors are typically slow to understand information, their active selling, after the selling by large investors, signals the selling is overextended, which indicates the low volume losers will rebound and earn high future return in general.

From the above discussion, positive new information from a company will produce a distinct cycle of market returns: At the first stage, share prices move up with low trading volume. At the second stage, share prices move up with high trading volume. At the third stage, share prices moves down with high trading volume, and at the last stage, share prices move down with low trading volume. Along the cycle, small traders are always at the wrong side of trading. This theory is supported by some recent empirical works (Lee and Swaminathan, 2000; Hvidkjaer, 2001).

Fourth, the differences in informational advantages can also explain some patterns in small investor trading behaviors. Odean (1999) documented that the shares individual investors sold outperform the shares they bought. He observed: “Return patterns after purchases and sales are more difficult to understand. It is possible that some of these investors are among the last buyers to contribute to the rise of overvalued momentum securities” (Odean, 1999, p. 1296). This observation is confirmed by the empirical work by Hvidkjaer (2001) on the trading behaviors of large and small investors and can be easily explained by the information theory, which indicates that small investors mainly depend on low cost and hence low value information, such as news and technical trends, which put them "among the last buyers to contribute to the rise of overvalued momentum securities". Another reason for the differential performance of shares bought and sold by individual investors may be due to informational advantages of some of their trading counterparties. Chen, Jegadeesh and Wermers (2000) documented that shares bought by mutual fund managers outperform shares they sold. Even if individual investors randomly select buy and sell orders, most trades that get executed are not in their favor, because some of their counterparties, such as mutual fund managers, have informational advantages and select to fill the orders that are in their favor. For example, if an individual investor randomly select two stocks to sell with limit orders, it is more likely that the stock which will have a higher rate of return in the future will be bought by more informed investors. This simple and intuitive understanding from information theory avoids the complex task of a behavioral explanation of: "What is more certain is that these investors do have useful information which they are somehow misinterpreting" (Odean, 1999, p. 1296).

Barber and Odean (2000) documented that there is very little difference in the gross performance of households that trade frequently and those that trade infrequently. On average, shares individual investors sold outperform the shares they bought. (Odean, 1999) If all else are same, more frequently traded investors should earn lower rate of gross return. So this indicates that more frequently traded investors have better investment skills than less frequently traded investors. This prediction can be directly verified from trading records. Barber and Odean cited Carhart (1997) to show that frequent trading also hurts the performance of mutual funds. However, a more recent and more detailed study by Wermers (2000) shows that the most active traded mutual funds outperform least traded mutual funds by a wide margin. Barber and Odean attribute trading activities to overconfidence, a behavioral explanation. It may be simpler and more consistent with empirical evidences to explain trading as a type of learning activity. Like all other kinds of learning, trading is costly. The cost of learning is compensated by knowledge gained from the experience.

Fifth, since insiders understand information much better than others, they can and will take advantage of this information asymmetry to pursue certain corporate activities. From what we discussed in the third point, shares of individual corporations and the whole market often go through undervaluation-overvaluation cycles. These under or over-valuation create opportunities for merger activities. Shleifer and Vishny (2003) develop a new model on corporate mergers, in which mergers are a form of arbitrage by rational managers operating in inefficient markets. They show that their model has better explanatory power than the behavioral based models and is supported by new empirical evidences (Ang and Zheng, 2002). While assuming managers are rational, they at the same time assume financial markets are less-than-rational. However, form the information theory, there is no need to assume financial markets are less-than-rational.

So this new theory of investment based on generalized entropy theory of information offers a unifying understanding of the documented empirical patterns in the financial markets.

6.5. Criteria for a Good Theory

Recent years, the standard theories in finance, such as Modigliani and Miller (1958) theory in corporate finance and efficient market theory in investment have retreated substantially. Many new theories have emerged and more new theories will emerge to fill the void left behind. With so many new theories around, it is time to consider some general criteria for a good theory.

First, a good theory should be consistent with empirical evidences. Not so long ago, empirical works on market behavior were so few that it was very difficult to distinguish the validity of different theories. In the past several years, however, we witness the emergence of a growing number of empirical works, which often call into the question of the ability of existing behavioral theories in explaining broader sets of empirical patterns. In the last section, we list these empirical results and show how the new information theory provides a simple and unified understanding of the empirical patterns. This theory resolves some puzzles about the market patterns raised in the recent literature that could not be answered by existing theoretical frameworks. In corporate finance, both tradeoff and pecking order theories can explain part of the empirical patterns but our new theory is consistent with broad range of empirical patterns.

Second, a good theory should provide more precise predictive power than the existing theories. Kahneman and Tversky's prospect theory is an improvement over Simon's bounded rationality theory because it offered some concrete patterns of irrationality beyond the general statement of bounded rationality. Empirical evidences indicate that the patterns of trading by small, individual investors differ systematically from that by large, institutional investors. Yet the existing behavioral theories do not offer any particular behavioral explanations to this systematic difference. (Hvidkjaer, 2001; Shanthikumar, 2004) The new information theory, however, provide very precise understanding of the systematic differences between trading patterns of large and small investors. This information theory states that more valuable information is more costly to obtain in general. For large investors, it pays to spend a lot of effort and money to research the fundamentals. For small investors, it doesn't pay to dig into the fundamentals. They depend on processed and easy to understand information that is readily available at low cost, such as news from popular media and price movement of the shares. In this work, we will show that patterns of returns by small and large investors and patterns of return and trading volumes of stocks can be easily understood from this information processing perspective.

