Essays in Economic and Financial History

Alexander Tepper
University of Oxford
Christ Church

A thesis submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Economics

July 2011

Approximate Word Count: 55,000.

Essays in Economic and Financial History

Alexander Tepper, Christ Church

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Economics, Hilary Term 2011.

Abstract

Division One: "Malthus Gets Fat" (Two Chapters)

Chapter One develops a simple dynamic model to examine the takeoff from a Malthusian economy to a modern growth regime. It finds that several factors, most notably the rate of technological progress and the economic structure, determine the fastest rate at which the population can grow without declining living standards; this is termed *maximum sustainable population growth*. It is only when this maximum sustainable rate exceeds the peak rate at which a society expands that takeoff can occur. I also investigate the effects of trade and international income transfers on the ability to sustain takeoff. It is also shown that present income growth is not necessarily indicative of the ability to sustain takeoff and that factors which increase current income growth may actually inhibit takeoff, and vice versa.

Chapter Two applies the sustainable population growth framework to Britain during the Industrial Revolution. The model shows a dramatic increase in sustainable population growth at the time of the Industrial Revolution, well before the beginning of modern levels of income growth. The main contributions to the British breakout were technological improvements and structural change away from agricultural production. At least until the middle of the 19th Century, coal, capital and trade played a minor role.

Division Two: "Leverage and Financial Market Instability" (Four Chapters)

Chapter One develops a model of how leverage induces explosive behavior in financial markets. I show that when levered investors become too large relative to the market as a whole, the demand curve for securities can suddenly become upward-sloping as levered investors are exposed to forced liquidations. The size and leverage of all levered investors defines the *minimum elasticity-adjusted market size for stability* or *MinEAMASS*, which is the smallest elasticity-adjusted market size that can support the group of levered investors analyzed. This gives rise to a measure of instability that can predict when markets become vulnerable to a leverage-driven market liquidity crisis.

Chapter Two iterates the model of Chapter One forward in time to generate an inflating bubble that suddenly bursts, reproducing many of Kindleberger's (1996) stylized facts about the dynamics of bubbles in a simple framework.

Chapter Three applies my measure of instability in a historical investigation of the 1998 demise of hedge fund Long-Term Capital Management (LTCM). I find that a forced liquidation of LTCM threatened to destabilize some financial markets, particularly for bank funding and equity volatility.

Chapter Four discusses how the model applied to the stock market crash of 1929. There the evidence suggests that a tightening of margin requirements in the first nine months of 1929 combined with price declines in September and early October caused enough investors to become constrained that the market was tipped into instability, triggering the sudden crash of October and November.

Contents

Preface 5

Division 1. Malthus Gets Fat: The Beginnings of Income Growth in Britain and Beyond 7

Introduction 9

Chapter 1. A Common Framework for Contributions to Takeoff 13

Chapter 2. Accounting for Breakout in Britain: The Industrial Revolution through a Malthusian Lens 35

Appendix A. The Model with Endogenous Capital 77

Appendix B. Notes on the Dutch Data 79

Appendix C. British Data Sources 83

Division 2. Leverage and Financial Market Instability: A Theoretical and Historical Investigation 87

Introduction 89

Chapter 1. A Leverage-Based Measure of Financial Instability 91

Chapter 2. An Application: A Model of Bubble Formation 129

Chapter 3. The Collapse of Long-Term Capital Management 161

Chapter 4. A Textbook Case: The Great Crash of 1929 185

Conclusions 205

Appendix A. 12 Bubble Simulations 209

Preface

This thesis consists of two sets of papers. The first, entitled "Malthus Gets Fat: the Beginnings of Income Growth in Britain and Beyond," was supervised by Professor Knick Harley and was initially written from November 2007 to May 2008 and submitted as my M. Phil. thesis. The most recent version was revised in March 2010. The second, entitled "Leverage and Financial Market Instability: a Theoretical and Historical Investigation," was jointly supervised by Professor Knick Harley and Dr Howard Jones. It was written from September 2008 to February 2011. Pages are numbered separately.

Thanks are due to many who have guided me along this journey.

Knick Harley, who has supervised my work for nearly four years, turned me on to economic history as an M. Phil. student and shared his deep expertise in the practices and principles of the discipline. His guidance was always precise, calm and constructive. I am especially grateful to him for helping to validate my own voice and path and for his commitment to seeing the project through despite his retirement.

Howard Jones, who co-supervised the second paper, has been a tremendous source of knowledge about the financial markets, a fountain of encouragement, and an absolute pleasure to work with.

Professor Richard Sylla at NYU has been an invaluable resource and professional mentor, introducing me to the American academic community and providing an academic home for me in New York. I gained a great deal from the opportunity to present and participate in the wonderful financial history seminar he chairs.

I benefited immensely from the support and deep market knowledge of Neil Grossman and Ken Tremain, the partners of TKNG Capital, who developed my

market intuition, engaged in vigorous, extensive and enjoyable discussions on the topics in my thesis and many others, and introduced me to many useful market contacts.

Raphael Espinoza and Jonathan Dingel have been great friends and served as a sounding board for all things economic, both on substance and process.

I also thank Bob Allen, Greg Clark, Bob Lucas, Matthias Morys, Dimitri Tsomocos, Paul Warde, and Jan Luiten van Zanden for their helpful consultations, Michael Best, Joe Rowsell, Charlie Brendon, Nick Howarth and Nick Juravich for their insightful comments, and the many market participants who wish to remain anonymous who offered commentary and source material.

I am grateful every day to my partner in adventure, Michelle, who offered support and encouragement in difficult moments, cheerleading in better times, welcome distraction when necessary, and love always.

Finally, I would like to thank my father for fostering in me a spirit of scientific inquiry, my mother for teaching me its limits, and both my parents whose unconditional support in this endeavor and in all past ones has made me one of the most fortunate people on Earth.

Division 1

Malthus Gets Fat: The Beginnings of Income Growth in Britain and Beyond

Introduction

Robert Malthus had possibly the worst timing of any economist in history. At the start of the 19th Century, he propounded an economic theory that accurately described practically all of human history. Yet within decades, the predictions of the theory had been spectacularly shattered and his name had become synonymous with dour, unimaginative pessimism.

Malthus argued that because natural resources, in particular land to grow food, were fixed, humanity was doomed to poverty and starvation unless strict population controls were implemented. Specifically, Malthus claimed that any increase in incomes would allow people to have more children and to keep them alive longer. This would increase the price of food, impoverishing people and leading to death, falling fertility and population decline, bringing the economy back to its "Malthusian" equilibrium. While this "dismal" prediction earned economics its most lasting moniker, the next century saw the first large sustained rise in living standards in history. What changed?

Since Robert Solow founded modern growth theory 50 years ago, the fashion in economics has been to ascribe growth in living standards to technological progress. While it seems indisputable that technology is an important ingredient in growth, it is not, as most growth models predict, sufficient for growth. Technological progress has been around since the dawn of civilization, but it is only in the last 200 years that we have been fortunate enough to experience a takeoff in living standards. Indeed, the ancient world was full of technological advances that not only improved productivity but, no less than the steam engine or the cotton gin, revolutionized the productive process and opened vast new possibilities to the people of their time. The advent of stone tools, agriculture, the aqueduct, irrigation, the wheel, the bow

and arrow, water transport, metal-working, the wheelbarrow, the windmill—none of these resulted in a sustained rise in average living standards. Perhaps a certain level of technology is required for takeoff, but in the modern world, poor countries adopt modern technologies like corrugated steel building materials, cell phones, cars and the use of fossil fuels, yet their economies remain as stagnant as ever. The inspiration for this paper is the natural question these observations invite: why, when both the theory and casual observation suggest that technology is the principal cause of growth, did technology advance for so long without growth, and why do some countries today remain poor despite having access to technology far in advance of that available during the Industrial Revolution?

In this paper, we will view a modern or growth economy as one where technological progress at society's normal rate implies rising living standards. In a Malthusian economy, technological progress at society's normal rate implies stagnant living standards. We refer to the transition between these two regimes as "takeoff" or equivalently "breakout."

This is a more precise statement of Rostow's (1960) notion that, "the takeoff is the interval when... the forces making for economic progress... expand and come to dominate the society. Growth becomes its normal condition." (p. 7) While this paper does not argue for a return to Rostow's "five stages" framework, which has fallen somewhat out of fashion, it does demonstrate that moments of "takeoff" can be identified, and that the concept remains illuminating in understanding the beginnings of rising living standards.

Economic historians have pointed to a number of factors to help explain the unique breakout that occurred in Britain in the 19th Century, yet the specific links between these factors and the beginnings of modern growth are often tenuous, and precisely which factors are important is the still the subject of considerable debate. This paper develops a model that allows us to draw clear, bright lines between some of the factors cited by economic historians and the beginning of growth, and provides a framework that permits a quantitative assessment of the relative contributions of

those factors. The model has insights not just for the Industrial Revolution, but also for modern development.

This paper is divided into two chapters. Chapter One develops a broadly applicable model for understanding the transition from a Malthusian regime to modern growth. It concludes that there are a variety of strategies that can effect this transition, although it is difficult without technological progress.[1] Chapter Two makes use of the model to investigate British growth during the Industrial Revolution, and finds that the model is consistent with and provides insights into the British experience. In the British case, I find that the primary contributions to takeoff were technological progress and structural change, with a more modest role played by capital deepening and coal.

[1]The only cases I am aware of where an argument can be made that this occurred are some of the smaller countries surrounding the Persian Gulf, which arguably loosed their Malthusian constraints by relying entirely on oil revenues.

CHAPTER 1

A Common Framework for Contributions to Takeoff

I have opened this paper by remarking upon the existence of both Malthusian and growth economies, and the inadequacy of technological progress as a defining difference between them. This chapter develops a model that accounts for the existence of both regimes, and allows an investigation of the contributions to the transition. The overall conclusion is that contributions to takeoff must increase the rate of population growth the economy can support without declining incomes.

The worldview taken in the model is Malthusian. The economy is partially dependent upon a fixed factor of production, usually viewed as land, so that in the absence of technological progress other factors of production are subject to diminishing returns as they accumulate. In particular, a growing population leads to falling per capita incomes, *ceteris paribus*. Moreover, as people get richer, they have more children, so that wealth leads to population growth which tends to be impoverishing. However, wealthier societies have undergone a demographic transition, so that after a point this trend reverses itself and fertility declines with income. Takeoff can only occur when the maximum rate of sustainable population growth rises above the peak fertility of a society.

There already exists a modest literature modeling growth in a unified way, to which the three most prominent recent contributions are Kremer (1993), Galor and Weil (2000) and Hansen and Prescott (2002). Kremer assumes and empirically tests a simplistic relationship between essentially global population size and the global rate of technological progress and claims this relationship generates emergence from a Malthusian trap. Galor and Weil construct a complex model that includes assumptions similar to Kremer's as well as Beckerian assumptions about child quality-quantity tradeoffs to generate a demographic transition as societies get

richer. Hansen and Prescott assume exogenous technical progress and fertility decisions but include shifting from a Malthus sector that relies on land for production to a Solow sector that does not. The shift is an equilibrium phenomenon driven by diminishing returns to land as population grows with technical progress.

While the present model was developed independently, it shares with those papers an emphasis on a phase change generated by a demographic shift and an increase in sustainable population growth, although this is not their terminology. Sato and Niho (1971) also employ a framework of sustainable population growth in a two-sector model of emergence from Malthusian stagnation in a closed economy, concluding that only technological progress in agriculture aids in achieving breakout.

This chapter makes four main contributions to the literature. First, it suggests that effective development strategies are likely to vary from country to country, and to rely not just on technology. It thus expands the sustainable population growth framework to show how it acts as a "common currency" to assess the relative and absolute importance of various contributions to takeoff. Second, it suggests that in theory, most of the underlying changes in the character of economic expansion need only be temporary to effect a phase change from the Malthusian to the modern epoch. In practice, however, "temporary" is likely to mean at least many decades if not centuries. Third, it offers a new framework for understanding how trade and colonialism fit into the breakout story, and that the effects of these are not what might be expected. Fourth, it demonstrates that structural change is an important *cause* of the transition to a modern growth regime, not just a symptom or result of it as has been previously been assumed.

The model presented here has several advantages over previous attempts to model the transition process. First, it is much simpler and allows easy access to the intuition that can be obscured by a more complex model. While Sato and Niho and and Hansen and Prescott require two sectors, which essentially doubles the number of variables, and Galor and Weil employ four simultaneous difference equations, the

present model combines one production function and one differential equation governing population growth in a simple, intuitive way. Second, unlike Galor and Weil and Kremer, this model's assumptions do not dictate that takeoff can only occur in societies of a certain size or level of technological advancement. Given that the first takeoff actually did occur in Britain, rather than much larger France or China, this result is to be viewed as an advantage. Third, in this model, most of the underlying causes of growth are exogenous; I view this as an advantage both because it avoids arbitrary and simplistic assumptions about social behavior and because many of the important factors for takeoff are driven by processes that are as much political, historical or social in nature as economic. By remaining agnostic about the relationships between the elements of breakout, the model can accommodate a rich array of development strategies and can apply to a wider range of societies.

In terms of its simplicity, this model could be viewed as a prequel both to the modern "Unified Growth" literature, which is reviewed by Ashraf and Galor (2008), and to the previous generation of literature of which Sato and Niho is an example. While each of these papers provides its own path to takeoff, the restrictive assumptions and complexities do not offer insight into which aspects of the takeoff are fundamental and likely to be common to all countries and models.

By contrast, this paper examines the most stripped down possible version of a unified growth model. It highlights the most important and fundamental dynamics of the transition from a Malthusian regime to a Solovian one, showing that the many assumptions and complexities of the literature are not necessary to generate both a qualitatively and quantitatively correct story. It clearly highlights the two fundamental causes of the transition to growth: an acceleration of progress in technology and other factors affecting productivity, and structural change that reduces the economy's dependence on fixed natural resources.

Section 1 of this chapter lays out the basic framework for modeling transitions from Malthusian regimes to modern growth. Sections 2 and 3 discuss the role played

by trade and colonialization. Appendix A presents a version of the model with endogenous capital and shows that this does not change the basic results.

1. The Basic Model

I employ a very simple Malthusian model, with production a function of population and resources. Resources are fixed and population is endogenous. Technology is exogenous, for we wish to examine the role of technology in the transition from a Malthusian economy to a growth economy, and to do this we would like to exogenously vary the rate of technological progress. Later, we will add other factors of production and generalize the functional form.

The economy has Cobb-Douglas production function, and the birth and death rates are taken to be exogenous functions of per capita income:

$$Y = AL^{\alpha}R^{1-\alpha} \tag{1}$$

$$\frac{\dot{L}}{L} = b - d \equiv g(Y/L) \tag{2}$$

The birth rate rises with income until income reaches some critical level, and then falls thereafter, while the death rate falls with income. Defining per capita income $y = Y/L$, the functions $b(y)$, $d(y)$ and $g(y)$ look like figure 1. The justification for the qualitative functional form of $g(y)$ in Britain is discussed in some detail in section 2. Lucas (2004) also plots population growth relative to per capita GDP for five regions of the world and finds that all of them have the form proposed here.

To solve the model, we rewrite it in per capita terms:

$$y = A\left(\frac{R}{L}\right)^{1-\alpha} \tag{3}$$

Taking logs and differentiating yields:

$$\frac{\dot{y}}{y} = \frac{\dot{A}}{A} + (1-\alpha)\frac{\dot{R}}{R} - (1-\alpha)\frac{\dot{L}}{L} \tag{4}$$

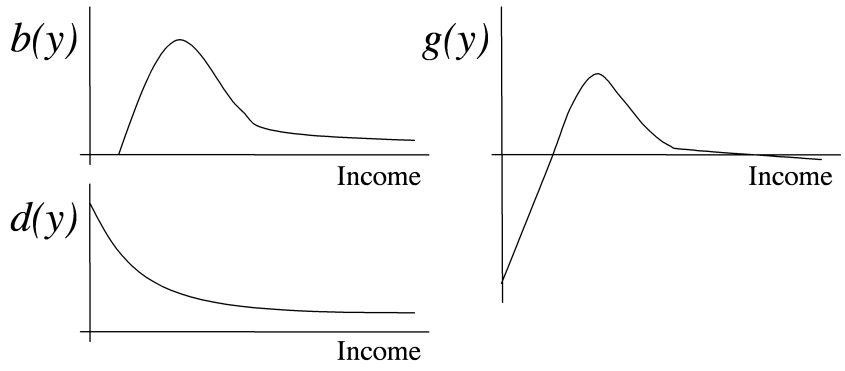

FIGURE 1. Functional Forms of Birth Rate, Death Rate, and Population Growth

Noting that resources are fixed and substituting for population growth, we have simply:

$$\frac{\dot{y}}{y} = \frac{\dot{A}}{A} - (1-\alpha)g(y) \tag{5}$$

This has a simple interpretation: if technology growth is faster than population growth times the resource share, per capita income is rising, otherwise it is falling.

$\frac{\dot{A}}{A}(1-\alpha)^{-1}$ defines the *maximum sustainable population growth* (MSPG), the fastest rate at which the population can grow without falling incomes.[1] MSPG can thus be thought of as the carrying capacity of the economy. If per capita income starts at a low level, then it will eventually rise to a Malthusian equilibrium at point M in figure 2, where the MSPG is precisely equal to $g(y)$. This equilibrium is stable: if per capita income is hit by a positive shock, the population will grow faster and income will fall again; if hit by a negative shock, population will grow more slowly and income will rise. The same analysis holds for population shocks.

The mechanism is classically Mathusian: a technology shock leads to higher incomes, causing population growth to rise and encountering diminishing returns to land. Diminishing returns combine with a population that is growing faster than the maximum sustainable population growth to cause falling incomes and a return to Malthusian equilibrium.

[1]If an economy's resource base is expanding, then $MSPG = \frac{\dot{A}}{A}(1-\alpha)^{-1} + \frac{\dot{R}}{R}$.

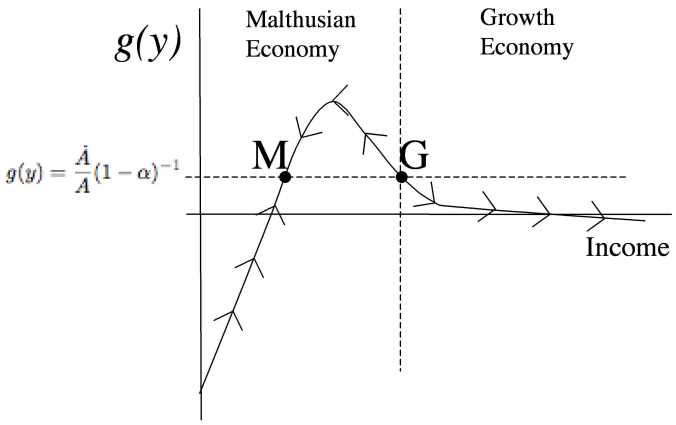

FIGURE 2. The Dynamics of the Economy

The first thing to notice about this economy is that the *level* of income is determined by the *growth rate* of technology. The situation is analogous to pouring water into a leaky bucket—the faster the water is poured in, the higher the steady-state level of water in the bucket, but if the faucet is turned off, the water leaks out to a lower level. This leads to a prediction that in pre-modern societies, higher income levels should be associated with periods of technological advancement (or high MSPG for other reasons), but should dissipate once the technological advancement ends.

In essence, maximum sustainable population growth is a reservoir that can be used to support a growing population or rising living standards. In a Malthusian society, the maximum sustainable population growth rate is below the peak population growth rate. All growth in productivity is used to support a larger population, fully exhausting the reservoir. But once MSPG rises above that peak, people do not want to reproduce fast enough to "use up" all the productivity advances being discovered in the economy. Some of these advances can be used to increase per capita income, effecting the transition to modern growth. Returning to the leaky bucket analogy, modern growth is equivalent to pouring water into the bucket so fast that despite the leaks, the bucket overflows.

Point G in the figure is also an equilibrium, albeit an unstable one. If the economy finds itself to the right of point G, income will rise, causing fertility to fall,

resulting in further income rises, fertility falls and sustained growth. If the economy is to the left of G, it will fall back to the Malthusian equilibrium at point M.

The transition from the Malthusian economy to the growth economy will happen if the economy can somehow get from point M to point G, the same level of technology and population growth with a much higher income. If the Malthusian population growth function g is fixed,[2] making this transition requires either a large positive shock to per capita income or an increase in the rate of technology growth, which, surprisingly, need only last long enough for the economy to get over the hump. Once it is over the peak and past point G, technology growth can fall back to its Malthusian level without jeopardizing growth. Alternatively, a temporary change in the link between income and population growth, such as an increase in the death rate or decrease in the birth rate for given levels of income, could bring the peak of g below sustainable population growth, the dotted line in figure 2, for long enough to effect the transition.

We can easily solve for how long the economy takes to transition to growth. In order to effect a transition, income must move from point M to point G. Denote the rate of TFP growth by π, and assume this temporarily increases to a rate π_H. Then equation (5) can be integrated to obtain:

$$\Delta \ln y = \pi_H T - \int_0^T (1-\alpha) g(y)\, dt \tag{6}$$

[2]As discussed in section 2, econometric evidence, while relying on somewhat dodgy data, indicates that $g(y)$ does vary with time. However, at least for Britain, the variation was nowhere near enough for exogenous demographic changes to effect a transition to growth without MSPG already near the peak in the population growth function.
I am also grateful to the many colleagues who have pointed out that the population growth function itself may be endogenous as people respond to societal changes correlated with wealth. As such, an exogenous increase in wealth may not be enough to effect a demographic transition, a point supported by the fact that contemporaneous functions relating wealth to fertility within Malthusian societies tend to be upward sloping throughout (see Clark and Hamilton, 2006). I acknowledge this point but reply in two ways. First, increases in MSPG tend to be permanent and far above the highest peaks in population growth. For a modern society with modest TFP growth of 1% and a high rent share of 5%, MSPG is 20% even without considering capital accumulation or other contributions. Second, I will assume that even if the fertility function does not operate in the way I assume on an individual level, it does operate that way on a societal level. I therefore assume that whatever causes the wealth increase also causes a demographic transition.

which after a change of variable solves to yield

$$T = \int_{y_M}^{y_G} \frac{y^{-1}}{\pi_H - (1-\alpha)g(y)}\, dy \tag{7}$$

This gives an expression for the transition time from a Malthusian trap to a growth economy in terms of TFP growth and the population growth function. The time decreases with the rate of technology growth and increases with $(1-\alpha)$, the resource share.[3]

If MSPG does not significantly exceed the top of the hump, the economy can theoretically escape but faces serious challenges to do so. First, the small denominator means that the transition will take a very long time. Second, even small fluctuations in MSPG can eliminate the ability to grow altogether. If the peak of the population growth function is 1.5%, and MSPG is "typically" 1.6%, then even one year in 15 of zero MSPG is enough to forever delay breakout.

Finally, it should be noted that none of the results in this section depend on the Cobb-Douglas functional form or having just two factors of production. Consider the production technology, or more generally an income function

$$Y = F(L, R_i, X_j; s_k) \tag{8}$$

which is assumed to have constant returns to scale in the extensive inputs L, R_i and X_j. Here R_i are different types of fixed resources, X_j are variable factors of production such as capital or human capital, and s_k are parameters of the economy, such as the terms of trade and level of technology. Then taking logs and differentiating, per capita income growth is given by:

$$\frac{\dot{y}}{y} = \sum_i \eta_i \frac{\dot{R}_i}{R_i} + \sum_j \eta_j \frac{\dot{x}_j}{x_j} + \sum_k \eta_k \frac{\dot{s}_k}{s_k} - \eta_R g(y) \tag{9}$$

where η_i and η_j are the elasticities of production with respect to each of the fixed resource and variable factors of production and η_k are the elasticities of income

[3]Note that in order for the time to be finite it is necessary that the denominator in the integral is always positive—that is, MSPG is at least temporarily above the top of the hump in $g(y)$.

with respect to the parameters s_k. Lowercase letters denote per capita amounts and $\eta_R = \sum_i \eta_i$ is the total share of output paid to fixed factors. (In a competitive economy, factors are paid their marginal products, so the elasticity of production with respect to a factor is the share of total income paid to that factor.) Maximum sustainable population growth is then determined as before as the rate of population growth giving rise to constant per capita income:

$$MSPG = \frac{1}{\eta_R} \cdot \left[\sum_i \eta_i \frac{\dot{R}_i}{R_i} + \sum_j \eta_j \frac{\dot{x}_j}{x_j} + \sum_k \eta_k \frac{\dot{s}_k}{s_k} \right] \tag{10}$$

$$= \frac{1}{\eta_R} \frac{\dot{y}}{y} + \frac{\dot{L}}{L} \tag{11}$$

The R_i terms in line (10) are familiar and represent the contribution from an expanding resource base. The terms in x_j are new. They represent the contributions of deepening in other factors of production, particularly capital. Growth in these factors contributes to MSPG proportionately to their shares in output, and inversely with the share of fixed resources in output. Thus structural change away from land-intensive production and into capital-intensive production raises MSPG by allowing capital to have a larger contribution.

Later, we will use the formulation in line (11), which offers a very general reduced form for MSPG that applies to any constant-returns-to-scale economy. Note that this is not a causal relationship but an observational one that allows us to calculate MSPG from economic observables.

Equation (11) highlights the difference between income growth and MSPG. The concepts are related but not identical, and changes that increase one may decrease the other. To see this, let us consider the effect of a small change in the structure of the economy by differentiating equation (11).

$$d(MSPG) = d\left(\frac{\dot{y}}{y}\right) \cdot \frac{1}{\eta_R} + d\left(\frac{1}{\eta_R}\right) \cdot \frac{\dot{y}}{y} + g'(y)dy \tag{12}$$

Equation (12) makes explicit that changes that increase instantaneous income or income growth may or may not increase MSPG depending on what happens to the

resource share. If a structural change increases the resource share, it may decrease MSPG even though it increases income or income growth. If it decreases the resource share, it may increase MSPG even in the face of a decline in income or income growth. The intuition is that a lower resource share raises the ultimate steady-state income level even as it decreases the rate of growth toward that equilibrium.

We shall explore this effect in more detail as we explore the effect of an opening to trade on the economy.

2. The Role of Trade and Structural Change

In a world with no trade, a poor country must produce its own food, which requires labor and land. It might have a production function like the one above:

$$Y = AL^{\alpha}R^{1-\alpha}$$

This production function represents the physical process of growing, harvesting and distributing food. Yet, if the country opens to trade, there is another way to "grow" food—it can, for example, open a call center, sell customer services to (increasingly frustrated) Western consumers, and use the proceeds to buy food. Since call centers do not require large amounts of natural resources, this process might be represented by the production function:

$$Y = (p_{services}/p_{food}) \cdot AL \tag{13}$$

where $p_{services}$ and p_{food} are the world prices of customer services and food. The effect of trade, from the point of view of the Malthusian country, is to change its production function.

As an abstraction, the economy's new aggregate production function might be

$$Y = AL^{\alpha'}R^{1-\alpha'} \tag{14}$$

where $\alpha' > \alpha$.[4] The process of opening to trade *raises the value of* α, reducing the resource share in the economy. This trade effect on the transition to growth can be seen in figure 3.

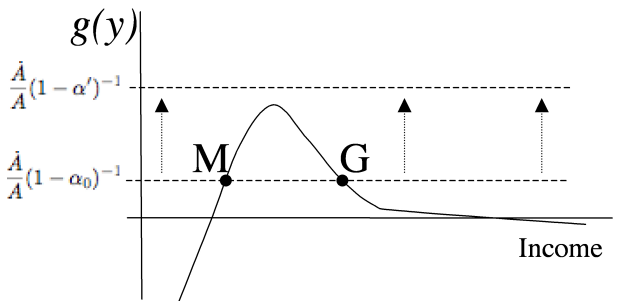

FIGURE 3. The Effect of Trade on MSPG

This increase in the value of α raises the level of population growth that can be sustained in a Malthusian equilibrium.[5] If the increase is large enough, as shown in the figure, the change can be enough to lift the resource constraint entirely and set off the transition to modern growth. As above, it is sufficient that this increase in α be temporary.

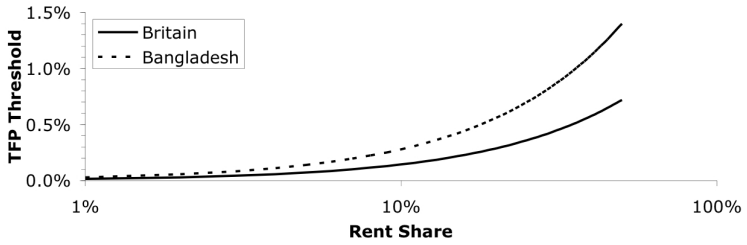

FIGURE 4. Ease of Takeoff at Different Rent Shares

To see how important the structure of the economy is, figure 4 plots the level of TFP growth necessary to escape the Malthusian trap as a function of the rent

[4]It is recognized that as a mathematical matter the production functions should simply be added. Given that Cobb-Douglas is already an abstraction that does not apply to the real world, it is hoped that the reader will accept the further abstraction that combining two industries with Cobb-Douglas production functions will be taken to yield a Cobb-Douglas with intermediate factor shares. Sticklers can imagine that the economy completely specializes into another good, but that all goods require at least some small share of non-tradeable resources, such as drinking water, living space or land on which to build factories and call centers.

[5]This depends on at least some part of technology being non-resource augmenting. Otherwise we have $A^{(1-\alpha)}$ in the production function and moving away from resource-intensive production hurts technological progress as well lifting the Malthusian constraint, and these effects offset each other.

share in the economy. These are plotted for Britain and for Bangladesh; the curves are different because Britain's population growth peaked at 1.44%[6] from 1800-1830 while Bangladesh's peaked at 2.8% in the 1970's.[7] These linear relationships are plotted on a log scale in order to better highlight the difference between economies with modern values for the rent share and those with Malthusian levels.

There are two things to notice about this chart. First, with rent shares at modern levels, virtually any TFP growth will allow the economy to escape a Malthusian trap, while at rent shares in the range of agricultural societies a significant level of TFP growth is needed. The second points to the importance of the fertility function; starting from a rent share of 30%, Britain would have required TFP growth of 0.43% for takeoff, within the range of that seen during the agricultural revolution (see chapter two). A country with Bangladesh's fertility profile, by contrast, would have required TFP growth of 0.84%, a level not seen by Britain until the mid-19th Century. It is thus no surprise that developing countries tend to be food importers.

The analysis so far applies only to small countries. Before the modern rise in living standards, however, all economies were resource constrained and had near-subsistence living standards. In that world, any country transitioning to modern growth would become a big economy if it were not so already, so its expanding exports would affect the terms of trade.

When the terms of trade depend on trade volumes, it is no longer the case that specialization away from (or into) resource-intensive production necessarily eases (or inhibits) the transition to growth. Despite this complication, the situation can be analyzed in the same basic manner. Trade is still to be thought of as an alternative production function

$$Y = \frac{p(\text{Home Production})}{p(\text{Consumption Basket})} \cdot AL^{\alpha'}R^{1-\alpha'} \tag{15}$$

The consumption basket will consist partially of home production and partially of imports. In a country like Britain that specialized in industry, imports will be

[6]Wrigley and Schofield (1981).
[7]http://www.populstat.info.

resource-intensive, meaning $\alpha' > \alpha$. For Britain's trading partners, exports will be resource-intensive, meaning $\alpha' < \alpha$. We also expect that the terms of trade will be worsening for the economy whose output is expanding faster.

Normalize the price of home produced goods to 1, and replace the price of the consumption basket with a weighted average of the prices of home production and imports:

$$p_C = \beta + (1 - \beta)p$$

β is the share of consumption produced at home, so lower β indicates a more open economy. p is the price of imports in terms of home production, so that increasing p implies worsening terms of trade. Taking logs and writing the production function in per capita terms, we have:

$$\ln y = -\ln[\beta + (1 - \beta)p] + \ln A - (1 - \alpha')\ln L + (1 - \alpha')\ln R \tag{16}$$

Differentiating and simplifying:

$$\frac{\dot{y}}{y} = -\frac{\beta(1-p)}{\beta + p(1-\beta)} \cdot \frac{\dot{\beta}}{\beta} - \frac{p(1-\beta)}{\beta + p(1-\beta)} \cdot \frac{\dot{p}}{p} + \frac{\dot{A}}{A} - (1 - \alpha')g(y) \tag{17}$$

implying that

$$MSPG = \frac{1}{1-\alpha'}\left[-\frac{\beta(1-p)}{\beta + p(1-\beta)} \cdot \frac{\dot{\beta}}{\beta} - \frac{p(1-\beta)}{\beta + p(1-\beta)} \cdot \frac{\dot{p}}{p} + \frac{\dot{A}}{A}\right] \tag{18}$$

There are thus three effects. The $\frac{\dot{p}}{p}$ term shows that importing products with declining relative prices increases per capita income growth and maximum sustainable population growth. Importing products with rising relative prices decreases per capita income growth and MSPG. The $\frac{\dot{\beta}}{\beta}$ term shows that if imports are cheap compared to home goods, more trade increases income and MSPG, while if imports are expensive, the opposite is true. The $(1 - \alpha')$ term shows that in the absence of a downward sloping demand curve for an economy's products, specializing in a resource-intensive industry exacerbates the resource constraint, decreasing per capita income growth and MSPG, while specializing in a resource-light industry increases

them. If all the effects are present, the price effect and the increasing trade effect will tend to work in the same direction, and the resource share effect will tend to work in the opposite direction. This is because we typically think of new industries as being the growth ones.

In order to simplify notation, we let

$$X = -\frac{\beta(1-p)}{\beta + p(1-\beta)} \cdot \frac{\dot{\beta}}{\beta} - \frac{p(1-\beta)}{\beta + p(1-\beta)} \cdot \frac{\dot{p}}{p} + \frac{\dot{A}}{A}$$

Roughly speaking, X represents the drivers of income growth: changes in trade volumes, terms of trade, and productivity. Equations (17) and (18) then become

$$\frac{\dot{y}}{y} = X - (1-\alpha')g(y) \tag{19}$$

$$MSPG = \frac{X}{1-\alpha'} \tag{20}$$

Note that these equations are dynamic and are true both in and out of steady state.

We can see that reduction of trade barriers will increase per capita income growth if

$$\Delta X > (1-\alpha')\Delta g(y) - (\alpha' - \alpha)g(y) \tag{21}$$

while the condition for it to increase MSPG is different:

$$\Delta X > -X \cdot \left[1 - \frac{1-\alpha'}{1-\alpha}\right] \tag{22}$$

Let us consider the simplified case where trade has no immediate effect on current income y. Conditions (21) and (22) then simplify to:

$$\frac{\Delta X}{\alpha' - \alpha} > -g(y) \qquad \text{(income growth)} \tag{23}$$

$$\frac{\Delta X}{\alpha' - \alpha} > -\frac{X}{1-\alpha} \qquad \text{(MSPG)} \tag{24}$$

The right hand sides of conditions (23) and (24) are only equal in steady state. (This can be seen by reference to equation 19.) If an economy is experiencing income gains,

MSPG may increase while income growth decreases. If an economy is experience income declines, MSPG may decrease while income growth increases. Thus, trade may increase income growth while making it harder to escape from the Malthusian trap, or vice versa. Later, we will see that this odd result is not just theoretical but in fact likely applied to Great Britain during the Industrial Revolution: Britain was experiencing income gains, but was nonetheless suffering immiserating trade that increased the sustainability of its growth.

The effect of trade on the Malthusian economy when terms of trade are not constant is therefore ambiguous. Trade may help or hurt an economy's ability to escape the Malthusian trap, and it provides a channel for breakout to be exported even without the prospect of technology transfer. Even more surprisingly, the analysis provides a graphic demonstration that factors increasing income growth may not necessarily make it easier to escape from Malthusian stagnation.

3. Colonialism, Development and International Transfers

Often mentioned in the same breath as trade are colonialism and international transfers such as foreign aid. This section examines income transfers from one country to another, which encompass some forms of foreign aid and colonialism. Some authors, such as Pomeranz (2000) have argued that colonies were essential to the Great Divergence by providing access to cheap natural resources. However, this misses the essence of colonialism. Colonialism at its core is not the mercantilist policy of importing raw materials and exporting manufactures; that is just trade. The essence of colonialism is that it allows the colonizing state to engage in trade at non-competitive prices and therefore to extract wealth. I therefore model the relationship as an income transfer from the colony to the colonizer. Income transfers are also a good model of the sort colonialism practiced by European empires in other Malthusian societies like China and India. In the ancient context, it might be applied to the Roman conquest of the Mediterranean. Counterintuitively, we will find that *ceteris paribus*, exacting wealth from protectorates or client states actually reduces

sustainable population growth in the conquering country, and aids breakout in the vanquished country. Similarly, receiving a fixed transfer of aid reduces maximum sustainable population growth in the country it is intended to help. This has some topical implications for development policy.

A country receiving an income transfer effectively has a production function:

$$Y = AL^{\alpha}R^{1-\alpha} + T \tag{25}$$

where T is the amount of the transfer from expropriation, tribute or foreign aid.[8] In per capita terms, this yields:

$$y = A\left(\frac{R}{L}\right)^{1-\alpha} + \frac{T}{L} \tag{26}$$

Taking logs, differentiating, and simplifying yields an expression for per capita income growth.

$$\frac{\dot{y}}{y} = \left[\frac{\dot{A}}{A} - (1-\alpha)g(y)\right] - \frac{T}{Y}\left[\frac{\dot{Y}_H}{Y_H} - \frac{\dot{T}}{T}\right] \tag{27}$$

where Y_H is home production (i.e., $Y - T$).

The first term is familiar from earlier. Looking at the second term, we can see that unless the amount of the transfer is growing faster than the home economy, incoming transfers such as tribute or aid actually *decrease* income growth and MSPG and so make it *harder* to escape the Malthusian trap. This counterintuitive result occurs because in order not to be a drag on income growth rates, the amount of the transfer must increase in line with home productivity and population growth.

It is difficult to imagine tribute income satisfying the conditions for increasing MSPG. This is because in order for the last term in (27) to be significantly positive, tribute income and hence the client economy must be both growing faster than the home economy ($\frac{\dot{T}}{T} > \frac{\dot{Y}_H}{Y_H}$) and be large in relation to it ($T/Y \sim 1$). If this is true, it begs the question as to how the colony got conquered in the first place! Exploitation

[8] I have abstracted from the fact that in reality collecting this income requires labor in a process that is subject to diminishing returns.

of another economy is thus unlikely to help and may even hurt the conquering country because it causes per capita income to depend in part upon tribute revenue, which is harder to grow in line with home productivity and population.

Unsurprisingly by now, the effect on the colony is the reverse. Per capita income still obeys equation (27), except that now the transfer is negative. If the tribute is fixed (in real terms), then $\dot{T}$ is zero and the second term is positive if the economy is growing. Thus, after an initial downward shock to income and reduction in population growth, paying a fixed tribute actually *increases* the growth rate of per capita income and MSPG for any given level of them, and makes it easier to cross the threshold out of the Malthusian trap. Even when the economy remains stuck in the Malthusian world, the tribute increases steady state population growth and living standards. The intuition for this result is that paying a fixed tribute adds a source of growth of per capita income: as the population rises, each person pays a smaller share of tribute and so sees his income rise.

Rather than a fixed tribute, it may be more realistic to think that a conquering state would extract tribute in proportion to the economy of the client state. But even then, $\frac{\dot{T}}{T}$ would be precisely equal to $\frac{\dot{Y}}{Y}$, the second term would be zero, and the conquest would have no effect on steady state per capita income in the client economy.

Caution is required with this conclusion, for there are other reasons to believe that being a client state rather than a free state may hinder economic growth. These include being subject to the patron's arbitrary authority and suffering from violent political repression. And if payment of a tribute truly is conducive to break-out, it is possible that the colonial authority would see that explosive growth as a threat and act against it. However, the conclusion that exaction of a fixed tribute from a client state by itself lowers per capita income in that state is not correct in a Malthusian world—surprisingly, it is just the opposite.

This analysis applies equally well to the provision of development assistance. Here, T would be interpreted as a transfer of finished consumer goods, such as food, from the developed world.

From equation (27), we can see that an annual fixed transfer of consumer goods will result in a temporary increase in income when the transfer is implemented, followed by population growth and a return of per capita income to its pre-transfer level. What is worse, the existence of the transfer will make it harder for the economy to grow out of its Malthusian trap—it creates what others have called a "culture of dependency" that drags down per capita income growth as population rises, because the transfer must be split between more people.

Thus, this analysis suggests that aid of finished consumer goods such as food may save lives now, but is counterproductive for takeoff in the future—whether this trade-off is worth it is a moral and ethical question that economists are no more or less qualified than others to answer.

The converse of this is that perversely, paying a tax to the developed world, or indeed simply burning a fixed amount of food each year, would ease the transition to modern growth. Such a policy seems ridiculous of course, but this is because it is not only an obvious a waste of resources, it would also result in an initial period of mass starvation. Only the surviving few would enjoy the subsequent easier transition to growth.

However, this insight does inspire some ideas for government policy in the developing world that do not result in such ruthless utilitarian trade-offs. One topical area is how developing country governments should react in the aftermath of a deadly epidemic, such as AIDS (or less topically, the Black Death), that kills a large portion of the population but leaves capital stocks intact. Diminishing returns means that such an epidemic results in temporarily increased income for the rest of society, as the economy moves to the (out of equilibrium) point D in the figure 5.

One option for the government is to prevent this temporary rise in income and fertility through the imposition of a tax that exactly matches the income increase.

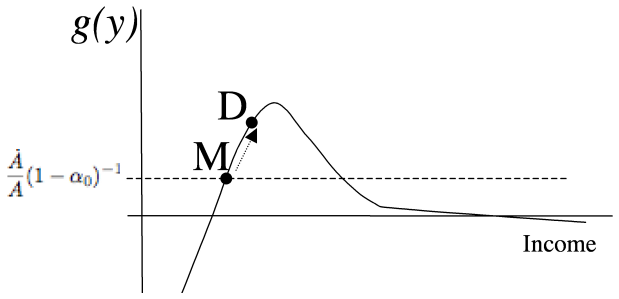

FIGURE 5. The Effect of an Epidemic on Population Growth and Income

This way, the survivors' income will be just what it was before, there will be no fertility increase, and the population growth rate will not increase. The economy will be back at point M, but with a smaller population and a government budget surplus. The government's strong fiscal position is then an asset that can be used to spur growth.[9]

A literature on poverty traps in a purely Solovian framework (*i.e.*, in models with no dependence on fixed natural resources) has led some modern development economists, such as Sachs et. al. (2004) to advocate a "big push," an effort to raise the capital stock of a less developed country to a high enough level that it is able to escape a poverty trap without further assistance. Thus, development assistance is conceived as transfers of capital goods rather than consumption goods.