Third, a good theory should help provide deeper understanding to existing models. For example, Hong and Stein's (1999) results are built on three key assumptions. The first two assumptions are that traders are classified as "newswatchers" and "momentum traders" according to their information processing abilities. They commented that, "the constraints that we put on traders' information-processing abilities are arguably not as well-motivated by the experimental psychology literature as the biases in Barberis et al. (1998) or Daniel et al. (1998), and so may appear to be more ad hoc." (Hong and Stein, 1999, p. 2145) These assumptions can actually be derived naturally from this new information theory. Depending on the value of assets under management, different investors will choose different methods of information gathering with different costs. "Newswatchers" are large investors who are willing to pay a high cost to collect private information and to make a deep understanding of public information. "Momentum traders" are investors who spend less cost or effort on information gathering and rely mainly on easy to understand low cost information such as coverage from popular media and price momentum signals. Cohen, Gompers and Vuolteenaho (2002) empirically confirm that institutional investors buy on fundamental news while individual investors buy on price trends. The third assumption of Hong and Stein (1999) is that private information diffuses gradually across the newswatcher population. The gradual diffusion of private information means that people gradually learn about the background knowledge of information and understand information better over time. This new information theory provides clear understanding to all three assumptions in Hong and Stein's model.

The new information theory provides a clear understanding of pecking order theory and the production theory developed in Chapter 3 provides a clear understanding of tradeoff theory in corporate finance theory.

Fourth, a good theory should be consistent with theories in other areas or at more fundamental level. The financial market is often said to adjust to new information instantly because of intense competition. Since the intellectual market is highly competitive as well, one would expect it will adjust to new information in several years. However, it is well known that theories of fundamental importance are often neglected for many decades. (Khun, 1996) For example, Bachelier's theory had to wait more than half century before it got the attention. This is hard to reconcile with efficient market theory. However, it is logical from information theory. If a research result is an addition to a well established theory, it will be noticed and accepted immediately because of low level of equivocation. When a fundamentally new theory appears, however, the level of equivocation is high and the theory may take great effort to get accepted.

From the efficient market theory, a company's stock have high return because of some unforeseeable events that cannot be predicted. However, from the researches in business strategy, a company does well often because it persists in a good strategy for a long time before the competitors and the stock market react (Collins and Porras, 1994**).** Take Wal-Mart as an example. One of its most important strategies is to set up large discount stores in small communities. The early entry of one large store in a small community preempts the entry of other big stores. The resulting local monopoly ensures high level of profit. Since the value of information is positively related to scarcity, a player adopting a superior strategy will keep quiet about it. To keep a low profile, Wal-Mart avoided opening new stores where Sears and K-mart already had existence. This gave other giant retailers the impression that Wal-Mart was not very competitive. Hence other retailers will be less likely to imitate the strategies of Wal-Mart. In fact, the strategy of local monopoly in small rural communities was not copied by other giant retailers such as Sears and Kmart for a long time for they thought small communities are too small markets for big players. The extensive time lag in adopting a superior strategy from a competitor is not consistent with efficient market theory, but is a natural result from the generalized entropy theory of information.

6.6. Concluding Remarks

In Section 6.2, we stated that the increasing role of finance in the business world in the last fifty years indicates that the relevance of financial structure can not be explained away as market imperfection. In the last fifty years, the fixed costs of businesses have increased tremendously. From Figure 3.1, high fixed cost systems are very sensitive to uncertainty. Since the choice of financial structures affects the fixed cost of businesses, it is highly relevant to the operation of business, especially those with high fixed costs. As the levels of fixed cost increase in the businesses, the financial structures also become more relevant.

Most of the research efforts in financial markets are to reduce information asymmetry as investors. Will all research efforts improve the efficiency of the overall financial markets over time? While the information asymmetry of a mature company or industry may reduce over time, the information asymmetry of the whole financial market may not. The utilization of natural resources generally follows from easy to difficult among living organisms (Atkins, 1995). Similarly, the development of human industries generally follows from easy to difficult. As the technologies of mature industries become widely known and less valuable, competition pressure drives innovations and new industries. At time passes by, the new industries are generally more and more difficult to develop and more and more difficult for most people to understand. So information asymmetry may grow instead of decrease. Around year 2000, the stock prices of many of the high tech companies dropped over ninety percent in a short period of time, signaling the difficulty of the general investment public to understand the intricacy of the complex technological systems.

While there are several existing theories on IPO underpricing, it can be a measure of information asymmetry. “IPO underpricing doubled from 7% during 1980-1989 to almost 15% during 1990-1998 before reverting to 12% during the post-bubble period of 2001-2003” (Loughran and Ritter, 2004, p. 5). Figure 6.1 shows the annual average IPO underpricing in the US market since 1980. This offers an empirical measure that information asymmetry does trend upward in recent decades.

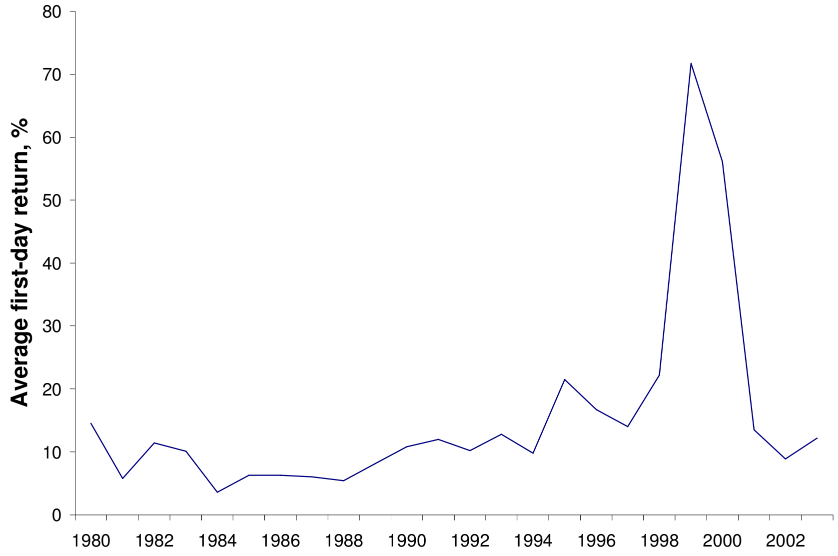

Figure 6.1 Average IPO underpricing in the US market since 1980. Source: Loughran and Ritter (2004)

This new theory of finance, like all other theories developed in this book, is still at an early stage of development. Many aspects of financial problems have not been discussed, which will be left to future research.

Epilogue

Mathematics, Beauty and Reality

When one explores wilderness uninhabited by human beings, it is very difficult to determine where and how to start. It is very difficult to go through forests or bushes. More importantly, it is almost impossible to know if one can get sufficient food supply along the way and what one can get at the end of the journey.

Facing great uncertainty and physical difficulty, early explorers often tried to find trails opened and used by large animals. These trails greatly expedite the exploration of otherwise impenetrable wilderness. More importantly, there are good reasons for animals to use the trails, which often lead to resource rich destinations.

When one explores intellectual wildernesses uninhabited by researchers, it is very difficult to determine where or how to start. More importantly, it is almost impossible to know if one can keep a job for financial survival along the way and what one can get at the end of the journey.

Facing great uncertainty and financial difficulty, early explorers of intellectual wildernesses often try to find and use intellectual trails in minds evolved from our animal ancestors. These intellectual trails greatly expedite the exploration of otherwise infinitely many leads. More importantly, there are good reasons for our animal ancestors to evolve these trails, which often help to obtain resources.

There are many intellectual trails developed in the mind. We shall only discuss two of them. The first is mathematics. If some decision making process is truly important and is needed again and again in life, it is highly economical that quantitative modules to be developed in the mind to expedite the process. For example, predators need routinely to assess their distance from the prey, the geometry of the terrain, the speed differential between itself and the prey, the energy cost of chasing down its prey, the probability of success of each chase and the amount of energy it can obtain from prey to determine whether, when and where to initiate a chase. There are many other sophisticated functions, such as navigation by migrating birds over long distances, that need sophisticated mathematical capabilities. Many animals need to make precise calculations of many of these quantitative problems many times in life. To reduce the cost of estimation, mathematical models must have evolved in their mind so that many decision making processes are simplified into parameter estimation and numerical computation. It is highly likely that, if some function is very important for the survival of the animal, in the process of evolution, this function will be genetically assimilated. By incurring a fixed cost of developing and maintaining such a mathematical function, it reduces the variable cost in each decision making. Therefore, it is natural that a fundamental understanding about life can be expressed as a mathematical theory. More precisely, since entropy is the only mathematical function to measure scarce resources, it is almost inevitable that a basic theory on life and social systems should be a mathematical theory based on entropy.

As we have discussed in this book, low entropy, information, economic value and resources are essentially the same thing looked from different angles. Because of the fundamental importance of entropy in life, the human mind, which is an evolutionary product, thinks around entropy. This is why an entropy theory based economic theory turns out to be so simple and universal. This is also why so many people have been engaged in the discussion to apply the concept of entropy in economics for many years, despite the stern reprimand from Paul Samuelson, the most powerful authority in economics:

> And I may add that the sign of a crank or half-baked speculator in the social sciences is his search for something in the social system that corresponds to the physicist's notion of "entropy". (Samuelson, 1972, p. 450)

The second intellectual trail is the sense of beauty. "The geometry of beauty is the visible signal of adaptively valuable objects: safe, food-rich, explorable, learnable habitats, and fertile, healthy dates, mates, and babies." (Pinker, 1997, p. 526) More generally, the sense of beauty is an evolved intuition about resources. Long ago, Eddington noticed the relation between entropy and beauty:

> Suppose that we were asked to arrange the following in two categories – distance, mass, electric force, entropy, beauty, melody. I think there are the strongest grounds for placing entropy alongside beauty and melody, and not with the first three. Entropy is only found when the parts are viewed in association, and it is by viewing or hearing the parts that beauty and melody are discerned. All three are features of arrangement. It is a pregnant thought that one of these three associates should be able to figure as a commonplace quantity of science. The reason why this stranger can pass itself off among the aborigines of the physical world is that it is able to speak their language, viz., the language of arithmetic. It has a measure-number associated with it and so is made quite at home in physics. (Eddington, 1958 (1935), p. 105)

Personally, for many years, I have been deeply attracted by the beauty of stochastic processes and their deterministic representations in partial differential equations. In the end, the theory developed in this book was germinated from the theory of stochastic processes and partial differential equations.

The above discussion indicates that beautiful mathematics often has deep connection with the real world. These connections, once established, are often plain and obvious. But the process of establishing the connections may be long and elusive. The understanding about the relation between information and physical entropy provides a good example.