This strategy shares with my approach a recognition that there is a certain threshold for assistance to make any difference beyond the near term; if this threshold is not reached, the temporary gains dissipate over time. However, in the Sachs et. al. story, the source of the poverty trap is insufficient savings at low levels of income. The aim of the big push is to suddenly move income levels to a high enough level that savings can occur and the economy can undertake a convergence path to the "good" Solow equilibrium rather than the poverty trap equilibrium.

[9]It should be noted that this policy is not politically robust, in that there would be a great temptation to use the resources for other purposes such as aid to the people or yachts for the rulers. However, even if the tax is used for yachts for the rulers, the lesson of this model is that while corruption may inhibit growth because of the institutional damage that may be done to innovation and enterprise formation, the income transfer alone does not hurt an economy's ability to escape from the Malthusian trap.

In the model presented here, the big push could potentially also work, but for a different reason: the increase in the capital stock could raise income enough to effect a demographic transition. It is likely, however, that a calibration would show a considerably larger push necessary in my model relative to Sachs et. al. The reason is that savings rates generally begin to rise at significantly lower levels of income than the demographic transition occurs.

Moreover, in my model, unless the higher savings rate increases the rate of TFP growth, something which Sachs et. al. do not address, even the gains of a big push will likely ultimately dissipate as the economy runs up against diminishing returns to fixed natural resources. Development strategies must take this dynamic into consideration.

Conclusion

So far, I have offered a simple framework for understanding the causes and elements of the transition from a Malthusian economy to a growth economy. Just as Solow's growth accounting allows us to break growth into its component parts, this analysis provides a similar common currency, sustainable population growth, for analyzing contributions to the takeoff to modernity. Unlike Solow's accounting exercise, however, in this paper the component parts can be interpreted as at least proximate causes of the transition.

The model also sheds light on how various factors interact to cause takeoff. Rather than a complex interaction among, for example, resource and technology growth, the model shows how the growth rates of various contributions to the takeoff can simply be added in the MSPG framework, and if the sum reaches a certain threshold, then breakout occurs.

An important corollary is that there is no one prescription for development; an economy need not necessarily look a certain way, nor have a certain quantity of capital, level of education, set of industries, nor arrangement of institutions to achieve takeoff. Rather, wherever the economy starts, modern growth is touched off

by sustained but potentially temporary rapid productivity gains or other increases in maximum sustainable population growth. These gains can be achieved by capital accumulation, technological transfer, education, exploitation of natural resources, structural change, or even, as shown in the appendix, tax cuts, in any combination sufficient to raise maximum sustainable population growth above the peak in the Malthusian population-income function. They can be achieved in combination with factors like population control that act to lower the threshold.

The model then identifies industrialization as a part of one possible path to modern growth. Specifically, the model suggests that productivity gains and structural change accompanying industrialization both raise MSPG. In the chapter two of the paper, we will investigate this in more depth and find that indeed this was the primary process by which the Industrial Revolution set off the sustained rise in living standards we are enjoying today.

While there exist many strategies to development, it bears mentioning that there is nothing in this paper to question Solow's original conclusion: indefinitely sustained income growth requires growth in total factor productivity. Rather, the conclusion is that even sustained growth in TFP is not sufficient to guarantee income growth if the economy is stuck in the Malthusian phase of development. It is the phase change itself that can be accomplished by a diversity of strategies, and does not necessarily require an increase in TFP.

The model also has some surprising implications about what will not touch off growth. While the transition to growth can occur in any size population and at any level of technology, an analysis that includes endogenous capital (see appendix A) shows that it is probably easier in larger, more advanced societies. Exploitation of and exaction of tribute from other Malthusian societies is actually likely to make takeoff harder. And critically, development strategies and forms of aid that do not raise MSPG above the peak in population growth will see temporary gains in income, but those gains will eventually be completely dissipated into population growth.

In short, chapter one of this paper has shown simply how the transition to growth is a binary phase change, and it provides a Solow-style accounting mechanism—the sustainable population growth framework—for the ingredients into that change. It naturally gives rise to an understanding of how both Malthusian and growth economies can exist in the same world, where in some countries economic growth is channeled solely into population growth while in others it results in a sustained increase in living standards. Most importantly, it provides a unified and intuitive framework in which to understand both the industrial revolution and modern development.

CHAPTER 2

Accounting for Breakout in Britain: The Industrial Revolution through a Malthusian Lens

Now that we have seen how the model works in theory, this chapter will apply the model to the breakout that occurred in Britain at the time of the Industrial Revolution. In doing so, we will see that Industrial Revolution was characterized by an unprecedented rise in maximum sustainable population growth from firmly Malthusian levels prior to the end of the 18th Century to levels that dwarfed not only the peak of British population growth but also any rate of population growth that has ever been recorded. The set of economic changes we refer to as the Industrial Revolution brought with it a vast increase in the amount of population growth the British economy could absorb without engendering declining living standards, firmly linking the Industrial Revolution to the beginning of a sustained rise in incomes. We will see that while many of the factors that have been cited by other authors—technology, structural change, coal and capital—made a contribution to lifting the Malthusian constraint, the process was dominated by technology and structural change.

In addition to solidifying the link between the Industrial Revolution and rising living standards, this chapter makes several additional contributions to the economic history literature. First, it reconciles the gradualist and limited Crafts-Harley view of the Industrial Revolution with a dramatic and rapid change in Britain's macroeconomic character. Second, it estimates, in an accounting sense, the contributions of various economic factors to the ability of the economy to sustain income growth. Third, it shows that the link between the Industrial Revolution and the economy's ability to sustain a rise in living standards does not depend on whether the "old" Deane and Cole view or the "new" Crafts-Harley view of the Industrial Revolution

is correct. Although the pace of transition is different in the two views, the general qualitative story is robust to a wide variety of estimates for the various components of economic growth. Fourth, along the way to compiling MSPG estimates I come up with new estimates for total factor productivity during the Industrial Revolution; these generally follow Crafts and Harley but employ improved estimates of factor shares and natural resource growth. The new estimates point to a somewhat larger role for TFP than the most recent estimates put forth by Crafts and Harley. Fifth, the MSPG estimates combined with population growth determine GDP per capita growth, so this paper contains new estimates for English GDP per capita growth from the 14th-17th Centuries.

The MSPG estimates presented here also carry two implications at variance with the conventional wisdom. First, the analysis shows that the effects of trade were complex and are likely to be misunderstood in the literature. In particular, the effect of trade on the ability of the British economy to transition to modern growth is probably exaggerated. Second, we will see that short- or even medium-term economic growth is a different phenomenon than the ability to sustain a breakout in living standards. If we want to understand the end of the Malthusian era, just asking which factors contributed to growth does not necessarily shed light on this question. Rather, the question we must answer is "What allowed the British economy to transition, possibly fairly abruptly, from a regime where per capita income was trendless to one where it was growing at 1% per year?" The difficult thing to understand is not the "growth" but the "transition."

Finally, it is important to recognize that this paper does not purport to investigate ultimate causation of the Industrial Revolution. Rather, it is an accounting exercise in the spirit of Solow, linking the economic changes that were part of the lndustrial Revolution to the accompanying rise in living standards. While it would be correct to consider the factors identified in this paper as the direct causes of breakout, they are likely to be linked to each other as part of the broader underlying

process of the Industrial Revolution. Although some cursory thought is given to these links here, a thorough examination of them is beyond the scope of this paper.

The first two sections of this chapter discuss some data issues surrounding the measurement of pre-1800 living standards and to estimating the existence of a Malthusian relationship between living standards and population. It is hoped that the calculations in this paper can be improved considerably once current research being undertaken by Broadberry *et. al* (constructing U.K. national accounts prior to 1700) and Taylor and Wrigley (on British occupational structure in the early modern period) is in more finished form. The next several sections explain data sources and estimation methods for the various components of MSPG and contain some discussion of their relative importance. Section 8 combines the data to estimate MSPG in Britain and offers some insights into the connection between the Industrial Revolution and the rise in living standards. Section 9 provides new GDP estimates for Britain from 1300-1700. Section 10 relates my findings to the findings of Broadberry et. al. (2010), who are currently developing new national accounts for Britain from 1270-1870. As a contrast to the British case, in section 11 I calculate MSPG for Golden Age Holland.

1. A Puzzle: Living Standards Before 1800

We aim to explain the rise in individual living standards before and during the Industrial Revolution. Considering economy-wide evidence, it appears convincing that beginning around 1650, productivity improvements and institutional change in agriculture combined with other factors to give rise to a sustained increase in British per capita output that has not ceased since. Yet, microeconomic evidence does not bear this out—real wages were roughly stagnant from 1550 through the first quarter of the 19th century, and even broader measures of personal incomes show only modest growth from 1688 to 1801. This paper moves part of the way towards resolving this discrepancy; however, its perniciousness shows just how serious the data issues still are.

Estimates of per capita GDP all show a considerable rise in living standards between 1500 and 1820. The longest available series, Maddison's (2001), shows that GDP per capita rose from $714 (1990 USD) in 1500 to $1705 in 1820, a 2.39-fold increase. This is based on a simple extrapolation of Crafts' growth rates prior to 1700. Van Zanden (2004), using a simulation-based approach, concludes that English per capita GDP grew 82% from 1500 to 1800. For the period from 1700-1800, Crafts (1985) estimates per capita GDP growth of 29%, presented as a revision of Deane and Cole's (1967) estimate of 50% over this period.

Puzzlingly, these significant increases in GDP per capita do not show up in data on personal incomes and wages. Averaging Allen's series on wages in different parts of the country (Allen, 2001) shows that real wages fell by a factor of three from 1400 to 1550 and did not begin to rise again until the second decade of the nineteenth century. Clark (2005) similarly finds that real wages of building craftsmen did not rise over the course of the 18th Century or the longer period from 1550 to 1800, and going back 50 more years shows that they fell 20% from 1500 to 1800. The Phelps Brown and Hopkins (1981) wage series is more volatile but also confirms these findings.

While it may be difficult to believe that GDP per capita roughly doubled over the three centuries to 1800 without any increase in real wages, it is likely that the period did see a rise in inequality. Labor's share of GDP fell from approximately 70% in 1500 (Allen, 2005) to approximately 55% around 1800 (Allen, 2007). This can account for about 1/3 of the discrepancy between estimates of per capita output growth and wage growth over the period. Maddison suggests that payments in kind could have accounted for some of this difference, and Clark and Allen do not appear to have taken this into account. Because the relative price of agricultural commodities rose compared to industrial goods (Allen, 2001), the failure to account for payment in kind would understate wage growth.

There are other good reasons to believe that wages may not be an appropriate measure of the rise in living standards. First, as has already been alluded to, wages

are not a complete measure of incomes. Second, Lindert and Williamson's (1982) revisions of social tables indicate that only about half the working population were primarily wage earners over the period from 1688-1801.

More troublingly, the Lindert and Williamson tables, which are an attempt to reconstruct GDP using the income approach and should therefore include not only wage income but also returns to capital and land rents, indicate that real incomes across social classes were less than 5% higher in 1801 than in 1688, and that in contrast to most wage and GDP estimates, incomes for most social groups *fell* in the first half of the 18th Century and then *rose* from 1759 to 1801.

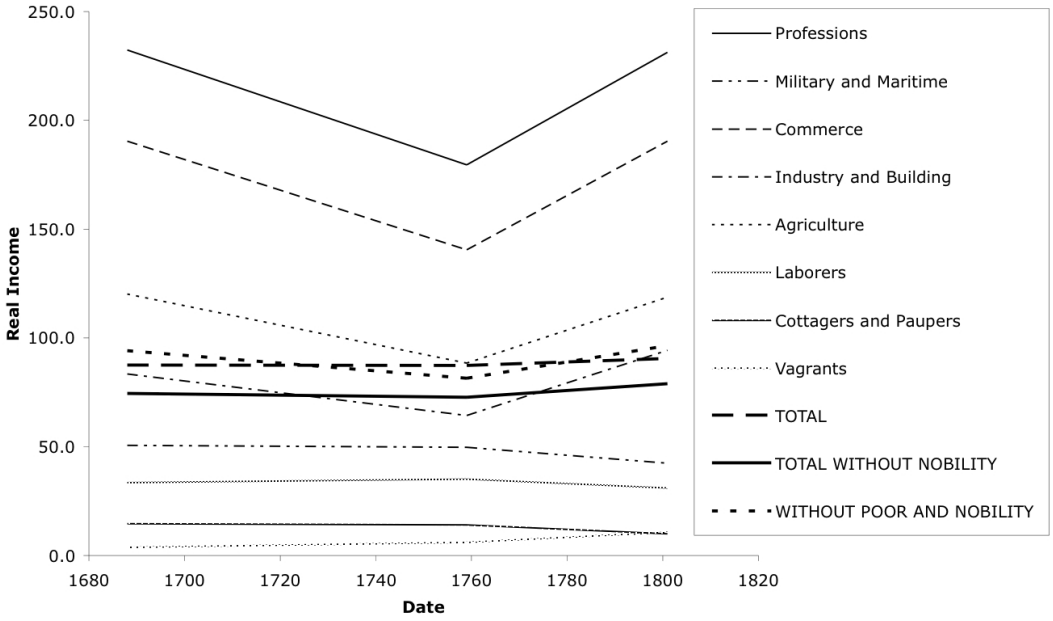

FIGURE 1. Real Family Incomes 1688-1801 (1801£ p.a.)
Sources: Lindert and Williamson (1982) for income, Allen (2001) for price index.

The wide discrepancy between various measures of income in pre-industrial England should cast serious doubt over any econometric exercise attempting to link living standards to population growth before the late 18th Century. As a result, the applicability to Pre-Industrial Britain of the theory presented in chapter one cannot be tested in an econometric sense until better data become available. Nonetheless,

it does provide an intuitively appealing parable that sheds light on what changed before and during the industrial revolution.

2. The Fertility Function

The theoretical section of this paper has proposed a definite functional relationship—$g(y)$—between per capita income and population growth, presented as a manifestation of the Malthusian mechanism by which higher incomes enable higher fertility and reduce mortality. Despite the data shortcomings, several authors have investigated whether a Malthusian link of this kind exists, as well as the reciprocal link by which higher population causes lower living standards. All of these investigations use wages rather than GDP per capita as a measure of living standards, despite the incompleteness of this approach. Unfortunately, however, before the beginning of the end of the Malthusian era in 1700, wages are at present the "only game in town;" data on per capita GDP simply are not available on any reasonable or consistent time scale prior to this.

The strength of the Malthusian links in Britain varies significantly depending on the timescales examined. However, all authors I am aware of who attempt to test a Malthusian framework[1] find a positive relationship between wages and population growth over the short term in Pre-Industrial Britain. Over the long term, however, population growth is dominated by factors other than wages. In other words, while $g(y)$ appears to be well defined over short timescales, the functional relationship itself changes over longer ones. However, despite the finding that population growth may vary exogenously to wages, the literature concludes that the "homeostatic" Malthusian feedback mechanism placed a constraint on the Pre-Industrial British economy that prevented a sustained simultaneous increase in both population and living standards. While I rely on previous findings as to the existence of the Malthusian causal relationship between wages and population growth, I also provide a "sanity check"

[1]Other authors such as Boserup instead consider a framework where a larger population drives technological progress, which may lead to anti-Malthusian conclusions. It has also been noted by many authors (e.g. Lindert, 1985) that nominal prices explain real wages as well as population does. There has yet to be proposed a convincing explanation for this.

by plotting a raw estimate of the $g(y)$ function based upon data from the 1541 to the present and find that it matches the qualitative form proposed.

Perhaps the seminal paper in the modern econometric literature testing Malthusian hypotheses is Lee (1973). As is standard throughout the literature, Lee uses wages as a measure of living standards. Lee tests both elements of the Malthusian hypothesis using data on wages and population at 25-year intervals for the population-wage link and 50-year intervals for the wage-population growth link. For the link this model depends upon, wages driving population growth, he finds zero correlation between the two variables, concluding instead that over time periods on the order of 50 years population growth is largely exogenous to wages. This may appear to deal a severe blow to the theory here outlined, but in fact it does not.

First, it is not altogether surprising that during a period that includes the Black Death, the dominant impact on mortality and population growth would not be the level of real wages but exogenous influences such as the plague. In our theory, events such as the plague should be viewed as shifts in $g(y)$. Yet even without the impact of the plague, the correlation between the two series is still essentially zero, with population growth driven by seemingly exogenous changes in mortality.

Second and more importantly, what Lee has really shown, as he later acknowledges (Lee, 1993), is that the equilibrium population level changes quickly compared to the speed at which the homeostatic process drives the population towards equilibrium. Indeed, the homeostatic process over the period from 1540-1870 had a half-life of 107 years (Lee and Anderson, 2002). Crafts and Mills (2007) split the period, arguing that there was strong homeostasis (half-life 19 years) prior to 1645 and extremely weak homeostasis (half-life 431 years) from 1645-1800. Lee's (1993) analysis concludes that "population size and wages are determined at any historical moment mainly by more recent shocks, as homeostasis slowly overcomes the influence of older ones."(p. 1) If only small and short-term movements were necessary to effect a breakout to a growth regime, this finding would negate the theory proposed in this

paper. However, although we found in chapter one that temporarily pulling population growth down below MSPG temporarily could effect the transition to growth, in practice with the very low capacity for sustainable population growth that existed prior to the Industrial Revolution the shift in the population growth function would have had to last for centuries. A shift of this durability never occurred. Because the peak in the population growth function is quite wide, as discussed below, horizontal shifts would be insufficient to push the economy over the hump unless MSPG was already quite close to the peak.

To put a finer point on it, Lee (1993) finds that in England, the standard deviation of wages was about 20% from its Malthusian equilibrium at any given time. In order to effect the transition to growth using changes in the $g(y)$ function, we will see that wages would have had to reach at least double but more likely several times their Malthusian equilibrium, an event that is dozens of standard deviations away from the mean and the probability of which is practically zero. Shifting the level of income at which peak fertility occurs thus will not effect the transition; the peak itself must be reduced.

We are thus left with three timescales. Over the very long term, the homeostatic force constrains population and wages to fall within a Malthusian range. In Lee's words: "Even weak homeostatic tendencies, quite possibly undetectable over decades or centuries of apparently random population movements, must in the long run be decisive in determining the general range in which population size and wages fell." (1993, p. 2)

However, even broad swaths of European history are likely to be too short for the homeostatic force to be dominant. Over the medium term, exogenous fluctuations in trend rates of mortality and fertility (due to environmental, institutional or other factors), what Lee and Anderson (2002) call the "intercept" in the Malthusian wage-population growth relation, determine population growth rates. The homeostatic mechanism is operative but its parameters are changing more quickly than it is operating so it is extremely difficult to detect.

Over the short term up to a couple of decades or so, mortality and fertility functions change slowly, so the parameters of the homeostatic mechanism linking living standards to population growth do not change, and the wage-population growth link can be detected again. This is shown, for example, by Wrigley and Schofield (1981) and by Weir (1984).

The key insight is that while the function linking income to population growth, $g(y)$, is changing over time, its contemporary peak must be exceeded by MSPG for a sufficiently long period of time to raise living standards enough that a transition to a modern growth economy occurs. As we continue with our analysis, it will become clear that this only occurred in Britain with the explosive increase in MSPG that accompanied the Industrial Revolution. While it is theoretically possible that a shift in $g(y)$ due to changing mortality trends or fertility practices could have increased living standards enough to short-circuit the "normal" breakout process, in practice such a shift would have had to be either implausibly large or favorable demographic trends would have had to persist unceasingly for several hundred years, neither of which occurred.

Despite the econometric limitations of attempting to develop a simple link between wages and population growth, it turns out the variation of fertility and mortality with income since 1541 is sufficiently stable to allow us to plot a raw estimate of $g(y)$, the relationship between income and population growth. Figure 2 plots population growth from Wrigley and Schofield (1981) against the real wage from Clark (2005) from 1541-2000.

This plot shows a relationship of the form proposed in chapter one.[2] The period from 1580-1650, which appears to depart from the functional relationship, is the exception that proves the rule; during this period, wages were particularly low relative to GDP (see section 5), so if fertility is related to total per capita income it should have been higher than what would be expected based on wages alone. The long tail

[2]As discussed, prior to 1541, this relationship or at least its parameters were likely different—Lee (1973) finds a correlation between wages and population growth of -0.03 using 50-year periods from 1250-1750.

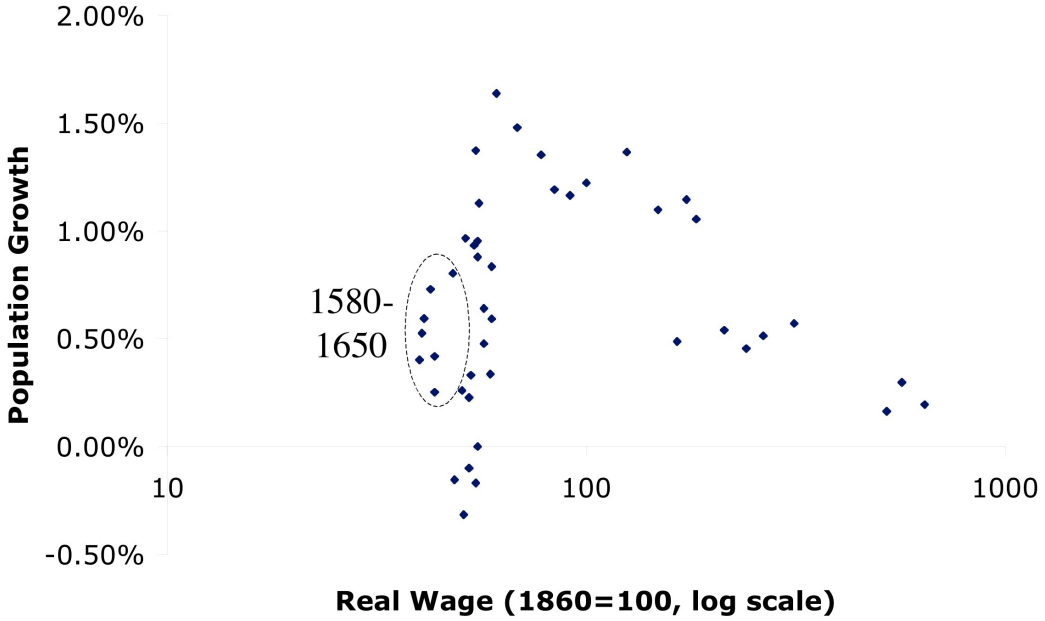

FIGURE 2. The Relationship Between Real Wages and Population Growth
Sources: see text.

indicates that if a country were stuck in the Malthusian era (at point M in figure 2 in chapter one), a simple shock to income would have to be very large before it pushed the economy over the hump and past point G into the modern growth regime. Indeed, figure 2 suggests that even with a relatively high Malthusian MSPG around 1% (which was not seen until the mid-17th century and certainly did not exist in the aftermath of the Black Death), per capita income would have to double to effect the transition to growth. For rates of MSPG that prevailed prior to this time, it was only with real wages at late-20th-Century levels—8-10 times what prevailed in the early modern period—that a growth economy could have been sustained.

3. Total Factor Productivity from 1300 to 1860

Total factor productivity, or A in the model, is one of the most important factors in allowing sustainable population growth, and along with structural change, it was one of the two largest contributors to the breakout that occurred at the beginning of the 19th Century. While TFP growth is quite difficult to measure precisely, the

available data indicate that prior to the scientific revolution in the mid-17th Century TFP growth was extremely slow and may have been dominated by exogenous events like climate change. The 150 years from 1650, including the agricultural revolution, saw modest but consistently positive rates of TFP growth in England. Then, beginning with the industrial revolution in the first half of the 19th Century, TFP growth slowly accelerated to modern levels.

Estimates of TFP growth prior to the advent of modern statistics are quite unreliable and difficult to come by. The estimates that have been used in this paper are based on those by Allen for the years prior to 1700 and by Crafts for the years after 1700.

Allen (2005) estimates total factor productivity in agriculture in 1300, 1500 and 1700, and he also estimates (Allen, 2000) labor productivity for 1400 and 1600. In order to estimate TFP for these years, I assumed TFP followed a similarly shaped path to labor productivity.[3] Table 1 shows Allen's estimates for TFP and labor productivity, and my interpolated estimates for 1400 and 1600. There is assumed to be no TFP growth outside of agriculture prior to 1700[4] so the agricultural TFP growth rate is multiplied by the share of GDP in agriculture, which is estimated as described in section 5. The top three lines of the table present index numbers at the turn of the centuries; the bottom three lines are averages over the century beginning with the date at the column heading.

Crafts (1985, 1995) and Crafts and Harley (1992) estimate real GDP growth for select periods from 1700-1860. I use their residual approach to calculate TFP from 1700-1860 (see Crafts, 1985, pp. 78 ff.), making use of Wrigley and Schofield's population estimates, Feinstein's (1988) estimates of capital stocks, and my estimates

[3]See appendix C for details of this calculation.

[4]This may appear to be an extreme assumption but in the absence of data I believe it to be the most appropriate. First, it avoids repeating Deane and Cole's error of assuming the sector with notable productivity growth (in their case textiles, in this case agriculture) was representative. Second, Crafts (1985) estimates industrial and commercial output grew at an annual rate of 0.70% from 1700-1760, while the population outside agriculture grew at a rate of 0.64% (Wrigley and Schofield, 1981). This leaves little room for productivity growth even in the 18th Century, and there was probably less in earlier times.

	1300	1400	1500	1600	1700
Allen TFP	0.83		1.00		1.38
Allen Labor Productivity	0.8	0.92	1.00	0.76	1.15
Interpolated TFP		0.94		0.83	
Estimated Ag TFP CAGR for Century	0.12%	0.06%	-0.18%	0.51%	
Proportion of GDP in Agriculture	77%	75%	70%	60%	
Estimated Economy TFP Growth	0.09%	0.05%	-0.13%	0.30%	

Sources: see text.

TABLE 1. Calculation of Total Factor Productivity pre-1700

of natural resource growth and factor shares as described in sections 5 and 6 below. Combining these figures with the Allen estimates gives rough estimates of TFP growth from 1300 to 1860. This time series is shown in table 2 and it will serve as our estimate of TFP growth for the rest of this paper.

	1300-1400	1400-1500	1500-1600	1600-1700	1700-1760	1760-1780	1780-1800	1800-1830	1830-1860
Allen/Tepper	0.09	0.05	-0.13	0.30					
Crafts/Harley/Tepper					0.32	0.03	0.40	0.55	0.86

Sources: see text.

TABLE 2. Annual Growth Rates of Total Factor Productivity

These estimates of TFP growth are close to Crafts' original (1985) estimates and generally somewhat higher than Crafts and Harley's revised figures. This does not stem from any fundamental disagreement with Crafts and Harley over the progress of the economy but rather reflects the need to treat land and capital as separate factors. The underlying GDP figures are the same.

4. Trade

Economic historians tend to agree that trade played an important role in the British Industrial Revolution, but there the agreement ends. Acemoglu, Johnson and Robinson (2005) argue that trade helped create and sustain the institutions necessary for economic growth. Harley and Crafts (2000, 2002) argue that trade was a critical ingredient in the exceptional structural change that took place in Great Britain away from agriculture. Pomeranz (2000) argues that trade with (or exploitation of) the New World provided the raw materials to ease the Malthusian

constraint. Davis (1979) argues that trade was not a primary factor in triggering the industrial revolution, but was critical in sustaining demand that allowed it to continue. This paper does not have implications for indirect effects of trade on growth such as that proposed by AJR, an effect that may coexist with those described here; the present theory examines the direct role of trade. Theoretically, the most likely role admitted for international trade in the British case is that it may have allowed specialization in non-resource-intensive activities, which would have raised maximum sustainable population growth and eased the transition to growth. In practice, however, trade probably did not have a large effect on MSPG until at least the second quarter of the 19th Century, well after the industrial revolution and the attendant rise in living standards was well underway.

Recall equation (17) from the first section of Chapter 1, governing income growth in an open economy, reprinted here for convenience:

$$\frac{\dot{y}}{y} = -\frac{\beta(1-p)}{\beta + p(1-\beta)} \cdot \frac{\dot{\beta}}{\beta} - \frac{p(1-\beta)}{\beta + p(1-\beta)} \cdot \frac{\dot{p}}{p} + \frac{\dot{A}}{A} - (1-\alpha')g(y)$$

Recall also that p is the price of imports and β is the share of imports in consumption. Thus, if a country's terms of trade worsen as trade increases, then trade may decrease income growth and *ceteris paribus* MSPG. Yet all else is not equal. Trade is associated with specialization, and in the case of Britain, this meant specialization in industry over agriculture. This kind of specialization reduces $(1-\alpha)$, the resource share, raising MSPG. There are thus two competing effects.

The contribution of trade to structural change in Britain was undoubtedly significant, a point that has been made many times in the literature. Figure 3 shows net exports as a share of GDP by sector from 1700 to 1855.[5] By the middle of the 19th Century, Britain was importing more than 15% of its GDP in resource-intensive food and raw materials, and using non-resource-intensive manufactures to pay for them. This likely resulted in a significant reduction in the all-important rent share in the British economy, although estimates of this are quite sensitive to assumptions

[5]Trade data are from Davis (1962, 1979). Export to GDP ratio from Crafts (1985), p. 131 with interpolation.

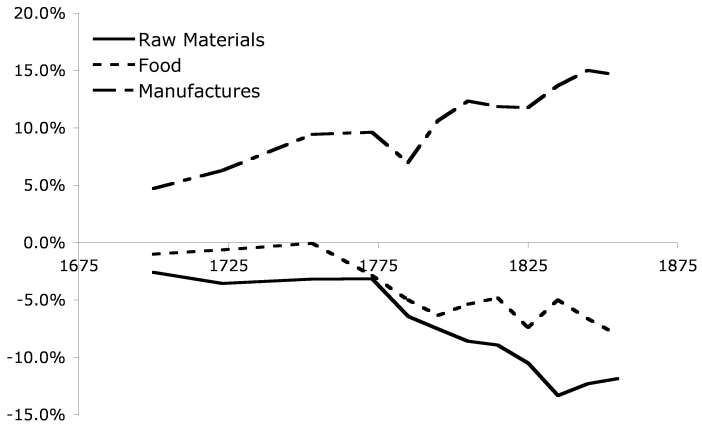

FIGURE 3. Net Exports / GDP by Sector

about the counterfactual situation. In the late 18th Century, this effect was smaller but likely still significant.

The simplest way to analyze the effect of trade on the rent share is to simply compare the British economy under autarky to the economy with trade, holding the consumption basket constant. While this makes a useful benchmark, it is not correct for two reasons. First, under autarky the price of agricultural goods would rise relative to that of manufactured goods. In response, consumers would substitute away from agricultural goods so that the consumption basket would consist of a higher proportion of manufactured goods. Second, a shift to more agricultural production at home would raise the marginal product of land and hence land rents. These two effects work in opposite directions, with the first tending to cause the simple comparison to overstate the effect of trade on the rent share and the second tending to cause the simple comparison to understate the effect of trade on the rent share. Table 3 shows the effect of trade on the rent share from 1700 to 1850.

Simply comparing trade with autarky holding the consumption basket constant suggests that trade did not have much of an effect on enabling takeoff in Britain. At the time the breakout occurred at the turn of the 19th Century, trade reduced the rent share in the British economy by an estimated 3-4 percentage points, enough to

increase MSPG from 3.8% to 4.3% in the first third of the 19th Century.[6] However, this was still a small proportion of MSPG at this time.

1700	1723	1753	1773	1785	1795	1815	1845
-0.8%	-0.7%	-0.4%	-1.8%	-3.1%	-3.9%	-3.3%	-4.5%

Sources: trade figures from Davis (1962, 1979). See text for calculation method.

TABLE 3. Effect of Trade on the Rent Share (Percentage Points)

For another approach to examining the effect of trade on MSPG, we can begin with Crafts and Harley's (2002) computable general equilibrium model, which simulates the effect of trade on the British economy from 1770-1841. This model takes into account both relative price effects and substitution effects described above, and when its results are put into the MSPG framework they are even more interesting.

The Crafts-Harley simulation considers the effect in 1841 of freezing agricultural imports at the 1770 level. Surprisingly, they find that when all effects are considered, the restriction of trade would have actually *increased* 1841 national income per capita by 8%, and hence its 1770-1841 growth rate by 0.1 percentage points per year. This is because the deterioration in British terms of trade over the period more than offset the production gains from specialization.[7] However, Crafts and Harley also find a large effect on factor prices, enough to increase the rent share in national income by a multiple of 1.55, from 12% to 18%. The net result of restricting trade would have been a decrease of 0.3 percentage points in MSPG to 4.0% for the first third of the 19th Century and a decrease of 3.6 percentage points to 9.6% for the second third of the century.[8] Nonetheless, in both periods MSPG with restricted trade would still have been well in excess of the peak in population growth.

We are thus left with the extremely odd result that freer trade from 1770-1841 decreased per capita income but nonetheless facilitated the transition to a modern

[6]Rents in agriculture are assumed to be 50% of trade values; rents in raw materials are assumed to be 10% of trade values; rents in industry are assumed to be zero. This is in line with Allen (2005) and Clark and Jacks (2007)

[7]See Harley (1994, p. 305) for British terms of trade during the 19th Century.

[8]These calculations are carried out using equation (11).

growth regime. The intuition for this is that what is ultimately important for escaping the Malthusian trap is reducing the economy's dependence on natural resources, combined with technological growth. Trade can ease the dependence on natural resources, but a large economy must pay a price for this in terms of worsening terms of trade. Ability to escape the Malthusian trap is inversely proportional to the rent share in production in the economy; income growth is not so simply related to the rent share.

In this section we have seen that in contrast to the conventional wisdom, the direct effects of trade were likely not a major factor in effecting the transition to a modern growth economy until well the middle of the 19th Century, by which time modern growth was well underway. Equally importantly, we have in trade an example of why it is important to ask precise questions about the industrial revolution: freer trade decreased per capita income in England but still contributed somewhat to its ability to escape the Malthusian trap. This is a graphic demonstration of the fact that "What factors increased economic growth?" and "What caused the economy to transition from a Malthusian regime to a growth regime?" are not really the same question. A tie between an economic phenomenon and general growth is not sufficient to establish that that phenomenon aids in shaking off Malthusian constraints.

5. Structural Change

Structural change, the movement of the labor force from agriculture into industry and from the country to the city, is often remarked upon as a feature of industrialization and the beginning of modern growth. Most of the literature, however, focuses on structural change as a *result* of the growth in living standards or of factors driving that growth.

The most obvious link between structural change and growth is that as agricultural productivity improves, fewer people are needed to produce the food supply for society, and there is surplus labor that must find other employment. As the

increased agricultural output can support a larger population, there is no need for that population on the land and it must migrate to cities and towns. While there is debate about what exactly drives people off the land (see, for example Weisdorf, 2006; Crafts and Harley, 2002 for different views on the topic), structural change is viewed as a consequence of or accompaniment to growth, rather than an important ingredient in itself.

While Mendels (1970) argued that "proto-industrialization," the move into handicrafts rather than (or in addition to) food production, set the stage for the growth of modern industry, his arguments are based on cost pressures spurring technological improvement. Mathias (1989) similarly argues that proto-industrialization allowed the development of industrial and entrepreneurial skills, a commercial infrastructure, and capital. Harley (1994) discusses how Britain's focus on rapidly-advancing industry became an engine of growth. These arguments, however, tend to be made in broad brush strokes and offer a more tentative rather than firm link between structural change and growth. They do not make it clear whether these channels were an idiosyncrasy of the British experience or whether they somehow represent an integral part of the process.

This paper views structural change as an integral part of breakout, as we saw in the discussion of trade in the previous chapter. Structural change is linked to breakout because it reduces the dependence of the economy on land and other natural resources that are constrained in the Malthusian sense.[9] While there are many measures of structural change, the one that matters for breakout is the factor shares in the economy. Specifically, the inverse resource share acts as a multiplier on MSPG, so that reducing the importance of land in the economy can have a large effect on sustainable population growth.

[9]Although it is not modeled as part of the main paper, it is easy to create a feedback loop so that structural change is endogenous in the model. Intuitively, if agricultural productivity increases slowly, then the entire productivity increase is absorbed into supporting a larger population. However, if agricultural productivity increases more quickly, then as incomes rise people wish to spend on goods other than agriculture (the income elasticity of food demand is less than one). This leads to people moving off the land, decreasing the rent share, which increases per capita income growth, which leads to a lower share of agriculture in GDP and more people moving off the land, which leads to a lower rent share, a virtuous circle.

Structural change may encompass not only a shift to less resource-intensive production but a shift to more capital-intensive production. Recall equation (10), which allows us to include capital in the model as follows:

$$MSPG = \frac{1}{\gamma}\frac{\dot{A}}{A} + \frac{\beta}{\gamma}\frac{\dot{k}}{k} + \frac{\dot{R}}{R} \tag{28}$$

where γ is the resource share in the economy and β is the capital share. As the resource share, γ, falls, if the move is into capital-intensive production, then β increases at the same time, magnifying the effect. A shift of this kind took place during the industrial revolution, with the capital share increasing from about 0.17 to 0.4 over the century and a half from 1700 to 1850, while the rent share decreased from about 0.26 to 0.08. $\frac{\beta}{\gamma}$, the impact of capital deepening on the ability to sustain population growth, thus increased by a factor of more than seven over this period.

Estimating the production elasticities β and γ is fraught with pitfalls. The simplest way to do so, and the approach taken by most authors, is to assume the pre-Industrial and Industrial Revolution British economy was approximately competitive and proxy the elasticity of production by the share of GDP paid to each factor. That is the approach taken here, although it is recognized this is a strong assumption.

For the rent share, I use Allen's (2007) estimates after 1760 and construct my own estimates prior to this date. These estimates, along with Allen's estimates for the capital share, are provided in table 4.

	1300-1400	1400-1500	1500-1600	1600-1700	1700-1760	1760-1780	1780-1800	1800-1830	1830-1860
Rent	0.22	0.14	0.26	0.28	0.24	0.22	0.19	0.16	0.11
Capital					0.17	0.19	0.20	0.32	0.39

Sources: see text.

TABLE 4. Factor Shares in National Income

The estimates prior to 1760 are based upon Allen's (2005) estimate of the share of agricultural income going to land and Wrigley's estimates of agricultural population (1987, p.170 and 2006, p. 468). Table 5 summarizes this calculation. I assume all land rents derive from agriculture prior to the 18th Century, as rents paid on coal

mines even in 1860 were at most 4% of agricultural rents, and were negligible before this time.[10]

Allen provides estimates of the distribution of agricultural income for 1300, 1500 and 1700. Van Zanden (2005) cites estimates implying that GDP per capita outside agriculture was 9% higher than in agriculture in 1290 and 16% higher in 1688. These figures are used to translate Wrigley's population share estimates into GDP share estimates. The share of GDP in agriculture is then multiplied by the share of agricultural production to obtain the share of GDP paid to land. This provides estimates for 1300, 1500 and 1700.

To get estimates for intervening dates requires some rather heroic assumptions. The rent share fell by half from 1300 to 1500. I assume this was due entirely to the Black Death in the middle of the 14th Century, so the rent share in 1400 was the same as the rent share in 1500. To determine the rent share over the 17th Century, I use Clark's (2002) rent index combined with the assumption that gross land area under cultivation was constant over the century. Combined with data on population growth, and the rent share in 1700, this uniquely determines GDP growth, which is estimated to have been 33% from 1600 to 1700.[11] This implies a rent share in 1600 of 37%.

Interpolations are made for the period from 1700-1760 to give the final rent share estimates for use in the MSPG calculation.

	1300	1400	1500	1600	1650	1700
Share of Population in Agriculture	80%		76%	70%		55%
Share of GDP in Agriculture	79%		74%			51%
Rent Share of Agricultural Income	0.39		0.19			0.51
Rent Share in GDP	0.30	0.14	0.14	0.37	0.25	0.26

Sources: share of population in agriculture from Wrigley (2006) and Wrigley (1987). Rent share in agricultural income from Allen (2005). See text for calculation methods. Share of population in agriculture for 1500 is Wrigley's 1520 estimate.

TABLE 5. Rent Share in National Income Estimates

[10]See Clark and Jacks (2007) for data on coal rents.
[11]See appendix C for the details of this calculation.

As a check on these figures, the implied rent share in 1700 is consistent with what be implied using a bottom-up estimate: multiplying the rent per acre by the number of acres (Clark, 2002) and dividing by nominal GDP (as implied by Crafts, 1985 and Davis, 1962) yields a rent share in GDP of 28%. My estimates for share of GDP in agriculture for 1700 are generally higher than those calculated by Crafts (1985, p. 16), but his estimates are based on income of workers and do not appear to include agricultural rents.

The estimates in table 4 show a large reduction of the rent share—by nearly a factor of three—from the 17th Century to the middle of the 19th Century. Because many components of MSPG are inversely proportional to the rent share, this structural change in the economy approximately doubled sustainable population growth over the period.

6. Coal and Land Improvements

Coal's importance in the Industrial Revolution remains a subject of debate. The New Economic History tends to view coal as playing a smaller role, in contrast to the earlier consensus that the move to an inorganic economy was a major driver of industrialization. Clark and Jacks (2007) have recently reviewed the current state of the debate and it does not bear repeating here. This analysis concludes that while coal could theoretically have played a large role in easing the Malthusian constraint, a quantitative investigation indicates that it probably did not do so, at least directly. If coal was a major force, this was because its location gave Britain a competitive advantage in industry that drove structural change and because fixed resources (coal mines or woodland) play a relatively smaller role in the coal production process than in the timber production process, rather than because of any vast expansion in the resource base.

Coal increased MSPG in two ways. Most significantly, it substituted for timber, which required land, the Malthusian fixed factor, for its cultivation.[12] Secondarily, because rents on coal producing land were proportionately less than rents on agricultural land, it reduced the factor share of land in the economy.[13] The substitution for timber meant that coal increased the effective amount of land that Britain had to support its population; land that previously was needed to grow timber could be used to grow food, and the demand for energy that would have needed to be met by additional timberland could be met by coal.