Shortly after Shannon (1948) developed the entropy theory of information, Weaver commented: "Thus when one meets the concept of entropy in communication theory, he has a right to be rather excited --- a right to suspect that one has hold of something that may turn out to be basic and important." (Shannon and Weaver, 1949, p. 13) This sense of excitement attracted a lot of attempts to apply the concept of entropy to many other areas. As it is often the case, earlier attempts to apply some promising intuition do not yield concrete results easily. In an editorial, Shannon tried to discourage the jumping on the bandwagon:

> Workers in other fields should realize that that the basic results of the subject are aimed at a very specific direction, a direction that is not necessarily relevant to such fields as psychology, economics, and other social sciences. Indeed, the hard core of information theory is

> essentially, a branch of mathematics, a strictly deductive system. (Shannon, 1956)

Recent authority reinforces the idea that information theory has only limited connection with physical and social sciences.

> The efforts of physicists to link information theory more closely to statistical physics were less successful. It is true that there are mathematical similarities, and it is true that cross pollination has occurred over the years. However, the problem areas being modeled by these theories are very different, so it is likely that the coupling remains limited.
>
> In the early years after 1948, many people, particularly those in the softer sciences, were entranced by the hope of using information theory to bring some mathematical structure into their own fields. In many cases, these people did not realize the extent to which the definition of information was designed to help the communication engineer send messages rather than to help people understand the meaning of messages. In some cases, extreme claims were made about the applicability of information theory, thus embarrassing serious workers in the field. (Gallager, 2001, p. 2694)

If Shannon's entropy theory of information is purely a mathematical theory with little connection with the physical laws, it would be a miracle that information defined as entropy turns out to have some magic technical properties in communication problems. However, once mathematical theories are thought to be a natural part of our evolutionary legacy, it would be natural for entropy theory of information to possess these properties.

> How can the independence of human volition be harmonized with the fact that we are integral parts of a universe which is subject to rigid order of nature's laws? (Planck, 1933, p. 107)

This question is called "one of man's oldest riddles". A major insight from this theory is that human mind, shaped by natural selection, is in tune with natural laws to lower the cost of information processing. Most of information we receive is processed unconsciously. It is only in rare occasions when decision making is needed that information processing becomes conscious. And in most situations, there are no real choices any way. For example, you have the choice to eat or not to eat. But if you decide not to eat, you will be wiped out by natural selection. In a competitive world, one has to follow "optimal" choice, which is not really a choice, on most important decisions to avoid being selected out.

In the following, we will discuss the general patterns of the evolution of scientific theories to understand the origin and process of scientific revolutions.

Scientific theories are developed to reduce the cost of understanding nature, which includes human society. Costs consist of fixed cost and variable cost. Fixed cost helps reduce variable cost. The basic set of fixed assets of a scientific theory is called paradigm (Kuhn, 1996).

> When the individual scientist can take a paradigm for granted, he need no longer, in his major works, attempt to build his field anew, starting from first principles and justifying the use of each concept introduced. (Kuhn, 1996, p. 20)

The establishment of a paradigm, by incurring a common fixed cost, reduces variable cost in scientific development and communication. As a theory matures, its fixed assets accumulated. For a high fixed cost system, its variable cost will be very low when uncertainty is small. This is why science education "is a narrow and rigid education, probably more so than any other except perhaps in orthodox theology" (Kuhn, 1996, p.166).

To further utilize the fixed assets that have been acquired by the scientific community, existing paradigm is being applied to broader and broader fields. "For normal-scientific work, for puzzle-solving within the tradition that the textbooks define, the scientist is almost perfectly equipped. Furthermore, he is well equipped for another task as well --- the generation through normal science of significant crises. When they arise, the scientist is not, of course, equally well prepared. Even though prolonged crises are probably reflected in less rigid educational practice, scientific training is not well designed to produce the man who will easily discover a fresh approach." (Kuhn, 1996, p.166) Therefore, somebody, who "appears with a new candidate for paradigm" is "usually a young man or one new to the field" (Kuhn, 1996, p.166). The following quote about the emission of light or heat has a parallel in scientific research:

> That the whole world is not aglow with radiation is a consequence of a competition between the discarding of energy as radiation and as heat. The products of most reactions are in such intimate contact with their surroundings that any excitation is quickly transferred to the neighboring molecules in the form of thermal motion. However, there are some reactions for which the contact is so weak that the excited state survives long enough for the relatively slow business of squeezing out a photon to occur. (Atkins, 1991, p. 206)

The above quote can be directly translated into a comment about research. That the whole world is not aglow with great idea is a consequence of a competition between the dissemination of information as great idea or as small idea. Most of us have intimate contact with our surroundings that any new idea is quickly transferred to the academic community in the form of thermal motion, or low impact research. However, there are some cases where the contact with the academic community is so weak that ideas hold long enough in the mind for the relatively slow business of squeezing out a truly fundamental theory to occur.

All practicing scientists are educated in a common paradigm, which make it easy for them to communicate with each other. The developers of new paradigms, however, have no such luxury. From information theory, equivocation is high in communication when the receiver of information does not share the same common background or paradigm with the sender of information. The promotion of a fundamentally new idea is in general so difficult that Wallace, the cofounder of the theory of evolution, gave much more credit to the promotion of new ideas over their creation. "No one deserves either praise or blame for the *ideas* that comes to him. … But the *actions* which result from our ideas may properly be so treated, because it is only by patient thought and work that new ideas, if good and true, become adapted and utilized." (George, 1964, p. 280)

Because of the harsh environment to new ideas, many pioneers in scientific research were able to develop and promote new theories only by sheer perseverance. "Neither by poverty, nor by incomprehension of the contemporaries who ruled over the condition of his life and work, did he allow himself to be crippled or discouraged." This is a comment about Kepler's life from Einstein. It is also a reflection of lives of many other pioneers. Their struggle provides a profound testimony that information is costly and information with high value is very costly to obtain.