The English and Welsh coal mining industry grew from virtually nothing in 1500 to an annual output of 75 million tons in 1860 (Church, 1986). The scale of this change is immense: to provide this much energy using timber would have required 93 million acres devoted to timber production,[14] more than twice the land area of Britain. Determining the effective easing of the land constraint, however, can be slightly more complicated. Coal energy is recognized to have been of a lower quality than charcoal, and coal mines are not as versatile as land in production of final goods. Additionally, because coal mining is a different production process than agriculture, rent was typically a much lower share of gross produce. Exactly what adjustments should be made for these differences is open to debate. This section conducts a number of estimates based on different methodologies, and then justifies the one that will be used for the remainder of the paper. The estimates vary widely, but we will consider that one acre of agricultural land is equivalent to about six tons of coal production per year.

[12]It may be argued that once nearly all energy came from coal, coal mines could no longer substitute for agricultural land because people cannot eat coal. However, coal could be and was exported, and thus held down prices of land-intensive goods on the world market, which Britain imported.

[13]Adam Smith comments on this in *The Wealth of Nations*: "Rent, even where coals afford one, has a generally smaller share in their price than in that of most other parts of the rude produce of land. The rent of an estate above ground commonly amounts to what is supposed to be a third of the gross produce ... In coal mines a fifth of the gross produce is a very great rent; a tenth the common rent ... Thirty years' purchase is considered as a moderate price for the property of a landed estate, [but] ten years' purchase is regarded as a good price for that of a coal mine." (p. 272) Clark and Jacks (2007) also figure that coal rents were typically just 7-10% of the pithead production.

[14]I use an estimate of 0.81 tons of coal per acre of woodland.

The simplest method is to simply calculate the amount of woodland that would be required to produce the same quantity of energy as British coal. One pound of air-dried wood (20% moisture content) contains about 6,400 BTU of energy, a figure which varies little by species of wood and can be found in a variety of sources.[15] British coppice was largely populated with oak, ash, beech, hornbeam and hazel, which are relatively dense woods ranging from 40-50 lbs./cu. ft. of solid wood when air dried.[16]

I assume that an acre of woodland yielded 78 cubic feet of dry wood per year. This is in line with the high end of modern coppice estimates (Crockford and Savill, 1991). Warde (2007) notes that the Forest of Dean in the 1690's yielded 4 cubic meters per hectare, which corresponds to 57.2 cubic feet of green wood per acre, while himself using a lower figure of 3.3 cubic meters per hectare. Hammersley (1973) claims that an acre of woodland could produce up to 100 cubic feet per year. Using Warde's yield figure would imply a produce of woodland of only about 15 s./year, which is much less than Allen's (2005) figure of £2 per year for agricultural produce and even less than his figure of £1 per year for agricultural rents. My figure implies that the gross produce of woodland was 31 s./year, which makes it less productive than intensively farmed agricultural land, but not too much so.

These figures imply that woodland produced about 22.5 MBTU of energy per year. I follow Warde in assuming that British coal, which was largely bituminous, yielded 27.7 MBTU per ton, in line with modern sources. An acre of woodland thus yielded as much energy as 0.81 tons of coal.

Other estimates of this figure varied widely. Hatcher (1993) estimates 0.5-1 ton per acre, Clark and Jacks (2007) estimate 0.9 tons per acre, and Allen (2009) estimates 1.3 tons per acre. Warde's (2007) figures imply an equivalence of 0.38 tons

[15]See, for example, the U.S. Department of Energy at http://bioenergy.ornl.gov/papers/misc/energy_conv.html.

[16]See, for example, Glover (2003), a British hardwood timber merchant at http://www.nichetimbers.co.uk/native-hardwood/british-oak/, http://www.simetric.co.uk/si_wood.htm, and the "Chimney Sweep Online" at http://www.chimneysweeponline.com/howood.htm.

per acre. All of these figures appear to have some problems in their approach. Their assumptions are presented in table 6 below.

	Allen (Green Wood)	Allen (Dry Wood)	Clark and Jacks (Dry Wood)	Warde (Green Wood)	Tepper (Dry Wood)
Wood Yield (cu. ft./acre)	128.0	107.6	97.5	47.2	78.1
Energy Content (BTU/cu. ft.)	140,000	166,500	251,000	185,000	288,000
Energy Content (BTU/lb.)	3,493	4,859	8,600		6,400
Density (lbs./cu. ft.)	40.1	34.3	29.2		45
Coal Energy Content (MBTU/cu. ft.)	23.6	23.6	26.9	27.7	27.7

Sources: Allen (2003), Allen (2009), Clark and Jacks (2007), Warde (2009).

TABLE 6. Assumptions Underlying Various Coal/Timber Equivalence Estimates.

Allen appears to have estimated timber yields in terms of stacked cubic feet per acre rather than cubic feet of solid wood per acre, as he has a much higher yield by volume but a lower density. Allen's energy content figures also appear to be about 25% too low for wood and 15% too low for coal, which may be due to an unstated assumption that these fuels does not burn completely efficiently. Finally, while Allen's (2009) text says that an acre of woodland is equivalent to 1.3 tons of coal (p. 96), his figures appear to imply the opposite: a ton of coal is equivalent to 1.3 acres of woodland. Adjusting for this error means that Allen's estimate of 1.3 becomes 0.76.

Warde appears to have understated both the yield of English woodland as well as the energy content of solid wood. His figures imply that the produce of English woodland in monetary terms was far below that of other agriculture.

Clark and Jacks' figures appear to be the least problematic. Their combination of energy content and density figures suggest they are using bone dry or oven dry wood, which is lighter and contains more energy on a per pound basis. However, their estimate of yields per acre is therefore at the extreme upper end of the range, especially as wood tends to shrink somewhat as it dries.

In any event, my figure of 0.81 tons of coal per year as equivalent to one acre of land implies that English and Welsh coal production by 1860 was equivalent to 93 million acres. Combined with advances in agriculture, coal could thus be argued

to have increased the effective land area of England and Wales from 17.5 million acres in 1300 to 110 million acres by 1860.[17] By this method, British effective land area increased faster than population over this period. At least at first glance, there is a case to be made that coal and land improvement alone lifted the Malthusian constraint facing Britain and allowed it to emerge as the first growth economy, even without technological progress.

Coal, however, was a less desirable fuel than charcoal and therefore sold at a discount. A better approach is therefore to determine the acreage necessary to produce timber of the same value as British coal output. This approach has the advantages that it implicitly takes into account a measure of land quality, and it reflects Britain's ability to trade coal for timber on the world markets; the acreage equivalence was not purely a theoretical matter.

According to Clark's price indexes, a ton of coal sold for about 4.1 s./ton in 1800, which agrees with Crafts' (1985) estimate that British coal production was worth £2.7 million in 1800. A cord[18] of firewood sold for 31 s. At 0.95 cords/acre, an acre's worth of firewood production cost as much as 7.2 tons of coal. This implies that coal was equivalent to an additional 10 million acres of land by 1860.

A more macroeconomic method is to compare the total output of the coal industry to the agricultural output. In 1860, coal output for Great Britain (including Scotland) was approximately £24 million[19], or 3.6% of GDP[20], compared to 18% of GDP in agriculture (Deane and Cole). Thus, English and Welsh coal mines produced as much additional GDP as did 7 million acres of agricultural land, implying a coal-land equivalence of 11 tons per acre. This method yields estimates of 8-10 tons per acre for earlier periods.

Finally, we can directly compare land rents in agriculture to site rents for coal mines. Agricultural rents per acre ranged from 35 times coal site rents per ton in the

[17]See Allen (2005, table 1) for effective land area calculations.
[18]I assume 82 cu. ft. of solid wood per cord after accounting for air space, which is what is implied by Allen's figures.
[19]This figure combines the pithead price of coal from Clark and Jacks (2007) with output figures from Church (1986).
[20]GDP is sourced from Deane and Cole (1967).

1710's to 70 times coal site rents in the 1860's (Clark and Jacks, 2007 for coal rents; Clark, 2002 for agricultural rents). This method would therefore imply an equivalence of 35-70 tons of coal per year per acre. This method could be justified by the principle that land value should be viewed as the best measure of land productivity in an efficient market. However, I use the sale price of coal for comparison, as it is the entire production process, rather than just the ownership of the coal, that is a substitute for the agricultural production process.

In this paper, I will use the figure of 7.2 tons of coal production per year as equivalent to one acre of woodland, which uses the result that a cord of firewood was 7.6 times as expensive as a ton of coal in coal in coal-producing areas, while an acre of woodland yielded 0.95 cords of wood. Combined with improvements in land quality as discussed by Allen (2005, table 1), coal increased the effective land area of England and Wales from 17.5 million acres in 1300 to 45 million acres in 1860. The table below shows the contributions of coal and land improvement to effective land area. The best estimate figure considers 7.2 tons of coal as equivalent to an acre of woodland, as described. The high estimate uses Warde's figure of 0.38 tons of coal as producing energy equivalent to an acre of woodland, without any adjustment for prices. The low estimate compares the land with coal mines on the basis of 1800 rents, with an equivalence of 50 tons per acre.

			Coal Equivalent Acreage			Total Effective Land Area			Effective p.a. Land Area Growth (following period)		
	Effective Land Area (Exc. Coal)	Coal Production (Tons MM)	(Low Est.)	(Best Est.)	(High Est.)	(Low Est.)	(Best Est.)	(High Est.)	(Low Est.)	(Best Est.)	(High Est.)
1300	17.5					17.5	17.5	17.5	0.04%	0.04%	0.04%
1500	18.9	0.1	0.0	0.0	0.2	18.9	18.9	19.1	0.15	0.16	0.26
1700	25.6	2.5	0.1	0.4	6.7	25.7	26.0	32.3	0.25	0.26	0.47
1750	28.9	4.5	0.1	0.6	11.9	29.0	29.5	40.8	0.33	0.38	0.98
1775		7.9	0.2	1.1	20.7	31.5	32.4	52.0	0.34	0.40	1.10
1800	34.0	13.0	0.3	1.8	34.3	34.3	35.8	68.3	0.04	0.19	1.48
1830		27.4	0.5	3.8	72.0	34.7	37.9	106	0.10	0.55	2.64
1860	34.2	75.1	1.5	10.4	198	35.7	44.7	232			

Sources: land area from Allen (2005). Coal production from Hatcher (1993), Flinn (1984) and Church (1986). See text for conversion factors.

TABLE 7. Coal Production and Effective Land Area in England and Wales

Effective land area growth contributes to sustainable population growth on a one-for-one basis, so it can be seen that while coal and land area growth contributed a moderately to MSPG, they were not in the best estimate case sufficient to cause a dramatic breakout. It is only at the very highest end of the estimates, and even then only beginning in the mid-19th Century, that coal could itself have been sufficient to raise MSPG above the threshold for breakout. Even then, other contributions to MSPG were of equal or greater magnitudes.

Coal may, however, have contributed to the Industrial Revolution in ways beyond the scope of this model. In addition to easing the land constraint, the discovery of coal provided the areas of Britain in the vicinity of the coalfields with extraordinarily cheap energy because of the lower transport costs (see Allen, 2009). Clark and Jacks find that in the absence of coal, the alternative would have been imported timber at about double the average domestic price; the multiple would likely have been significantly higher in the industrial centers located near coalfields. Perhaps the cost of energy to industry would have quadrupled without coal. Coal thus gave those parts of Britain with easy access to it a competitive advantage in energy-intensive industry, so it is no surprise that such industry did in fact grow up around the coalfields. Independent of its contribution to increasing effective land area, economic growth, or import revenues, coal thus contributed to structural change in the British economy. This structural change was another significant factor in raising MSPG. Quantitatively, however, we have only been able to estimate the amount coal added to Britain's natural resources base. In this respect, it was at most a modest factor in enabling the rapid takeoff in living standards attending the Industrial Revolution.

7. Capital Deepening

The technological change of the Industrial Revolution brought with it an increase in investment and therefore significant capital deepening. I use estimates of capital stocks from Feinstein (1988) post-1760 and follow Crafts (1985) in assuming that

prior to 1760 the capital to GDP ratio was constant. This assumption still allows some role for capital deepening, however, since GDP generally grew faster than the population in the 18th Century. Table 8 shows the figures on capital growth, both gross and per capita, that are used in MSPG calculations.

	1700-60	1760-80	1780-1800	1800-30	1830-60
$\Delta K/K$	0.7%	0.63%	1.30%	1.73%	2.48%
$\Delta k/k$	0.37%	-0.05%	0.26%	0.29 %	1.30%

Sources: see text.

TABLE 8. Growth in Capital Stock and Capital Per Worker

Prior to 1700, capital per worker is assumed to be constant outside of agriculture. In agriculture, any GDP changes due to capital per worker show up as increases in TFP.

8. Putting It All Together

One of the great questions about the Industrial Revolution is how, or whether, British industrialization relates to the beginning of the sustained rise in living standards that occurred at approximately the same time. To the extent that the Industrial Revolution refers to a shift in production away from agriculture into industry, and into more technologically advanced industrial production processes requiring higher capital intensity and relying on coal for energy, we are now able to draw a clear link between these economic changes and the beginning of sustained growth in living standards.

Figure 4 plots MSPG in Britain compared with population growth, calculated according to equation (28) using the estimates made in this chapter. In post-medieval British history, we can see that it was not until the last two decades of the 18th Century that sustainable population growth significantly exceeded the peak in the population growth function for any sustained period. Indeed, by the middle of the 19th Century, the economy would have been able to sustain a population growing

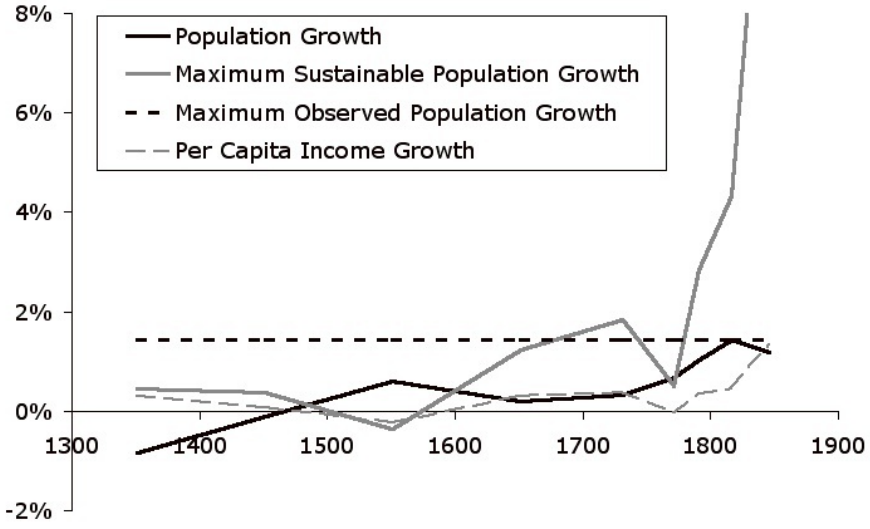

FIGURE 4. Maximum Sustainable Population Growth from 1300 to 1860

at more than 13% per year—nearly ten times the maximum observed rate of population growth and seven times the rate that could have been sustained at any time prior to the Industrial Revolution.

These results naturally explain Crafts and Mills' (2007) finding that the British economy exhibited strong homeostasis prior to 1645 and extremely weak homeostasis thereafter. Crafts and Mills tested an equilibrium version of a Malthusian model where wages and population are always in their Malthusian equilibrium. By contrast, the present model is a dynamic one that explains the behavior of the economy even in out-of-equilibrium states. Prior to 1645, MSPG was very nearly zero, so a traditional Malthusian model linking wages to population should perform relatively well. After 1645, MSPG was very nearly at the threshold for breakout, meaning the traditional Malthusian mechanism operated only very weakly.

Although higher frequency data would be needed to confirm this, the chart also suggests that from the 17th Century onwards the British economy was slowly and smoothly building toward the ability to achieve breakout, with the "knee" of the curve reached at the Industrial Revolution. While at first glance the chart shows a dramatic change at the end of the 18th Century, it could just as easily be interpreted

as a smooth curve from 1600 onwards with technological progress in agriculture flowing into technological progress in industry as part of a single phenomenon, save only for a depression from 1760-1780. That depression was characterized by a slowdown in TFP growth that was likely related both to the Seven Years War from 1756-1763 and the American Revolution from 1775-1781.

In any event, by the beginning of the 19th Century, the Malthusian constraint had been completely and—depending on which view of the 1760's and 1770's one takes—suddenly eliminated. This finding is consistent with Lee and Anderson (2002) who find (p. 217) a "very sharp and discontinuous rise in the rate of increase [of labor demand] starting around 1810." Their definition of the rate of increase of labor demand is the rate at which new labor can be absorbed by the economy; this is MSPG viewed through the econometrician's lens.

Maximum sustainable population growth is also the only measure of capacity I am aware of that shows such a dramatic change over precisely the time period of the Industrial Revolution and the beginning of the increase in living standards. TFP growth, for example, did not reach truly unprecedented levels until the middle of the 19th Century. The MSPG framework thus reconciles the Crafts-Harley limited view of the Industrial Revolution with a large and discontinuous change in the carrying capacity of the economy, which was linked to the beginning of a sustained increase in living standards.

Viewing the takeoff through the lens of MSPG also reconciles the timing of the economic change and expansion in output—the last two decades of the 18th century—with the timing of real wage growth, which did not begin in earnest until a few decades later. The model naturally produces an Engels Pause as the economy moves over the hump in the population growth function, which causes rapid growth in output but much lower growth in output per worker.

The estimates upon which MSPG is based are open to debate and measurement error. Fortunately, the broad, qualitative story of breakout and how and when it occurred does not change dramatically if different estimates are used. Figure 5 plots

a low, high and best estimate of MSPG from 1300 to 1860. The high estimate uses the highest estimates of GDP growth (Deane and Cole's), assumes that productivity growth in the non-agricultural part of the economy mirrored productivity growth in agriculture prior to 1700, employs the energy-equivalent method for coal mining with Warde's figure of 0.38 tons of coal per acre, and uses the lowest rent share estimates available. The low estimate uses the lowest estimates of TFP growth I found (from Antras and Voth, 2000), the rent-equivalent method for coal mining (implying an equivalence of 50 tons per acre), and the highest rent share estimates available. It should be recognized that the range presented does not represent a confidence interval in any sense. Rather, it is intended to show that the overall story is not sensitive to which "view" of the Industrial Revolution one takes. The narrow interval in the early years of the chart is indicative not of strong confidence in the figures but of a paucity of varying estimates in the literature.

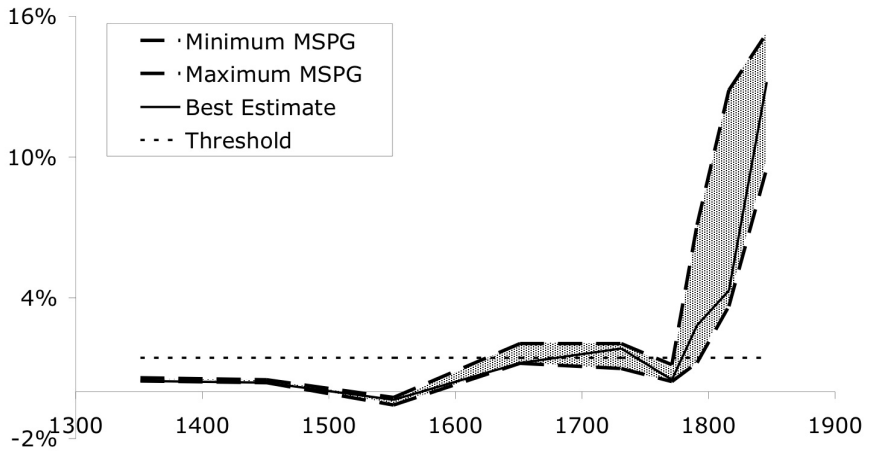

FIGURE 5. Range of MSPG Estimates, 1300-1860

Figure 6 plots the components of MSPG. These are also shown in Table 9. The figure shows that during the Industrial Revolution, MSPG was largely attributable

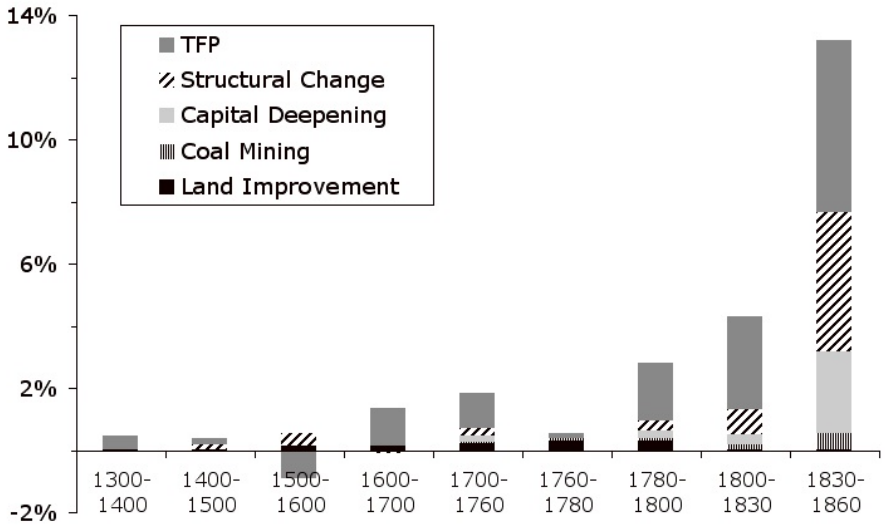

FIGURE 6. The Components of MSPG

to technological advances, structural change in the economy, and by the middle of the 19th Century, a shift to more capital-intensive production. The shift to more capital-intensive forms of production played a more minor role prior to 1830; to the extent that a capital story is to be told about the Industrial Revolution the chart leaves open the possibility that an upgrading of the quality of capital may have been an important part of enabling breakout. At least in an accounting sense, coal played only a minor part in the takeoff.

	1300-1400	1400-1500	1500-1600	1600-1700	1700-1760	1760-1780	1780-1800	1800-1830	1830-1860
TFP	0.40	0.28	-0.49	1.06	1.27	0.16	1.94	3.01	6.04
Structural Change	0.02	0.06	-0.01	0.00	0.05	0.01	0.25	0.70	3.73
Capital Deepening					0.26	-0.04	0.24	0.41	2.87
Coal Mining	0.00	0.00	0.01	0.01	0.02	0.06	0.08	0.18	0.54
Land Improvement	0.04	0.04	0.15	0.15	0.24	0.32	0.31	0.01	0.01
MSPG	0.46	0.38	-0.35	1.22	1.83	0.51	2.83	4.31	13.19

Sources: see text.

TABLE 9. The Components of MSPG (Percentage Points Per Year)

We finally have, therefore, a clear link between the changes of the Industrial Revolution and the beginnings of sustained income growth. The key to breakout is the ability to sustain non-immiserating population growth; MSPG is the measure

of this ability. We have seen that the Industrial Revolution, with its technological improvements and growth in less resource-dependent sectors, caused a clear and unprecedented increase in maximum sustainable population growth as early as 1780, well before a large effect was seen in wages and GDP per capita.

The MSPG framework explains how such a phase change can take place without unprecedented levels of per capita income growth, so that even though the rates of income growth seen from 1780-1830 were at most a tenth of a percentage point greater than those seen a century earlier, the massive change in MSPG shows that the growth in the Industrial Revolution period was part of a new and unprecedented process of breakout.

9. An Ancillary Benefit: New British GDP Estimates for 1300-1700.

Maximum sustainable population growth estimates can be combined with population estimates to produce estimates of per capita GDP growth. This is equivalent to building GDP estimates based on growth in factors of production, growth in TFP, and factor shares. MSPG can be converted into per capita GDP growth using equation (11), implying GDP per capita growth for the 14th through 17th Centuries as shown in table 10.

	1300-1400	1400-1500	1500-1600	1600-1700
MSPG	0.46	0.38	-0.35	1.23
Population Growth	-0.83	-0.11	0.60	0.21
Per Capita GDP Growth	0.29	0.07	-0.24	0.29

Sources: population growth from Hatcher (1977) and Wrigley and Schofield (1981). Calculations as described in text.

TABLE 10. Per Capita GDP Growth, 1300-1700, Percentage Points per Year

These estimates imply significantly slower growth and hence higher per capita GDP in earlier years than either the Maddison or the Van Zanden figures. According to my estimates, per capita GDP in England grew just 38% from 1500 to 1800, compared with Van Zanden's simulation estimate of 82% growth. Maddison's estimate of 139% growth from 1500 to 1820 compares with my estimate of 52%. Estimates

after 1700 will be identically equal to the Crafts-Harley estimates, which were used to determine TFP for the period as a residual.

These estimates go most of the way towards resolving the puzzle of falling wages against rising GDP per capita I described in section 1, although there is still some explaining left to do. After accounting for the change in factor shares, Van Zanden's estimates of GDP per capita are 82% above what would be expected based on the Clark (2005) wage series; mine are only 37% above what would be expected. Thus the new estimates explain more than half the discrepancy.

10. Relation with Ongoing Research

This paper has been written concurrently with a large research project: "Reconstructing the National Income of Britain and Holland, c.1270/1500 to 1850," which is being led by Steven Broadberry and others. That project has aimed to create annual GDP estimates by sector in England from 1270 to 1850, and will also include new information on occupational structure. Much of this project has now been completed in working paper form, and while sufficient data are not yet available to warrant a full revision of the estimates in this paper, it is possible to combine GDP estimates from Broadberry et. al. (2010) with estimates of the share of agricultural income that accrued to land from Allen (2005) to obtain headline estimates of MSPG based on the new figures in the framework of equation (11). This is shown in Table 11.

	1400-1450	1450-1480	1480-1553	1553-1600	1600-1650	1650-1700	1700-1760	1760-1780	1780-1801	1801-1830	1830-1861
Per Capita Income Growth	0.10	-0.44	-0.07	0.12	0.05	0.56	0.31	0.19	0.56	0.25	1.22
Rent Share	8	7	7	7	19	16	14	14	15	14	11
Population Growth	-0.14	0.29	0.54	0.67	0.45	-0.08	0.32	0.62	0.97	1.43	1.18
MSPG	1.09	-6.15	-0.48	2.36	0.71	3.48	2.61	2.01	4.74	3.24	12.63
Addenda:											
Sectoral Share in Agriculture	42	36	36	37	38	31	27	27	30	28	22
Share of Agricultural Income to Land	19	19	19	19	50	50	50	50	50	50	48

Sources: Broadberry et. al. (2010), Broadberry et. al. (2011), Allen (2005), author's calculations.
All figures in percentage points or percentage points per year, as appropriate.

TABLE 11. MSPG Using Broadberry et. al. Data, 1400-1860.

The central finding of this paper, that the decades around the turn of the 19th Century saw a dramatic lifting of the Malthusian constraint, reflected as an explosion of MSPG, is unaltered. However, Broadberry et. al. find considerably less change in the sectoral breakdown of GDP over the period from 1380-1800, with the share of GDP in agriculture estimated to be as low as 42% as early as 1380 and falling only somewhat to 31% of GDP by 1800. This is due both to a finding of a smaller share of the workforce in agriculture and a finding that labor productivity in agriculture was much lower than in services and industry. This leads to lower estimates of the rent share in national income, and as a result, implies MSPG that is both higher and more volatile than my estimates.

Broadberry et. al. also report significantly higher nominal GDP estimates than had been found previously, revising Deane and Cole's estimates up by 20% even as late as 1870. This contributes to lower rent share estimates as well.

Figure 7 shows MSPG estimates using the new figures compared to my estimates. I also plot an estimate that uses Broadberry et. al.'s GDP figures with my estimates of the rent share: this shows that the rent share accounts for most of the difference between the two series.

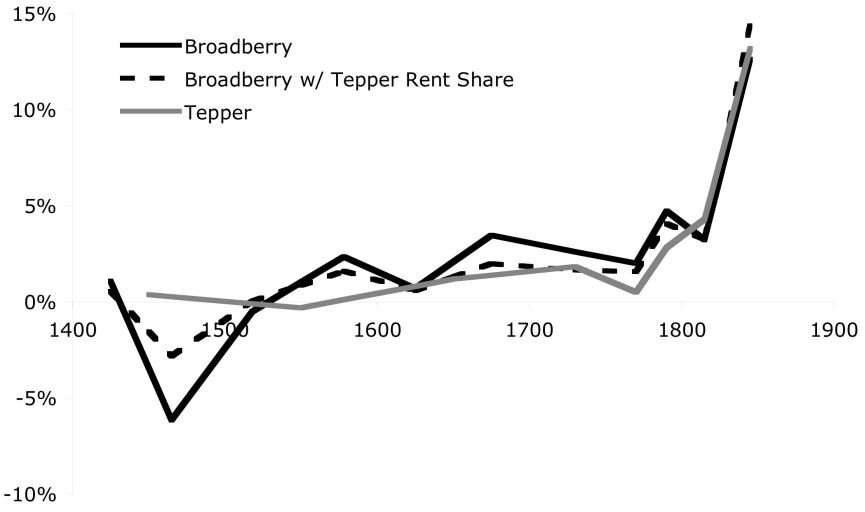

FIGURE 7. MSPG Using Figures from Broadberry et. al.

The other important source of variance between Broadberry et. al.'s estimates and mine is that Broadberry et. al. take a more optimistic view of the 16th and 17th Centuries than I do. I find that living standards fell during the 16th Century and then retraced these losses during the 17th Century, while Broadberry et. al. find flat living standards during the 16th Century and a rise of nearly 40% during the 17th Century. This is somewhat difficult to reconcile with evidence that real wages fell 25% or more from 1500 to 1700, although the discrepancy between wage growth and GDP growth is an old puzzle.

11. Holland as a Foil

Almost as old as the question, "What caused the English Industrial Revolution?" is "Why didn't it happen elsewhere first?" Perhaps foremost among these "elsewheres" is 17th Century Holland.[21] Dutch living standards in 1650 had reached as high as English ones in 1800. Its agricultural productivity was high; it exported a variety of industrial goods and international services; it had just come off three quarters of a century of sustained per capita income growth. Then, around 1650, this growth came to a fairly abrupt end. If early 17th Century Holland was, as De Vries and Van der Woude (1997) argue, the "first modern economy," where did it all go wrong? While this brief analysis cannot answer this question, it can give some clues, not only into why Holland did not continue its growth but into the crucial ways that its economy differed from that of England a century and a half later.

Prior to the Industrial Revolution in England, improvement in living standards was the result of the accumulation of factors of production, typically land but also capital, in relation to labor. While Holland did enjoy a "Golden Age" of improving living standards, and the 17th Century Dutch economy was, as De Vries and Van der Woude note, "modern" in its structure and institutions, the features of Holland's exceptional *growth* during the period were not different from those of other pre-Industrial booms. Rather, Holland's growth in the first half of the 17th Century was only the most spectacular example of growth based on factor accumulation.

[21]For clarity, the analysis in this section is for Holland, a province of the United Netherlands.

Herein lies Holland's difficulty, for even in a society unbound by Malthusian constraints, growth based on factor accumulation can only be temporary; it must eventually run into diminishing returns. This is the key insight from Solow's seminal 1956 paper. For growth to be sustained, it must be based on growth not just in labor productivity but in total factor productivity: the ratio of outputs of the production process to inputs. Holland's growth was based almost entirely on capital accumulation, and even at the height of its advance there was little or no role for TFP.[22] Thus, we should not view it as a surprise that early-modern Holland's growth came to an end, for that growth was nothing more than a classic adjustment to a new Solovian equilibrium based on a higher savings rate.

As a contrast to England, table 12 presents MSPG and its components in Holland from 1500-1805. Data on growth in GDP and factors of production, as well as factor shares, comes from various publications by Van Zanden and De Vries and Van der Woude, as detailed in Appendix B. Total factor productivity is computed as a residual in the usual way, and MSPG is calculated according to equation (11).[23]

The table shows clear differences between the Dutch and English patterns of development. In the Dutch case, there was never any TFP growth, even during the

[22]While De Vries and Van der Woude argue that there was substantial innovation in Holland, this innovation did not show up in TFP using their figures and Van Zanden's. This is not to say, however, that some innovation was not necessary to move to more capital-intensive forms of production, even if TFP was not increased. It is rather that the innovation did not enable Holland to make more efficient use of capital and labor generally.

[23]In preparing this analysis, I have relied on the data from Van Zanden and De Vries and Van der Woude, essentially without adjustment. However, while doing the analysis it became clear that these data are not fully consistent. I have fixed several of the inconsistencies, although there is still one that is pernicious. First, rent indexes prior to 1580 do not match the path of the rent share in the Dutch economy. The indexes appear to be overstated by a factor of two to three and they have been altered to ensure that a rent share of 6.4% in 1510/14 is consistent with a rent share of 4.9% in 1812/13. Second, Van Zanden's (1992) implied price level for 1500 is not consistent with other estimates, including his own later ones, and has been adjusted downwards by 40%. Third, the population levels implied by Van Zanden's capital stock and capital stock per capita figures are not consistent with current best estimates. I chose whether to rely on the aggregate or per capita figures based on which Van Zanden appeared to originally estimate.
Finally, Van Zanden's figures indicate that the total land value of Holland increased by a factor of 2-3 from 1650 to 1800, while comparing the rent share with GDP would imply an increase of only 1/3, even after accounting for the decline in interest rates. I have been unable to resolve this discrepancy, and it could have a potentially large impact on the estimated rate of capital deepening.

	England 1600-1700	England 1800-1830	Holland 1500-1580	Holland 1580-1650	Holland 1650-1805
GDP per Capita Growth	0.29%	0.46%	0.0%	0.7%	-0.06%
Population Growth	0.21%	1.44%	0.5%	1.0%	-0.08%
TFP Growth	0.30%	0.55%	0.0%	0.0%	-0.67%
Capital Deepening		0.29%	0.1%	1.4%	1.0%
Natural Resource Growth	0.16%	0.27%	0.06%	0.3%	0.0%
Capital Share		0.32	0.36	0.51	0.62
Rent Share	0.28	0.16	0.07	0.07	0.05
MSPG	1.22%	4.3%	0.6%	10.6%	-1.4%

See Appendix B for Dutch data sources. The English data are sourced from the preceding analysis. English TFP growth for 1600-1700 includes growth due to capital deepening, which was approximately 0.02% per year in agriculture. I do not have data for English capital deepening outside agriculture for this period.

TABLE 12. MSPG in England and Holland.

so-called "Golden Age" from 1580-1650.[24] Dutch growth was never "modern" in the sense of being driven primarily by growth in the Solow residual. Perhaps the real story, though, is the more than 60% decline in total factor productivity from 1650-1805. This probably reflects declining rates of return on foreign capital as well as the collapse of Dutch industry over this century and a half. Had Holland managed not to suffer this long decay in TFP, it would have enjoyed an MSPG of 12%.[25] Under the theory outlined in this paper, this would have resulted in income gains of 2.3 times, or 2.5 times if fertility did not respond, giving it a level of wealth in 1800 that was not reached by the UK until the turn of the 20th Century. Even had this occurred, however, there is no guarantee that growth could have been sustained. To do so, the Dutch would have had to continue to deepen the capital stock against ever-diminishing returns until fertility fell to modern levels.

This is not to say that the modernity, in the sense that De Vries and Van der Woude mean it, of the Dutch economy is not relevant to the ability to sustain growth. In terms of the statistics we have been examining, modernity as they see it means

[24]The capital figures used here are for private capital. If we subtract off Dutch government debt (see Van Zanden, 1992), the figures tell a slightly different story; TFP growth in Holland from 1580-1650 increases to 0.3% and the rate of capital deepening falls to 0.9%. While this shows a role for TFP growth, it is the same as English TFP growth over the same period and thus does not mark Holland out as being remarkable. Dutch growth is still a story of capital accumulation.
[25]MSPG is calculated according to equation (10) with zero TFP change and other figures unchanged.

a low rent share and high capital share. This implies that even modest growth in the capital stock or TFP is sufficient to lift the Malthusian constraint, an effect that can be seen formally from equation (28). Indeed, the capital deepening from 1580 to 1650 caused a temporary but complete elimination of the Malthusian constraints binding Dutch society; its contribution to MSPG is given by

$$\frac{\eta_K}{\eta_R}\frac{\dot{k}}{k} = \frac{0.51}{0.07} \cdot 1.4\% = 10.2\% \tag{29}$$

As a result, the Dutch population tripled over the century and half from 1500 to 1650, an achievement that was unmatched in England at any time prior to 1700 to 1850, in many ways an analogous period.

The behavior of Holland's economy from 1650 to 1805 is open to interpretation. Population growth was virtually zero despite the high level of income. Despite the lull in fertility, whether the country had truly undergone the demographic transition we typically think of as accompanying development is uncertain; Holland followed this century and a half of slowly declining population by increasing its population from 783,000 at the turn of the 19th Century to about 6 million today,[26], with population growth rates well in excess of 1% through the middle of the 19th Century.[27]

Ultimately, however, whether Holland had or had not undergone a demographic transition by the mid-17th Century may be immaterial. Because the structural change in the Dutch economy was largely complete, the stagnation of the post-Golden-Age period looks much more like a modern stagnation than a Malthusian one. Had Holland not suffered such a drastic decline in TFP, its low rent share would have ensured that a large proportion of the capacity gains from the growing capital stock would have been reflected in income even had the population been growing at the peak British rate of 1.5%.

The growth accounting figures we have examined suggest that Holland's early-modern Golden Age was destined to be temporary for Solovian rather than Malthusian reasons. The role played by TFP is a fundamental difference between Holland's

[26]See Statistics Netherlands, http://www.cbs.nl.
[27]http://www.populstat.info.

growth and England's, although if Holland had not been so unlucky as to be accumulating capital at a time of declining world interest rates, it could have fared much better in the century and a half following its Golden Age. While Holland, like any economy, needed efficiency improvements to sustain *indefinite* growth in living standards, the application of the MSPG model suggests that—due to the low level of land dependence in the economy—its *Malthusian* shackles were already quite loosely fitting even at the beginning of the 16th Century.

Conclusions

It seems obvious that there was in some sense a fundamental change in the British economy at the turn of the 19th Century, a change that at least in the popular imagination, has long been characterized by a small number of trends and inventions: the steam engine, cotton gin, coal, the reorganization of production into factories, migration off the land and into cities. Yet one of the great mysteries of the New Economic History is that as prime causes of wealth, these icons of the Industrial Revolution have proved ethereal when investigated quantitatively.

Here, we were able to use the framework of sustainable population growth to reconcile the older and popular notion of a cataclysmic, revolutionary change in the economy with the relatively limited size of the changes in most economic variables. During the Industrial Revolution, the rate of population growth that the British economy could sustain without declining living standards increased from less than 2% to more than 13%. This change meant the virtually complete elimination of Malthusian constraints, so that technological advances and capital investments could be used to increase incomes rather than population. Paradoxically, however, contributions to income growth and contributions to takeoff are not necessarily equivalent, which points to the importance of asking the right questions when we seek to understand what changed in the British economy.

In an accounting sense, the proximate causes of the increase in the British economy's carrying capacity were the twofold increase of total factor productivity growth

combined with an even greater decline in the economy's dependence on land from 1780-1860. Technology, particularly technology that improves total factor productivity, emerges as an especially important factor when England is compared to Holland. The commonly cited factors of coal and trade, despite the drastic increases in volumes, do not appear to have nearly the same immediate impact on MSPG. Trade does not seem to have had a large effect until the middle of the 19th Century, by which time the process of modern growth was well underway, while coal's direct contribution to the economy never appreciably increased sustainable population growth. Those wishing to argue for a strong role for coal must therefore argue that coal somehow had an impact far in excess of what it was compensated for.

While the accounting exercise demonstrates which proximate causes of breakout were most important and how those causes contributed to the economy's ability to sustain population growth, it tells us little about the underlying causes of the broad change. It is important, therefore, to view Industrial Revolution as a set of interrelated changes. The takeoff of coal mining, increases in capital per worker, rise in international trade, changing structure of the economy and increases in TFP all occurred roughly contemporaneously; it is hard to think of this as a coincidence. This observation points to some obvious next questions to take our understanding of the Industrial Revolution one level deeper:

- What is the connection between the changes in the economy that occurred at the time of the Industrial Revolution and allowed the transition to a sustained growth in living standards? If they were part of a single underlying process, what set this process off?
- As the agricultural revolution enabled more efficient food production, what caused the new commodities not also to be land-intensive?
- What caused the acceleration of TFP growth around 1800, or 1650 if the process is viewed as having been continuous with the agricultural revolution?

Many authors have spent careers examining questions related to these, but the analysis above brings new clarity to precisely which facts we are attempting to explain when we examine the sustained rise in incomes that began with the Industrial Revolution.

APPENDIX A

The Model with Endogenous Capital

We can also explore the effect of including capital accumulation in the model. However, this turns out not to make a big difference. Consider the model:

$$Y = AL^{\alpha}K^{\beta}R^{\gamma},\ \alpha+\beta+\gamma=1 \tag{30}$$

$$\frac{\dot{L}}{L} = g(y) \tag{31}$$

$$\dot{K} = sY \tag{32}$$

This model can be solved exactly as before to yield:

$$\frac{\dot{y}}{y} = \frac{\dot{A}}{A} - (1-\alpha)g(y) + \beta\frac{\dot{K}}{K} \tag{33}$$

or substituting for $\dot{K}$ and K in terms of y:

$$\frac{\dot{y}}{y} = \frac{\dot{A}}{A} - (1-\alpha)g(y) + s\beta\frac{A^{1/\beta}(R/L)^{\gamma/\beta}}{y^{(1-\beta)/\beta}} \tag{34}$$

Let y^* be the level of income corresponding to the peak of $g(y)$. If the last term in (34) is increasing with time (the A in the numerator dominates the L in the denominator), then eventually MSPG will exceed $g(y^*)$ and the economy will escape the Malthusian trap. If the last term is decreasing with time, then as time progresses the dynamics of the economy will approach those in the case with no capital.

Since the last term is increasing if its logarithm is, it is always increasing if:

$$\frac{d}{dt}\ln\left[s\beta\frac{A^{1/\beta}(R/L)^{\gamma/\beta}}{y^{(1-\beta)/\beta}}\right] > 0 \tag{35}$$

or, since we are interested in the steady-state value of y ($\dot{y}=0$):

$$\frac{\dot{A}}{A} > \gamma g(y^*) \tag{36}$$

Thus the economy will escape its Malthusian trap if this equation is satisfied—technology growth is greater than the resource share times maximum population growth. Since $\gamma < 1 - \alpha$, $(\alpha + \beta + \gamma = 1)$, if this condition is not satisfied then

$$\frac{\dot{A}}{A} < (1 - \alpha) g(y^*) \tag{37}$$

and the economy will not escape its Malthusian trap. So making accumulation of capital or other productive factors endogenous does not materially change the qualitative results of the model.