Bibliography

Acs, Z. and Audretsch, D. (1988). Innovation in large and small firms, *American Economic Review*, 78, 678-690.

Akerlof, G. (1970). The Market for 'Lemons': Quality Uncertainty and the Market Mechanism**,** *Quarterly Journal of Economics* 84, 488-500.

Applebaum, D. (1996). *Probability and Information, an integrated approach*. Cambridge: Cambridge University Press.

Arnott, R. and Casscells, A. (2003). Demographics and Capital Market Returns, *Financial Analysts Journal*, 59, Issue 2, p. 20-29.

Arrow, K. (1983 (1973)). Information and economic behavior, in Collected Papers of Kenneth J. Arrow, Volume 4, Cambridge, MA: The Belknap Press of Harvard University Press.

Arrow, K. (1999). Information and the organization of industry, in: Chichilnisky, G. (eds.), *Markets, information, and uncertainty*. Cambridge: Cambridge University Press.

Arthur, B. (1989). Competing technologies, Increasing returns and Lock-in by Historical Events. *The Economic Journal*. 99, 116-131.

Atkins, P. (1991). *Atoms, electrons, and Change*, Scientific American Libraray, A division of HPHLP, New York.

Atkins, P. (1994). *The second law*, New York : Scientific American Library : distributed by W.H. Freeman.

Atkins, P. (1995). *The periodic kingdom : a journey into the land of the chemical elements*, New York: Basic Books.

Auerbach, A. (2004) How much equity does the government hold? *American Economic Review*, 94, No. 2, p. 155- 160.

Bairoch, P. (1993). *Economics and world history: myths and paradoxes*, Hempstead, Hertfordshire: Harvester Wheatsheaf.

Barber, B. and Odean,T. (2000). Trading is hazardous to your wealth: The common stock investment performance of individual investors, *Journal of Finance* 55, 773-806

Barberis, N., Shleifer, A. and Vishny, R. (1998). A model of investor sentiment, *Journal of Financial Economics,* 49, 307-345.

Barberis, N. and Thaler, R. (2003). A survey of behavioral finance, in *Handbook of the economics of finance*, edited by G. Constantinides, M. Harris and R. Stulz.

Barzel, Y. (1982). Measurement cost and the organization of markets, *Journal of Law and Economics*, 25, 27-48.
Becker, G. (1993). A Treatise on the Family, Enlarged Edition. Harvard University Press.
Bennett, C. H. (1988). Notes on the History of Reversible Computation. *IBM Journal of Research and Development.* 32: 16-23.
Berk, J. and Green, R. (2004). Mutual fund flows and performance in rational markets, *Journal of Political Economy* 112, 1269-1296.
Bernstein, P. (1992). *Capital ideas : the improbable origins of modern Wall Street,* New York: Free Press
Black, F., and Scholes, M. (1973). The Pricing of Options and Corporate Liabilities, *Journal of Political Economy*, 81, 637-659.
Blaug, M. (1980). *Economic theory in retrospect*, Cambridge University Press, New York.
Bledsoe, C. (1994). "Children are like young bamboo trees": Potentiality and reproduction in Sub-Saharan Africa. in *Population, economic development, and the environment,* Edited by K. Lindahl-Kiessling and Landberg. New York: Oxford University Press,
Boulding, K. (1981). *Evolutionary Economics*. Sage Publications, Beverly Hills, Calif.
Brooks, D. and Wiley, E. (1988). *Evolution as entropy: Toward a unified theory of biology*, 2nd edition, The University of Chicago Press, Chicago.
Brown, J. and Maurer, B. (1986). Body size, ecological dominance, and Cope's rule. *Nature*, 324, 248-250.
Buffet, W. (2001). Warren Buffett on the stock market, *Fortune Investor's Guide*, with Carol Loomis, Dec 10, 80-94.
Camus, A. (1955). *The Myth of Sysiphus*, Penguin Books, London.
Carhart, M. (1997). On persistence in mutual fund performance, *Journal of Finance* 52, 57-82.
Chan, W., Frankel, R. and Kothari, S. (2002). Testing Behavioral Finance Theories Using Trends and Sequences in Financial Performance, Working paper.
Chen, J. (1999). When the Bubble is Going to Burst, *International Journal of Theoretical & Applied Finance*, 2, No. 3, 285-292
Chen, J. (2000). Economic and Biological Evolution: A Non-equilibrium and Real Option Approach, Working paper.
Chen, J. (2002a). An Entropy Theory of Value, Working paper
Chen, J. (2002b). Economic and Biological Evolution: A Non-equilibrium Thermodynamic Theory, Working paper
Chen, J. (2002c). An Analytical Theory of Project Investment: A Comparison with Real Option Theory, Working paper.
Chen, J. (2003). An Entropy Theory of Psychology and its Implication to Behavioral Finance, *Financiële Studievereniging Rotterdam Forum*, 6, 26- 31
Chen, J. (2003b). Derivative Securities: What They Tell Us? *Quantitative Finance*, 3, No. 5, C92 - C96.
Chen, J. (2004). Generalized Entropy Theory of Information and Market Patterns, *Corporate Finance Review*, 9, No. 3, 23- 32.
Chen, H., Jegadeesh, N. and Wermers, R. (2000). The Value of Active Mutual Fund Management: An Examination of the Stockholdings and Trades of fund managers, *Journal of Financial & Quantitative Analysis* 35, 343-368.