Endogenous capital does, however, add an interesting dynamic. Because the last term in equation (34) can be increasing with time, the threshold rate of technology growth for transition changes over time. If $\gamma g(y^*) < \frac{\dot{A}}{A} < (1 - \alpha) g(y^*)$, the economy will be Malthusian but with slowly rising income standards as A rises, until the last term gets large enough to push the economy over the hump into growth. This means that with capital accumulation, takeoff can occur without any underlying change in the economy.

APPENDIX B

Notes on the Dutch Data

All data in this section are deflated using Allen's price index except where otherwise noted. The real economy is assumed not to change from 1790-1820.

Population figures come from De Vries and Van der Woude (1997, p. 52). These are reproduced in table 1; I use the 1514 estimate for 1500, the 1680 estimate for 1650, and the 1795 estimate for 1790-1813. This seems reasonable since most of the growth between 1622 and 1680 was prior to 1650; the population may have even declined after 1650. The 1795 estimate is identical to the 1750 estimate so there was little population change at this time. For 1580 I use the 1514 figure combined with Van Zanden's (1992) estimate that population increased at a rate of 0.5% p.a. from 1500-1580.

	1514	1580	1622	1680	1750	1795
Population	275	447	672	883	783	783

TABLE 1. Population of Holland (thousands)

Capital figures are from Van Zanden (1992). Following Van Zanden's calculation method, I used his per capita capital figure for 1500 and his total capital stock figures for 1650 and 1790. I adjusted his real per capita capital figure from 1500 upward to account for what appears to be an error in his price index. He uses a factor of five in the piece cited while in other work he uses a factor of eight for these dates, which is also in line with Allen's price index (2001).

Because land is dealt with separately, I also subtract off estimates of agricultural land values from the capital stock. I take Van Zanden's (1992) estimate that in 1500 half the capital stock was agricultural land; at a 6.25% interest rate as he assumes, this accords with the value of land that would have been necessary to

accord real rents equivalent to Van Zanden's (2002) estimate for 1510/14. For 1790, Van Zanden's (1992) estimates include real property of $fl.$ 480 million. Using his interest rate of 3% implies that $fl.$ 311 million of this was agricultural land in order to yield real rents that agree with his 1812/13 estimate for Holland adjusted for 1790. For 1650, Van Zanden estimates real property of $fl.$ 180 million, and I assume the share of agricultural land in real property did not change from 1650-1790, implying land values of $fl.$ 116 million. To get capital stocks in 1580, I assume that real capital deepening occurred at a rate of 0.10% from 1500 to 1580. This is in line with Van Zanden's statement (1993, p. 278) that most of the capital growth in the 1500-1650 period occurred after 1580, as well as his estimate that per capita GDP was constant from 1500-1580.

	Capital Stock ($fl.$ m.)	Land Value ($fl.$ m.)	Capital Stock per Capita ($fl.$)	Real Capital per Capita	Real Non-Land Capital per Capita
1500	11	5.5*	40*	320	160
1580					173
1650	530*	116	600*	600*	468
1790	1712*	311	2186	2186	1790

TABLE 2. Dutch Capital Stocks

The capital figures I use are presented in table 2. Those figures taken directly from Van Zanden (1992) are starred. The rest are calculated as described.

Holland was also unique in significantly expanding its land area, which it did by more than 1/3 from the 16th–18th Centuries. My estimates for Dutch land area are constructed from De Vries and Van der Woude's figures on land reclamation (1997, pp. 31-32). Because of the very low rent share in the Dutch economy, these figures do not have much impact on the calculation either of per capita income or MSPG. Nonetheless they are included for completeness. I assume that peat extraction grew in the same proportions as land reclamation from 1500-1650, and that peat extraction declined from 1650-1800 at a rate to offset land reclamation. Although there are no hard figures to fully justify these assumptions, the are broadly

in line with the story told by De Vries and Van der Woude (pp. 37–40). The effective land area figures I use are presented in table 3.

	Land Index	Effective Land Index
1500/1514	1.00	1.00
1580	1.05	1.05
1650	1.23	1.23
1790/1812	1.36	1.23

TABLE 3. Dutch Land Area

National accounts for 1510/1514 and accounts for the agricultural sector for 1812/13 are taken from Van Zanden (2002). Using Van Zanden's (1993, p. 271) reckoning that GDP per capita of the Netherlands in 1805 was 77% of the province of Holland's allows us to convert the Netherlands' per capita GDP reckoned for 1807 into a figure for Holland. Doing so implies that real GDP per capita in Holland rose by 46% from 1510/14 to 1812/13. Combining this with Van Zanden's estimates that GDP per capita was flat from 1500 to 1580 and fell 9% from 1650-1805, we can deduce that GDP per capita rose 60% from 1580 to 1650. Van Zanden (1993) had estimated this at about 20-50%, although the 60% figure is consistent with his 2005 simulation. The real GDP figures used in this paper are presented in table 4.

	1500	1510/14	1580	1650	1790	1812/13
GDP per Capita	209	196	209	316	286	286

TABLE 4. Dutch Real GDP per Capita (1805/14 Guilders)

Van Zanden's national accounts can also be used straightforwardly to determine the rent share in 1510/14 and 1812/13, by multiplying the rent per hectare by the number of hectares under cultivation and dividing by GDP. This yields a rent share of just 6.4% of GDP for 1510/14 and 4.9% of GDP for 1812/13. The real rent indexes that have been constructed for the period 1500-1800 (e.g., De Vries and Van der Woude, 1997 and Van Zanden, 2002) are generally not consistent with these figures and imply that the rent share at the beginning of the 16th Century was 3-4

times what it was in 1650 and at the beginning of the 19th Century. I have chosen here to trust the national accounts, which state that real rents in 1812/13 were 2.5 times their 1510/14 levels, and to adjust the rent indexes for 1510/14 downward to account for this. These indexes combined with the land index from De Vries and Van der Woude and the GDP figures from Van Zanden allow estimates of the rent share. These are shown in table 5.

	1500	1510/14	1580	1650	1790	1812/13
Rent Index (Adjusted)	23	25.5	50	94	60	60
Rent Share	5.4%	6.4%	8.2%	5.5%	4.8%	4.9%

TABLE 5. Rent Share in the Dutch Economy

From this table I interpolate a rent share of 7% for 1500-1580 and 1580-1650; I use 5% for 1650 to 1790.

The capital share is computed as a residual from the rent share and labor share. The labor share is based on De Vries and Van der Woude's real wage indexes (1997, p. 629) together with per capita GDP. This is shown in table 6.

	1510/14	1500-1580	1580-1650	1650-1805
Wage Index	1	0.85	0.8	0.8
Labor Share	0.67	0.57	0.42	0.34
Capital Share	0.26	0.36	0.51	0.62

TABLE 6. Labor and Capital Share in the Dutch Economy

Finally, TFP is computed as a residual.

APPENDIX C

British Data Sources

Where available, I have used data sources for England and Wales rather than only England or Great Britain as a whole. However, because I am primarily interested in growth rates rather than absolute levels, I have endeavored to find the longest time series possible and have tried hard to avoid switching geographical coverage in the middle of a series. I have therefore sacrificed some geographical exactness in the name of achieving estimates that are comparable across time, on the belief that growth rates for Great Britain depend mainly on England and the effects of geographical mismatch are not likely to be large compared to the already tentative nature of the data.

Population data post-1541 are from Wrigley and Schofield (1981, p. 208). Pre-1541 data are from Hatcher (1977, p. 71). Both are for England only.

Except where otherwise noted, real wage data are from Clark (2005) and are the average of the two wage series he computes for England.

Total factor productivity in agriculture for 1300, 1500 and 1700 comes from Allen (2005, table 17) and is for England and Wales. I used TFP that was adjusted for labor inputs and land quality but not for capital. Thus prior to 1700 any productivity increases due to capital are included in TFP. I assumed that the course of TFP followed labor productivity (from Allen, 2000) in order to get estimates for 1400 and 1600. The calculation for 1400 is as follows:

$$\frac{TFP_{1400}}{LP_{1400}} = \sqrt{\frac{TFP_{1300}}{LP_{1300}} \cdot \frac{TFP_{1500}}{LP_{1500}}} \tag{38}$$

where LP_i is labor productivity in year i. The calculation for 1600 is analogous.

Proportion of the English workforce in agriculture comes from various sources. For 1300 I use the estimate from Wrigley (2006, p. 468). The estimates for 1500 and later are from Wrigley (1987), except that I use Wrigley's 1520 estimate for 1500.

I use Van Zanden's (2005) estimates of how much more productive non-agricultural production was than agricultural production in England for 1290 and 1688, and I interpolate for years in between. This is combined with proportion of the population in agriculture to yield share of output due to agriculture.

To get averages over a period for share of output in agriculture, I simply average the endpoints of the period, except for the 17th Century which (if AS is the share of agriculture in GDP) is given by

$$AS_{\text{17th Century}} = 0.25 \cdot AS_{1600} + 0.5 \cdot AS_{1650} + 0.25 \cdot AS_{1700}$$

Estimates of British GDP per capita growth for 1700-1760 and 1831-1860 are from Crafts (1985). Estimates for 1760-1831 are from Crafts and Harley (1992).

I use net capital stocks for Great Britain from Feinstein (1988, Table XIII).[1]

Effective land area for English and Welsh agriculture is from Allen (2005, table 1), which gives estimates for 1300, 1500, 1700, 1750, 1800 and 1850. The estimate for 1850 is used for 1860; estimates for other dates are interpolated.

Estimates of factor shares post-1760 come from Allen (2007) and are for Great Britain. The rent share for 1300, 1500 and 1700 is calculated from the rent share in English and Welsh agriculture from Allen (2005) times the proportion of GDP in agriculture as calculated above. The rent share for 1400 is assumed to be the same as the rent share for 1500, as all the change in share over this period is assumed to be due to the Black Death.

The method to calculate the rent share for 1600 was rather involved. Consider first the equation for income growth:

$$\frac{\dot{y}}{y} = \frac{\dot{A}}{A} - \eta_R \frac{\dot{L}}{L} \tag{39}$$

[1]Crafts and Harley (1992) say they take their figures from Feinstein's Table XVI. However, the growth rates they use come closest to matching the rates from Table XIII.

which, consistent with my earlier assumptions, absorbs any capital effects into TFP over the period from 1600-1700. η_R is the rent share and is given by $(\rho R/Ly)$ where ρ is rent per acre, which I take from Clark (2002). Thus, we have data on all variables in this differential equation at both ends of the 17th Century except GDP per capita y in 1600. There is thus a unique level of y in 1600 that is consistent with this differential equation, which in turn determines the rent share in this year. This implies that GDP per capita grew 33% over this period and the rent share in 1600 was 37%.

The capital share for 1700 is calculated by assuming the labor share was constant from 1700 to 1760 so that any decreases in factor payments to land were absorbed by capital.

Data on trade volumes post-1780 are for Great Britain and are from Davis (1979). Pre-1780 data are for England only and are sourced from Davis (1962). Because I only consider trade on a proportionate to GDP basis this does not lead to severe data inconsistencies. The export to GDP ratio for Great Britain is from Crafts (1985, p. 131) and is interpolated and rounded to the nearest half a point for intervening years. This ratio is then used with Davis' import data to calculate the trade balance in each commodity.

Coal production for England and Wales comes from Hatcher (1993) for 1560 to 1700, from Flinn (1984) for 1700 to 1830, and Church (1986) from 1830 onwards. Pithead prices are from Clark (2007), and industry revenue is given as pithead price times production.

1861 GDP, which has been used to determine the share of coal in GDP for this date, is from Deane and Cole (1967).

Division 2

Leverage and Financial Market Instability: A Theoretical and Historical Investigation

Introduction

Financial crises are perhaps the most common "unthinkable" event in modern history. Despite the widespread assumption in markets that the liquidity characteristics of investments will not suddenly change, financial crises and asset price crashes have happened with startling frequency across cultures, continents, and centuries. Yet, economists who have correctly predicted financial crises are in a small minority: even Nobel laureates such as Krugman and Stiglitz who foresaw the collapse of the housing market and who have made a significant part of their career studying financial crises admit that they did not see the attendant financial meltdown coming.

Financial crises—large, sudden breaks downward in the price of assets accompanied by the evaporation of liquidity—are difficult to understand using the standard assumptions of elementary graduate economics courses, for these permit neither incorrect valuation of assets nor large price changes without changes in expectations, nor do they even countenance a role for liquidity, money or banks. As a result, economists tend to view financial crises as arising from some "market imperfection"—language that suggests that the phenomenon is not a result of behavior or conditions that are deep or fundamental.

In this paper, I shall explore the role of leverage in determining asset prices and in building the conditions for financial crisis. Leverage, meaning the use of borrowed funds to purchase assets, has been a mainstay of virtually all parts of our financial markets for centuries.

The aim of this paper is to develop a simple model that gives leverage a starring role in linking bubbles, crashes and financial crises, and in so doing move towards a model for asset prices that encompasses these phenomena and reproduces many stylized facts about them. Given the ubiquity of bubbles, crashes and crises, we

would like to have a model that produces them from simple and fundamental dynamics. The results of my model are not dependent upon assumptions about the intricacies of behavior: blunt assumptions are sufficient to produce rich dynamics. The model naturally gives rise to a measure of financial stability, which can be used by systemic regulators to determine how susceptible a market is to a crash and crisis. This measure of financial stability, which I term an *instability ratio*, is perhaps the most important contribution of the paper.

Chapter 1 presents a model of asset pricing that focuses on the behavior of levered investors and on the constraints they face. It includes my measure of financial instability that can give regulators and policymakers advance warning of financial crises. Chapter 1 examines the behavior of the model and asset prices at a point in time, although the model is necessarily intertemporal because in the presence of leverage the equilibrium price of assets is path-dependent. I focus on, and examine in some detail, the specific moment when equilibria become unstable.

The model of chapter 1 is a static model that relates today's price to yesterday's. Chapter 2 iterates that model forward in time to examine the impact of leverage on the evolution of asset prices over time. It includes a computer simulation, showing that the model reproduces typical patterns of bubbles followed by price crashes as well as an association of high volatility and trading volumes with high leverage.

Chapter 3 applies my measure of financial instability to the 1998 collapse of hedge fund Long-Term Capital Management, while chapter 4 examines the 1929 stock market crash.

CHAPTER 1

A Leverage-Based Measure of Financial Instability

In this chapter, I build a simple model of asset pricing with leverage and collateral constraints. The model naturally gives rise to a measure of financial instability. We use this measure to derive a quantitative condition for when markets will become unstable, and it is sufficiently general that central banks could use it as a guide to determine how near a real-world asset market is to becoming unstable and triggering a crisis.

I define a financial market equilibrium as unstable when the process of *tatonnement* pushes the system away from, rather than towards, equilibrium. In the context of a limit order book, this corresponds to a situation where demand rises with with price, and does so faster than supply. It is well-known that levered investors may have upward-sloping demand curves, and that this causes shocks to asset prices to be amplified via what is known in the literature as a "financial accelerator" mechanism. In this chapter, however, we shall see that if assets become too concentrated in the hands of levered investors, not only does the demand curve of levered investors become upward sloping, the aggregate demand curve for assets across all investors may become upward sloping. The market becomes unstable, and a feedback loop causes a large and rapid change in the asset price. This change can occur without any change in expectations, news, the real economy, or underlying market architecture.

In recent financial crises, sharp asset price declines have often been accompanied by tightening margin requirements that force liquidations. Probably as a result of this regularity, much of the public policy debate around financial reform has focused on proposals that would weaken this feedback loop, such as counter-cyclical capital requirements for banks and other financial institutions. There is also a small

but growing literature, discussed later, showing that high leverage combined with procyclical margin requirements causes precisely the kind of instability I have alluded to. By contrast, the model I present does not require endogenously tightening margin requirements to produce similar underlying dynamics. I show that even when margin requirements are constant and exogenous, an overly leveraged system is susceptible to crisis. This suggests that the increasing margin-forced liquidations feedback loop is an amplification of the underlying mechanism of the crisis, rather than the essence of the mechanism itself. As a result, policy proposals that aim at mitigating the impact of increasing perception of risk on margin tightening do not address the fundamental mechanism of the crisis and are more likely to alter the timing and severity of a crisis than to prevent it outright.

Our model has two sorts of investors: fundamentals-based investors, who invest with no leverage, and speculators, who use leverage. Fundamentals-based investors are assumed to have downward-sloping demand for assets as a function of price, both because they are sensitive to estimates of value and because when assets are expensive they cannot afford to buy as many of them.

Speculators invest with leverage, but face a margin requirement, which limits their total investment position. At the margin, therefore, speculators' behavior is governed not by their expectations but by the need to partly fund their investments with their own equity. Asset prices are thus invariant to a wide range of speculators' expectations of fundamental value. Instead, asset prices are determined at the margin by structural considerations such as leverage and the distribution of wealth.

The model's collateral constraint causes speculators' demand curve for assets to be upward sloping. If asset prices appreciate, speculators get richer, reinvest their profits, and accumulate a larger share of the total assets in the economy. If speculators hold a large enough share of the assets, their upward-sloping demand curve overwhelms the downward-sloping demand of the fundamentals-based investors. The *total* demand curve for securities then becomes upward-sloping, and the equilibrium

that was previously stable suddenly becomes unstable, resulting in sudden explosive behavior: a price spike or (more likely) a crash. I save the modeling of the longer-term dynamics for chapter 2 on bubbles. In this chapter, we will explore the dynamics at the moment where the demand curve shifts from being stable to being unstable.

Section 1 relates the approach I take to the previous literature. Section 2 discusses my reasons for taking an approach where agents do not formally optimize expected utility in a rational expectations-based framework. Section 3 develops a simple, intuitive version of the model. Section 4 fully generalizes the model to a degree that permits it to be used in the real world. As part of the generalization, I show that the use of financial derivatives is equivalent to leverage and is usually more likely to create the conditions for instability. We will also see that when levered investors play in multiple markets, a crash in one market can lead to contagion into other markets. Section 5 concludes.

1. Previous Literature

Broadly speaking, models where limitations on leverage and collateral play an important role have been referred to in the literature as models with "net worth effects." These models fall into three main categories.

The first are generally known as financial accelerator models. They focus on the impact of credit constraints tied to entrepreneurs' net worth on the real economy, although they also have implications for asset prices.

Bernanke and Gertler (1989) and Kiyotaki and Moore (1997) have made the seminal contributions to this literature. In both papers, the authors consider models where entrepreneurs must contribute a certain proportion of capital in the form of equity to undertake real projects. As a result, the price of capital becomes correlated with the net worth of entrepreneurs, because when entrepreneurs are poor they can afford less investment. This leads to a financial accelerator effect, where declines in asset prices reduce the net worth of entrepreneurs, which in turn reduces their

demand for capital, which reduces asset prices, and so on. While shocks are amplified relative to the canonical RBC model, equilibria in these models are stable in the sense that small shocks lead to first-order changes in asset prices.

While the financial accelerator literature generally focuses on implications for the real economy, a second strand of the literature evaluates the effect of collateral requirements with a focus on financial markets.

Geanakoplos (2003) develops a model where agents express varying degrees of optimism about an asset. Optimists purchase the asset using borrowed funds with the asset as collateral, while pessimists lend. When bad news arrives, optimists are forced to liquidate. The price crashes, both because expectations about the assets' ultimate value decline and because after the liquidation assets are held by investors who are intrinsically less optimistic and so have a lower estimate of fundamental value. Additionally, Geanakoplos' model includes endogenous margin requirements, which may provide another source of forced liquidations and price declines. Feedback between these effects amplifies the price decline in a similar manner to the large literature on financial accelerators.

Fostel and Geanakoplos (2008) present an approach to asset pricing in an equilibrium framework when assets can be used as collateral. If more optimistic agents are also more impatient, then asset prices exceed the marginal buyer's fundamental value because they include a *collateral value*, which is the additional value holding the asset imparts to its owner by enabling him to borrow at secured rather than unsecured interest rates. If, instead, optimists were more patient, they would bid up asset prices to the point where they would be indifferent between using their savings to buy assets and lending their savings in the money market, and the collateral value would be zero.

Like Geanakoplos (2003), Fostel and Geanakoplos' (2008) model employs a financial accelerator effect so that the effect of bad news is amplified and leads to contagion to other assets, as increasing uncertainty in their 'anxious economy' reduces collateral values.

In all of the papers discussed so far the financial crisis remains an equilibrium phenomenon. Demand curves are downward-sloping and all equilibria are stable. A handful of papers have attempted to deal with equilibria that are unstable, meaning that even infinitesimal shocks to demand or expectations can lead to a financial crisis. In these models, a financial crisis is a departure from equilibrium.

Yuan (2005) builds a two-period model where an unstable market results from the interaction of borrowing constraints and information asymmetries. She constructs a model of informed but potentially credit constrained investors together with uninformed investors. Uninformed investors take the price signal as informative about fundamental value, but when the price is low, these signals are less informative because the informed investors may be credit constrained. Lower prices thus lead uninformed investors to be more uncertain about fundamental value, which can cause them to have a backward-bending demand curve. As a result, at low price levels the function relating price to fundamental value can be discontinuous. Outcomes below this discontinuity are described as a crisis. However, because trading only occurs in one period of Yuan's model, the crisis outcome does not represent a price decline from one period to the next, but rather simply a too-low price. Adding a third period to the model would potentially change this, but would seriously complicate the modeling exercise because Yuan's rational expectations framework creates the need to compute optimal strategies for all investors.

Xiong (2001) builds a model where risk-averse "convergence traders" arbitrage mispricings. Convergence traders generally try to trade against mispricings, but because they are risk averse (Xiong employs logarithmic utility) their optimal strategy when they are poor may cause them to liquidate as prices decline due to a wealth effect. Here, there is no overcollateralization requirement. Rather, the upward-sloping demand of these traders comes from their risk aversion and desire to maintain positive wealth. Xiong explicitly assumes that there are sufficient long-term investors so that aggregate demand remains upward sloping and that equilibria are stable. However, this is not a critical component of the model.

Adrian and Shin (2008) admit the possibility of upward-sloping demand leading to a feedback effect when markets are hit by a shock, but do not construct a model.

The previous work most closely related to the current paper is Chowdhry and Nanda (1998). They build a three-period model where instability results solely from the requirement for collateralized borrowing, rather than an interaction between a collateral constraint and some other departure from perfect neoclassical markets. Chowdhry and Nanda define markets as unstable "if prices can move rationally even in the absence of changes in economic fundamentals." (p. 180) This definition of instability is related though not identical to the definition of instability I use here, which is that a market is unstable if the process of *tatonnement* pushes the system away from, rather than towards, equilibrium. The importance of this distinction becomes clear upon a close read of Chowdhry and Nanda's model, for even in the initial periods of the model, when the market is supposed to be functioning normally, the equilibria the authors compute are generally unstable in that a slightly higher price would create excess demand while a slightly lower price would create a demand shortfall.

While Chowdhry and Nanda come at the problem from a utility-maximizing perspective, Thurner, Farmer and Geanakoplos (2009) construct a model where investors follow behavioral heuristics, as in the current paper. They have two types of investors, hedge funds, which act as value investors and invest with leverage, and noise traders, who have weakly mean-reverting but random demand. They explore their model with Monte Carlo simulations and find that the model generates financial crises where large portions of the more highly levered hedge funds are bankrupted as a result of small shocks and forced liquidations. The present model is similar in spirit and dynamics to Thurner et. al, but takes the inquiry in a complementary direction, pursuing mainly analytical rather than simulation-based results and attempting to define the conditions for instability. By contrast, Thurner et. al mainly examine the statistical properties of returns.

In addition to forced liquidations resulting from net worth effects, other authors have examined feedback loops where small shocks cause cascading liquidations through other channels.

Shleifer and Vishny (1997) develop a model where there is no leverage but poor performance causes investors to withdraw assets from their portfolio managers, creating a feedback loop where temporary shocks reinforce themselves.

Diamond and Verrecchia (1980) produce instability using a mechanism that is based on information aggregation rather than leverage. In their model, investors with imperfect information assume that others who are liquidating assets have better information, and price movements therefore become self-reinforcing because they have information content. Gennotte and Leland (1990) employ a similar mechanism with similar results.

Brunnermeier and Pedersen (2009) combine many of these aspects and construct a model where financiers uninformed about fundamental value take price fluctuations as new information about fundamental value. They assume that volatility of fundamentals follows an ARCH process, so that large price changes signal large changes in fundamental value which signals future volatility. As a result, lenders respond to price falls by tightening margin requirements, leading to a feedback loop. While Brunnermeier and Pedersen acknowledge the possibility of unstable equilibria, they view these as "uninteresting" (p. 16) and restrict their attention to stable equilibria only. In their model, financial crises can arise from small changes in fundamentals because there are multiple equilibria, and a small change in fundamentals can result in switching between equilibria.

There is also a large literature on banking crises, which is dominated by framework of Diamond and Dybvig (1983) and reviewed by Allen and Gale (2007). In the Diamond-Dybvig framework, banking crises are triggered because assets are too long lived to support short-term demands for consumption and precautionary liquidity. A crisis occurs when demand for short-term assets forces banks to call loans, forcing

the liquidation of real projects at a loss, rendering the banks insolvent or bankrupt. Assets are always valued at their (exogenously given) fundamental value.

The present model also gives rise to a new measure of financial instability that can be applied empirically, a task to which a few authors have taken a wide variety of approaches.

Adrian and Brunnermeier (2008) develop a measure of systemic risk, which they call "CoVaR." Conceptually, CoVaR for a financial institution measures the Value at Risk of the entire financial system conditional on that institution being in distress. The authors then discuss "ΔCoVaR," the difference between the VaR of the financial system and the CoVaR for the institution. This measures the increase in risk in the financial system that would occur from a particular institution becoming distressed.

CoVaR is useful because it can be used as an indicator of the systemic threat of any given financial institution, which in turn may inform systemic capital charges or other regulations authorities wish to impose. Implicit in the idea of CoVaR is that an increase in perceived risk can cause systemic instability. This is related to common results in the literature that increases in risk perceptions can cause increases in margin requirements that force liquidations.

Kamada and Nasu (2010) use a so-called "asset quality index," which combines the average risk of an institution's assets with its leverage to give an overall risk measure. They use their risk measure to compare various banking systems during the current crisis, but do not find an obvious connection between countries that score poorly on their measure and countries with weak banking systems. An important insight from their paper is that naive limitations on leverage are likely to encourage banks to hold riskier assets because if the limit is binding then optimizing banks will try to increase risk in other ways. Thus leverage limits may not improve stability.

Borio and Lowe (2002) simply use the credit-to-GDP ratio, and find that it can predicts 80% of financial collapses with a 4:1 signal-to-noise ratio, a remarkable performance for a single variable.

Finally, Tsomocos (2003) proposes that instability be measured jointly by low banking profitability and a high number of defaults, a measure that focuses on the banking system.

2. Methodological Considerations

The agents in my model have what Farmer and Geanakoplos (2009) call "zero intelligence," an approach to economic modeling they attribute to Becker with roots as early as 1962. These models are particularly useful to investigate problems where "structure dominates strategy," or where the institutional arrangement is a greater determinant of the economic outcome than the strategy of the economic agents. Here, I claim that there is value in viewing asset pricing as such a problem. (In chapter 3 we shall apply the model to the collapse of an arbitrage-focused hedge fund run by Nobel laureates in economics, perhaps the most classically rational of all possible investors. Even there, we shall see that viewing investors as behaving according to rules of thumb yields significant insight into the event.) For those with a more traditional view, the assumption effectively being made is that investors are always at a so-called "corner solution" of any optimization, and that their behavior is therefore determined by binding constraints on behavior rather than a first order condition.

By employing a rule of thumb model, I have surely sacrificed many of the advantages of fully optimizing models. However, there is also much to be gained from this approach. No ungrounded assumptions are required about the functional forms of preferences or processes. I am able to incorporate heterogeneity without having to worry about strategic or game theoretic effects. Avoiding the mathematical complexities of optimization means that the model is far more tractable; with this comes improved transparency about what is driving the results.

Ad hoc assumptions on behavior do have some precedent in the literature, for example in Thurner *et. al.* (2009) Even in papers which attempt to hew closely to rational expectations, authors tend to make *ad hoc* assumptions. For example,

Kiyotaki and Moore (1997) essentially arbitrarily set farmers' savings rates by designating a portion of output as non-tradable and by making assumptions about the relationship of different agents' discount rates to each other. These assumptions play a central role in the model. In another example, Brunnermeier and Pedersen (2009) employ the device of assigning vanishingly small probability to the state they wish to study in order to avoid having to consider complicated forward-looking strategic behavior by financiers and speculators. Xiong (2001) includes in his model long-term investors with an *ad hoc* behavioral rule. The simpler framework means my model can accommodate a richness of financial instruments and heterogeneous beliefs that makes it operationalizable by policymakers. Simplicity also brings robustness, so that the results are not sensitive to small changes in model specification. The simplicity of the model allows its non-linearities to be fully explored in an intuitive way, which is beneficial now that financial crises are widely acknowledged to be non-linear events.

Finally, the key variables in the model are *observable* and *measurable*: leverage, margin requirements, the interest rate, and the net worth of levered investors versus unlevered investors. There is a minimum of reliance on unobservable, psychological variables such as utility, expectations, and subjective probability distributions.

3. The Basic Model

We shall consider a model with one asset, four kinds of agents, and at first, no short sales.[1] The four agents in our model are:

- *Speculators*, who are sufficiently confident in their views that they lever their positions to the maximum level permitted by their lenders
- *Retail or real-money investors*, who have a downward sloping demand curve for the asset, and who deposit any excess funds in a bank account

[1]Short selling restrictions are widely known to permit persistent overvaluation of assets (see Sheinkman and Xiong, 2003, for a review of this literature), but this is different from the phenomenon of a bubble or financial instability.

- *Banks*, who provide credit to the speculators at the market interest rate and a fixed margin requirement (the reasons for these assumptions are discussed later)
- A *central bank*, whose sole function is to hold interest rates fixed in the near term[2] by providing credit to the market against sound collateral.

Although the model has all of these agents, it is driven almost exclusively by the speculators. It is to the speculators' behavior that we now turn.

3.1. Behavior of the Speculators. Suppose an investor is extremely enthusiastic about an asset, so that he wishes to purchase as much of it as he is able with as much leverage as his lenders permit.[3] Viewed narrowly, these investors act according to a naive behavioral rule. However, there is a small literature investigating optimal portfolio choice in the presence of net worth constraints and credit constraints,[4] which shows that sufficiently optimistic rational agents do indeed lever up to the maximum degree permitted by their lenders. Moreover, even if real-world speculators have some slack and cushion built in for the short term, investors still tend to target a certain leverage ratio over the medium term: they voluntarily liquidate when their net worth declines in order to avoid forced liquidations later, so they face what is effectively a "soft" margin requirement. At the very least, the

[2]It is understood that there is a long-standing disagreement among economists of the proper instrument of monetary policy. This paper takes no stand in that dispute. Instead, I use this formulation because at least until the financial crisis, the interest rate was the instrument of monetary policy used by the vast majority of the developed world's central banks.

[3]Formally, we could consider this as an investor whose estimate of the fundamental value of an asset is v_π far above the market price, and is sufficiently risk-neutral that he is willing to speculate with leverage. However, it is often the case that market participants' behavior is totally unanchored in fundamentals, so that some investors will continue to buy as the market rises. Some reasons for this include:

- They may simply extrapolate past performance;
- They may believe that prices will continue to rise in the short term;
- They may be responding to skewed incentives, such as the "trader's put," where gains are partly shared by the money manager but losses are fully borne by the investor.

Viewing the speculators as having a very high estimate of fundamental value, rather than no estimate of fundamental value, is more in keeping with the traditional assumptions of the literature. However, this distinction is not important for the key results and insights of the model. Different readers will have different views about which is more intellectually appealing. In our exposition, we will jump between the two for the sake of convenience.

[4]See, for example Grossman and Vila (1992) and Liu and Longstaff (2004).

assertion that some small subset of investors should be wildly optimistic ought not to be controversial.

Formally, we let:

- p_t be the price of the asset at time t
- m_t^{lv} be the quantity of the asset held by the leveraged investor (the "lv" superscript and t subscript will sometimes be left off to reduce clutter and will only be included when necessary to avoid ambiguity)
- λ be the permitted leverage ratio, the maximum proportion of the investor's assets that may be funded by debt
- r be the interest rate charged on margin loans
- d be the dividend paid by each unit of the asset

The investor's net worth is thus given by:

$$\text{Net Worth} = (\text{Margin Percentage}) \cdot (\text{Assets}) \tag{40}$$

$$= (1 - \lambda) \cdot pm \tag{41}$$

provided that, as we have assumed, the speculator is fully invested. Each period, the investor will reap the benefit of all price appreciation and dividends from the assets and pay margin interest, so the change in net worth will be given by:

$$\Delta NW = \text{Appreciation} + \text{Dividends} - \text{Margin Interest} \tag{42}$$

$$= (p_t - p_{t-1})m_{t-1} + d\Delta t \cdot m_{t-1} - \lambda r \Delta t \cdot p_{t-1} m_{t-1} \tag{43}$$

Adding (41) and (43) and simplifying, we have:

$$NW_t = m_{t-1} \cdot [p_t + d\Delta t - \lambda(1 + r\Delta t)p_{t-1}] \tag{44}$$

This equation says simply that net worth is given by the current value of last period's assets, plus any dividends received on those assets, minus the value of debt (with interest) funding those assets.

The enthusiastic investor puts his entire net worth back into the bubble asset with leverage. Combining (41) and (44), we have:

$$(45) \qquad (1-\lambda)p_t m_t = m_{t-1} \cdot [p_t + d\Delta t - \lambda(1+r\Delta t)p_{t-1}]$$

which simplifies to:

$$(46) \qquad m_t^{\text{lv}} = \frac{m_{t-1}^{\text{lv}}}{1-\lambda} \cdot \left[1 + \frac{d\Delta t - \lambda(1+r\Delta t)p_{t-1}}{p_t}\right]$$

This gives the speculators' demand for assets m_t^{lv} as a function of the price p_t; it is their demand curve for assets, and it is almost certainly upward sloping.[5] Investors targeting a specific leverage ratio demand more of an asset as its price increases. Finally, leverage introduces path-dependent demand; levered investors' demand depends on their holdings and the price in the previous period.

Equation (46) fully characterizes the behavior of the speculators.

3.2. Retail or Real-Money Investors. In addition to the speculators, there are also some investors who invest without leverage, whom I term retail or real-money investors. These investors are not modeled in detail, but are assumed to have a downward-sloping demand curve for assets. This downward-sloping (rather than horizontal) demand may be for a variety of reasons, including heterogeneity of opinion about the value of the asset, relative value considerations and the desire for portfolio diversification. One simple example is investors who wish to keep a fixed proportion of their portfolio in different assets, such as the orthodox portfolio split of 60% stocks and 40% bonds. As the price of stocks rises, investors need to hold fewer shares to account for 60% of their portfolio. More simply, investors who eschew the use of leverage are limited in their asset purchases by their equity, and so the maximum number of shares they are able to purchase is a declining function of the share price.

[5]The only way it is not is both if the time step is large and the dividend is very large compared to the margin interest rate. To see how unlikely this is, consider the case where $\Delta t = 1$, that is, portfolio reallocations take place only once a year. Consider a modest leverage ratio of 20% (one dollar of debt for every four dollars of equity), and margin interest rate of just 5%. The dividend yield would then have to be greater than 21% in order for the demand curve to be downward-sloping.

Let:

- m_t^{gp} be the total demand for assets by the general population of retail investors
- $D(p)$ be the number of assets that the average retail investor demands as a function of price. We shall assume that $D'(p) < 0$ so that demand is downward sloping, and that demand does not depend upon the investor's net worth[6]
- N be the total number of investors in the economy
- μ be the proportion of investors that are levered

Then we have:

$$\text{Total Retail Demand for Assets} = (\text{Proportion of Retail Investors}) \times (\text{Population}) \times (\text{Demand per Investor}) \tag{49}$$

or:

$$m^{\text{gp}} = (1-\mu) N \cdot D(p) \tag{50}$$

This fully defines the behavior of the retail investors.

3.3. Banks. Banks are assumed to be conduits that lend to all comers against collateral at the prevailing interest rate, which is fixed by a central bank, and with

[6]This is an abstraction, and it is unlikely to be correct in the real world. However, the dependence on price can capture this effect at each moment in time, since we have not specified a functional form. At first glance, it might appear that wealth effects could be large enough that they make retail demand upward sloping. Closer examination reveals that this is not the case. Suppose that retail investors hold a proportion of their wealth $\beta(p)$ in an asset, and that they have net worth y. Then the number of assets each retail investor demands is given by

$$D(p) = \frac{y\beta(p)}{p} \tag{47}$$

Differentiating in logs gives:

$$\frac{d \log D(p)}{dp} = \frac{1}{y}\frac{dy}{dp} - \frac{1}{p} + \frac{\beta'(p)}{\beta(p)} \tag{48}$$

The last term in this equation is negative. After examining the first term, we shall see that it is always outweighed by the second term, so that demand remains downward sloping. Suppose that the retail investor owns m assets at price p, and has additional assets y_0. Then

$$y = pm + y_0$$

and

$$\frac{1}{y}\frac{dy}{dp} = \frac{m}{pm + y_0} < \frac{1}{p}$$

for $y_0 > 0$. So the demand curve remains downward sloping.

a fixed haircut.[7] At first glance, such behavior seems irrational. Why should we assume banks behave in this way, rather than optimizing in some sense?

In the real world, risk-free overnight interest rates are controlled by the central bank, with credit spreads set by the market. But if each overnight interest rate set by the central bank comes with a credit spread that is the result of optimizing market participants' reactions to that interest rate, then it is not incorrect to say that the central bank indirectly controls the interest rate at which banks lend to their customers. It might even be argued that real central banks actually do set the risk-free interest rate in order to target private-sector credit conditions. The supply curve for funds is thus horizontal.

The assumption that margin requirements are fixed exogenously requires a bit more justification. I make this assumption for three reasons. First, much of the recent literature and public debate about financial crises has focused on the role played by the endogenous tightening of margin requirements in triggering liquidations.[8] This has led to proposed policy interventions to impose countercyclical capital requirements and other restrictions to prevent the tightening of liquidity just when it is most damaging. Investigating a model with fixed margin requirements allows us to disentangle the effect of liquidations due to tightening margins from liquidations due to losses, the difference between what Brunnermeier and Pederson (2009) call "margin spirals" versus "loss spirals". As we shall see, tightening margin requirements are not an important part of the story of rapidly evaporating liquidity; the results are qualitatively the same with fixed margin requirements. In 1929, in contrast to the stories told by Geanakoplos and Brunnermeier and Pedersen, lenders actually tightened margin requirements during the summer as stock prices spiked, and then eased them in the later part of the crash in the autumn.

Second, in many cases, banks' margin requirements are constrained by regulation and are thus effectively fixed.

[7] In financial parlance, the "haircut" is the margin requirement for a specific asset. For example, if a bank will lend against a high-grade corporate bond with a 10% haircut, it means that the bank will lend \$90 for each \$100 worth of the bond posted as collateral.

[8] See Geanakoplos (2003) for an approach to modeling endogenous margin requirements.

Third, if we do not take on faith that banks act perfectly rationally, we would like to investigate the impact of changes in margin policy by the banks. To do so, we view such changes as essentially exogenous choices of experimenting bank management, and it is easier to do so in a framework where we can vary the margin requirements directly.

A fully rational model would have banks optimize over the interest rate and haircut on collateral. Such a model would be tremendously complex, with banks needing to take into account forward-looking liquidity conditions, fundamental valuations, and investors' behavior. But even this is a gross oversimplification of the factors real-world banks must take into account, of the institutional arrangements within which they must operate, and of the dizzying array of potential contract terms beyond just the interest rate and collateral haircut. Examples include PIK versus cash interest, various remedies in case of defaults, loan covenants, term of the loan, etc. It is far from clear that a model where banks optimize over credit spreads and collateral requirements is an appreciably better approximation to reality than one where they simply lend out all available reserves on given terms. Instead, the "structure" of the lending arrangements dominates the "strategy" used by individual banks in determining the dynamics of the system.

An obvious question is why real-world banks do not necessarily tighten credit as asset prices rise further above their assessment of fundamental value.[9] One important reason is that banks do not care about fundamental value; they care about getting their money back on the loan they have made. Thus, in the event of a default by the investor, they care only about the market price of the asset at the time of default and what discount they must accept when liquidating the collateral quickly. The bank's estimate of this discount is usually based on the analysis of stress scenarios drawn from history of how fast the market can move at any given time, as well as current market conditions. These estimates will change slowly. Further, different

[9]Historically, banks have not acted consistently—in the 1928-29 stock market boom, they tightened margin requirements as the bubble grew; during the 2000's housing bubble in they loosened requirements as the bubble grew. There was also considerable heterogeneity across banks.

banks may have different estimates. Those banks who are perhaps more rational and do raise margin requirements aggressively as prices rise will be the most conservative precisely at times when markets are most vulnerable, and thus will not have many customers and will be less economically significant.

Empirically, while specific banks certainly have exposure limits to specific asset classes, in the aggregate banks do lend out all available liquidity, at least in normal times. That behavior is reflected in the negligible level of excess reserves that commercial banks typically hold at central banks. In fact, it *must* be this way: the total level of reserves in the banking system is the instrument central banks typically use to control the interest rate, and the tight link between the level of reserves and the interest rate is dependent upon all reserves being levered to the maximum degree permissible.

This behavior of banks was perhaps most succinctly and famously summarized by Citigroup CEO Charles Prince, who told the *Financial Times* in July 2007, "When the music stops, in terms of liquidity, things will be complicated. But as long as the music is playing, you've got to get up and dance. We're still dancing."[10]

While the model could work just as well without banks, we include them because they allow us to examine explicitly the growth of credit in the financial system that accompanies speculative excesses. The existence of banks links the money supply to credit growth and hence links monetary policy to credit provided to speculative endeavors.

The inclusion of banks allows us to answer the question: "Where do the speculators borrow all the money to buy more assets?" The answer is that the speculators borrow the money from the banks, the banks get the money by accepting deposits from the retail investors, and the retail investors have money to deposit because they have just sold the assets to the speculators. Actual money flows in a circle, while at each stage either credit is created or assets change hands. When speculators buy

[10] "Citigroup Chief Stays Bullish on Buy-Outs." *Financial Times.* July 9, 2007.

from retail investors, credit is created and bank balance sheets expand; when speculators liquidate, credit is destroyed and bank balance sheets contract. Taken as a whole, the group of transactions is self-funding: the money borrowed or withdrawn from the bank to buy the assets is immediately deposited again by the sellers of the assets.

The circular flow of money is shown in figure 1 below.

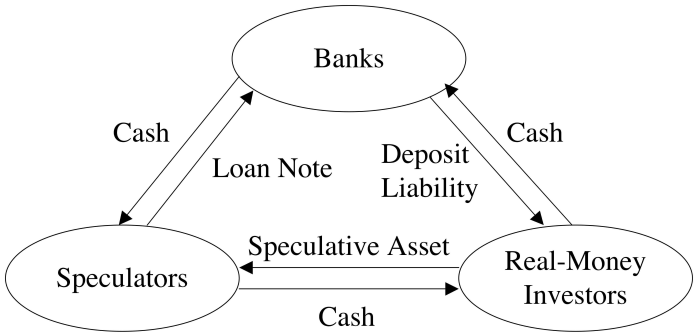

FIGURE 1. The circular flow of cash surrounding the debt-funded purchase of an asset by a speculator from a real-money investor.

3.4. The Central Bank. The central bank simply supplies liquidity to the banking system in order to keep the interest rate fixed at r, which is given exogenously but which we may vary in order to examine the impact of central bank policy.

3.5. Model Dynamics. The model's dynamics are governed largely by the behavior of the speculators and real-money investors. We shall investigate how the model behaves over short periods of time, where interest and dividend payments can be neglected.

Equations (46) and (50), giving the demand curves for levered and real-money investors, respectively, then become:

$$m_t^{\mathrm{lv}} = \frac{m_{t-1}^{\mathrm{lv}}}{1-\lambda} \cdot \left[1 - \frac{\lambda p_{t-1}}{p_t}\right] \tag{51}$$

$$m^{\mathrm{gp}} = (1-\mu)N \cdot D(p) \tag{52}$$

The total demand for the asset, m_t, is the sum of the demand by different investors:

$$m_t = m_t^{\text{gp}} + m_t^{\text{lv}} \tag{53}$$

$$= (1-\mu)N \cdot D(p_t) + \frac{m_{t-1}^{\text{lv}}}{1-\lambda} \cdot \left[1 - \frac{\lambda p_{t-1}}{p_t}\right] \tag{54}$$

It is clear that if $(1-\mu)N \cdot D$ is large compared to m_{t-1}^{lv}, the retail investors will dominate and the demand curve will be downward-sloping, with the supply-demand diagram for securities appearing like panel A of fig. 2. However, if levered investors begin to do well, reinvest their proceeds and accumulate the asset, m_{t-1}^{lv} will begin to grow large relative to $(1-\mu)N \cdot D(p)$.

Eventually, the market reaches a tipping point. If levered investors get infinitesimally richer, the demand curve deforms itself from a curve like D into a backward-bending curve like D' in panel B of fig. 2. This infinitesimal change in the levered investors' fortunes changes the character of the system. Point E remains a sort of "equilibrium," but it is now unstable: any slight positive demand shock (to curve D'') would result in excess demand but a lower equilibrium price at point E''.

The demand shock thus would result initially in a price that was above equilibrium but with excess demand. The excess of buyers over sellers would push the price up further, resulting in still greater excess demand: the price would explode upward until it reached a price so high that even speculators were not willing to buy, and this sky-high price would be the new equilibrium.[11]

Similarly, a slight negative demand shock when the market reaches its tipping point (for example from curve D'' inward to D') will result in excess supply but a higher equilibrium price, with no mechanism to pull the price back up to equilibrium. Supply exceeds demand, resulting in price declines, causing supply to exceed demand further. A slight increase in demand therefore results in the price exploding to a

[11]I show in chapter 2 that this new equilibrium may be only temporarily stable; after a sufficiently large temporary shock or the passage of a moderate amount of time, the equilibrium may collapse and never return to its high level, even after the shock has dissipated.

very high level; a slight decrease in demand results in the price crashing to the point where retail investors can absorb all the assets.

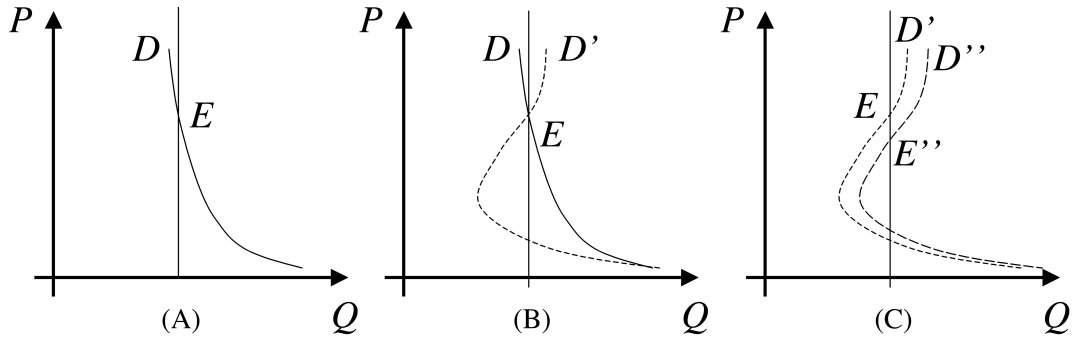

FIGURE 2. Supply and Demand with Levered Buyers

Operationally, we can examine how this process would unfold in a market where, like many financial markets, the price-discovery mechanism is a limit order book.

For illustrative purposes, consider the closest situation to the Walrasian case: a market where (unrealistically) the full retail demand curve is expressed in the form of open buy and sell orders in the limit order book.[12] The levered investors' demand curve below the market price is expressed in terms of *stop-loss* orders: if the market price falls, these stops convert automatically into sell-at-market orders that are matched against the highest bid. Their demand curve above the market price is expressed as a series of stop-buy orders, which convert into buy orders as the price rises.[13]

Figure 3 illustrates the case where the initial shock is negative. Once the tape prints a downtick, the first stop-loss order is activated. It is matched against the highest bid, but because the demand curve is upward-sloping the highest bid is not for a large enough size to fully fill the sell order. The part of the sell order that could not be filled jumps to the next lowest price and executes against the next highest bid. When these orders are matched, the tape prints another downtick, the next

[12]To be clear, these orders tied to the derivative of the demand curve; they are the *changes* in quantity that would be required for each change in price level.

[13]I assume the levered investors already have finance lined up to purchase additional units when their equity increases.

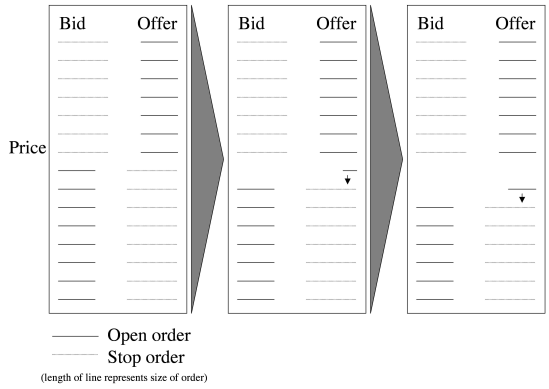

FIGURE 3. Evolution of the Limit Order Book after an Initial Price Decline with Upward-Sloping Demand

stop-loss order is activated, and the process cascades as all the stops are ultimately activated and quickly execute against progressively lower bids as the market crashes to a level where retail investors are able to fully absorb the liquidations by the levered speculators.

Once the liquidations begin, rational non-levered investors, if they understand the process underway, will respond by withdrawing their bids and waiting for the dust to settle at the new equilibrium, rather than supporting the price all the way down. Some of the more sophisticated of these investors will even engage in predatory selling, temporarily worsening the collapse, along the lines of what is modeled by Brunnermeier and Pederson (2005). The price will gap downwards rather than fall smoothly. If the initial shock is upward, the result is similar. Theoretically, the upward cascade continues to infinity; in practice it continues until the speculators have fully exhausted their credit lines and must raise more finance in order to continue to buy assets. This event is usually referred to in the literature and the popular press as "panic" selling, which seems to imply that it is somehow irrational or incomprehensible. It is neither; in this model the mass liquidations are a precise rational or required response to a decline in asset values.

After the fall, the model produces a condition of depressed asset prices. Because the speculators lose their entire net worth, the stock of assets must be held

by a smaller group of unlevered investors. These investors have downward sloping demand, so the asset price must be less than it was before demand became concentrated in a small group of speculators, as it might in a bubble.

These adjustments are not simply an ordinary process of market adjustment to a new, efficient equilibrium. Rather, as in much of the literature on currency crises (for example, Krugman, 1999), multiple equilibria are possible and prices can move discontinuously in response to small shocks.

More fundamentally, efficiency in this model is not well-defined, because we have not introduced any notion of welfare. The notions of "better off" and "worse off" upon which most definitions of efficiency depend are not meaningful.

To the extent that we define the efficient outcome as one where those who value assets the most hold those assets, there is no guarantee that the institutional setup we have considered will achieve that outcome even temporarily. And to the extent that the efficient outcome is achieved because the price spikes instead of crashes, the optimistic investors may not be able to hold on to their positions indefinitely; they may be forced by their lenders to liquidate even if their optimism would have ultimately proven well-founded.

Additionally, "efficient" in this sense means that *ex ante*, it is not possible to make any agent *believe* he is better off without making another agent *believe* he is worse off. By contrast, the typical definition of Pareto Efficiency is that it is not possible to make any agent *actually* better off without making another agent *actually* worse off. But, metaphysical debates about the nature of objectivity aside, reality is different from beliefs. The assets in our model are not conceived as real assets that have a different productivity in the hands of different entrepreneurs, nor are they different consumption goods where there are potential welfare gains from trade. Rather, they are financial assets with a definite if uncertain value. In a model like this one with heterogeneous beliefs, some agents must regret their earlier decisions to buy or sell. Just because an outcome is efficient in the sense we have discussed does not mean it is good or welfare-maximizing.

Finally, economists generally view efficient pricing in financial markets as important because the relative and absolute prices of financial assets are a key determinant of the allocation of capital in a market economy. If the adjustment process of prices in financial markets to equilibrium is slow, as it is in this model, then capital may be misallocated among real economy projects for quite some time, resulting in real production losses. This outcome is certainly not an efficient one by any definition.

In any event, equation (54) determines when the system will exhibit explosive behavior. If its derivative is positive, demand slopes upward at the previous period's equilibrium. The system is unstable and a crash or price spike will result. Otherwise it is stable.

Let us examine this formally. Differentiating (54), we have:

$$\frac{dm_t}{dp_t} = (1-\mu)N \cdot D'(p_t) + \frac{\lambda m_{t-1}^{\text{lv}}}{1-\lambda} \cdot \frac{p_{t-1}}{p_t^2} \tag{55}$$

If we know the demand response of the public to small changes in price, we can determine whether the system is stable.

Define $A \equiv (1-\mu)NpD(p)$ to be the total dollar amount unlevered investors in the aggregate wish to hold of the asset. Substituting into equation (55) and rearranging, the condition for stability ($\frac{dm_t}{dp_t} < 0$) then becomes:

$$\frac{NW_{t-1}}{(1-\lambda)^2} < \eta_D A + p_{t-1} m_{t-1}^{\text{lv}} \tag{56}$$

where η_D (a positive number) is the real-money investors' price elasticity of demand $-pD'(p)/D(p)$. We define the right hand side of the equation as the *elasticity-adjusted market size* or *EAMS*; if $\eta_D = 1$ then the EAMS is just the total assets in the market. In general, the EAMS is the total assets held by constrained investors, plus the total assets held by real-money investors adjusted for their elasticity of demand.

The left-hand side of equation (56) is the net worth of the speculators times the square of the leverage ratio. This quantity defines the *minimum elasticity-adjusted*

market size for stability of the market, which I shall term *MinEAMASS*. It is the smallest elasticity-adjusted market size that is consistent with stability.

We can form a ratio of MinEAMASS to EAMS; we call this the *instability ratio*. If the instability ratio is greater than one, the market is unstable; if it is less than one, the market is stable. We can use the instability ratio in a straightforward way to determine how close the market is to an instability point. The closer is the instability ratio to one, the closer is the market to becoming unstable.

We can see that lower leverage ratios support stability with a higher share of levered investors. Interestingly, a relatively small number of levered investors can create instability if the leverage ratio is high. For example, suppose speculators are levered 9:1 ($\lambda = 0.9$) and the elasticity of demand η_D is 1. Then speculators will only need net worth of 1% of the total demand for the asset to create an unstable situation. This suggests that instability in markets may not be a particularly rare state of affairs. On the other hand, if leverage is moderate, say 1:1, speculators will need net worth of 25% of the total demand for the asset to create an unstable situation. Higher leverage is so dangerous because, as equation (56) shows, MinEAMASS goes as the *square* of the margin requirement $(1 - \lambda)$.

Once the instability ratio exceeds one and the market becomes unstable, we have seen that one of two events occurs. Either there is an immediate crisis where the price crashes and levered speculators are wiped out, or the price explodes up to v_π, the maximum the speculators will pay. The speed of this explosion prevents speculators from fully levering their net worth to buy the assets without driving the price up past v_π. After such an explosion, a small shock to retail demand does not change the price at all; speculators simply adjust their holdings to absorb the shock. However, if the shock to retail demand is large enough, the speculators will be forced to fully lever to absorb the shock. At this point, any additional shock will reduce the price of the asset below v_π and will force liquidations. The speculators will be constrained again and the demand curve below v_π again will be given by equation (54). This means we can still use equation (56) to assess the stability of

the market. However, it is important to note that the relevant leverage ratio λ is the maximum leverage ratio permitted by lenders (or targeted by investors), rather than the actual leverage ratio observed in the market.[14] These dynamics are discussed in more detail in chapter 2.

This section has developed the basic concept of MinEAMASS and the instability ratio, showing that there is a minimum elasticity-adjusted market size relative to leverage in the system that is required for asset markets to be stable. The instability comes from excessive leverage: if leveraged investors grow in the market, their perverse demand curves overwhelm the downward-sloping demand from more prudent investors, eventually causing total demand to become upward sloping. This makes the typical equilibrium between supply and demand an unstable "knife edge" with no mechanism to force a convergence to that equilibrium. The higher leverage is permitted, the more fragile the market in the sense that it takes a smaller share of speculators to cause an unstable situation.

4. An Operationalizable Version of the Model

In the name of expositional clarity, we have so far restricted ourselves to a market with only one asset, no short-selling, and only one class of speculator. Can we use this model to predict when instability will occur in the real world?

In this section, we shall expand the MinEAMASS framework to incorporate markets where different investors lever to differing degrees, markets where investors sell short and take positions using derivatives, and markets with multiple assets, which will give rise to contagion. We shall combine these to produce a measure that accounts for sufficient diversity that it might be used by macroprudential regulators to use it as an early warning sign against financial crises.

[14]This maximum leverage ratio is more difficult to observe, but fortunately for our model (though perhaps unfortunately for the world), fund managers with large amounts of excess capital tend to either return that capital to investors or put it to work in other strategies, particularly during periods of euphoria. We sometimes will use this fact to justify approximating the maximum leverage by the actual leverage at the top of bubbles. In general, this will bias our estimate of MinEAMASS downward because, because higher maximum leverage leads to greater MinEAMASS, *ceteris paribus*.

4.1. Demand of Speculators. Let us suppose that there are many assets, indexed by j, and that each asset has some derivative contracts associated with it, indexed by δ. Speculative investors are indexed by i and have net worth NW_i.

Each asset has a *collateral value*, the maximum amount that can be borrowed against it, which may vary over time and by investor. If the price of asset j at time t is p_t^j, then its collateral value for investor i is defined to be $\lambda_i^j p_t^j$. Each derivative contract δ on asset j also must be collateralized.

At each time t, investor i allocates a proportion of his net worth $\pi_{it}^{j\delta}$ (which may be a function of the price vector $\mathbf{p_t}$) to collateralize each asset and derivative contract in which he invests. We have assumed speculators are leverage-constrained, meaning they use all their capital: $\sum_{j\delta} \pi_{it}^{j\delta} = 1$.

Each investor's direct demand for the asset is given by the quantity he can buy with the share of his net worth he devotes to that asset:

$$\text{Direct Demand} = \frac{\text{Capital Devoted to Asset}}{\text{Collateral Requirement per Unit}} = \frac{\pi_{it}^j NW_{it}}{(1-\lambda_i^j)p_t^j} \equiv m_{it}^{jd} \tag{57}$$

Derivatives are treated as follows. We shall consider contracts with single-period margining, meaning that any changes in fair value of the contract are paid or received each trading period. The most common examples of such contracts are exchange-traded futures contracts, although credit default swaps (CDS) and interest rate swaps have similar features.

The value of each derivative contract δ on asset j is a function of the price of the asset, which we shall denote $f_{j\delta}(p_t^j)$. We treat a short sale as the special derivative contract where $f(p) = -p$. For each investor i, asset j, derivative contract δ, and time t, we shall say that he holds $C_{it}^{j\delta}$ derivative contracts. For each of these contracts, he must post a fixed dollar amount of collateral $\chi_{it}^{j\delta}$ as *initial margin* with the exchange or counterparty.[15]

[15] *Initial margin* is collateral deposited to guard against the possibility that a counterparty fails to deliver the daily margin payment required if the position moves against him. In the event of such a failure, the initial margin is used to make the daily margin payment and the contract is automatically closed.

We shall assume that there are speculators on both the long and short sides of the derivative contracts, so that C may be positive or negative. Of course, every derivative contract has two sides, so the net supply of derivative contracts must be identically equal to zero. Therefore, we assume the existence of *market makers* to absorb any disparity in demand between long and short speculative positions. For each derivative contract, the position $C_{ht}^{j\delta}$ held by market makers is just the inverse of the net position of the speculators:

$$C_{ht}^{j\delta} = -\sum_i C_{it}^{j\delta} \tag{58}$$

However, market makers do not wish to speculate. Instead, they hedge their position in the cash market. In order to be hedged, a market maker wishes to be indifferent to price changes in the underlying asset. He therefore demands assets m_t^h according to the condition

$$\frac{d}{dp_t}[C_{ht}f(p_t) + m_{ht}p_t] = 0 \tag{59}$$

or

$$m_{ht} = -C_{ht}f'(p_t) = \sum_i C_{it}^{\delta}f'(p_t) \tag{60}$$

Readers familiar with the practice of market making in options will recognize this condition as basic so-called "delta hedging" with $\Delta \equiv f'(p_t)$. Each investor i makes a contribution to the market maker's delta hedging activities for derivative contract δ in proportion to the investor's holdings. It therefore makes sense to refer to this contribution as the investor's indirect demand for the asset:

$$\text{Indirect Demand} = C_{it}^{j\delta}f'_{j\delta}(p_t^j) = \frac{\pi_{it}^{j\delta}NW_{it}}{\chi_{it}^{j\delta}} \cdot f'_{j\delta}(p_t^j) \tag{61}$$

Investor i's total effective demand for asset j is the sum of direct and indirect demand:

$$m_{it}^{j} = \frac{\pi_{it}^{j} NW_{it}}{(1-\lambda_i^j)p_t^j} + \sum_{\delta} \frac{\pi_{it}^{j\delta} NW_{it}}{\chi_{it}^{j\delta}} \cdot f'_{j\delta}(p_t^j) \tag{62}$$

4.2. "The Greeks": Delta and Gamma. At this point it is helpful to bring in two concepts from the options pricing literature and practice, delta (Δ) and gamma (Γ). Delta and gamma will be the building blocks of our stability analysis. For each investor i and asset j, his Δ_{it}^{j} is the change in his net worth for every dollar increase in the price of asset j. That is:

$$\Delta_{it}^{j} \equiv \frac{\partial NW_{it}}{\partial p_t^j} \tag{63}$$

To get an explicit expression for Δ, we write out each investor's net worth this period as his net worth last period, plus his profit from owning assets, plus his profit from the change in price of his derivative contracts:

$$NW_{it} = NW_{i,t-1} + (\text{Profit from owning asset}) + (\text{Profit from derivatives}) \tag{64}$$

$$= NW_{i,t-1} + \sum_{j} m_{i,t-1}^{jd}(p_t^j - p_{t-1}^j) + \sum_{j\delta} C_{i,t-1}^{j\delta}[f_{j\delta}(p_t^j) - f_{j\delta}(p_{t-1}^j)] \tag{65}$$

Differentiating equation (65) with respect to the price of a specific asset j':

$$\Delta_{it}^{j'} = \frac{\partial NW_{it}}{\partial p_t^{j'}} \tag{66}$$

$$= m_{i,t-1}^{j'd} + \sum_{\delta} C_{i,t-1}^{j'\delta} f'_{j'\delta}(p_t^{j'}) \tag{67}$$

$$= m_{i,t-1}^{j'} \tag{68}$$

Not coincidentally, the investor's Δ is his total net effective demand for the asset.

We will also be interested in gamma (Γ). For each investor i and asset j, Γ_{it}^{j} measures the change in the investor's exposure to the asset as its price changes, keeping his positioning constant. It represents the convexity of his net worth relative

to the price of asset j. Gamma is the price derivative of delta:

$$\Gamma_{it}^{j} \equiv \frac{\partial \Delta_{it}^{j}}{\partial p_t^j} \tag{69}$$

$$= \frac{\partial}{\partial p_t^j}\left[m_{i,t-1}^{j'd} + \sum_{\delta} C_{i,t-1}^{j'\delta} f'_{j'\delta}(p_t^{j'})\right] \tag{70}$$

$$= \sum_{\delta} C_{i,t-1}^{j\delta} f''_{j\delta}(p_t^j) \tag{71}$$

$$= \sum_{\delta} \frac{\pi_{i,t-1}^{j\delta} NW_{i,t-1}}{\chi_{i,t-1}^{j\delta}} \cdot f''_{j\delta}(p_t^j) \tag{72}$$

These definitions of delta and gamma are precisely the same as those in the options literature.

4.3. Stability Analysis. As with all our previous analyses, the market for each asset j will be stable if demand is downward sloping. As before, we add retail investors with a downward-sloping demand curve:

$$m_t^{j,\mathrm{gp}} = (1-\mu) N D_j(p_t^j) \tag{73}$$

The total demand curve is just the sum of the demand of all the investors:

$$m_t^{j,TOT} = \sum_i m_{it}^j + m_t^{j,\mathrm{gp}} \tag{74}$$

The slope of the demand curve is:

$$\frac{dm_t^{j,TOT}}{dp_t^j} = \sum_i \frac{dm_{it}^j}{dp_t^j} + (1-\mu) N D'_j(p_t^j) \tag{75}$$

$$= \sum_i \frac{dm_{it}^j}{dp_t^j} - \frac{\eta_j A}{p_t^{j2}} \tag{76}$$

where η_j (a positive number) is the elasticity of demand of retail investors and A is the total value of the assets they hold, as before. Expanding the total derivative in terms of partial derivatives, we have

$$\frac{dm_t^{j,TOT}}{dp_t^j} = \sum_i \left[\frac{\partial m_{it}^j}{\partial p_t^j} + \frac{\partial m_{it}^j}{\partial NW_{it}} \frac{\partial NW_{it}}{\partial p_t^j} + \sum_{\delta} \frac{\partial m_{it}^j}{\partial \pi_{it}^{j\delta}} \frac{\partial \pi_{it}^{j\delta}}{\partial p_t^j}\right] - \frac{\eta_j A}{p_t^{j2}} \tag{77}$$

We will aim to rewrite this equation in terms of Δ and Γ. Working just with the first term, we have:

$$\frac{\partial m_{it}^j}{\partial p_t^j} = \frac{\partial}{\partial p_t^j}\left[\frac{\pi_{it}^j NW_{it}}{(1-\lambda_i^j)p_t^j} + \sum_\delta \frac{\pi_{it}^{j\delta} NW_{it}}{\chi_{it}^{j\delta}} \cdot f'_{j\delta}(p_t^j)\right] \tag{78}$$

$$= -\frac{\pi_{it}^j NW_{it}}{(1-\lambda_i^j)p_t^{j2}} - \sum_\delta \frac{\pi_{it}^{j\delta} NW_{it}}{\chi_{it}^{j\delta 2}} \cdot f'_{j\delta}(p_t^j)\frac{\partial \chi_{it}^{j\delta}}{\partial p_t^j} + \sum_\delta \frac{\pi_{it}^{j\delta} NW_{it}}{\chi_{it}^{j\delta}} \cdot f''_{j\delta}(p_t^j) \tag{79}$$

If the initial margin on derivative contracts χ is proportional to the price, as is usual, then we can simplify further:

$$\frac{\partial m_{it}^j}{\partial p_t^j} = -\frac{1}{p_t^j}\left[\frac{\pi_{it}^j NW_{it}}{(1-\lambda_i^j)p_t^j} + \sum_\delta \frac{\pi_{it}^{j\delta} NW_{it}}{\chi_{it}^{j\delta}} \cdot f'_{j\delta}(p_t^j)\right] + \sum_\delta \frac{\pi_{it}^{j\delta} NW_{it}}{\chi_{it}^{j\delta}} \cdot f''_{j\delta}(p_t^j) \tag{80}$$

Substituting equation (62) in the first term and equation (72) in the second term gives:

$$\frac{\partial m_{it}^j}{\partial p_t^j} = -\frac{1}{p_t^j}\left[m_{it}^j\right] + \Gamma_{i,t+1}^j \cdot \frac{f''(p_{t+1}^j)}{f''(p_t^j)} \tag{81}$$

We can now substitute equation (81) into the first term in the demand curve (77):

$$\frac{dm_t^{j,TOT}}{dp_t^j} = \sum_i \left[-\frac{m_{it}^j}{p_t^j} + \Gamma_{i,t+1}^j \cdot \frac{f''(p_{t+1}^j)}{f''(p_t^j)} + \frac{\partial m_{it}^j}{\partial NW_{it}}\frac{\partial NW_{it}}{\partial p_t^j} + \sum_\delta \frac{\partial m_{it}^j}{\partial \pi_{it}^{j\delta}}\frac{\partial \pi_{it}^{j\delta}}{\partial p_t^j}\right] - \frac{\eta_j A}{p_t^{j2}} \tag{82}$$

Working now with the third term, we can substitute equation (63) to give:

$$\frac{\partial m_{it}^j}{\partial NW_{it}}\frac{\partial NW_{it}}{\partial p_t^j} = \left[\frac{\pi_{it}^j}{(1-\lambda_i^j)p_t^j} + \sum_\delta \frac{\pi_{it}^{j\delta}}{\chi_i^{j\delta}} \cdot f'_{j\delta}(p_t^j)\right] \cdot \Delta_{it}^j \tag{83}$$

$$= \left[\frac{m_{it}^j}{NW_{it}}\right] \cdot \Delta_{it}^j \qquad \text{by (62)} \tag{84}$$

$$= \left[\frac{\Delta_{i,t+1}^j}{NW_{it}}\right] \cdot \Delta_{it}^j \qquad \text{by (67)} \tag{85}$$

Substituting (85) into (82) gives the slope of the demand curve in terms of Δ and Γ:

(86)
$$\frac{dm_t^{j,TOT}}{dp_t^j} = \sum_i \left[-\frac{m_{it}^j}{p_t^j} + \Gamma_{i,t+1}^j \cdot \frac{f''(p_{t+1}^j)}{f''(p_t^j)} + \frac{\Delta_{i,t+1}^j \Delta_{it}^j}{NW_{it}} + \sum_\delta \frac{\partial m_{it}^j}{\partial \pi_{it}^{j\delta}} \frac{\partial \pi_{it}^{j\delta}}{\partial p_t^j} \right] - \frac{\eta_j A}{p_t^{j2}}$$

Note that while some of the subscripts in this equation have the value $t+1$, these values are nonetheless all known at time t.

To find MinEAMASS and evaluate the stability of an equilibrium, we will be interested in the sign of this derivative in steady state, *i.e.* when $p_{t+1}^j = p_t^j$. In other words, if p_t^j is an equilibrium, is it stable?

The condition for stability in asset market j is that $dm^{TOT}/dp < 0$. Imposing this condition, dropping the now superfluous subscripts j and t, and rearranging terms gives another form of the stability condition:

(87)
$$\sum_i NW_i \cdot (p\Delta_i/NW_i)^2 + p^2 \left\{ \sum_i \Gamma_i + \sum_{\delta i} \frac{\partial \Delta_i}{\partial \pi_i^\delta} \frac{\partial \pi_i^\delta}{\partial p} \right\} < \sum_i p m_i + \eta A$$

The left side of equation (87) is MinEAMASS; the right side remains the EAMS or elasticity-adjusted market share, given by the total assets of the speculators plus the elasticity-adjusted assets of the retail investors.[16] There are three contributors to the minimum elasticity-adjusted market size for stability, corresponding to the three terms on the left side of equation (87).

In the first term, each investor makes a contribution to MinEAMASS that is proportional to the square of his Δ relative to his net worth. Recall that Δ tells us the dollar amount that an investor's net worth changes for each dollar change in

[16]Equation (87) gives the stability condition where the independent variable is the price of an asset. However, many fundamentals-based investors consider relative value in their asset allocation decision, so that their demand for an asset depends not only on the price p^j of asset j but also on the price level of assets generally. It is easy to incorporate this into equation (87) via a change of variable. For example, we might let P be the general level of asset prices and $q^j = p^j - P$ be the idiosyncratic component of the price of asset j. Then $\partial/\partial q^j = \partial/\partial p^j$, so the equation does not change, but on the right side the elasticity η has a slight definitional change: it becomes

$$\eta = \frac{(q^j + P) \cdot [\partial D^j(P, q^j)/\partial q^j]}{D^j(P, q^j)}$$

the price of an asset. $(p\Delta/NW)$ is therefore a measure of leverage. Indeed, in the simple case where there are no derivative contracts and an investor is invested in only one asset, $(p\Delta/NW)$ is precisely equal to the conventionally-defined leverage ratio.

Equation (87) therefore shows that each speculator makes a contribution to MinEAMASS in proportion to his net worth and the *square* of his leverage. This non-linearity means that a single investor can have a large impact on the market. The fact that contribution to MinEAMASS is proportional to delta squared also means that a levered investor always makes a destabilizing contribution, whether he is long or short. If two levered investors enter into a futures contract, taking offsetting positions, both investors increase their squared delta and thus both contribute positively to MinEAMASS. MinEAMASS and instability increase with the *gross* notional exposure of speculators, not their aggregated net position, meaning that derivatives used for speculative purposes increase instability in proportion to the *square* of the *open interest* in the contract.

Finally, this term also contains a contagion effect. If the price of another asset falls, so that the net worth of an investor decreases while Δ^j stays constant, the first term in equation (87) will increase, MinEAMASS will rise, and the market will move closer to instability.

The second term of equation (87) is more subtle, but it will be intuitive to derivatives market participants as representing *convexity* or *gamma hedging*. This term arises because the value of a derivative may not be linear in the price of the underlying asset. As a result, if the price of the underlying asset changes, $f'(p)$ will change, and the market maker will no longer have a neutral stance with respect to the price of the asset (that is, equation 60 will no longer be satisfied). The market maker will therefore have to adjust his position in the underlying asset to remain hedged. This change in hedging demand in response to price is what is described by the second term, and it shows that the speculators contribute to instability and MinEAMASS in proportion to their net gamma position. Because gamma only

arises in relation to open derivative contracts,[17] the total gamma in the market is zero. However, we need to exclude from our calculation the gamma position of market makers, so this term will not in general be zero, although it is likely to be quite small compared to the first term. This leads to the interesting prediction that when speculators sell volatility, market stability is enhanced. This phenomenon was confirmed in a conversation with a market participant who managed a large portion of the derivatives portfolio at a megabank: he told me that when a large investor writes options, market makers delta-hedging their positions are forced to buy when the market falls and sell when it rises, reducing volatility. In practice, however, this convexity effect is highly likely to be destabilizing. This is because real-world market makers are almost always short convexity/volatility, so that other investors in the aggregate have positive Γ.

The third term in equation (87) captures the contribution to MinEAMASS from speculators rebalancing their portfolios in response to price movements. In practice, this term is likely to be negative, though this is not certain; sufficiently inelastic substitution away from appreciating assets can cause speculators actually to increase the proportion of net worth held in an asset as its price rises.

In this section, we have seen how to generalize the model to accommodate a wide variety of investors with different strategies and risk parameters. Each levered investor contributes to instability in proportion to the square of his leverage (Δ) and in proportion to his net volatility position (Γ). In theory, all the information necessary to evaluate the stability condition (87) could be collected by a systemic regulator and straightforwardly aggregated to evaluate the stability of the market. While such an undertaking may sound daunting, in fact the information is simply a standard set of summary statistics of the portfolios of investors and is already compiled daily (or even more frequently) by all sophisticated investors in their risk

[17]Recall that if the value of an investors' portfolio is a function $v(p)$ of the underlying price of the asset, then $\Gamma = v''(p)$. If the investor holds only outright long or short positions and has not entered into derivatives contracts, then the value of his portfolio is simply proportional to the price p, and $\Gamma = v''(p) = 0$.

reports. Many of these reports are already reviewed by the Federal Reserve. Aggregating these data is eminently achievable, and could allow authorities to have early warning of financial crises.

5. A Statistical Strategy for the Estimation of η_D

The elasticity of real-money demand, η_D, is a crucial parameter in the model. Here I propose a possible econometric strategy to estimate this parameter.

I propose first to capture order flow data from a sample of real-money investors over periods when there is no news, so that price movements cannot be attributed to changing fundamentals. Then, these can be combined with price movements on those days to obtain estimates of demand elasticity for each class of real-money investor. Finally, these disaggregated estimates of demand elasticity can be weighted by the share of the market to which they apply to obtain a total elasticity of demand for real-money investors.

The buy-side universe consists broadly of the following classes of investor:

- individual investors with cash accounts
- individual investors with margin accounts
- mutual funds
- pension funds
- insurance companies
- banks
- sovereign wealth funds
- hedge funds and private equity funds

Which of these should be counted as real-money investors probably varies slightly according to which market we are analyzing. Broadly, however, it makes sense to think of real money investors as including individual investors with cash accounts, mutual funds, pension funds, sovereign wealth funds, and those hedge funds that operate without leverage. It is also worth noting that some of these investors (like individual investors with cash accounts) are principals, while others (like mutual

funds) are agents acting on behalf of the principals. However, we are ultimately interested in the net demand by real-money investors, which is obtained by adding their demand through all channels through which they invest.

Today, individual investors with cash accounts trade in part through online brokerages and in part through full-service broker-dealers. The other investor classes trade largely through full-service broker-dealers.

According to classical asset pricing theory, asset prices depend on the risk free rate, a risk premium, and expected future dividends. To these variables, our theory adds a measure of downward-sloping demand and imperfect substitutability. We could thus estimate a real-money demand function for each investor type i with the following specification:

$$\Delta \log m_{it} = \eta_{Di} \Delta \log p_t + \alpha_{1i} \Delta \log P_t + \alpha_{2i} \Delta r_t + \alpha_{3i} \Delta VIX_t + \Delta \mathbf{X_t} \cdot \beta_\mathbf{i} + \epsilon_\mathbf{it} \tag{88}$$

Here, m_{it} is the quantity of the asset demanded by investor class i, p is the price of the asset, P is the price of close substitutes for the asset, r is the risk-free rate of interest, VIX is the level of the VIX index for the asset class in question (a proxy for the risk premium), $\mathbf{X}$ is a vector of fundamentals that inform expected future dividends, and ϵ is an error term.

One way to estimate this equation would be using order flow data from a large broker-dealer. Such an institution would have records of every order received, the client who placed the order, whether the order was executed, and at what price. Each executed order represents a change in positioning of an investor. A complete set of orders in a single security could be gathered for a subset of clients and analyzed.

Also needed are the underlying position data. These could be gathered either from the broker-dealers' counterparties directly, or from custodians such as the Depository Trust and Clearing Corporation which keep records of the holders of record of securities for the purposes of dividend and interest payments. To the extent that securities are held in accounts with the broker-dealer itself, positioning data might also be available through the broker-dealer.

Pricing data is easily obtained from sources such as Bloomberg or Yahoo! Finance.

Finally, we need to estimate the change in dividend expectations underlying the asset. Perhaps the easiest way around this problem is to only estimate the regression on days where there is no news related to those future dividends. Alternatively, one might use analysts' price targets on the stock as a proxy for fundamental value, which would allow the regression to be run over longer time frames.

Once we have estimated the η_{Di} for each class of investor, these can be simply added in a weighted-average summation according to the market share of each investor class to provide an estimate of η_D for real money investors in aggregate. To see this, we simply write out η_D as follows:

$$\eta_D = -\frac{pD'(p)}{D(p)} \tag{89}$$

$$= -\frac{p\sum_i D_i'(p)}{D_(p)} \tag{90}$$

$$= -\sum_i \frac{pD_i'(p)}{D_i(p)} \cdot \frac{D_i(p)}{D(p)} \tag{91}$$

$$= \sum_i \eta_{Di} \frac{D_i(p)}{D(p)} \tag{92}$$

This provides an estimation method for the elasticity of real-money demand.

6. Conclusion

In this chapter, we have developed a simple model of asset pricing that focuses on leverage-induced constraints on investors, rather than expectations, as the primary driver of asset prices. The model shows that if levered investors accumulate too much of any particular asset, the market becomes unstable and vulnerable to collapse from even small shocks. The most important contribution stemming from the model is that it gives rise to a leverage-based measure of financial instability, enabling a quantitative answer to the question of how much leverage is too much. The data

necessary for the measure could easily be gathered by systemic regulators, so that the measure could be made operational.

In our model, we have combined speculators and retail investors who suffer from limits on the size of positions they can take in markets. Retail investors invest without leverage and have downward-sloping demand curves for assets; speculators invest with leverage but are limited by their lenders or the authorities in the maximum permitted leverage ratio. The maximum leverage ratio is exogenous to the model, and this permits the modeler to investigate how the behavior of the market changes when lenders change their maximum leverage ratio.

While no model can be completely divorced from expectations, the result of this simple setup is that expectations play only a minor role in determining asset prices, rather than being the central ingredient. The asset prices calculated in the model are robust to moderate random changes to expectations. Specific expectations play a role in just one place: the model requires that there be a small group of speculators who believe the asset will appreciate in the near term. This is to be expected in a large population, and the dynamics of the model are robust to changes in those expectations provided they remain optimistic.

The model's speculators also leverage themselves to the absolute maximum extent possible, so that any negative price movement results in forced liquidations. To follow a strategy like this in the real world would be folly; real-world investors aim to keep some "dry powder" to allow them to withstand an adverse move in asset prices.

However, even with the cushion a large enough decline in asset prices still will bring an investor up against the hard constraint on leverage and force immediate liquidations. The possibility of an instability-induced crash in prices that wipes out speculators thus remains an important part of our story, although actual liquidations are likely to be "sticky" in a way that they are not in this model. Systems with such sticky dynamics have been studied in biology and physics,[18] among other places,

[18]See, for example, Koch (1999) for a study of such behavior in neurons.

and exhibit quite complex behavior when changes are small. At times of collapse, however, the stickiness becomes much less relevant. The behavioral rules in this model thus are reasonable approximations to real-world behavior.

In the interest of empirical relevance, I have taken as given some of the critical institutions and policies in the market, such as banks, an interest-rate-targeting central bank, margin requirements, and barriers to investing with leverage. By assuming incomplete markets, we also prevent recourse to arguments that eliminate crisis outcomes because investors are fully hedged.

The model also permits investigation of contagion in a less complex framework than is usually possible, and shows that contagious instability is much more likely to occur during crashes than during spikes, because in a crash net worth declines, exacerbating constraints on levered investors.

This chapter has used the model to define financial instability as instability of supply-demand equilibria, and has investigated the moment when a stable market becomes an unstable one. This moment may give rise to a severe crash in asset prices accompanied by a sudden evaporation of bids and offers from the market, an event that we have interpreted as a liquidity crisis.

We have seen that markets tend to become unstable when speculators accumulate too large a share of the assets. The total net worth and the distribution of net worth assigned to speculation together determine a minimum market size for the market to be stable. The ratio of this minimum market size to actual market size defines an instability ratio which determines how close the market is to an instability-induced crisis. We have seen that instability rises with the square of speculators' leverage and with their gross notional exposure, rather than net exposure.

The measure we have presented is sufficiently general and simple that it could be calculated and applied by central banks to provide advance warning of a crisis.

CHAPTER 2

An Application: A Model of Bubble Formation

The existence of asset bubbles has perplexed economists for as long as they[1] have been around. Since the time of Adam Smith we economists have tended to view individuals as more-or-less rational acting in their own self-interest, but euphoric asset bubbles followed by despondent crashes are, at least on their face, difficult to reconcile with anything approaching rational behavior. While the literature (for example, Blanchard and Watson, 1983) has produced bubbles consistent with rational behavior under certain conditions, in this paper there is no discussion of rationality and the pattern of the bubble to be modeled is consistent with either rational or irrational behavior.

In this chapter, we shall see how economies with investors who behave as we have modeled naturally develop asset price bubbles that spontaneously burst. The model presented here robustly produces many of the dynamics associated with bubbles that many other models fail to explain.

While there is no precise agreed-upon definition, the economics literature generally views a bubble as any persistent deviation of asset prices from their fundamental values—that is, the net present value of future cash flows.[2] However, such a definition is not very satisfying. In popular parlance, an "asset bubble" seems to refer to specific and much narrower phenomenon, a specific pattern of euphoric, often debt-based asset price appreciation followed by a crash that has been repeated with some regularity. In just the last decade, we have seen housing and technology stock bubbles in the United States and a stock price bubble in Shanghai, and that is just in the $1 trillion-plus category. As Kindleberger would no doubt point out if he

[1]Economists, not asset bubbles. Asset bubbles have been around longer than economists.
[2]See, for example, Blanchard and Fisher (1990), pp. 213-224.

were still alive, these modern bubbles followed similar patterns to bubbles throughout history, from the Dutch Tulip Bulb Bubble to the South Sea Bubble to the 1929 stock bubble, and many smaller ones along the way.

The pattern is set out in some detail in Kindleberger (1996), who applies, in narrative form, what he calls the "Minsky Model" of financial commitments and instability (see Minsky, 1986) to speculative manias and asset bubbles. While I will not repeat Kindleberger's generic narrative in full here, it is worth recalling a few of his stylized facts that apply to all or nearly all bubbles—facts that simple formal models have had difficulty explaining or replicating simultaneously.