Clark, J. (1990). *The political economy of world energy: A twentieth-century perspective*, The University of North Carolina Press, Chapel Hill.

Coase, R. (1937). The nature of the firm, *Economica n.s.*, 4, 386-405.

Cohen, R., Gompers, P. and Vuolteenaho, T. (2002). Who underreact to cash-flow news? Evidence from trading between individuals and institutions, *Journal of Financial Economics* 66, 409-462.

Colinvaux, P. (1978). *Why big fierce animals are rare: an ecologist's perspective.* Princeton: Princeton University.

Colinvaux, P. (1980). *The fate of nations*. New York: Simon and Schuster.

Colinvaux, P. (1986). *Ecology*, Wiley, New York.

Collins, J. and Porras, J. (1994). *Built to last: successful habits of visionary companies*, New York : HarperBusiness.

Cope, E. (1896). *The primary factors of organic evolution*. Chicago: The open court publishing company.

Coval, J. and Moskowitz, T. (2001). The geography of investment: Informed trading and asset prices, *Journal of Political Economy,* 54, 1-39.

Cronqvist, H. and Thaler, R. (2004). Design choice in privatized social-security systems: learning from the Swedish experience, *American Economic Review*, 94, No. 2, p. 424-428.

Daly, H. (1991). *Steady-state economics*, 2nd ed.. Island Press, Washington, D.C.

Daniel, K., Hirshleifer, D. and Subrahmanyam, A. (1998) Investment psychology and Security market under and overreactions, *Journal of Finance* 53, 1839-1885.

Dasgupta, P. (1995). The population problem: Theory and evidence. *Journal of Economic Literature* 33, 1879-1902.

Dawkins, R. (1999). *The extended phenotype: The long reach of the gene*, Oxford University Press, Oxford, New York.

DeBondt, W. and Thaler, R. (1985). Does the stock market overreact? *Journal of Finance,* 40, 793-805.

Debreu, G. (1959). *Theory of value; an axiomatic analysis of economic equilibrium*. New York: Wiley.

DeYoung, R., Hunter, W. and Udell, G. (2003). The past, present, and future for community banks, Working paper.

Diamond, J. (1997). *Guns, Germs, and Steel, The Fates of Human Societies*. W.W. Norton, New York.

Diamond, J. (2005). *Collapse: How societies choose to fail or succeed*, Viking, New York.

Dixit, A. and Pindyck, R. (1994). *Investment under uncertainty*, Princeton University Press, Princeton, New Jersey.

Dixit, A. and J. Stiglitz. (1977). Monopolistic competition and optimum product diversity. *American Economic Review* 67, 297-308.

Durham, G., Hertzel, M. and Spencer, M. (2005). The market impact of trends and and sequences in performance: new evidence, *Journal of Finance*, Forthcoming.

Eddington, A. 1958 (1935). *The nature of the physical world*. Ann Arbor: The University of Michigan Press.

Fama, E. (1998). Market efficiency, long-term returns and behavioral finance, *Journal of Financial Economics,* 49, 283-307.

Fama, E. and French, K. (2002). Testing Tradeoff and Pecking Order Predictions About Dividends and Debt, *Review of Financial Studies*, 15, Issue 1, 1-33.

Frank, S. (1998). *Foundations of Social Evolution*, Princeton University Press, Princeton, New Jersey.
Friedman, M. (1953). *Essays in positive economics*, The University of Chicago Press, Chicago.
Fujita, M., Krugman, P. and Venerables, A. (1999). *The Spatial Economy: Cities, Regions, and International Trade*. Cambridge, MA: The MIT Press,
Gallager, R. (2001). *Claude E. Shannon: A Retrospective on his life, work, and impact*, IEEE Trans. IT, vol 47,#6,
Galor, O. and Weil, D. (2000). Population, Technology, and Growth: From Malthusian Stagnation to the Demographic Transition and Beyond. *American Economic Review*, 90, Issue 4, p. 806 – 828.
George, W. (1964). *Biologist philosopher, A study of the life and writings of Alfred Russel Wallace.* London: Abelard-Schuman.
Georgescu-Roegen, N. (1971). The entropy law and the economic process**.** Harvard University Press, Cambridge, Mass.
Ghiselin, M. (1974). *The economy of nature and the evolution of sex***,** University of California Press, Berkeley
Gould, S. (1996). Full house : the spread of excellence from Plato to Darwin**.** Harmony Books, New York.
Gould, S. and Eldredge, N. (1993). Punctuated equilibrium comes of age, *Nature*, 366, 223-227.
Grossman, S. and Stiglitz, J. (1980). On the impossibility of informationally efficient markets, *American Economic Review*, 70, 393-408.
Haigh, M. and List, J. (2005). Do professional traders exhibit myopic loss aversion? An experimental analysis, *Journal of Finance*, Forthcoming.
Hamilton, B. and Whalley, J. (1984). Efficiency and distributional implications of global restrictions on labour mobility. *Journal of Development Economics*, 14, 61-75.
Hardin, G. (1993). *Living within limits: ecology, economics, and population taboos*. New York : Oxford University Press.
Hong, H., Lim, T. and Stein, J. (2000). Bad news travel slowly: Size, analyst coverage, and the profitability of momentum strategies*, Journal of Finance,* 55, 265-295.
Hong, H. and Stein, J. (1999). A unified theory of underreaction, momentum trading and overreaction in asset markets, *Journal of Finance* 54, 2143-2184.
Hong, H. and Stein, J. (2003). Differences of opinion, short-sale constraints and market crashes, *Review of financial studies*, 16, 487-525.
Hull, D. (1988). *Science as a process: An evolutionary account of the social and conceptual development of Science*, The University of Chicago press, Chicago.
Hvidkjaer, S. (2001). A trade-based analysis of momentum, Working paper.
Isihara, A. (1971). *Statistical physics*, New York: Academic Press.
Ivkovic, Z. and Weisbenner, S. (2005). Local does as local is: Information content of the geography of individual investors' common stock investments, *Journal of Finance*, forthcoming.
Jegadeesh, N. and Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency, *Journal of Finance*, 48, 65-91.
Jevons, W. (1957). *The theory of political economy*. New York: Augustus M. Kelley.
Jorion, P. and Goetzmann, W. (1999). Global Stock Markets in the Twentieth Century, *Journal of Finance*, 54, 953-980.