- A common feature of a great many bubbles and crashes is the use of high leverage to purchase speculative assets, causing speculators to quickly lose all their net worth (and sometimes more) in the ensuing crash. The growth of leverage is often, though not always, associated with easy monetary policy and credit conditions.
- Bubbles tend to begin with a small group of well-informed insiders. These insiders profit inordinately as the universe of investors expands into ever-wider circles, driving the price up.
- As the bubble begins to grow, there is often an identifiable "knee" in the graph of price against time. Before this point, the growth rate of prices is slow and steady; after it the growth rate markedly increases. However, an econometric identification of this concept has been elusive.
- As the bubble gathers steam, the bubble asset appears to be in shortage, with offerings oversubscribed, prices bid up well above initial offering prices, and investors disappointed by unsuccessful attempts to buy assets.
- Volatility and trading volumes both tend to increase more than proportionately with the price. At the top of the bubble, volatility and volumes explode to unprecedented levels.

- Just before the bubble bursts, there is often a final, large price spike that dwarfs previous moves. The spike is then quickly reversed in the ensuing crash.
- In the eventual crash, buyers for the assets disappear suddenly and almost completely. There is panic as everyone attempts to sell at once, the price gaps downward as demand appears to become discontinuous. Often, market participants complain that "there are no bids at any price." (Such a lament is particularly perplexing to economists as we are taught that there is always a market-clearing price for any good. Usually, we react to statements like this by claiming they simply are not true—what the market participant means is that there are no bids at any (in his view) *reasonable* price.)

These stylized facts about bubbles occur with sufficient regularity that a theory of bubbles ought to be able to explain or produce at least a large majority of them as an integral part of the theory, and to demonstrate how they relate to each other. In what follows, we shall see that the model of the previous section naturally produces bubbles and generates these dynamics from a single, deeper, underlying process that accords with a stylized description of micro behavior.

The modern formal literature on bubbles, reviewed in an excellent survey article by Brunnermeier (2008), has had some difficulty replicating key aspects of this story. Brunnermeier divides the literature into four main strands.

Most closely associated with the neoclassical school is the literature on rational bubbles, seeded by Blanchard and Watson (1983). The rational bubbles literature explains bubbles as the result of self-fulfilling expectations of rational actors: if everyone expects asset prices to continue to increase, it is rational to pay continually increasing prices. Blanchard-and-Watson-style models leave much to be explained. Because they are representative agent models, there is no trading so they do not address trading volumes. There is no role for debt, nor is there any explanation of why or when the bubble bursts. A stochastic version of the model, where the bubble has a small probability of bursting each period, does not remedy this problem, as

ultimately changes in price must be reflected in changes in demand or supply; there is no mechanism in the model for this.

Further, although Blanchard and Watson do not make this point, their model can actually justify any price path simply by assigning a positive probability to that price path. Using their notation, if c_t is the bubble component of the asset price and θ^{-1} is the required gross return, then rational expectations requires that the expected price next period must be the price this period plus the required return:

$$E_t(c_{t+1}) = \theta^{-1} c_t$$

If expectations are that c_{t+1} is stochastically distributed with probability density function $\psi_{t+1}(c_{t+1})$, then any given realization c_{t+1} is consistent with rational expectations if the *ex ante* probability of it occurring was small. This argument can be iterated one period forward, so that *any* price path $\{c_t, c_{t+1}, c_{t+2}, \ldots\}$ is consistent with rational expectations provided it is in the support of expectations, whether or not the path has anything to do with the fundamental value. The only requirement is that the *expected* rate of return is θ^{-1}, or formally:

$$\forall t, \int c_{t+1}\, \psi_{t+1}(c_{t+1}|c_t)\, dc_{t+1} = \theta^{-1} c_t$$

This is not a particularly stringent condition. To summarize then, while the rational bubbles class of models is consistent with any particular realized price path, it does not explain one path over another, nor does it address volume or volatility behavior. There is also no role for debt. Finally, there is no explanation for what, if anything, eventually causes the bubble to burst.

Brunnermeier's second strand of the literature concerns the possibility that while all investors may be aware of the bubble, and all are rational, there is no common knowledge of awareness of the bubble. As a result, investors hold assets they know are worthless in the hope of being able to offload them at a higher price later. Allen *et. al.* (1993) and Abreu and Brunnermeier (2003) are emblematic of this literature. While this mechanism is able to explain both a precise path and the end of the

bubble in a plausible way, its predictions are very difficult to test in a quantitative way because they require a detailed and quantitative knowledge of agent's beliefs and information sets. This is particularly challenging since this information often is not expressed quantitatively even by those making judgments on it. However, these models have a lot of support in anecdotal evidence from traders describing their own behavior. Although it is not discussed in their paper, Abreu and Brunnermeier's model at least sometimes produces a spike in volumes at the moment of the crash.

The third strand of literature takes as its driving force the limitations on arbitrage. In this literature, rational investors would like to arbitrage away mispricings, but are unable to do so because their ability to arbitrage is limited by their capital. These investors face a similar constraint to that of the speculators in chapter 1 of this paper. Unarbitraged deviations from fundamental value are driven by noise traders (De Long *et. al.*, 1990 and Sheifer and Vishny, 1997). This literature takes what should be relatively uncontroversial assumptions—capital constraints on arbitrage and noise in securities demand—and uses these assumptions to show that mispricings can persist. While the conclusion is compelling, it remains to explain the type of price path that bubbles often take, the patterns of volumes and volatility typically observed, and the relationship of credit growth to the growth of the bubble. Indeed, in these models, loose credit should expand the capital available to arbitrageurs and therefore damp the bubble, rather than fuel it.

Finally, Brunnermeier provides some examples of a literature demonstrating how heterogeneous expectations combined with short-sale constraints can cause a bubble. At its most basic, this phenomenon is simply the winner's curse applied to stock markets, where those with the highest expectations purchase the assets so that they are systematically overvalued. If heterogeneous expectations evolve, this generates an option value that further raises the price. While some specific arrangements (Scheinkman and Xiong, 2003) can associate volumes and volatility, the growth path and association with leverage remain unexplained, and the bubble does not burst until expectations are updated, perhaps because of news.

A final strand of the literature not reviewed by Brunnermeier is a class of models demonstrating how agency problems may cause investors to overpay for assets (Allen and Gale, 2000). Here, investors who buy assets with non-recourse loans essentially receive a free put option, suffering only part of the loss of an assets value but getting to keep all of any gain. It is clear how this will cause investors to overpay for assets, and it is also clear that the more leverage a bank is willing to provide, the less is the investor's maximum loss and the more he will be willing to pay for the asset. However, these models do not imply an accelerating price path of bubbles or a relationship of volumes to volatility and debt.

The model in this paper borrows from many of these strands of literature, and it is consistent with all of the literature in the sense that any of the effects discussed could be overlaid on it. We will use the model of the previous chapter with two modifications. First, speculators will recruit retail investors into the bubble, which will provide a source of increasing demand for the asset. These new speculators are necessary to provide an increasing source of demand for the asset before the market becomes unstable. Second, there will be some "producers" who manufacture new assets provided the price is sufficiently high. This refinement is not strictly necessary but is meant to capture the idea that when financial assets are overpriced, agents will attempt to increase issuance.

Put into the context of the literature, the mechanisms fueling the bubble are limits to arbitrage (speculators are prevented from bidding the asset up to their estimate of value) and heterogeneous beliefs (in the form of downward-sloping demand and the divide between speculators and retail investors). However, in the present model there is no discussion of rationality or fundamental value. The model is consistent with either rational or irrational behavior, and is consistent with a bubble that is fundamentally justified or one that is not. The wildly optimistic speculators may be correct or incorrect.

The model we will explore naturally and robustly produces Kindleberger's stylized facts even without the complicated effects that could be added based on the

earlier literature. A slow build occurs as awareness of the bubble asset broadens from insiders to a wider audience. Eventually, the bubble asset comes into shortage when new assets cannot be created fast enough to meet demand, and the price is bid up. As debt-ridden speculators accumulate more and more of the bubble asset, the price becomes more volatile and trading volumes increase because speculators actively manage their exposure in response to fluctuations in price, which is driven by random changes to retail demand. As accumulation by the speculators continues, the market approaches the instability threshold, leading to a final, self-fueling spike before a crash that the model predicts and explains. After a crash, the asset price is below its pre-bubble value for a time because the most optimistic investors have lost their capital, removing a source of demand from the market.

In the spirit of Minsky (1986), the growth of a bubble creates the specific, fragile financial conditions that both pave the way for a crash and determine its dynamics. It therefore examines the bubble and crash phase in concert, beginning with a fairly simple natural model of the growth of bubbles, demonstrating how it creates conditions for a crash, and then exploring the dynamics of the crash.

We have just discussed some stylized facts about the growth and bursting of bubbles and provided a brief overview of the literature that attempts to explain them. Section 1 applies the model of the previous chapter to explore the formation of a bubble and crash. Section 2 addresses the case where the instability in the market initially results in a price spike instead of a crash. Section 3 discusses some policy implications.

1. A Model of Bubble Formation

Thus far, we have seen how, if levered speculators accumulate a large enough position in an asset, total demand for that asset may become upward-sloping. Once this occurs, the price of the asset may either crash back to a level at which unlevered investors are are able to absorb all the assets, or it may rise rapidly as demand outstrips supply in ever-greater quantities until the price reaches a level so high that

even the speculators no longer wish to hold the asset or some other factor results in the bubble bursting. In this section, we will initially examine a rather natural way that a bubble might grow, driven by levered speculators.

We will use the model from the previous chapter, with speculators, retail investors, banks and a central bank, with two modifications.

First, we introduce "producers" who manufacture new assets at a rate $\dot{m}(p)$ which depends on the price, so that the quantity of assets available increases from period t to period $t+1$ by $\dot{m}(p)\Delta t$. $\dot{m}(p)$ may be thought of as a kind of supply curve for assets.

Second, we permit speculators to "evangelize," telling others that they meet about the tremendous opportunities in the asset. A certain share of these contacts become speculators themselves and sink their entire net worth, with leverage, into the asset.

The speculators, as before, are extremely enthusiastic about the asset, so they buy as much of it as their lenders will permit. They initially represent a small portion of the population. However, as they grow wealthy and evangelize, they eventually accumulate a large share of the asset. Their upward-sloping demand curve then overwhelms the downward-sloping demand curve of the retail investors and the market becomes unstable, marking the end of the appreciation phase.

Formally, speculators and retail investors have demand curves exactly as described in the static analysis earlier:

$$m_t^{\text{lv}} = \frac{m_{t-1}^{\text{lv}}}{1-\lambda} \cdot \left[1 + \frac{d\Delta t - \lambda(1+r\Delta t)p_{t-1}}{p_t}\right] \tag{93}$$

$$m^{\text{gp}} = (1-\mu_t)N \cdot D(p_t) \tag{94}$$

Additionally, there are $\Delta\mu \cdot N$ new speculators who are recruited by the existing speculators, with demand curve:

$$m_t^{\text{ns}} = \Delta\mu_t \cdot N \cdot \frac{y_t}{(1-\lambda)p_t} \tag{95}$$

where y is the net worth of the new speculators.[3]

It is not particularly important for the qualitative results of the model how new speculators are recruited, provided there is a steady stream to provide increasing demand for the asset. However, I will model new speculators as being recruited via a "handshake" model as follows. Each speculator goes out into the market and meets random people, at a rate $1/\tau$. If the new contact is not already a speculator (probability $1-\mu$), he is converted into one with probability α. This yields a differential equation for the recruitment of new speculators:

(100)
$$\frac{d\mu}{dt} = \frac{\text{(Odds of Converting)} \times \text{(Prop. of Speculators)} \times \text{(Odds of Meeting Non-Speculator)}}{\text{Average Time to Meet New Person}}$$

or:

$$\frac{d\mu}{dt} = \frac{\alpha\mu(1-\mu)}{\tau} \tag{101}$$

which solves to yield

$$\mu(t) = \frac{\mu_0}{\mu_0 + (1-\mu_0)\exp(-\alpha t/\tau)} \tag{102}$$

Equation (102) is a sigmoid function that sensibly yields $\mu(0) = \mu_0$ and $\mu(\infty) = 1$. I assume that once speculators are recruited they remain speculators for the remainder of the simulation.[4]

[3]In continuous time, equations (93) and (95) can be combined to give

$$\frac{dm^{\text{lv}}}{dt} = \frac{m^{\text{lv}}}{1-\lambda} \cdot \left[\frac{d+\lambda\dot{p}}{p} - \lambda r\right] + \frac{\dot{\mu}Ny}{(1-\lambda)p} \tag{96}$$

The system to be solved or simulated is then this equation, equation (102) and the following:

$$m^{\text{gp}} = (1-\mu)N \cdot D(p) \tag{97}$$

$$m = m^{\text{lv}} + m^{\text{gp}} \tag{98}$$

$$\frac{dm}{dt} \equiv \dot{m}(p) \tag{99}$$

[4]As a possible refinement to the model, one could assume that speculators return to being retail investors with some probability, which might be dependent on the asset price. While this would not make a difference in the initial qualitative dynamics of the growth of the bubble, if the exit of speculators was sufficiently price-sensitive then this might put an endogenous limit on the size of the bubble. In other words, there would be a threshold level of rationality sufficient to prevent the bubble from reaching the point of instability.

Equilibrium is just given by the condition that supply equals demand:

$$m_t = m_t^{\text{lv}} + m_t^{\text{gp}} + m_t^{\text{ns}} \tag{103}$$

This system has six equations (93-95, 102 and 103 plus the rate of creation of new assets $\dot{m}(p)$) and six unknowns ($m, m^{\text{lv}}, m^{\text{gp}}, m^{\text{ns}}, \mu$ and p). It can therefore be solved or simulated.

Figures 1 and 2 show a simulation of the system for a non-stochastic case and a stochastic case ($D(p)$ is buffeted with small random shocks each period). Appendix A shows 12 consecutive simulations of the stochastic case to give a flavor of the variability in the results. Parameters are given in the footnote.[5]

The simulations reproduce many of the important features of bubbles that were noted in the introduction.

Most importantly, leverage and credit, and the ability to obtain it, are the driving forces behind the inflation of the bubble. This occurs in a natural way and it does not depend on difficult-to-verify changes in expectations, the public mood, or other emotional variables that contain a significant arbitrary component. To the extent that expectations play a role, they do so only in the decision to become a speculator. The force restraining the bubble from immediate explosion is the existence of margin requirements—the ability to speculate is limited by net worth and thus the bubble builds slowly over time.

---

[5]The parameters of the model are as follows:
Initial proportion of the population that are speculators: $\mu_0 = 0.04$
Population size: $N = 10000$
Time constant for how fast agents meet each other: $\tau = 0.9$
Probability of converting a handshake into a new speculator: $\alpha = 0.4$
Number of time steps: 2000
Size of each time step: $dt = 0.005$ (approx. 1 trading day)
Initial wealth of all agents: $y = 1$
Retail demand curve: $D(p) = \phi_t/p, \phi_0 = 0.5$
Shock to retail demand curve: $\phi_t = \phi_{t-1}\exp(\epsilon_t), \epsilon \sim U[-0.007, 0.007]$
Initial quantity of bubble asset: $m_0 = 5200$
Leverage: $\lambda = 0.3$
Rate of production of new assets: $\Delta m_t/\Delta t = 0.02$ if $p_t > 1$, 0 otherwise
Margin interest rate: $r = 6\%$
Dividends per share: $d = 0.02$

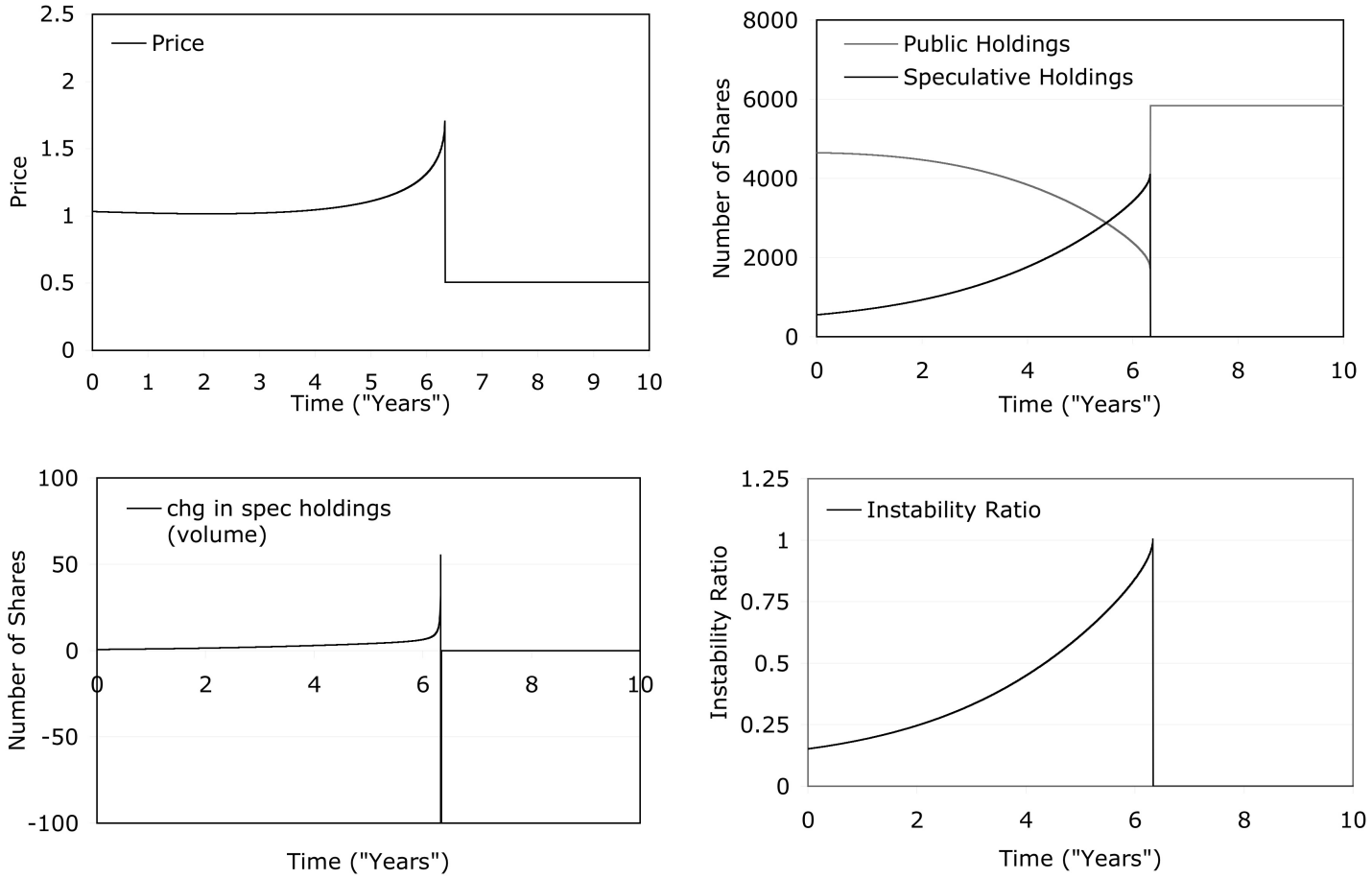

FIGURE 1. A Bubble with Deterministic Retail Demand

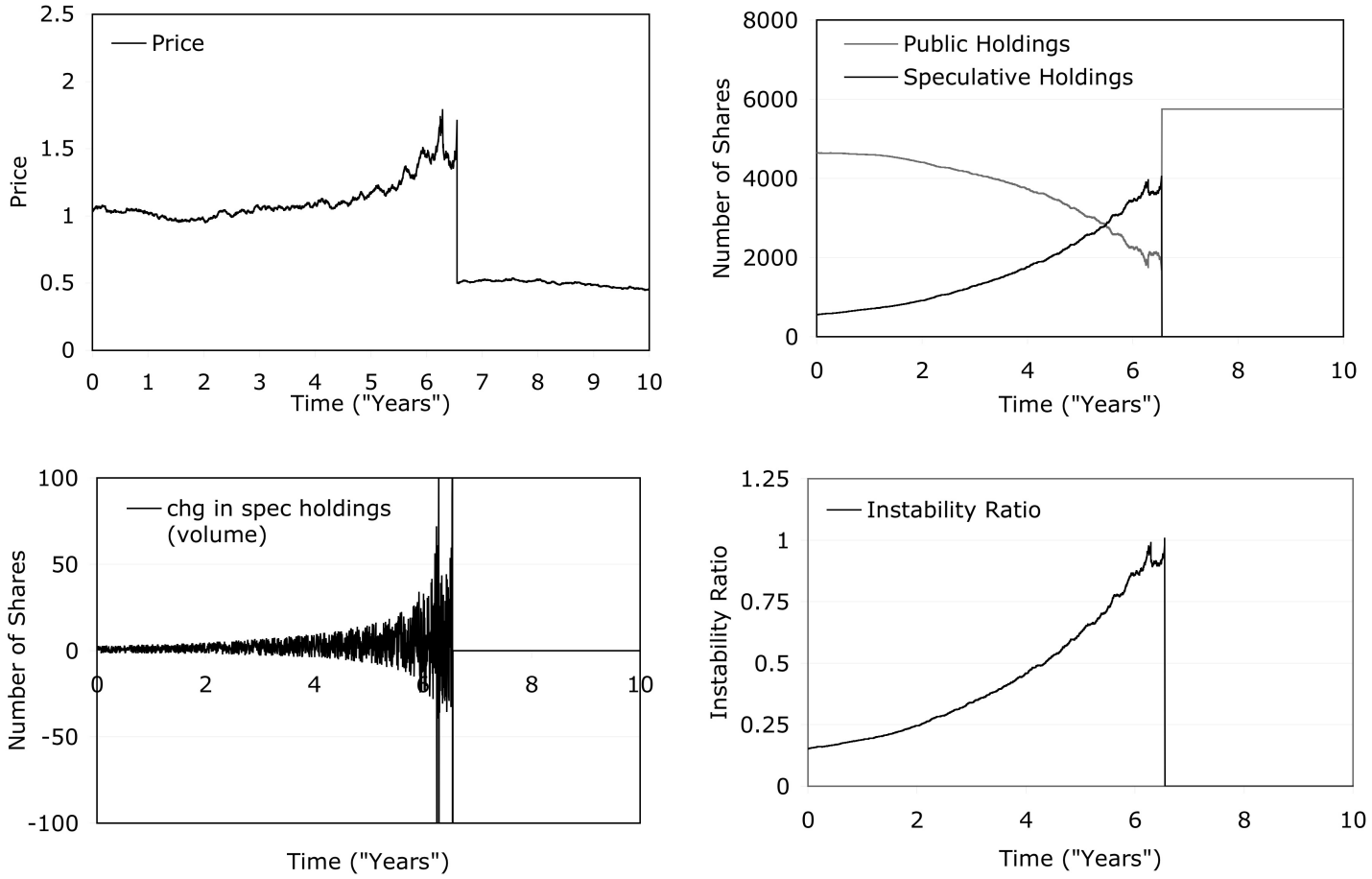

FIGURE 2. A Bubble with Stochastic Retail Demand

Second, the model produces the two important associations in trading statistics. As the price rises and the bubble inflates, volatility and volume also increase, driven by two different forces. Volatility increases because as the bubble inflates the demand curve becomes closer to vertical (the influence of leveraged speculators is growing). With a supply curve that is also vertical in the short run, small shocks to demand have large price effects.[6]

Volume is driven by speculators adjusting their portfolios in response to price changes. As the price increases and speculators accumulate a larger share of the assets, a given price movement causes more assets to be traded. The increasing volatility also drives volume because a given shock to demand for assets has a larger effect on price, causing larger required portfolio adjustments.

To see this formally, we consider the price effect of a small shock to retail demand. Let ϵ be a small random shock to retail demand, so that total demand is given by

$$m_t = (1-\mu)N \cdot (1+\epsilon)D(p) + \frac{m_{t-1}^{\text{lv}}}{1-\lambda} \cdot \left[1 + \frac{d\Delta t - \lambda(1+r\Delta t)p_{t-1}}{p_t}\right] \tag{104}$$

We wish to examine how price changes with the shock: $dp_t/d\epsilon$. Again taking the limit as $\Delta t \to 0$ and differentiating with respect to ϵ, we have:

$$0 = (1-\mu)N \cdot (1+\epsilon)D'(p_t)\frac{dp_t}{d\epsilon} + (1-\mu)N \cdot D(p_t) + \frac{m_{t-1}^{\text{lv}}}{1-\lambda} \cdot \frac{\lambda p_{t-1}}{p_t^2}\frac{dp_t}{d\epsilon} \tag{105}$$

which simplifies to

$$\frac{d\log p_t}{d\epsilon} = -\left[(1+\epsilon) \cdot \frac{p_t D'(p_t)}{D(p_t)} + \frac{\lambda}{1-\lambda} \cdot \frac{p_{t-1}m_{t-1}^{\text{lv}}}{p_t m_t^{\text{gp}}}\right]^{-1} \tag{106}$$

We examine the log of price because we are looking at volatility, which is properly viewed as the *percentage* change in price for a given shock. Note that the first term of the bracketed expression is negative, and that stability requires that it dominates the second term for ϵ close to 0 (by eq. 56). Thus, $d\log p_t/d\epsilon$ is positive, meaning

[6] In real bubbles, it is often the case that volatility first declines as the bubble inflates and then rises again as the bubble nears its peak and begins to burst. This more nuanced pattern may be accounted for by a growing willingness of real-money investors to provide liquidity as an asset becomes more well-known, a dynamic that is outside the scope of this model, but which fits anecdotal evidence well.

that a positive demand shock leads to a higher price as expected. Looking at the second term, we can see that a shock has a larger price effect if the speculators hold a larger share of the assets relative to the retail investors and if the permitted leverage ratio is larger. Additionally, as the bubble nears its peak, the bracketed term approaches zero and volatility increases without bound. All of these match the stylized facts discussed in the introduction.

Volatility drives volumes, which are the result of speculators readjusting their portfolios. Neglecting any new speculators, and subtracting m^{lv}_{t-1} from both sides of equation (46) and simplifying gives:

$$\Delta m^{\text{lv}} = m^{\text{lv}} \cdot \frac{\Delta p}{p} \frac{\lambda}{1-\lambda} \tag{107}$$

Because investors in this model only trade due to random shocks or because speculators are rebalancing their portfolios. Δm^{lv} is thus the entire volume of traded shares. It increases with the number of shares held by the speculators, so that volume naturally increases in line with the size of the bubble. However, it is also proportional to the change in price, which in the presence of shocks is determined by volatility. Thus, volume increases more than proportionately with the size of the bubble.

Third, the dynamics of the system tend to produce a price path that grows faster than exponentially. This can be seen most easily in the case without the shock in figure 1. When we introduce stochastic shocks to retail demand, we can often get something that looks like a "knee" of the curve, where the rate of appreciation and volatility regime appear to change sharply, although this is not econometrically identified. This does not occur in figure 2 but can be seen in some of the simulations in Appendix A. Typically, this sharp uptick in capital gains is caused by a small random shock in demand for the assets on the part of retail investors. That small increase causes the price to increase, providing a windfall to the speculators who then reinvest their profits. Issuance is unable to keep pace with demand. We are, so to speak, "off to the races."

The bubble finally bursts when speculators accumulate too large a share of the assets, so that the instability ratio exceeds one. At this point, the price may crash as speculators are forced to liquidate in a vicious circle, or it may spike sharply and discontinuously before ultimately crashing, a dynamic which we shall examine in the next section.

If the price crashes at this stage, the model generates what is truly a "pop" in that it is a large and rapid price decline. The speed and the severity of the decline is not the result of changing expectations, but rather of the market institutions and terms of credit. It is a structural feature of the model. Indeed, the model replicates a very interesting and much glossed-over aspect of real-world bubbles: as the bubble begins to burst, the initial crash in prices happens *in spite of* the optimistic expectations of the speculators, who may view price declines as "crazy" or "not reflective of the fundamentals," but are being forced to liquidate and are unable to express their views in the market. In other words, as bubbles burst, expectations follow prices, not the other way around.

Prices are determined not by value, but by the availability of funds to make purchases, availability that vanishes remarkably quickly as equity is depleted and risk-taking capacity is sharply reduced. The collapse is therefore irreversible, another important feature of real-world bubbles. This is an important test: if the collapse of bubbles were driven primarily by expectations, there is in principle no reason why the sudden pessimism should not be reversed, reinflating the bubble. In this model, the initial collapse bankrupts the levered investors; quick reinflation is not possible because the optimists have lost all their capital.

Some authors, such as Galbraith (1954), Minsky (1986) and Koo (2008), have taken the failure of bubbles to reinflate as evidence that bubbles are secular phenomena. They argue that once a bubble bursts, the same asset is unlikely to experience another bubble in the same generation, because investors who have already been burned are likely to be more cautious in the future. This is especially true when asset bubbles grow economy-wide, as they did in the U.S. Great Depression, Japan's

Great Recession, and the most recent U.S. housing and stock bubbles. The first two cases were accompanied by long-term declines in agents' leverage ratios to levels that had not been seen for decades; whether this occurs in the current cycle remains to be seen.

While this temporary immunity in the aftermath of a bubble bursting is not an integral part of the present model, it can be easily included. We might think of this Galbraith/Minsky/Koo story as arguing that when investors have learned from the collapse of the bubble, they are less likely to be convinced to become leveraged speculators. In our model, that is equivalent to a reduction in α in equation (102), reducing the rate of recruitment of new speculators into the bubble. As we shall see in equation (132) below, this reduction in recruitment can be enough to prevent price appreciation from getting off the ground to begin with. So it is easy to incorporate the Galbraith/Minsky/Koo arguments into this model.

In short, this section has shown how leverage can produce steadily growing bubbles that spontaneously become unstable. The model naturally produces the dynamics of the inflation and bursting of the bubble that appear in the real world, and that other, expectations-based models have had difficulty replicating. From just a few simple equations, we have been able to produce suggestive behavior for volatility and volumes, to explain the crucial importance of leverage and credit, to generate a large price spike followed by a truly irreversible "pop," to produce the proper relationship between prices and expectations, and to maintain consistency with many earlier writings on the secular nature of bubbles. The model explains why these features are integral to the particular phenomenon we call an asset price bubble.

2. The Case of a Price Spike

Up until now, we have generally assumed that when equilibrium in an asset market becomes unstable, a crash results. However, in principle, an alternative result may be an upward explosion in price. In chapter 1, we saw how this was

less likely to be the case in a multi-asset world. Here, however, we include that possibility for completeness. We shall see that even if the price spikes at first, it may ultimately collapse endogenously as the speculators become more levered and are forced to liquidate.

Let us suppose that at the moment the demand curve becomes upward sloping so that the market is unstable and the price action is explosive. The price will then explode upward to the highest price the speculators are willing to pay.

Consistent with our model of chapter 1, we shall assume that the speculators believe the asset will be worth a very high price v_π at some future date T, and that they require a return of ρ (greater than the rate on margin loans r) in order to hold the asset. Thus, at any time $t < T$, the speculators will want to maximally lever provided the price is less than $v_\pi(1+\rho)^{t-T}$.

We shall also assume that unconstrained investors who do hold the asset demand a similar risk premium, so that if the asset appreciates as expected by $r + \rho$ then their demand does not change. Formally, this corresponds to the assumption that:

$$D_T(p) = D_t[p(1+\rho)^{t-T}]$$

Finally, we assume as before that new assets are created at rate $\dot{m}$, or ΔM per period.

Let us consider the demand curve for assets at the moment the bubble becomes explosive. At prices above $v_\pi(1+\rho)^{t-T}$, the demand comes only from a small number of retail investors, equal to $(1-\mu)ND(p)$. At the other end of the spectrum, where the price is so low that speculators are bankrupt, demand also comes only from retail investors and is again equal to $(1-\mu)ND(p)$. In the intermediate range, demand is upward-sloping as derived in section 3. The supply-demand diagram for the asset is thus shown in figure 3.

The new equilibrium after explosion is at point E. It is easy to see that at this new equilibrium, the demand of retail investors is fully satisfied, for if it were not they would bid up the price ever so slightly and buy assets from the speculators. The

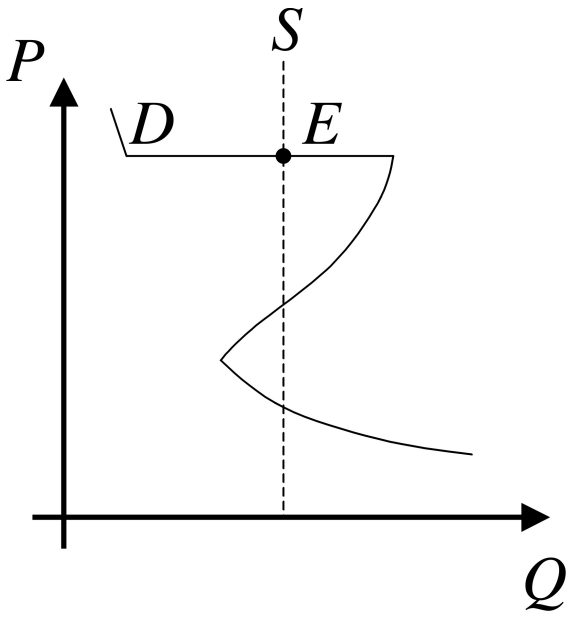

FIGURE 3. Supply-Demand Diagram After the Bubble Becomes Explosive

speculators, however, do not have their demand fully satisfied as they are willing and able to absorb more assets at the prevailing price. They are thus less levered than they would ideally like to be: if the maximum permitted leverage is λ^*, and the speculators' actual leverage at time t is λ_t, then $\lambda_t < \lambda^*$. This occurs because there is a shortage of assets, but price increases do not help because speculators' demand is discontinuous at the equilibrium price.

As time passes, this equilibrium evolves. The speculators earn a return on the assets they hold, pay margin interest on their borrowings, and absorb any new assets that are created. Retail investors do not absorb new issuance because their demand is always equal to $(1-\mu)ND_T(v_\pi)$. In other words, the supply curve moves to the right with new issuance, and the upward-sloping part of the demand curve moves to the right as speculators' net worth increases.

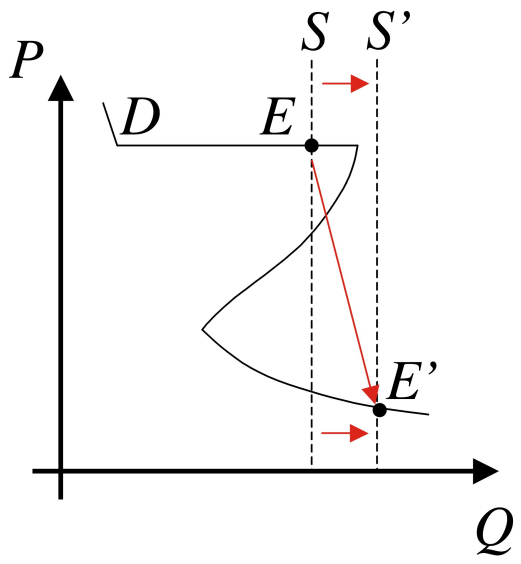

FIGURE 4. A Bubble Bursts After the Explosive Phase

Ignoring dividends, the speculators' net worth evolves as follows:

$$\Delta NW_t = \text{Asset Appreciation} - \text{Margin Interest} \tag{108}$$

$$= (p_t - p_{t-1})m_{t-1} - r(\lambda_{t-1}p_{t-1}m_{t-1}) \tag{109}$$

Noting that $p_t = v_\pi(1+\rho)^{t-T}$ from above, we can substitute to obtain:

$$\Delta NW_t = m_{t-1}v_\pi(1+\rho)^{t-T-1}(\rho - r\lambda_{t-1}) \tag{110}$$

Meanwhile, we also have the definition of leverage:

$$NW_t = (1-\lambda_t)p_t m_t \tag{111}$$

Finally, we have the condition that all newly issued assets are absorbed by speculators:

$$m_t = m_{t-1} + \Delta M \tag{112}$$

We can combine these three equations and take the limit as the time step is small. Doing so and simplifying yields an expression for how the leverage ratio λ evolves with time:

$$\frac{d\lambda}{dt} = -\lambda(\rho - r) + (1 - \lambda)\frac{\dot{m}}{m} \tag{113}$$

Thus faster new issuance puts upward pressure on leverage as speculators absorb new issuance, while a higher risk premium puts downward pressure on leverage as speculators reap the benefits of waiting for convergence of the price of a risky asset.

There are thus two possibilities: either

(1) leverage will steadily increase until it reaches its maximum, at which point the bubble will burst[7], or

(2) time T will arrive before the bubble bursts, forcing a reappraisal of estimates of value

[7]The bubble is still unstable at this point. The proof is as follows:
Let the moment the bubble initially becomes unstable be designated time t_1, and the moment when leverage reaches its maximum after the explosion in price be designated t_B.
Then we immediately have three results:
First, since $\dot{M} > 0$ by assumption, the total number of assets outstanding can only have increased:

$$M_B > M_1 \tag{114}$$

Second, since $\dot{\mu} > 0$ by equation (102):

$$\mu_B > \mu_1 \tag{115}$$

Third, by assumption there was a price spike:

$$p_B = v_\pi(1+\rho)^{t_B - T} > v_\pi(1+\rho)^{t_1 - T} > p_1 \tag{116}$$

which, because demand is downward sloping implies

$$D_B(p_B) < D_1(p_1) \tag{117}$$

Combining these three equations immediately gives:

$$\frac{M_B}{(1-\mu_B)ND_B(p_B)} > \frac{M_1}{(1-\mu_1)ND_1(p_1)} \tag{118}$$

$$\Leftrightarrow \frac{M_B}{(1-\mu_B)ND_B(p_B)} - 1 > \frac{M_1}{(1-\mu_1)ND_1(p_1)} - 1 \tag{119}$$

$$\Leftrightarrow \frac{M_B - (1-\mu_B)ND_B(p_B)}{(1-\mu_B)ND_B(p_B)} > \frac{M_1 - (1-\mu_1)ND_1(p_1)}{(1-\mu_1)ND_1(p_1)} \tag{120}$$

$$\Leftrightarrow \frac{m_B^{\text{lv}}}{(1-\mu_B)ND_B(p_B)} > \frac{m_1^{\text{lv}}}{(1-\mu_1)ND_1(p_1)} \tag{121}$$

$$\Leftrightarrow \frac{p_B m_B^{\text{lv}}}{A_B} > \frac{p_1 m_1^{\text{lv}}}{A_1} \tag{122}$$

In other words, even in the event of a price spike, new issuance can still eventually cause the bubble to suddenly burst. In this case, rather than a bubble that immediately bursts after a so-called "blow-off top," the model produces a period of rapid price appreciation followed by very low volatility as speculators can absorb new issuance with virtually no change in price. This regime then suddenly comes to an end when new issuance satiates the speculators' demand. This is illustrated graphically in figure 4.

3. Effect of Interest Rate and Margin Policy on the Bubble

It is common in financial circles to hear market participants complain about loose monetary policy as enabling the growth of bubbles. Commentators in the economic, financial and political communities have battered the Greenspan/Bernanke Fed for having enabled or (in the more extreme view) caused both the tech bubble and housing bubble. However, despite this near consensus, the economic literature has produced few if any convincing models that explains how precisely this mechanism operates.[8] As a result, those who argue that central banks have a duty to "lean against" or "prick" asset price bubbles are left without a framework to determine precisely how this ought to be done. The lack of such a framework is one reason the mainstream view remains that monetary policymakers should not consider asset prices directly (Bernanke, 2002), although there has been some movement in this view (Bernanke, 2010). In this section, we shall explore the relationship between

---

Now, equation (56) can be rearranged to give the condition for instability which held at t_1:

$$\frac{p_1 m_1^{\text{lv}}}{A_1} > \frac{1-\lambda}{\lambda}\eta_D \tag{123}$$

Combining the previous two equations gives

$$\frac{p_B m_B^{\text{lv}}}{A_B} > \frac{1-\lambda}{\lambda}\eta_D \tag{124}$$

which, by equation (56), shows that the bubble is unstable provided that η_D is not too much greater at price p_B than at price p_1.

[8]Goodhart, Osorio and Tsomocos (2010) build a model with 10 asset markets that is meant specifically to explain the U.S. subprime crisis of 2007-2009, and this model does have a channel whereby excessively loose monetary policy causes expansion of credit that worsens a crisis if bad states of the world occur. However, the model does not explain how the monetary policy influences whether the bad state occurs or not.

monetary policy and asset price bubbles, and how central bankers might prick bubbles should they choose to do so.

The most extreme form of the argument that the Federal Reserve has been responsible for the recent asset price bubbles is often made by those in the financial community, where the Austrian view of business cycles still has some credence. In that view, when the central bank prints excess money, more money chases the same quantity of assets, causing asset price inflation and investment in uneconomic projects. When it becomes clear that these projects are unprofitable, or when interest rates rise again, the bubble bursts. This view lost currency in academia around the time of the rational expectations revolution, when it was argued that it required businessmen to be irrationally optimistic about the projects they invested in, and to fail to learn that they would lose money when the central bank ultimately raised rates again.

The Austrian view has some support in empirical research that shows the importance of bank credit in diagnosing asset price bubbles and predicting financial crises, results that also support the explanation of bubbles and instability proposed in this paper. Borio and Lowe (2002), for example, find that a high credit-to-GDP ratio predicts 80% of financial collapses with a 4:1 signal-to-noise ratio, a remarkable performance for a single variable. Iso and Iwaisako (1995) also estimate that bank credit was a contributing factor to the Japanese stock and real estate bubble.

The Austrian view, however, neglects the source of money supply growth. In spite of monetarist efforts, central banks typically target interest rates, not the money supply. During a period of increasing demand for speculative credit, the money supply will have to grow in order to keep interest rates constant. These "Wall Street Neo-Austrians" therefore have their causality reversed. The central bank does not create the bubble by pushing new money into the financial system. Rather, speculators pull new money into the system by demanding credit; the central bank's failure to feed this demand would result in a spike in interest rates and recession.

In the more mainstream view, stated by Fed Governor Mishkin (2001), overly accommodative monetary policy can cause asset price bubbles in three ways, all linked to fundamentals. First, it decreases the discount rate for future cash flows from an asset, so that other things equal an investor will place an increased value on those cash flows and hence the asset. Second, it will temporarily stimulate the economy, boosting sales and profits and (it is argued) asset prices.[9] Third, the inflation attendant to unexpectedly low interest rates debases the value of debt on corporate balance sheets and raises the relative value of equities and other real assets financed by debt.[10] These effects also appear to be what Taylor (2008) has in mind in his criticism of the Fed's policies.