Johnson, G. (2000). Population, Food, and Knowledge. *American Economic Review*, 90, Issue 1, p. 1-14.

Kacperczyk, M., Sialm, C. and Zheng, L. (2005). On the industry concentration of actively managed equity mutual funds, *Journal of Finance*, forthcoming.

Kahneman, D. and Tversky, A. (1979). Prospect theory: A analysis of decision under risk, *Econometrica*, 47, 263-291.

Kaldor, N. (1972). The Irrelevance of Equilibrium Economics. *the Economic Journal*, 82, 1237-55.

Karlin, S. and Taylor, H. (1981). *A second course in stochastic processes,* New York: Academic Press.

Krugman, P. and Venerables, A. (1995). Globalization and the Inequality of Nations, *Quarterly Journal of Economics,* 110, 857-880.

Kuhn, T. (1996). *The structure of scientific revolutions*, 3rd edition. University of Chicago Press, Chicago.

Lee, C. and Swamminathan, B. (2000). Price momentum and trading volume, *Journal of Finance,* 55, 2017-2069.

Loughram, T. and Ritter, J. (2004). Why has IPO underpricing changed over time? *Financial management*, 2004, p. 5-37.

Lovelock, J. (1988). *The ages of Gaia: a biography of our living earth*. Oxford: Oxford University Press.

Lucas, R. (1988). On the mechanics of economic development, *Journal of monetary economics*, 22, 3-42.

MacArthur, R and Wilson, E. (1967). *The theory of island biogeography,* Princeton University Press , Princeton, N.J.

Malloy, C. (2005). The geography of equity analysis, *Journal of Finance*, forthcoming.

Odean, T. (1999). Do Investors Trade Too Much, *American Economic Review* 89, 1279-1298.

Margulis, L. (1998). *Symbiotic Planet, A new view of evolution*. New York: Basic Books.

Martin, P. And C. Rogers. (2000). "Long-term growth and short-term economic instability," *European Economic Review*, 44, 359-381.

Marx, K. (1844). *Critique of Hegel's philosophy of right*.

Maxwell, J. (1871). *Theory of heat.* Longmans, London: Green & Co.

Mazzucato, M. (2000). *Firm size, innovation and market structure : the evolution of industry concentration and instability*. Edward Elgar, Cheltenham, UK ; Northampton, MA.

Mehra, R. and Prescott, E. (1985). The Equity Premium: A puzzle? *Journal of Monetary Economics*, 22, 133-136.

Mill, J. (1985(1871)) *Principle of political economy*, London : Penguin books

Miller, G., 2000, *The mating mind, how sexual choice shaped the evolution of human nature*. Doubleday, New York.

Miller, M. (1991). Leverage, *Journal of Finance*, 46, p. 479-488.

Miller, M. (1998). The M&M propositions 40 years later, *European Financial Management*, 4, No. 2, p. 113-120.

Mirowski, P. (1989). *More heat than light, Economics as social physics: Physics as nature's economics.* Cambridge: Cambridge University Press.

Modigliani, F. and Miller, M. (1958). The cost of capital, corporate finance, and the theory of investment, *American Economic Review*, 48, p. 261-297.

Morowitz, H. (1992). *Beginnings of cellular life: metabolism recapitulates biogenesis*, New Haven: Yale University Press
Moses, J. and Letnes,B. (2004). The Economic Costs to International Labor Restrictions: Revisiting the Empirical Discussion. *World Development*, 32, No. 10, p. 1-18.
Myers, S. and Majluf, N. (1984). Corporate Financing and Investment Decisions When Firms Have Information That Investors Do Not Have, *Journal of Financial Economics*, 13 (2) p. 187-221.
Nelson, R. (2001). *Economics as religion, from Samuelson to Chicago and beyond*, The Pennsyvania State University Press, University Park, Pennsylvania
North, D. (1981). *Structure and change in economic history,* Norton, New York
North, D. (1990). *Institutions, institutional change, and economic performance***.** Cambridge University Press, Cambridge ; New York.
Øksendal, B. (1998). *Stochastic differential equations: an introduction with applications, 5th edition*. Springer, Berlin ; New York.
Peters, R. (1983). *The ecological implication of body size*, Cambridge University Press, Cambridge.
Pinker, S. (1994). *The language instinct*. New York : W. Morrow.
Pinker, S. (1997). *How the mind works***,** W. W. Norton. New York.
Planck, M. (1933). *Where is science going*? Translated and edited by Murphy, J. London: George Allen & Unwin.
Ponting, C. (1991). *A green history of the world,* Sinclair-Stevenson, London.
Prigogine, I. (1980). *From being to becoming : time and complexity in the physical sciences*. W. H. Freeman**,** San Francisco.
Raimirez, M. (2003). Mexico under NAFTA: A Critical Assessment. *The Quarterly Review of Economics and Finance* 43 863-892.
Ricardo, D. (1817). *The principles of political economy and taxation***,** London : J. Murray.
Roll, R. (1986). The hubris hypothesis of corporate takeover, *Journal of Business,* 59, p. 197-216.
Romer, P. (1986). Increasing returns and long-run growth, *Journal of Political Economy*, 94, 1002-1037.
Romer, P. (1994). New goods, old theory, and the welfare costs of trade restrictions, *Journal of development economy*. 43: 5-38.
Runnels, C. (1995). Environmental degradation in Ancient Greece, *Scientific American*, 272, March, 96-99.
Rushton, P. (1996). *Race, evolution, and behavior: a life history perspective***,** New Brunswick, NJ: Transaction Publishers.
Samuelson, P. (1972). *The collected scientific papers*, Vol. 3, Cambridge, Mass.: MIT Press.
Samuelson, P. and Nordhaus, W. (1998). *Economics*, 16th edition, McGraw Hill.
Schrodinger, E. (1944). *What is Life?* Cambridge University Press, Cambridge.
Schorn, R. (1935). *Planetary astronomy : from ancient times to the third millennium***,** College Station: Texas A&M University Press.
Schumpeter, J. (1954). *History of economic analysis*. New York: Oxford University Press.
Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal,* 27, 379-423, 623-656.

Shannon, C. (1956). The Bandwagon (Editorial)*, IEEE Transactions Information Theory,* volume 2, p.3

Shannon, C. and Weaver, W. (1949). *The mathematical theory of communication*, Urbana: The University of Illinois Press.

Shanthikumar, D. (2004). Small Trader Reactions to Consecutive Earnings Surprises, working paper.

Shefrin, H. (2005). *A behavioral approach to asset pricing*, Elsevier Academic Press, Amsterdam.

Shiller, R. (2000). *Irrational exuberance***,** Princeton, N.J. : Princeton University Press.

Shleifer, A. and Vishny, R. (2003). Stock market driven acquisitions, *Journal of Financial Economics* 70, 295-311.

Simpson, G. (1944). *Tempo and mode in evolution*. Columbia university press, New York.

Smil, V. (1999). *Energies*, The MIT Press, Cambridge, Mass.

Smil, V. (2003). *Energy at the crossroads: Global perspectives and uncertainties*, The MIT Press, Cambridge, Mass.

Smoller, J. (1994). *Shock waves and reaction-diffusion equations,* New York : Springer-Verlag.

Spengler, J. (1938). *France faces depopulation***.** Durham, N.C. : Duke University Press.

Stearns, L. and Allan, K. (1996). Economic Behavior in Institutional Environments: The Corporate Merger Wave of the 1980s, American Sociological Review, 61, p. 699-718.

Stearns, S. (1992). *The evolution of life histories*, Oxford University Press, Oxford.

Stebbins, G. (1982). *Darwin to DNA, molecules to humanity*. W.H. Freeman and Company, San Francisco.

Tainter, J. (1988). *The collapse of complex societies*. Cambridge University Press, New York.

Theil, H. (1967). *Economics and information theory*. Amsterdam: North-Holland Publishing Company.

Titman, S. (1984). The effect of capital structure on a firm liquidation decision. *Journal of Finance*, 13, 1-19.

Toye, J. (2000). *Keynes on population*, Oxford University Press, Oxford.

Toynbee, A. (1972*). A study of history***.** London : Thames & Hudson.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases, *Science* 185, 1124-1131.

Voelkel, J. (2001). *The composition of Kepler's Astronomia Nova*, Princeton University Press, Princeton, New Jersey.

Wallis, J. and North, D. (1986). Measuring the transaction sector in the American economy, 1870-1970, in Long-term Factors in American Economic Growth, Edited by S. Engerman and R. Gallman. The University of Chicago Press, Chicago and London.

Walras, L. (1954). *Elements of pure economics*. Translated by William Jaffe, Homewood, Illinois: Richard Irwin, Inc.

Weitzman, M. (2000). Economic Profitability versus Ecological Entropy. *The Quarterly Journal of Economics*, 115, p. 237 – 263.

Wermers, R. (2000). Mutual fund performance: An empirical decomposition into stock-picking talent, style, transaction costs, and expenses, *Journal of Finance* 55, p. 1655-1703.

West, R. (2005). A real options approach to understanding international relations, Working paper.
Wiener, N. (1948). *Cybernetics*. MIT Press, Cambridge.
Wilson, E. (1975). *Sociobiology: the new synthesis*. Belknap Press of Harvard University Press, Cambridge.
Withers, P. (1992). *Comparative animal physiology*, Fort Worth : Saunders College Pub.
Zingales, L. (2000). In Search of New Foundations, *Journal of Finance*, 2000, 55, Issue 4, 1623-53

Index