Unfortunately, neither the neo-Austrian view nor the mainstream view can explain some critical facts about bubbles. First, neither clearly produces the trademark accelerating path of asset prices and explosive volumes followed by the abrupt crash. Second, and more importantly, in both theories bubbles are economy-wide phenomena, and monetary policy should affect all assets more-or-less equally. The models cannot explain why loose monetary policy should have created a bubble in tech stocks in the 1990s but a bubble in real estate in the 2000s. While bubbles may be accompanied by generally excessive asset values, they are primarily characterized by a misallocation of capital among different assets. The problem is real, not monetary: a group of speculators control a large quantity of resources and concentrate them in one asset in the belief that the asset will pay large returns. If this belief is mistaken, a great deal of wealth will be destroyed because expensive investments will be made that ultimately prove to have little value.

This insight does not, of course, absolve the central bank of some blame for the bubble. While the "driving force" for the bubble is the enthusiasm of a small group of levered speculators, policy may still have a role if it is able to prevent the expression

[9]Note, however, that if investors are fully rational, overly accommodative monetary policy should hurt the economy and profits in the long term and hence this part of the effect should actually be negative.

[10]However, asset bubbles and imbalances can still build in low-inflation environments, as Borio and Lowe point out.

of this minority view from dominating the market. Policy might be effective in doing so either by inhibiting the formation of the bubble, or by "pricking" it once its growth is already underway. Often, the central bank's policy rate is the tool that is argued for, but as we shall see, the adjustment of margin requirements also may be effective. As has been noted by Galbraith (1954) and others, the 1929 stock bubble inflated despite rates of margin interest as high as the mid-teens, so that high interest rates may not actually be effective in pricking or restraining bubbles.

Central bankers have argued that prudential regulation is a better tool than interest rates. Intuitively, we would expect that a central bank or other policymaker could use an increase in margin requirements or interest rates to prick an asset bubble. Our model confirms this, as we shall see.

We shall use the model of section 1 with one small change. Because we are now allowing a change in the interest rate, we assume that the retail investors demand function $D(p)$ will now also depend negatively on the interest rate r. The obvious intuition is that interest rate hikes reduce the net present value of the future cash flows from the asset and thus its fundamental value.

We will consider an instantaneous change in the interest rate or leverage ratio, so that Δt and $\Delta\mu$ in equations (93) and (95) are both zero. We will also need to consider that the leverage ratio λ may vary with time. Equations (93 ff.) therefore become:

$$m_t^{\mathrm{lv}} = \frac{m_{t-1}^{\mathrm{lv}}}{1-\lambda_t} \cdot \left[1 - \frac{\lambda_{t-1} p_{t-1}}{p_t}\right] \tag{125}$$

$$m^{\mathrm{gp}} = (1-\mu) N \cdot D(p, r) \tag{126}$$

$$m_t = m_t^{\mathrm{lv}} + m_t^{\mathrm{gp}} \tag{127}$$

It is clear that an increase in either the margin requirement $(1-\lambda_t)$ or the interest rate r will reduce the demand for the bubble asset at any given price. This reduced demand means the price must fall.

The monetary authority's success in orchestrating an initial price decline does not guarantee that the price decline will persist. We can see this by examining the conditions for renewed growth of the bubble after the initial decline. We shall consider the bubble as growing as long as the price of the bubble asset is increasing.[11] For ease of exposition we will examine the continuous time version of the model presented in footnote 3.

Differentiating equation (98) gives

$$\dot{m}(p) = \dot{m}^{\text{gp}} + \dot{m}^{\text{lv}} \tag{128}$$

Substitution from equations (96) and (97) into the above yields

$$\dot{m}(p) = \frac{m^{\text{lv}}}{1-\lambda} \cdot \left[\frac{d + \lambda\dot{p}}{p} - \lambda r\right] + \frac{\dot{\mu} N y}{(1-\lambda)p} - \dot{\mu} N \cdot D(p) + (1-\mu) N \cdot D'(p)\dot{p} \tag{129}$$

which can be solved for $\dot{p}$ to give:

$$\dot{p} = \frac{\frac{m^{\text{lv}}}{1-\lambda}[\lambda r - d/p] + \dot{\mu} N [D(p) - \frac{y}{(1-\lambda)p}] + \dot{m}(p)}{(1-\mu) N \cdot D'(p) + \frac{m^{\text{lv}}}{1-\lambda} \cdot \frac{\lambda}{p}} \tag{130}$$

$$= \frac{\frac{m^{\text{lv}}}{1-\lambda}[\lambda r - d/p] + \dot{\mu} N [D(p) - \frac{y}{(1-\lambda)p}] + \dot{m}(p)}{(MinEAMASS - EAMS)/p^2} \tag{131}$$

This looks like quite a mess, but is really not so bad. The denominator of the fraction is just the slope of the instantaneous demand curve. It is negative because EAMS is greater than MinEAMASS in the stable phase of the bubble. Thus $\dot{p}$ will be positive only if the numerator is negative. In order to prevent the growth of a bubble, therefore, we require that

$$\dot{\mu} N \cdot \left[\frac{y}{(1-\lambda)p} - D(p, r)\right] - \frac{m^{\text{lv}}}{1-\lambda}\left[\lambda r - \frac{d}{p}\right] \leq \dot{m}(p) \tag{132}$$

[11] An alternative and in many ways more sensible definition is that the bubble is growing if the speculators are accumulating a growing share of the assets. While this may be the case even if the price is falling (because of new entrants), speculators probably will not have much luck recruiting new entrants to invest in a depreciating asset. By contrast, the downward-sloping demand curve of the retail investors means that a rising price guarantees that speculators are accumulating the bubble asset. We will therefore consider the bubble as growing if and only if the price of the bubble asset is increasing.

Equation (132) has a simple interpretation. The first term is the increase in demand for assets as a result of some retail investors becoming speculators. It is positive because $D(p)$, the quantity of the bubble asset demanded by the average retail investor, is always less than the maximum he could possibly demand y/p. The second term is the decrease in demand as a result of the speculators having to liquidate some assets to pay margin interest. It is negative (including the preceding minus sign) unless the asset is providing enough cash flow to service the debt of the speculators, which is not how we usually imagine a bubble. Recall that $\dot{m}(p)$ is exogenous and meant to represent the production of new assets in response to their high price. This equation therefore says that the price of the bubble asset will increase if demand is increasing faster than supply at the current price.

Let us consider the effect of an increase in the margin interest rate r on condition (132). We shall show that the left hand side of the inequality is decreasing in r, so that an interest rate increase make the condition for a bubble to grow more stringent for any given price of the asset.

Consider the first term on the left hand side. We shall presently show that it is invariant to a change in r (assuming no change in $\dot{\mu}$). To see this, we note that because the retail investors are the depositors and owners of the banks, their claim on the banks must be equal to the banks' total assets, that is $\lambda p m^{\text{lv}}$. Their own holdings of assets are just $D(p)$ per investor. Thus, the net worth y of each individual retail investor is:

$$y = [(1-\mu)N]^{-1}\lambda p m^{\text{lv}} + D(p,r)p \tag{133}$$

Substituting for y in the first term of the left-hand side of equation (132) gives:

(134)
$$\dot{\mu}N \cdot \left[\frac{y}{(1-\lambda)p} - D(p,r)\right] = \dot{\mu}N \cdot \left[\frac{D(p,r)p + \lambda p m^{\text{lv}}[(1-\mu)N]^{-1}}{(1-\lambda)p} - D(p,r)\right]$$

(135)
$$= \dot{\mu}N \cdot \left[\frac{\lambda(1-\mu)ND(p,r) + \lambda m^{\text{lv}}}{(1-\mu)N \cdot (1-\lambda)}\right]$$

(136)
$$= \frac{\dot{\mu}\lambda}{(1-\lambda)(1-\mu)} \cdot [m^{\text{gp}} + m^{\text{lv}}]$$

(137)
$$= \text{Constant with respect to } r$$

The total change of the left hand side when the interest rate changes is thus the change in the second term:

(138)
$$\Delta LHS = -\Delta\left[\frac{m^{\text{lv}}}{1-\lambda}\left(\lambda r - \frac{d}{p}\right)\right]$$

(139)
$$= -\frac{\Delta NW^{\text{lv}}}{p(1-\lambda)^2}\left[\lambda r - \frac{d}{p}\right] - \frac{NW^{\text{lv}}}{p(1-\lambda)^2}\left[\lambda \Delta r\right]$$

Because an increase in the interest rate reduces retail demand and the total number of assets is constant, it must necessarily increase speculative demand and therefore speculators' net worth at any given equilibrium price. Thus ΔLHS is negative, and an increase in the interest rate must make the no-price-increase condition (132) easier to satisfy for any given price. An increase in the interest rate, if it is large enough, may therefore halt the growth of a bubble. Intuitively, it does so first by concentrating assets in the hands of speculators, and second by increasing the cost to the speculators of holding the assets, restraining the growth in their net worth.

The effect of a reduction in leverage is more complex, because leverage appears in many places in equation (132). An increase in margin requirements has two effects for a given price of the bubble asset. First, it reduces the impact of the entry of new speculators on the price, because these new speculators have less buying power. This is reflected in the first term in equation (132) and reduces price appreciation. Second, because the speculators are less levered, less of the additional debt capacity they obtain from appreciation of the asset will need to be used to pay margin interest,

and more can be used to buy additional assets. This is reflected in the second term in equation (132) and increases price appreciation. The larger the share of assets held by speculators, the larger is the second effect compared to the first.

Higher margin requirements are therefore a double-edged sword. While they reduce price appreciation in the early phases of a bubble and may prevent it from getting underway, once the speculators accumulate a large enough share of the assets, the lower leverage actually makes the price appreciation more robust. This is consistent the earlier finding of equation (56) that lower leverage can actually support a larger bubble before instability sets in.

Figure 5 shows simulated bubble with a series of negative shocks to the permitted leverage ratio of speculators. Each time there is a shock, speculators are forced to liquidate some assets and the price declines, sometimes sharply. In the early stages of the bubble, the leverage shocks restrain price appreciation, but as long as the speculators are still successful in evangelizing, the bubble still eventually reaches its point of instability. However it requires a much larger percentage of the investing population to be speculative—in this case an implausibly high 77%, as compared to 31% without the leverage shocks. The time to reach the point of instability has approximately doubled. In short, while these leverage shocks may delay the growth and crash of the bubble, they will not prevent it unless they somehow interrupt the flow of new speculators into the bubble or if there is an upper limit on the number of investors that may become speculators. This is entirely plausible, but outside the scope of this model.

Both interest rates and margin requirements have benefits and drawbacks as tools to manage bubbles.

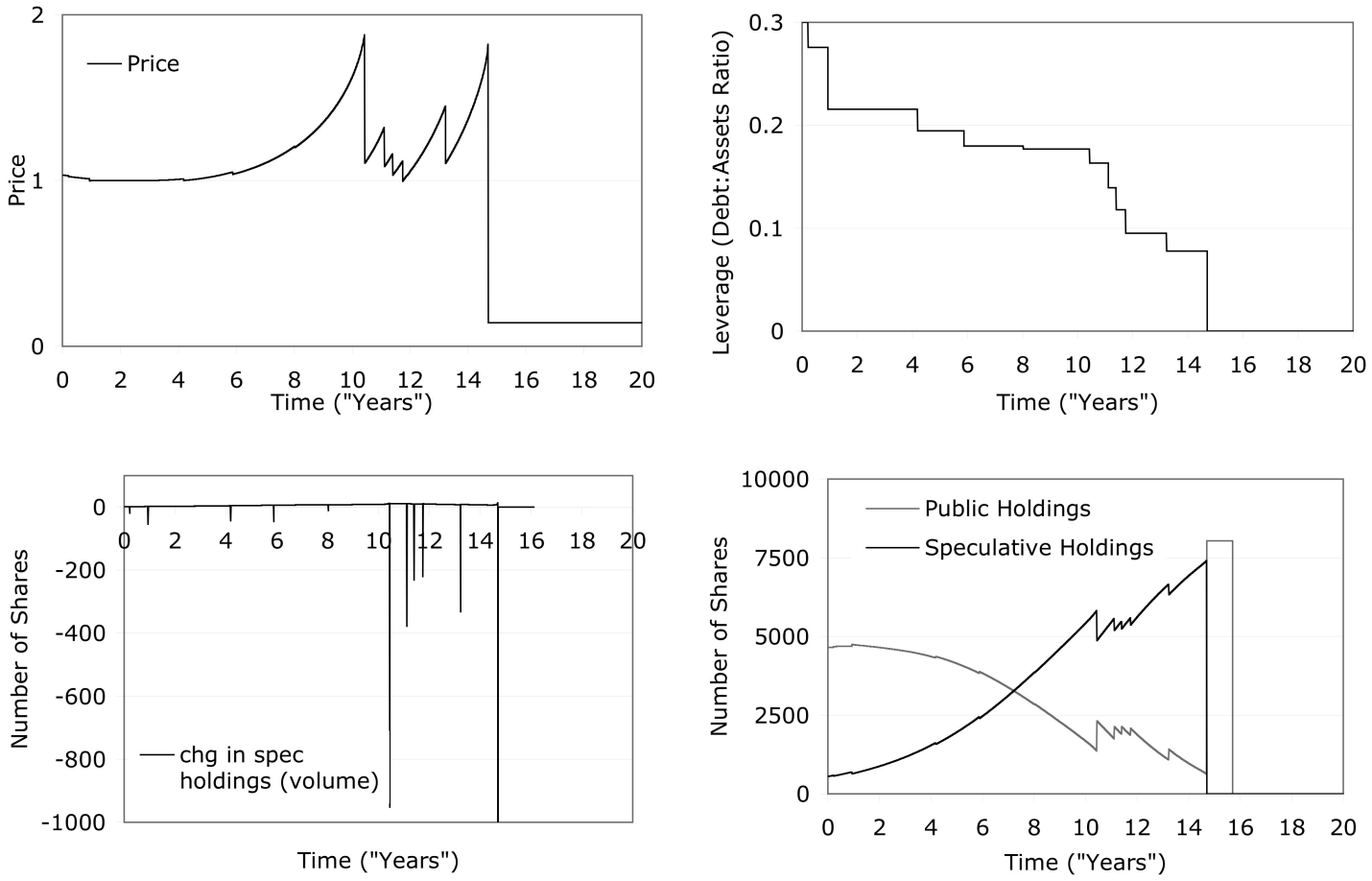

FIGURE 5. A Simulated Bubble with Periodic Negative Leverage Shocks

A small interest rate increase will unequivocally slow the bubble's growth, but may not be sufficient to permanently prick the bubble. While high enough interest rates will always be successful in preventing the leveraged expression of views and thus the growth of asset bubbles, the cure may be worse than the disease in that the required increase in interest rates may be very large, plunging the entire economy into recession and reducing asset prices across the board.

By contrast, a small increase in margin requirements may actually be counter-productive—while the bubble will grow more slowly it will ultimately grow larger than it would have with lower margin requirements, and the resulting crash will be more painful. In spite of this danger, margin requirements can always be raised again if they are not proving effective in restraining price increases, and policymakers can make this assessment in real time. Unlike interest rate increases, however, margin requirements can be narrowly targeted at the asset in question rather than the entire economy.

In this section, we have seen that both margin requirements and interest rates can be used by policymakers to prick bubbles. Whichever they choose, their action will not be effective unless it is drastic enough to alter the dynamics of the bubble. In particular, one of the following must hold:

- The initial decline in asset prices must be large enough to completely bankrupt the speculators;
- A rate hike must continuously drain enough net worth from the speculators to offset new capital from entering speculators;
- The initial decline in asset prices must be large enough to cause the speculators to change their expectations and behavior (which is outside the scope of this model); or
- The decline in leverage or rise in interest rates must be sufficient to suppress the rate of demand growth below the rate of supply growth.

If one of these conditions does not hold, then the bubble will suffer only a temporary price decline and then resume growing.

4. Conclusion

This chapter has applied our earlier model to explore the formation of bubbles. In contrast to most of the literature, changing or self-fulfilling expectations take a back seat. Instead, the driving force is credit constraints on speculators, so that overly optimistic views cannot be immediately expressed in the market but gain expression slowly as optimistic speculators accumulate an increasing share of society's wealth. Whether a bubble develops depends on the speed with which new speculators are recruited or enter the bubble relative to the production of new assets, the leverage ratio and the interest rate in the ways that would be expected. The bubble and its growth are robust to small changes in expectations.

In this model, unlike a representative agent model, the bubble is driven by the optimistic few. Even these optimistic few need not be infinitely optimistic. In the simulations I ran, the bubble rarely grew by more than a factor of 3 or 4. These kinds of factors are also typical of real-world bubbles. Thus, the model does not require that speculators be willing to overpay by a factor of 50 or 100 for an asset; erring by a factor of 5 is quite sufficient, and this is both perfectly plausible and consistent with historical experience for a small, optimistic group claiming a new investment paradigm.

The bubble reaches a critical point when speculators accumulate too large a share of the bubble asset. Then the minimum elasticity-adjusted market size for stability grows too large relative to the actual market size and the previous equilibrium suddenly becomes unstable. When speculators do reach this critical mass, any small negative shock will cause a crash, wiping out the speculators. This shock could be anything, including an increase in the supply of the asset, credit or monetary tightening, and a demand shock on the part of retail investors. The collapse in asset prices occurs in spite of the optimistic expectations of the speculators, not because the speculators suddenly realize their earlier expectations are mistaken. Speculators liquidate only because they are forced to do so. The central bank may also prick the

bubble earlier, either by use of margin requirements or interest rates, but the action must be sufficiently drastic.

Despite the extremely simplistic behavior of investors, the model presented produces many of the stylized facts characteristic of bubbles and crashes that earlier models have had difficulty explaining, including most critically the critical role played by credit both as the fuel for the bubble. The association of volatility and volumes with prices results from the large and frequent portfolio adjustments speculators must make at the top of the bubble. When the bubble pops, the pop is irreversible, not because previously irrational investors suddenly become intelligent, but because the optimists no longer have the capital to express their views in the marketplace.

Finally, the model treats market liquidity crises—a sudden evaporation of bids and offers—and bubbles in a unified framework, showing how each is a consequence of the same investor behavior, where winners double down and losers exit the market. Ironically, economists have traditionally argued that this kind of evolutionary algorithm produces rational behavior. Here it does the opposite, driving bubbles and instability.

CHAPTER 3

The Collapse of Long-Term Capital Management

In this chapter, we shall apply the model to a complex case with sophisticated strategies and financial instruments. We will explore the 1998 collapse of hedge fund Long-Term Capital Management (LTCM), an episode that ranks among the two or three most severe financial crises that most active Wall Street professionals have experienced in their careers. The potentially destabilizing consequences of the crisis eventually prompted the Federal Reserve to step in and coordinate a private-sector bailout.

1. Background on LTCM and Its Markets

[1]Long-Term Capital Management was a large relative value hedge fund that was in business from 1994 until 1999. The fund was founded in 1994 by John Meriwether, a star bond trader and later vice chairman at Salomon Brothers. LTCM employed a strategy of *relative value arbitrage*, in which it bought some assets it considered to be relatively cheap while selling short other, very similar assets it considered to be relatively expensive. Relative value hedge funds are, by necessity, very highly levered institutions, and LTCM was perhaps the archetypal such fund. LTCM was also very large: at its peak in April 1998 it had $4.87 billion in capital, $125 billion in assets, and another $115-125 billion of notional value in off-balance sheet derivatives.[2]

[1]The historical facts in this narrative are taken largely from Lowenstein (2000) and MacKenzie (2003), as well as conversations with former LTCM principals who wished to remain anonymous.
[2]Other sources generally report a figure of $1.25 trillion in off-balance sheet derivatives. However, this figure fails to take into account that many of these positions were not just negatively correlated but in fact perfectly offsetting, that is, long and short exactly the same instrument. This is because LTCM, like other market participants, typically engaged in an offsetting trade when it wanted to close out a swap contract, rather than ending or selling the original contract. The so-called "replacement value" of these swaps was $80-90 billion (Interview with LTCM Principal, 2010), and equity derivatives accounted for an additional $35 billion (Dunbar, 2000).

An example of a typical trade would be for LTCM to buy a 29 1/2-year Treasury bond and sell short a 30-year Treasury bond. LTCM made money because the 30-year bond was more liquid, so it traded with a slightly lower yield (i.e., higher price). After six months, the Treasury would issue a new 30-year bond, and the 30-year bond LTCM had sold short would become a 29 1/2-year bond while the 29 1/2-year bond it owned would become a 29-year bond. Because the 29 1/2-year bond has similar liquidity characteristics to the 29-year bond, the yields would converge, and LTCM could liquidate the trade at a profit. More examples of LTCM's trades can be found in Perrold (1999).

Trades such as these require very high leverage. The typical yield differential between the 30-year bond and 29 1/2-year bond might be 6 basis points (0.06%). This would allow LTCM to realize a profit from the trade of about 1% over six months if the yields fully converged, rather paltry for a hedge fund. However, if LTCM levered this trade 50 times, it could make 50% in six months. This is a much more exciting prospect.

Of course, there is no free lunch. Other things equal, higher leverage is associated with higher risk. LTCM justified its high leverage in trades such as this one by arguing that the price of a 30-year Treasury is very highly correlated with the price of a 29 1/2-year Treasury, as indeed it is. As long as markets are guided primarily by fundamentals, such a position has very little risk. However, when markets become constrained, so that insufficient capital is available to arbitrage price differences between similar assets, the force that ensures that high correlation is no longer operative, and prices can diverge substantially. This is what happened to LTCM. Its argument and belief that its offsetting trades carried very little risk was only true when markets had sufficient liquidity. When that liquidity dried up, the assets that LTCM had assumed would be very highly correlated ceased to be so, and as a result the firm's risk skyrocketed.

For the first several years of its existence, Long-Term's results were spectacular. However, the fund grew as it succeeded, and as it grew it attracted imitators both in

the hedge fund community and among the trading desks of the Wall Street banks. Returns and opportunities began to dwindle.

Date	Beginning Assets under Management (Net Capital)	Annualized Return	End of Period Leverage (Excluding Derivatives)
3/94-2/95	$1.1 Billion	25%	16.7
3/95-2/96	$1.8	50%	27.9
3/96-2/97	$4.1	34%	22.8
3/97-2/98	$5.8	11.5%	26.8
3/98-7/98	$4.7	-35%	31.0

FIGURE 1. LTCM Performance and Leverage Ratio, Excluding Derivatives. Source: Perrold, 1999

LTCM's troubles began in late spring of 1998 and continued into the summer 1998, especially when Salomon Brothers, one of LTCM's imitators, began to close down its arbitrage desk, both to reduce risk and in response to poor results. Salomon's liquidation of a similar portfolio put pressure on LTCM's trades, for as we have suggested, the willingness and ability of other market participants to engage in LTCM-like arbitrage strategies was not perfectly elastic.

The situation turned dire on August 17, 1998, when Russia announced a default on its ruble-denominated debt and forbade banks from honoring certain foreign exchange contracts, while continuing to honor its foreign-currency debt, a move that few, if any, had anticipated. The default caused the bankruptcy of one of Long-Term's competitors, an $850 million fund known as High Risk Opportunities (HRO). Many of Long-Term's trades began to move against it as HRO and other competitors liquidated similar positions, reducing the demand for assets LTCM was long and increasing demand for assets it was short.

As August and September wore on, LTCM began to lose money in dramatic fashion. On two occasions, it lost $550 million in a single day, about 12.5% of its pre-crisis capital. However, LTCM's principals found themselves unable to liquidate to reduce risk at anything close to what they viewed as a reasonable price. Other

market participants moved to liquidate ahead of LTCM, pushing prices against it and causing even deeper distress. Partner Eric Rosenfeld compared LTCM to a large ship in a small harbor in a storm—it was too large to maneuver, and all the other boats were just trying to get out of its way (Rosenfeld, 2009).

By the end of September, LTCM had barely been able to reduce its risk at all, but it was nearly out of capital. The Federal Reserve, fearful that a default would result in a sudden liquidation of a portfolio that included $125 billion in assets and $1.25 trillion in derivatives, and that this would destabilize markets, stepped in to orchestrate a bailout by LTCM's counterparties.

The Fed justified its action to organize a bailout on the grounds that a default by LTCM would have put the entire financial system at risk of serious disruption. As then-Chairman Alan Greenspan put it, "our sense was that the consequences of a fire sale... should LTCM fail on some of its obligations, risked a severe drying up of market liquidity." (Greenspan, 1998) In the somewhat starker language of New York Fed President McDonough, "there was a likelihood that a number of credit and interest rate markets would experience extreme price moves and possibly cease to function for a period of one or more days and maybe longer." (McDonough, 1998)

We will never know for certain whether the Federal Reserve's fears of a severe financial disruption would have been realized had it not acted. Indeed, even at the time, the Fed's own public and private statements show that it was unsure its controversial action was necessary. We can, however, examine how the MinEAMASS framework could have been used to assess whether financial markets would have suffered from an episode of instability had LTCM been forced to liquidate. We shall find that the instability might have occurred in at least a few of the markets in which Long-Term played. One of these markets was the bank funding market, which is of particular systemic importance because a dysfunctional bank funding market may cause contagious bank failures.

2. Relation with Previous Literature

There is a modest literature on the near-collapse of LTCM. Two full-length books (Dunbar, 2000, and Lowenstein, 2000) offer accounts in a journalistic style. These, as well as a number of contemporary press articles, tend to focus on the personalities involved, rather than providing an analytically rigorous narrative. Despite their analytical shortcomings, Dunbar and especially Lowenstein still provide the most authoritative and complete account of the actual day-by-day events of LTCM's demise. They include some important data, as well as some information on what drove the decision-making of the major players.

Lowenstein and Dunbar conclude that LTCM collapsed because the principals of the firm were greedy, took recklessly excessive risk, and blindly relied on models, and that these failings exposed LTCM to extreme liquidity events. Their work, as well as reports from the Bank for International Settlements (BIS) and the President's Working Group on Financial Markets, argues that LTCM was not equipped to handle a so-called "flight-to-quality" trade, where market participants express increased preferences for safe and liquid instruments. They argue that this flight to quality is what occurred in August and September of 1998 following Russia's default on its domestic debt.

Two academic studies examine the failure of LTCM through a business framework (Perrold, 1999) and a sociological framework (MacKenzie, 2003). Perrold's Harvard Business School case study presents the facts as they were seen by LTCM's management, and does not offer an explicit conclusion on the cause of the collapse. MacKenzie's qualitative account argues that LTCM's trading strategies were widely mimicked. Rather than a flight to quality, he concludes, there was a coordinated liquidation of similar trades held by a number of institutions, not just LTCM. MacKenzie terms the set of these trades a "superportfolio" and argues that the evidence supports this explanation over the flight-to-quality explanation.

MacKenzie's explanation is echoed in a presentation given by LTCM partner Eric Rosenfeld to an MIT business school class in 2009 (Rosenfeld, 2009). Rosenfeld

argues that LTCM was undone in part by what he terms *endogenous risk*. That is, by its very existence, LTCM altered the risk profile of its trades. Because LTCM could be forced to liquidate when its trades soured, the added selling pressure would push trades even further against the fund. In the language of this paper, Rosenfeld argued that too many levered investors held the same positions: MinEAMASS was too large for the markets.

My own results are the most in line with the analyses of MacKenzie and Rosenfeld. The argument of Lowenstein, Dunbar and the government reports discussed above, that LTCM was felled by a flight to quality for which it was unprepared, is in large part rebutted by MacKenzie and Rosenfeld. They show that the flight to quality and liquidity was actually a flight from LTCM's trades specifically. The flight to liquidity was concentrated in those liquid instruments in which LTCM had a position, while it was much less severe in markets where LTCM was not active, and less severe still in markets where LTCM owned the more liquid instrument. In addition to serving as proof of concept for the MinEAMASS framework, this analysis deepens and refines the analyses of MacKenzie and Rosenfeld by demonstrating the potential instability of the specific markets in which LTCM suffered its most severe losses and quantifies Rosenfeld's concept of "endogenous risk" as it applies to LTCM.

3. Stability of LTCM's Markets

I now apply the model of instability that I have developed in order to examine the stability of the global equity, US Treasury, and bank funding markets as well as the market for equity volatility in the late summer of 1998.

Hard portfolio data on LTCM and its competitors is very difficult to come by because there were no public reporting requirements and the funds were very secretive while they were trading. Because of the media scrutiny to which LTCM was subject after the crash, some of the partners in the f und were more forthcoming than they had been previously, and some information is available on its portfolio.

For LTCM's competitors and copycats, however, the situation is worse. They fall into two categories: large investment banks and hedge funds.

Most of the investment banks that were competing with LTCM remain going concerns today and try hard not to share the positions in their proprietary trading portfolios. They were required to publicly report losses at the time of the collapse of LTCM, but these were generally only on an aggregate basis. Their arbitrage portfolios were only a part of the proprietary trading businesses.

LTCM's hedge fund competitors were under no obligation to disclose their holdings and took care not to do so. In some cases, the total capital of these funds is available, but that is generally as far as it goes. However, these funds tended to be both significantly smaller and less levered than LTCM (Anonymous, 1998), meaning that for our theory they contribute far less to MinEAMASS. (Recall that contributions to MinEAMASS are proportional to fund capital and to the square of leverage.)

The crucial data, therefore, relate to LTCM's portfolio, as well as the size of the markets in which it played. Information on LTCM's portfolio has been compiled from a number of media and academic sources, as well as a discussion with former LTCM principals. Information on the size of markets has been compiled from public sources such as the flow of funds accounts from the Federal Reserve.

We will examine LTCM's impact on the bank funding markets and the equity markets. These were chosen because they are two of the most economically significant and transparent markets in which it played, and accounted for a significant portion of the fund's risk.

In general, I shall assume that the reader is familiar with the basic architecture of the financial markets and the main instruments that are traded—the sort of knowledge that would be gained from an introductory undergraduate finance course or an educated layperson's guide to the financial markets, such as *The Wall Street Journal Guide to Understanding Money and Investing*. These include simple options such as *puts* and *calls* and the practice of *selling short* to bet on stock price declines.

Virtually all of the investment instruments in which LTCM took positions are commonplace in the financial industry and are understood by most people at hedge funds, investment banks, and institutional investors. However, many of these instruments are not well understood by individual investors or outside the financial industry. I shall discuss these instruments as they occur in the narrative for readers without financial market experience.

One of the fund's most important trades was in equity volatility. While it might not seem that one could bet on the volatility of the stock market, the market's forecast of the volatility of an asset or portfolio of assets can in fact be calculated from options prices. This is because the expected volatility of the asset price is a key input into the widely-used Black-Scholes option pricing formula (see Hull, 2005 for details). Thus, if an investor knows the price of an option and the price of the underlying asset, a no-arbitrage condition allows him to calculate the market-implied volatility of the asset's price. In financial markets, the standard unit of volatility is the one-standard-deviation percentage movement in the price of an asset over a year, which is typically shortened simply to *percent*.[3] Thus, if an option "trades with a 20 percent vol", this means that the implied one-standard-deviation annual movement in the price of the underlying stock would be 20 percent. Alternatively, participants will say that 20 is the *price of volatility*.

Retrospectively, the price of an option will be proven to have been either too high or too low, depending upon the actually realized volatility of the price of the underlying asset. If the *ex ante* implied volatility from the option price was greater than the subsequently realized volatility, then the price of the option was too high, and an investor who bought the option and engaged in the arbitrage underlying the Black-Scholes formula (known as *dynamic hedging* or equivalently *delta hedging*) would have lost money. The opposite would be true if the implied volatility was too low.

[3]Technically, this is usually calculated as the daily standard deviation times $\sqrt{252}$, because there are usually 252 trading days in a year. This incorrectly assumes that the daily price moves are independent and normally distributed, but it is the standard practice for quoting volatility.

Long-Term Capital engaged in volatility trades because it noticed that the implied volatility on five-year stock index options was out of line with historically realized volatility. In particular, the implied volatility on five-year stock options was generally about 20 percent,[4] while the actual realized recent volatility had been 10-13 percent (Perrold, 1999). In fact, in the previous 50 years, five-year realized volatility on the S&P 500 had approached 20 percent only once, in the five years from 1987-1992 that included the 1987 stock market crash.[5] LTCM thought it understood the reason for this mispricing: large insurance companies and pension plans that had promised their participants a guaranteed rate of return needed to buy puts to hedge those promises, and there was no so-called "natural seller."

So LTCM stepped into the breach, selling *straddles*, which are compound options of one put and one call with the same strike price and the same maturity. LTCM would then dynamically hedge the options as the price moved. Each time the price moved, LTCM would lose a little bit of money on its dynamic hedge, so the more volatile the price of the stock, the more money the fund would lose. LTCM was paid for this loss up front in the form of the premium on the options it had sold.

This strategy is known as "selling volatility," because LTCM would show a mark-to-market loss if implied volatility rose. As long as realized volatility over the life of the option was less than implied volatility at the time LTCM entered the trade, it would make money eventually provided it stayed in the trade. However, if implied volatility (and therefore the market price of the options LTCM had sold short) rose too much, LTCM's balance sheet would show financial distress, and it might be forced to buy back the options at a much higher price than it had sold them, thus losing money. On the other hand, if the price of volatility fell, LTCM could exit the trade early, buying back the options at a lower price, and making even more profit than they initially anticipated.

[4]Lowenstein (2000, p. 126) cites a figure of 19 percent, while Perrold (1999) cites a figure of 20 percent

[5]As it turned out, S&P 500 volatility in the five years from late 1998-2003 averaged upwards of 22 percent.

LTCM's largest equity trades were sales of volatility (usually at-the-money straddles) on broad stock indexes in the US and Europe.[6] According to Dunbar (2000), by January 1998 LTCM's 5-year equity option position was about $100 million per percentage point of volatility.[7] This implies (using the Black-Scholes option pricing formula) that LTCM had written options with a notional value of about $11.5 billion.[8]

The consequences of this for stability in the equity market can be examined with respect to the stability condition (87), reprinted below for convenience.

$$p^2 \left\{ \sum_i NW_i \cdot (\Delta_i / NW_i)^2 + \sum_i \Gamma_i + \sum_{\delta i} \frac{\partial \Delta_i}{\partial \pi_i^\delta} \frac{\partial \pi_i^\delta}{\partial p} \right\} < \sum_i pm_i + \eta A \tag{141}$$

Recall that the left side of this equation is MinEAMASS, the right side is EAMS, the elasticity-adjusted market size.

---

[6] More background on these trades, and the reasons behind them, are described in Perrold (1999), Dunbar (2000), and Lowenstein (2000).

[7] This means, for example, that if volatility was priced at 20%, an increase to 22% would have cost LTCM $200 million. See Hull (2005) for an explanation of option pricing.

[8] In standard option pricing, an investor's exposure (how much he will gain or lose) to changes in volatility is denoted by the (made up) Greek letter *vega* (ν). Vega is the derivative of the option price with respect to the implied volatility, or annualized standard deviation of the price of the underlying instrument. Thus, LTCM's vega was $100 million per percentage point or $10 billion. In standard option-pricing notation (given expository treatment in Hull (2005), the vega of a put or call option is given by:

$$\nu = S_0 \sqrt{T}\, N'(d_1) \tag{140}$$

where

- S_0 is the price (or for a portfolio, total notional value) of the underlying asset
- T is the time to expiry
- $N'(d_1)$ is the probability density function for the standard normal distribution, $N'(d_1) = \frac{1}{2\pi} \exp(-d_1^2/2)$
- $d_1 = \frac{\ln(S_0/K) + (r - q + \sigma^2/2)T}{\sigma\sqrt{T}}$
- K is the strike price of the option
- r is the risk-free rate of interest
- q is the dividend yield on the asset
- σ is the implied volatility, in percentage points per year

LTCM traded at-the-money-forward options, which means that $\ln(S_0/K) = -(r-q)T$, simplifying the expression for d_1. LTCM's vega is thus given by:

$$\nu = S_0 \sqrt{T}\, N'(\sigma\sqrt{T}/2)$$

$$(\$10 \text{ billion}) = S_0 \cdot \sqrt{5} \cdot \frac{1}{2\pi} \exp[-(20\%)^2 \cdot 5/8]$$

Solving for S_0, the notional value of the outstanding options, thus gives $S_0 = \$11.5$ billion.

For stability of the equity market as a whole, the consequences of LTCM's distress were small. LTCM's hedges meant that it had no exposure to outright stock price movements—it had a Δ of zero. Its Γ, however (or $p^2\Gamma$) was about \$10 billion.[9] Plugging these figures into equation (141) and assuming no changes in LTCM's portfolio distribution gives a MinEAMASS in the equity markets of \$10 billion. The US and European equity markets at the time were larger than this by a factor of over a thousand. Clearly, a forced liquidation of LTCM was nowhere near enough to destabilize the equity markets. Indeed, while equity markets declined along with most risk assets during the summer of 1998, they never ceased to function in an orderly manner.

Nonetheless, LTCM's sales of volatility were one of the trades that hurt it the most, with losses of \$1.3 billion. To see why this was the case, we can note that volatility is traded on Wall Street as a separate asset class. With volatility priced around 20%, LTCM's \$100 million exposure per point of volatility translates to having sold short \$2 billion of volatility.[10]

We can use equation (141) to determine the stability of the market for volatility. The parameters are as follows:

- Δ = \$100 million per volatility point
- p = 20 points

[9]Recall that Γ is the derivative of Δ with respect to the price. The Γ of an option is given by Hull as:

$$\Gamma = \frac{N'(d_1)}{S_0\sigma\sqrt{T}} \tag{142}$$

where the variables are as defined in the previous footnote. We care about the $p^2\Gamma$ of the portfolio. If the portfolio includes options on m shares of the underlying asset, then the total $p^2\Gamma$ in equation (141) is:

$$\begin{aligned} p^2 m\Gamma &= p^2 m \frac{N'(d_1)}{S_0\sigma\sqrt{T}} \\ &= (p/S_0)\cdot pm \cdot \frac{N'(d_1)}{S_0\sigma\sqrt{T}} \\ &= 1\cdot(\$11.5\text{billion})\cdot\frac{N'(20\%\cdot\sqrt{5}/2)}{20\%\sqrt{5}} \\ &= \$10 \text{ billion} \end{aligned}$$

[10]The exposure is quoted as $p\Delta$, where $p = 20\%$ and Δ = \$100 million.

- NW = \$2.1 billion[11]
- $\Gamma = 0$

We again assume no changes in LTCM's portfolio balance during a forced liquidation. We can then calculate MinEAMASS from the left hand side of equation (141):

$$\text{MinEAMASS} = p^2 \cdot \left[NW \cdot \left(\frac{\Delta}{NW} \right)^2 \right] \tag{143}$$

$$= (20)^2 \cdot \left[\$2.1b \cdot \left(\frac{-\$100m}{\$2.1b} \right)^2 \right] \tag{144}$$

$$= \$2 \text{ billion} \tag{145}$$

Recall that the elasticity-adjusted market size is given by equation (141) as

$$EAMS = pm + \eta A \tag{146}$$

where pm is LTCM's position (negative \$2 billion), η is the elasticity of demand of the other investors, and A is the assets held by other investors.[12] According to Dunbar (2000) and Lowenstein (2000), LTCM was responsible for about a quarter of the long-term volatility sales while the investment banks were responsible for the rest, meaning that the demand for volatility on the part of unlevered investors, A, was about \$8 billion. This volatility was mainly sold to pension funds and unit trusts that had promised their owners a minimum rate of return.

EAMS is therefore $(-\$2b + \eta \cdot \$8b)$, and the instability ratio, given by the ratio of MinEAMASS and EAMS, simplifies to:

$$\text{Instability ratio} = (4\eta - 1)^{-1} \tag{147}$$

[11]While LTCM's equity in early summer was \$4.5 billion, only \$2.1 billion of this was actually required as margin for trades (MacKenzie, 2003), with the rest as risk capital intended to absorb losses. By early September, LTCM had lost its entire risk capital cushion. The \$2.1 billion is the correct number to consider because it was here that liquidations would have been forced.

[12]As in the rest of this paper, we exclude the banks' short positions in volatility from our calculation of market size as these institutions are assumed not to be price sensitive in their sales, which are determined by risk management considerations. In other words, the banks are treated as issuers, rather than traders.

If the elasticity of demand on the part of unlevered investors was less than 1/2, then EAMS would have been less than $2 billion (the MinEAMASS) and the market for volatility was unstable. This is not entirely implausible, because the volatility was sold to insurance companies and pension funds that were using it to hedge guaranteed returns on their policies and would not have been inclined to sell their options to take advantage of short-term price movements. This shows that LTCM was of the right order of magnitude to destabilize this market.

These results are confirmed by the behavior of the equity market during the late summer of 1998. Despite significant declines, cash equity markets continued to function normally with reasonable liquidity. The market for long-dated volatility in equities, however, was thrown into disarray and became almost completely illiquid. Trading became very sparse and price quotes spiked upward and became divorced from fundamentals, according to market participants quoted in Dunbar (2000) and MacKenzie (2003).

As one banker said:

> When it became apparent that they [LTCM] were having difficulties, we thought that if they are going to default, we're going to be short a hell of a lot of volatility. So we'd rather be short at 40 than 30, right? So it was clearly in our interest to mark at as high a volatility as possible. That's why everybody pushed the volatility against them, which contributed to their demise in the end. (in MacKenzie, 2003)

This quotation demonstrates that in the long-dated volatility markets, prices were not clearly defined. LTCM's counterparties had considerable discretion in what prices to place on the options that LTCM was short. This is only possible in a market that is not liquid. If the market had been active with many participants ready to buy and sell at their estimate of fundamental value, such discretion would not be possible because prices would be determined by the intersection of supply and demand. LTCM was such a big player that the possibility of it being forced to

liquidate was enough to prevent prices from being well-defined. This is a hallmark of instability.

	Equity Volatility	Bank Funding	US Treasury
LTCM Net Notional Exposure ($p\Delta$)	-$2 billion	$20 billion	$20 billion
LTCM Net Worth	$2.1 billion	$2.1 billion	$2.1 billion
Notional Position of Unconstrained Investors	$8 billion	$618 billion	$5.5 trillion
MinEAMASS	$2 billion	$200 billion	$200 billion
EAMS (Assumes $\eta = 1$)	$6 billion	$638 billion	$5.5 trillion
Instability Ratio	0.33	0.31	0.04

TABLE 1. Stability Analysis for Selected LTCM Markets

A similar analysis is possible for another of LTCM's trades, a bet on bank funding costs.

LTCM made its bet using a financial instrument known as an interest rate swap, which is ubiquitous in financial markets although not well-known to those outside. In an interest rate swap, LTCM would "swap" one interest rate for another, agreeing to receive a fixed interest rate in exchange for paying a variable (or "floating") interest rate. For example, if the 10-year swap rate was 5 percent, LTCM could receive this fixed rate every six months for 10 years and pay whatever the six-month LIBOR rate happened to be at that time. This whole transaction is referred to simply as "receiving in swaps."

The fixed swap rate includes a measure of bank credit risk, because in exchange for receiving the fixed rate, LTCM would pay LIBOR. LIBOR is the London Interbank Offer Rate, the interest rate that large banks charge each other for unsecured lending in London. Six-month LIBOR is thus generally higher than the six-month rate on Treasury bills, because somebody lending in the interbank market is taking bank credit risk. As a result, the fixed rate on 10-year swaps should be higher than the 10-year rate on Treasury notes.

The difference between these two interest rates is called the *10-year swap spread*, and it is a measure of the market's expectation of the creditworthiness of banks over a 10-year time horizon.[13] If the swap spread is *wide*, this means that the market

[13]In practice, there are also market and funding liquidity considerations that go into determining the swap spread.

expects that LIBOR will be significantly higher than the rate on Treasury bills in the future. If the swap spread is *narrow*, it means that the market expects that LIBOR will be close to the rate on Treasury bills in the future.

In the spring and summer of 1998, with swap spreads at the wider end of historical norms, LTCM bet that swap spreads would narrow. They did this by receiving fixed in swaps and going short Treasury notes. LTCM would thus receive the fixed rate in swaps and pay the 10-year Treasury rate.[14] If the swap spread narrowed, LTCM could put on the same trade in the opposite direction but with a lower swap spread, locking in a profit.

By betting that swap spreads would narrow, LTCM was effectively buying a basket of 10-year bank bonds and selling 10-year Treasury bonds as a hedge, only it was doing so in a more liquid and efficient way by using swaps.[15] Comparison across sources suggests that LTCM's exposure to US swap spreads in the late summer of 1998 was about $16 million per basis point of swap spread,[16] which corresponds to an exposure of about $200 million per point of the price of a 10-year bond, and

[14]In practice, there is also a spread between the floating "reverse repo" rate earned on the cash received from shorting the Treasury note and floating LIBOR rate received on the swap.

[15]In reality, the 10-year swap rate is a bit different than the rate on a 10-year bank bond. This is because the floating rate on a swap is LIBOR, which embeds a measure of credit risk of the largest money-center banks at the time of the LIBOR fixing. If such a bank suffers a deterioration in its credit risk, its 10-year bond would sell off, but LIBOR might not, because the bank would likely no longer be included in the panel used to calculate LIBOR. However, during a systemic crisis, the risk of immediate default dominates: a bank will likely default soon or not at all. This means that during a crisis, the swap rate and bank bonds become fundamentally much more similar. In any event, a basket of bank bonds probably still the best real asset proxy for the swap rate.

[16]Lowenstein (2000, p. 187) implies that LTCM's exposure to a 15 basis point adverse move in swap spreads was $240 million, which implies an exposure of $16 million per basis point. This would imply that the trade lost $160 million for LTCM on August 21, 1998, a day when swap spreads moved 10 basis points (although as many authors note, they moved up to 20 basis points intraday). Overall, LTCM lost $550 million on August 21, of which $160 million was due a merger arbitrage trade gone wrong and perhaps a bit more was due to increases in equity volatility. This would mean that losses on US swap spreads were a substantial portion of the remaining losses, which is likely given that this was by all accounts one of LTCM's core trades (see, for example, Perrold, 1999). Additionally, a former LTCM principal told me that $10 million per basis point was a plausible estimate, which I take to mean that it is within a factor of two. Finally, the same principal told me that the so-called replacement notional value of LTCM's interest rate swaps was about $80-90 billion. This would imply that slightly over $20 billion of this was in US swap spreads. This is plausible because other spreads used swaps much more intensively for example a single relative value trade between the UK and Germany would have used 4-6 times as much swap notional for each basis point of risk. (Trade #5 in Perrold, 1999, page C6, for example.) Trades in Japan (trade #8, page C7) would have used twice as much swap notional for each basis point of risk.

therefore a notional exposure of $20 billion both to Treasury bonds and to bank credit.

We can again use equation (141) to determine the MinEAMASS associated with this position.

The relevant parameters are:

- Δ = $200 million per bond point
- p = 100 points
- NW = $2.1 billion
- $\Gamma = 0$

Plugging these into equation (141) gives a MinEAMASS of $200 billion. The Treasury market was nearly 30 times larger than this, or $5.5 trillion, thus leading to an instability ratio of less than 5%. LTCM was nowhere near big enough to destabilize the Treasury markets.

For the bank funding market, however, we have:

- pm = $20 billion
- A = $618 billion[17]

This gives EAMS of the bank funding market as $638 billion if the elasticity of unconstrained demand was equal to one. This leads to an instability ratio of about 0.3. However, if the demand for bank credit had been inelastic ($\eta <$ about 1/3), then EAMS would have moved down towards MinEAMASS and the instability ratio would have approached one. In this case, LTCM alone could have been enough to destabilize the market for bank credit.

In this analysis, we have considered the stability of the equity and Treasury markets. While LTCM was not large enough to destabilize equity or Treasury markets in general, it could, in an inelastic demand environment, have been large enough to destabilize the markets for bank funding and equity volatility.

[17]Federal Reserve Flow of Funds, Table L.3, "Credit Market Debt Owed by Financial Sectors." $618 billion is the sum of $188.6 billion owed by commercial banks, $193.5 billion owed by bank holding companies, $212.4 billion owed by savings institutions, $1.1 billion owed by credit unions, and $42.5 billion owed by broker-dealers, minus the $20 billion held by LCTM.

4. A Comment on the Choice of Markets

We have examined LTCM's impact on the stability of the global equity markets, global equity volatility markets, US Treasury markets, and US bank funding markets. The choice of these markets has been informed by the positions LTCM held, and in some sense seems obvious. But a skeptical reader could easily object that we have chosen the wrong markets. On the one hand, why consider the stability of the global equity markets as a whole rather than the US and German markets separately? On the other hand, why consider the global equity markets in isolation from the global debt markets?

In the context of the analysis we have presented, there is no assuredly right or wrong answer. We can use the MinEAMASS framework to examine the stability of any market segment we can define. However, the narrower the definition of the market, the more elastic is the demand of unconstrained investors, and thus the larger is the elasticity-adjusted market size[18] relative to the unadjusted market size. When EAMS is larger, the levered speculators such as LTCM must control a larger overall proportion of the market to destabilize it.

For example, in our analysis of the effect of LTCM's swap spreads exposures on market stability, we have examined the impact on bank funding markets, because (especially during crisis scenarios) swap spreads are a measure of bank funding costs. We could equally have considered the effect of the trades on stability in the funding market for only the 16 banks whose funding costs are considered in the calculation of LIBOR. However, bonds of these banks are highly substitutable for bonds of other banks in investors' portfolios, so the elasticity of retail demand that is an input into calculating EAMS would be very high. Similarly, we could have considered the effect of LTCM's swap spreads trades on the high-grade corporate funding markets generally, because if swap spreads widened, investors could sell their holdings of highly-rated non-financial corporate bonds to buy bank bonds. However, the elasticity of retail demand for corporate bonds as a whole is less than that of

[18]Recall that EAMS is given by $pm + \eta A$.

bank bonds in isolation because there are fewer close substitutes. This would tend to reduce EAMS, other things equal. We have considered the bank funding market as rather than the general corporate bond market because instability in the bank funding market is linked to financial crisis scenarios, and in these scenarios, banks (such as Citigroup) become markedly more risky than high-grade corporations (such as Toyota), and their bonds cease to be good substitutes.

Regardless, the framework I have proposed can be used to evaluate the stability of any of these market subdivisions; it is possible for a market subdivision to be unstable while the broader market is stable. Undoubtedly, there will always be some substitution between assets and an attempt by some investors to arbitrage price discrepancies between similar assets, which will lessen instability somewhat. This tendency is captured by the parameter η, the elasticity of retail demand.[19]

5. Corroborating Evidence

Real-world behavior is always more complex than economic models, but the narrative of the rise and fall of LTCM generally corroborates the key behavioral assumptions and predictions of the model. LTCM acted through most of its history as though it was constrained by its capital; asset prices became driven by credit and net worth constraints on levered investors rather than their expectations; LTCM's leverage and size relative to the most important markets in which it played made it difficult or impossible for the fund to liquidate without causing instability.

Perrold (1999) has obtained directly from LTCM the total assets and equity of the fund, reported monthly from inception to mid-1998 (see figure 2). It is clear

[19]In the canonical, representative-agent models of asset pricing, if an asset class is small enough that changes in its price do not change aggregate consumption, then its price is determined only by expectations, so that $\eta = \infty$.
Specifically, the price of the asset must cause the rate of return to satisfy

$$u'(c_t) = \beta E_t[u'(c_{t+1})R_t]$$

using the usual notation.
However, once the assumption of representative agents is violated, then different agents can have different information sets and different correlations between asset returns and consumption, so a change in the distribution of wealth produces changes in relative asset prices. The current model is explicitly not a representative agent model, and thus there is no reason to believe that demand for a given asset class should be perfectly elastic.

from even a casual glance that the fund's equity is an important determinant of its assets. As LTCM ramped up its operations and began both to raise and earn capital, it was able to increase its assets as well, although as Perrold points out, the growth in assets outpaced the growth in capital for a time as LTCM built its operations. This point, that LTCM's ability to take positions was constrained by its capital, seems so obvious that it should escape mention. Yet it is an important point, because LTCM was precisely the kind of rational arbitrageur whose ability to take large positions is said to produce efficient pricing. In fact, LTCM's balance sheet size was, in the early stages of its existence, more governed by right-sizing its assets to meet its capital base than by its assessments of the available arbitrage opportunities in the marketplace.

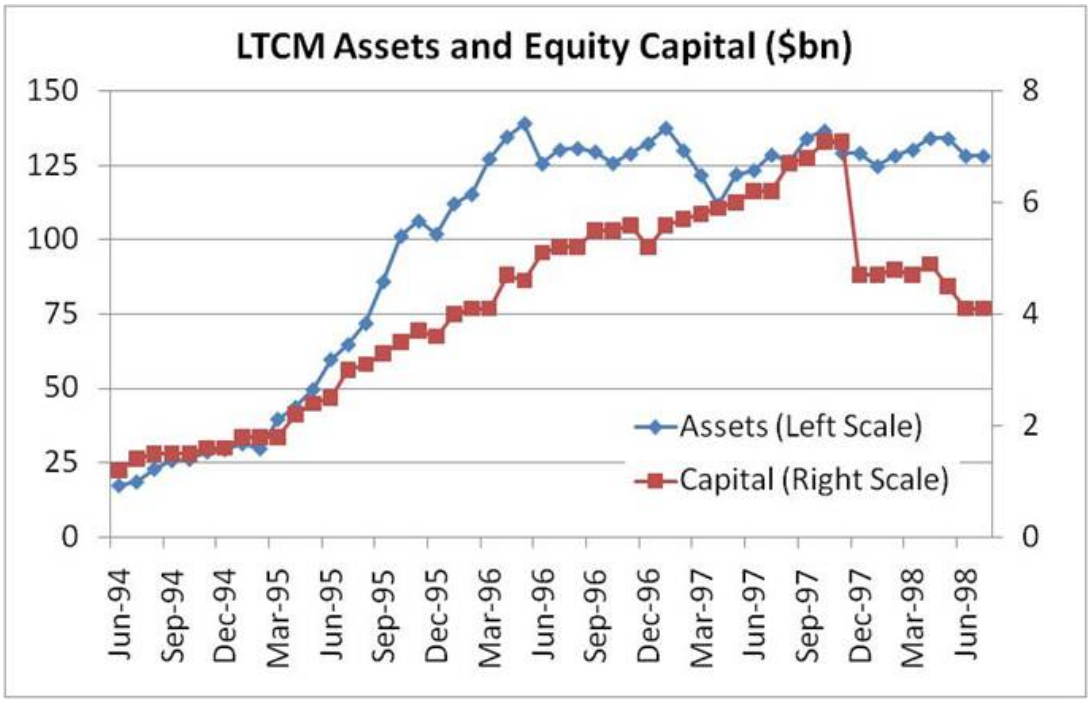

FIGURE 2. Source: Perrold (1999)

LTCM's behavior in this early stage of its existence is closer to what is proposed in this paper, or earlier models of capital-constrained arbitrage and investing behavior such as Grossman and Vila (1992), Shleifer and Vishny (1997) or Liu and Longstaff (2004), than it is to an unconstrained neoclassical agent.

As LTCM grew and was successful, its strategies were noticed and widely imitated. MacKenzie (2003) explains that many of the large investment banks, such as Salomon Brothers, Deutsche Bank, Goldman Sachs, and others, began trading similar strategies. New hedge funds, such as Convergence Asset Management, also grew up to compete with Long-Term. This process, which MacKenzie views as a sociological one, corresponds in my model to the recruitment of new speculators into a bubble. Of course, LTCM was trying to discourage rather than encourage new entrants into its markets, but the mathematical description of the behavior is the same, with success being imitated.

LTCM was not completely price-insensitive in its buying, contrary to what the model assumes about speculators. Instead, as it grew and was imitated, LTCM noticed that favorable opportunities were "drying up big," as Eric Rosenfeld put it (quoted in MacKenzie, 2003). LTCM reacted by returning capital to investors, putting a ceiling on its willingness to continue its strategy at less and less attractive pricing. This behavior prevented the kind of instability in the upward direction that is contemplated as a possibility in the model. Instead, the decision to return capital to investors kept LTCM's capital at a level where margin constraints had to be considered. Instability in the downward direction was thus still very much a possibility, because as all the narrative histories of LTCM that I am aware of point out, capital cannot be easily raised by a fund once it begins to lose money and approaches its credit or net worth constraint.

This instability was brought into play in the summer of 1998, when prices began to move against LTCM. Even at the time, sophisticated market participants understood and commented that asset prices were moving away from fundamental value. As William Winters, head of J.P. Morgan's European Fixed Income business, put it at the height of the crisis, "any concept of long-term or fundamental value disappeared." (Coy, 1998)

In this situation, a value-oriented, unconstrained investor would be betting on prices to converge. So, too, would a speculator who saw prices moving against his

long-term view, and indeed, this is precisely what LTCM *wanted* to do. LTCM principals continued to have confidence in their trades even as the market moved further and further against them. As Meriwether put it in his August 1998 letter to investors (reproduced in Perrold, 1999):

> With the large and rapid fall in our capital, steps have been taken to reduce risks now... On the other hand, we see great opportunities... The opportunity set in these trades at this time is believed to be among the best that LTCM has seen... LTCM thus believes it is prudent and opportunistic to increase the level of the Fund's capital to take full advantage of this unusually attractive environment.

Rosenfeld put it more directly: "We dreamed of the day when we'd have opportunities like this." (Lowenstein, p. 166)

This was not just talk. No source disputes that Meriwether was actively trying to raise additional capital. As Lowenstein (2000), Dunbar (2000), and others note, the partners' faith in the their trades was ultimately proven correct. The consortium of investment banks that took over the fund was left with a double-digit profit one year later.

Yet as markets were moving against it, creating more attractive opportunities, LTCM was liquidating some trades, adding to the price pressure against it. According to Dunbar (2000, p. 194), LTCM decided at the end of June to reduce its daily value-at-risk (VAR)[20] from $45 million to $35 million.

The market moves that undid LTCM were not driven by fundamentals. Instead, they appear to have been driven by other savvy traders' anticipation of the instability that would have been caused had LTCM been forced to liquidate. Investment banks were open about this. As one Goldman Sachs trader put it, "If you think a gorilla has to sell, then you sure want to sell first." Goldman CEO Jon Corzine did not deny that the firm "did things in markets that might have ended up hurting LTCM. We had to protect our positions. That part I'm not apologetic about." (Lowenstein,

[20]VAR is a measure of risk, typically quoted as 95% confidence interval of daily profit and loss.

2000, p. 175.) Lowenstein cites similar sentiments from executives at other banks, in particular Salomon Brothers, which had a portfolio of a similar size to LTCM (Dunbar, 2000).

Given that the market moves were driven by fear of instability, it is no surprise that LTCM's losses were concentrated in the areas where it had most destabilized markets. Lowenstein (2000, p. 234) provides a breakdown, showing that of the $4 billion lost by Long-Term in 1998, $1.6 billion was in swaps and another $1.3 billion in equity volatility. No other category of losses even tops $500 million.

6. Conclusion

Most accounts of the demise of Long-Term Capital Management argue that the fund's fundamental failure was that it was too highly levered and took too much risk. Lowenstein, Dunbar, the President's Working Group on Financial Markets, and the Basel Committee all accept this hypothesis. Yet the LTCM principals point out that the fund was no more levered than an investment bank, and had long-term financing in place that made this comparison at least superficially reasonable. The firm's risk measures proved conservative for years before failing amid the regime shift of the financial crisis. The principals rightly point out that in normal times, LTCM's leverage massively overstated its risk because the assets it owned were highly correlated to the assets it sold short.

Yet leverage is clearly part of the story. The stability analysis we have conducted makes clear that LTCM's problem was not that it was simply too highly levered. Had it been smaller it might have survived. Rather, the problem was that it was *both* highly levered *and* large relative to the markets in which it invested. LTCM's very existence destabilized the markets, creating the potential for much larger price moves than would have been possible in the absence of the fund's existence. Instead of the risk being mitigated by LTCM's hedges, as it would have been had the fund been smaller, the crisis risk became governed by the theoretical notional exposure.

All of a sudden, the correlations between the assets LTCM was betting on changed, precisely because LTCM was betting on them.

As LTCM partner and Nobel Laureate Myron Scholes put it: "As a result of the financial crisis, LTCM was forced to switch from being a large supplier to being a large demander of liquidity, at a cost that eliminated its capital." (Scholes, 2000)

Rosenfeld hinted at this in his presentation to business school students at MIT in 2009, as he argued that LTCM failed to take into account what he termed "endogenous risk." Rosenfeld's endogenous risk is precisely what is being explored in our stability analysis.

A fuller analysis of the collapse of LTCM will have to wait for a day when more data on LTCM's portfolio are made public. However, we have shown how at least in some of the markets that LTCM participated in, it was approximately the size necessary to destabilize those markets and cause a crisis.

CHAPTER 4

A Textbook Case: The Great Crash of 1929

1929 was the climax of a decade of unbounded optimism, reflecting unprecedented prosperity and a rate of sustained economic progress that had not been seen for at least a generation. The dissemination of new technologies such as electricity and radio spurred contemporary analysts to speak and write of a "New Era" of plenty. As Galbraith (1954) put it:

> American capitalism was undoubtedly in a lively phase. Between 1925 and 1929, the number of manufacturing establishments increased from 183,900 to 206,700; the value of their output rose from $60.8 billions to $68.0 billions [amid slightly declining prices]. The Federal Reserve index of industrial production which had averaged only 67 in 1921 (1923-1925=100) had risen to 110 by July 1928, and it reached 126 in June 1929. In 1926, 4,301,000 automobiles were produced. Three years later, in 1929, production had increased by over a million to 5,358,000, a figure which compares very decently with the 5,700,000 new car registrations of the opulent year of 1953.

It was not just production but productivity that advanced. Manufacturing output per worker, having been essentially flat for two decades, rose by about half in the 1920's (Fisher, 1930). Inflation-adjusted earnings per share for S&P 500 companies rose from $9 per share in 1922 to $20 per share in 1929[1]. Yet in the fall of 1929, stocks crashed, losing nearly half their value in two months. Ultimately, stocks bottomed out in 1933 about 90% below their 1929 peaks.

[1]http://www.multpl.com/s-p-500-earnings/table

This section will apply the MinEAMASS framework to attempt to elucidate why stocks crashed with such ferocity during the two months from mid-September to mid-November 1929. The historical record is clear that forced selling was abundantly evident during the crash, but economic historians have not, as yet, produced a quantitative explanation of why the crash happened when it did.

In this section I shall argue that crash was the result of the interaction between three principal dynamics. In the months leading up to the crash, banks and brokers tightened margin requirements to unprecedented levels (Smiley and Keehn,1988), which had the effect of sharply reducing the excess collateralization in many margin accounts. Then, the economy and market encountered a series of modest negative shocks, which began to force weak accounts to reduce exposure. The selloff turned into a rout because of the degree of margin borrowing, which created a situation where MinEAMASS was dangerously close to the threshold for instability.

1. Previous Literature

There is obviously a large literature on the Great Depression that followed the stock market crash of 1929. However, despite the linkage in the popular imagination, economists have tended not to view the crash as having played a major role in the onset of the Depression. (Exceptions are Romer, 1988, who cites the increase in uncertainty resulting from the crash in depressing consumption, and Kindleberger, 1973, who cites a liquidity crunch resulting from the crash as having led to cancellations of industrial orders, calling of loans, and deflation.) As a result of this, the academic literature on the anatomy of the crash itself is rather thin, and surprisingly, there is no comprehensive source of an academic standard that attempts an integrative analysis of the causes and mechanics of the crash.

The first systematic study of the crash was undertaken by Irving Fisher (1930). Fisher evaluated a number of potential causes of the crash that had been proposed by contemporary financiers, politicians, and economists. Among these were overvaluation of stocks, changes in economic fundamentals, excessive speculation, excessive

margin borrowing and forced liquidation, excessive new issuance and undigested securities, stock pools and manipulation, foreign selling, and capital gains tax laws. While economists have debated the various potential causes of the crash in the decades since Fisher's work, virtually no new potential causes have been proposed.

Fisher's list of explanations can be divided into four main categories:

- Previously existing unsustainable or unstable conditions that were due to be corrected, for which the crash was the mechanism (overvaluation, excessive speculation, excessive margin borrowing);
- Economic or financial shocks that were sufficiently large and rapid to cause a crash (changes in economic fundamentals, stock pools and manipulation, excessive new issuance);
- Small shocks that acted as a triggering mechanism for a larger crash (various news items, foreign selling); in conjunction with—
- Amplifying mechanisms (capital gains tax laws, forced liquidation).

These causes are, for the most part, not mutually exclusive, and it is possible, perhaps probable, that many of them were operating in concert.

Perhaps the most hotly contested issue is whether stocks were, indeed, overvalued in 1929. The issue was debated at the time, and there is still no consensus among economists today. As has been documented by Fisher, Allen (1931), Friedman and Schwartz (1963) and others, many market observers, regulators, and large and sophisticated investors in 1929, including (contrary to popular belief) Fisher himself, believed that stocks were overvalued. However, aside from the doomsayers such as columnists Alexander Dana Noyes and Roger Babson, the consensus view among the informed was that the overvaluation was mainly speculative froth: stock prices required a correction to more reasonable levels before continuing to grow with the economy. As Fisher put it, "My own impression has been that the market went up principally because of sound, justified expectations on earnings, and only partly because of unreasoning and unintelligent mania for buying." (p. 50) Stocks had their cheerleaders, too, of course—John J. Raskob, a senior executive of General

Motors and Democratic National Committee Chairman from 1928-1932, gave his famous "everybody ought to be rich" interview to Ladies' Home Journal just months before the crash.

Time has done little to shed light on the question. Galbraith (1954) scoffed at the idea that stocks were anything other than a massive bubble, and Shiller (2000) provided quantitative evidence.[2] Rappoport and White (1994) go so far as to argue, based on rising margin requirements at broker-dealers, that the crash was actually expected by market insiders months in advance. By contrast, McGrattan and Prescott (2001) conclude that in light of the information available at the time, stocks were appropriately valued for the future economic prosperity that was rationally anticipated, and even Kindleberger (1973) does not claim that stock prices were obviously excessive.

A more careful examination by Wigmore (1985) demonstrates that stock prices were not monolithic, with some industries such as utilities, investment trusts and banks showing absurd valuations of as much as 100 times earnings while valuations in other industries remained quite reasonable. Stocks traded on the Curb Exchange (forerunner of the American Stock Exchange) generally also reached higher valuations than NYSE stocks, although this varied by industry (see Kuvin, 1930 for price indexes for the Curb).

In any event, the inability of economists and investors to agree, even decades later and with ample time to study all the facts, on whether the state of the world in 1929 justified the high level of stock prices at that time makes it difficult to rely on overvaluation of stocks as an explanation for the crash. If we believe that prices are determined by expectations of fundamentals, and time has not appreciably changed opinions on whether those expectations were rational, then the crash was probably not due to a realignment of prices with rational expectations.[3]

[2]At the peak, stocks were valued at about 20 times 12-month earnings and more than 30 times ten-year trailing inflation-adjusted earnings. (http://www.multpl.com/) The Shiller P/E multiple crossed 20 in April of 1928 and 25 in November. The long-term average has been approximately 16.

[3]An exception to this argument is if the mechanism is as described in Abreu and Brunnermeier (2003), whereby a bubble can persist even if all investors recognize that prices are out of line with

Of course, rational expectations may be revised in response to deteriorating fundamentals, which brings us to another potential cause of the crash—a revision of expectations to reflect the worsening of the economy and the onset of the Depression. Industrial production peaked in June and declined at an annual rate of more than 10% over the next three months before falling off a cliff in November and December. The pace of deflation, which had been ongoing at a modest rate in response to the Fed's tight monetary policy, accelerated before and during the crash as well. But, as Kindleberger (1973) pointed out, the worst of these data were not public at the time of the crash, although certainly the nation's top businessmen would have had a feel for the state of business activity independent of government data. Temin (1989) investigates actual expectations at the time of the crash and finds that neither a severe deflation nor a deep depression was anticipated. Perhaps the strongest evidence that rational changes in expectations could not justify the crash is the failure of perhaps the most respected American economist of the day, Irving Fisher, to attribute the crash primarily to a weakening economy, even after the fact. Fisher also notes that several large corporations actually increased their dividends during the crash in a show of confidence about the future, but that "*the market had no ear for such news*, because it was deafened by the stentorian voices calling upon individuals and brokers to repay their loans." (Emphasis added.)

Some authors, most notably Allen (1931) and Galbraith (1954), point vaguely to stock pools and manipulation as a cause of the crash and bubble, with large investors manipulating stock prices up and down for profit. Fisher, too, addresses this assertion, without a firm conclusion. While it is well-documented that such pools were in operation, Wigmore (1985) explains that the pools were not omnipotent and suffered severe losses during the crash. If bear-raiding stock pools engineered the

---

fundamentals because speculators prefer to ride the bubble rather than bet against it. In such cases, any piece of news may act as a coordination device that causes speculators all to liquidate at once, leading to a crash. While Abreu and Brunnermeier do not apply their model to the Great Crash of 1929, their story does fit a big part of the narrative quite well. In particular, despite many traders in the industry believing that stocks were overvalued, these traders appear to have attempted to trade on their short-term beliefs about the path of stock prices, rather than long-term fundamental value.

crash, then they defeated well-financed and apparently incompetent bullish pools as well. Mahoney (1999) conducts a systematic study and finds that pools were not successful at manipulating prices for profit, arguing that the primary purpose of pools was for market making and improving liquidity. However, his analysis focuses narrowly on the issue of pools buying and selling stocks for profit and does not address many of the schemes described in the sources he cites, such as profits on options. Further, he makes a few assumptions that may weaken the power of his tests,[4] so that a negative result should not necessarily be taken at face value. In any event, the Pecora Commission (1934)[5] transcripts contain explicit admission by pool participants that some practices of pools were "bad" and designed to create enthusiasm by exaggerating trading volumes.

Overall, the literature contains some evidence that pools manipulated the prices of individual stocks over periods of a few days or potentially weeks, mainly through "pump and dump" style schemes. Pools were also formed to facilitate insider trading, which was legal at the time. However, the primary purpose of the pools appears to have been to engage in distribution of large blocks of securities and to stabilize markets during those distributions. There is no evidence that pools were responsible for sustained excess valuations of stocks generally, or for the broad-based declines of the autumn of 1929. Indeed, during the crash itself, some prominent pools attempted to stabilize the market by publicly buying stock, and so were a mitigating, rather than aggravating, factor.

1929 also marked a high point for new issuance of securities. Galbraith (1954) and Wigmore (1985) claim that this produced significant selling pressure, while White (1990) conjectures that it could not have because it was too small to do so. In one sense, White is clearly correct in that new issuance of more than 1% of stock

[4]For example, he restricts his attention to pools that operated pursuant to a written agreement; presumably pools that were engaged in untoward activities would have been less likely to have a written agreement. He also excludes from his analysis pools in more thinly traded stocks not found in the CRSP database, which would have been easier to manipulate because there was less liquidity.

[5]Directed by Ferdinand Pecora, this commission was chartered by Congress to investigate causes of the crash.

market capitalization, as occurred in September 1929,[6] was clearly not enough to account for the full measure of decline in stock prices. However, this was a sizable shock to supply, and if (as I shall argue) the market was already on the verge of instability, this could have pushed it over the edge. Indeed, newspaper accounts of the time cited a glut of undistributed securities being carried by dealers as a source of selling pressure.

Foreign liquidation in September has also been cited as one cause for the crash, most notably by Fisher (1930). Fisher and others after him, including Bierman (1991), argue that the discovery of Clarence Hatry's fraud in London on September 21, and the subsequent bankruptcy of his companies, caused London financiers to sell their American holdings to cover some £12.5 million in unsecured credit exposure. To support this claim, Fisher notes that the dollar-sterling exchange rate moved rather rapidly from the British gold export point to the gold import point (a dollar depreciation) as foreigners sold their stock holdings and converted the proceeds into sterling for repatriation.

However, despite the splash that Hatry's fraud made in both the London and New York papers, the economic impact was relatively small. Relative to the capitalization of the stock markets, Hatry's £12.5 million bankruptcy was comparable to the collapse of two subprime-focused Bear Stearns hedge funds in June 2007, not the thirty-fold larger bankruptcy of Lehman Brothers in September 2008. A more likely candidate for the London liquidations was the Bank of England's hike in Bank Rate of one full percentage point on September 26. This increase was anticipated for the same reason it occurred: the Bank of England was losing gold reserves at an alarming rate, having suffered outflows of nearly a quarter of its reserves in the preceding 12 months. The Bank of England raised rates to stop gold outflows and encourage investors to invest in the UK instead of abroad. It worked, at least temporarily, and some other European central banks followed.

[6] *Banking and Monetary Statistics.*

There are no solid data on international capital flows sufficient to trace the link between this foreign selling and the stock market decline. However, even in a world with no frictions, this 1 percentage point rise in interest rates would have been responsible for a several percentage point selloff in equity markets, simply by lowering the net present value of future cash flows. While this is clearly insufficient to account for the full measure of the crash, it could certainly have been a triggering factor in an unstable market.

The Hatry affair, together with the October 11 refusal of the Massachusetts Department of Public Utilities to allow Boston Edison to split its stock on the grounds it was wildly overvalued, have been cited by Fisher as shaking investors' confidence, and thus contributing to the crash. Yet the Massachusetts decision did not even trigger a selloff: the Dow Jones Industrial Average was down just 0.05% on Friday, October 11 and 0.49% on Monday. The selloff accelerated on Tuesday and Wednesday, but the Wall Street Journal attributed this to worsening economic data and triggering of stop-loss orders.

Ultimately, the picture that emerges from a close look at September and October 1929 is that markets were buffeted by a series of modest negative shocks, including interest rate hikes, the Hatry affair and Massachusetts decision, and worsening economic data.

Together, these sent the Dow Jones Industrial Average down 12.5% from its high on September 3 to the end of the orderly part of the market decline on October 18. Yet, over the next month, to November 14, the DJIA fell a further 40%, with little news aside from price declines. Over this period, I shall argue, declines became self-fulfilling as they forced liquidations in an unstable manner. In the remainder of this section, I shall demonstrate that selling from this point on was accompanied by heavy margin calls, and I shall apply the MinEAMASS framework to show that the market was near a point of instability, suggesting the possibility of a self-fulfilling downward spiral.

2. Forced Liquidation During the Crash

Virtually all discussions of the 1929 stock market crash, including Fisher (1930), Allen (1931), Galbraith (1954), Sobel (1965) and Wigmore (1985), give a starring role to margin calls and forced liquidations. A reading of the *Wall Street Journal*'s "Abreast of the Market" column, which summarized each day's market developments and the drivers of stock price movements, during the period of the crash gives a similar impression. Using the market color provided in the *Journal*, at least seven of the ten down days between the beginning of the crash proper on October 24 and the local minimum of stock prices on November 15 were accompanied by forced liquidations. In the two months from September 16 to November 14, the average day when forced liquidations were reported saw the Dow Jones Industrial Average decline 2.8% for the day, with an average intraday low of 5.6% below the previous day's close. By contrast, days with no forced liquidation reported saw an average gain of 0.5%. While declines clearly cause forced liquidations, some inferences may also be made as to causality from margin calls to price declines, because margin calls often went out overnight and then margin selling overwhelmed the market the next morning.

In addition to margin selling, the *Journal*'s narrative provides a window into another source of forced selling: stop-loss orders.[7] The less severe declines of late September were accompanied by triggering of stop-loss orders, and the period prior to the primary crash from October 24-29 saw first a buildup of stop-loss orders and then their liquidation as short sellers reportedly sold in the hope that these orders would be triggered. The declines triggered by the stop-loss orders then triggered truly forced liquidation. These stop-loss orders reflect a sort of soft margin call. According to the *Journal*, these orders were employed to "protect" weak margin accounts by reducing leverage as account values declined. The idea was to ensure that accounts would not be forcibly liquidated and to control risk. Accounts using such a strategy, however, behaved as the speculators in my model—they liquidated

[7] A stop-loss order is an order to sell at the market if the price falls below a certain level.

as positions moved against them in order to maintain a targeted leverage ratio. Whether this behavior was forced by external forces or not hardly matters for the result.

The *Journal*'s account of the crash also makes clear that the forced liquidation of speculators took place over a period of several weeks, as compared with the model's prediction that the crash should be instantaneous in the absence of frictions. The explanation for this is likely complex, but there appear to be three major themes.

First, the actual act of liquidation took time. Each undermargined account required a determination by a clerk that the account was undermargined, a decision of the broker to sell out the account, and a formulation, transmission and execution of the orders to do so. The back offices had to perform the clearing and settlement functions. Margin determinations were not automatic (there were no computers back then), so that these calculations were generally (though not always) undertaken daily, with margin calls sent out by telegram after the market close. During the crash, the *Journal*'s reports indicate that the sheer volume of this work overwhelmed the infrastructure of the financial system, and many firms fell days behind in their work of actually executing the forced liquidations.

Second, insiders appear to have engaged in the kind of predatory trading that Brunnermeier and Pedersen (2005) described, selling short along with the forced liquidations and then covering their positions when the forced liquidation had temporarily completed. This resulted in several days where forced selling in the morning led to sharp losses, and short covering in the afternoon led a recovery. Examples include Black Thursday, October 24,[8] when the Dow Jones Industrial Average was at one point down 11% intraday before recovering to close down only 2%; Black Tuesday, October 29, when the Dow was down a whopping 18.5% intraday before closing down about 12%; and Thursday, November 7, when the Dow fell more than 6% intraday before closing up 2.6%.

[8]Although here the bankers' pool intervention also had an effect.

Third, information dissemination and processing was not instantaneous, and it took some time for smaller investors to step in and provide liquidity, as they eventually did.

In any event, it does seem clear that margin calls and forced liquidations were the major driving force behind the crash. In the remainder of this section, we shall use the MinEAMASS framework to examine why the market was as vulnerable as it was to a spiral of forced liquidations.

3. Leverage in the Stock Market in 1929

Leverage in the stock market in 1929 came essentially from two sources. The first was investment trusts, and the second was orthodox margin borrowing.

Investment trusts were essentially mutual funds with a tiered liability structure. The trusts' assets consisted of common stocks, including other investment trusts. Their liabilities consisted of senior debt and preferred stock, which entitled the holder to fixed interest or dividend payments, and of common stock, which held the residual value. The investment trusts thus embedded leverage, typically of about 1:1 (see, for example, Wigmore, 1985, or De Long and Shleifer, 1991), because the debt and preferred stock would be entitled to the same payments as long as the value of the assets was sufficient to cover them, and any change in the value of the assets would be fully gained or lost by the common equity. Many investment trusts were pyramided, in that they held shares of other investment trusts, magnifying the leverage.

It would seem that we should simply increase our estimates of leverage to take account of investment trusts. The correct treatment is a bit more complicated, however. First, the debt instruments issued by investment trusts were term instruments, which means that investment trusts would not be forced to liquidate simply because the value of their assets declined. Instead, an investment trust would enter forced liquidation only when it could no longer raise sufficient cash to meet current contractual commitments. This means that the trust would not behave like a

levered investor in the model. Holders of investment trust equity, however, could be required to liquidate their holdings faster during a crash if they had bought on margin, because the percentage decline in investment trust equity would be greater than the percentage decline in the underlying stocks in the investment trusts' portfolio. But again, there are caveats. Brokers typically required different margin for different stocks, and they presumably would have had higher margin requirements for the riskier investment trusts, in effect offsetting the embedded leverage. Further, it is unclear whether there were rational arbitrageurs ensuring that the value of the investment trusts accurately reflected the value of the underlying portfolio. Indeed, De Long and Shleifer (1991) claim evidence of a divergence between the value of investment trust shares and the underlying portfolios.

Investment trusts dominated the issuance of new stocks for much of 1929. Despite the prominence of the explosion of investment trusts in late 1928 and 1929 in many narratives of the crash, however, these narratives have generally not drawn clear causal lines from the investment trusts to the crash. Rather, they have cited investment trusts as one manifestation of increased leverage and financial chicanery of the period leading up to the crash. What role they played in actually causing the crash is left unsaid. More than likely, the trusts, which traded at a premium to the value of their underlying assets (De Long and Shleifer, 1991), contributed somewhat to the overvaluation of stocks, although the macro effect of this would have been small because the total capital raised by investment trusts through 1929 was only \$3.4 billion (*ibid.*), or less than 3% of total U.S. market capitalization.

A second source of leverage is traditional margin borrowing in the call money market, where banks and corporations lent money overnight to investors, secured against their stock holdings. Most discussions of the crash cite the vast increase in margin borrowing that occurred in 1928 and 1929. Yet while it is true that margin borrowing on the New York Stock Exchange rose sharply during the second half of the 1920's, so did the market capitalization of the exchange. In fact, the percentage of the market capitalization of the NYSE that was funded with broker loans (which

were primarily used for margin credit, and were its largest single source) stayed roughly constant from 1926 until the crash in 1929, at which point it fell precipitously to its lowest level since data began to be collected.[9]

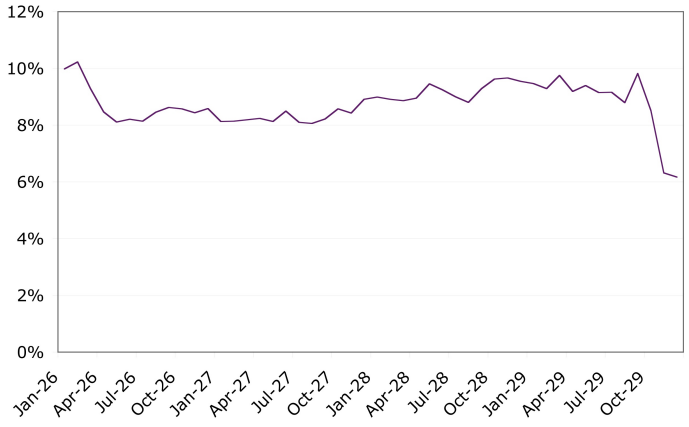

FIGURE 1. Share of Market Cap on the NYSE Financed with Broker Loans

This does not tell the full story, however: the institution of margin investing did undergo at least one unprecedented change prior to the crash. From late 1928 through the summer of 1929, banks and brokerage firms tightened margin requirements from about 10-20 percent to about 50 percent, levels that were without precedent.[10]

While good data are lacking, this doubling or tripling of margins would probably have made more investors act in a constrained manner—that is, more like the speculators than the retail investors in our model—for two reasons. Most obviously, the tightening margins may have forced some investors to post collateral and squeezed them financially. Secondly, for investors who would have wished to be more levered

[9] *Banking and Monetary Statistics*, Smiley and Keehn (1988).

[10] There is no comprehensive data on margin requirements or funding constraints during the 1920's. However, a few authors have investigated the issue by compiling selected primary sources. During peacetime, margin requirements during the 20th century prior to 1928 generally ranged from 10 to 30 percent. (Smiley and Keehn, 1988) However, in late 1928, banks grew concerned about valuations and began increasing margin requirements (or haircuts) on loans to brokers, and brokers passed this increase on to their customers. Margin requirements rose from as low as 10-20 percent up to 40-50 percent by the middle of 1929. At the tail end of the crash, margins were quickly loosened, beginning in late October and November. See, for example, Wigmore (1985), Rappoport and White (1994), Smiley and Keehn (1988).

than the 1:1 ratio subsequently permitted, the collateral constraint became binding and they would rationally have behaved more like the speculators in the model.

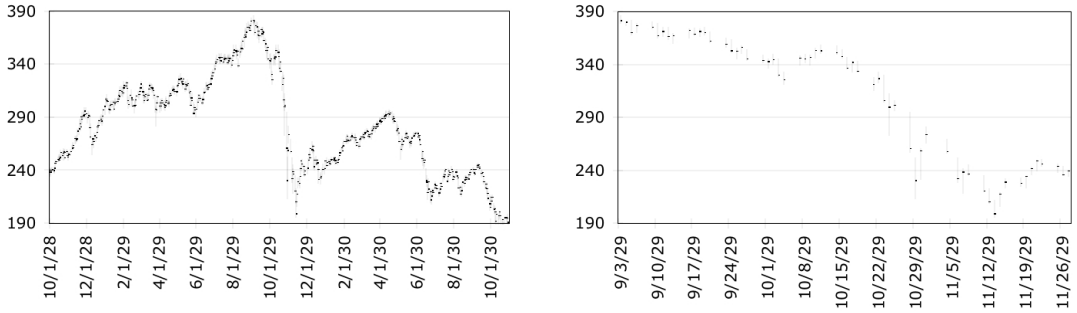

FIGURE 2. The Dow Jones Industrial Average, 1928-1930

It seems almost obvious that such a severe tightening in margin requirements would lead to a sell-off amid forced liquidations. Yet just a few years later, on April 1, 1936, the Federal Reserve used its new-found powers to tighten margin requirements generally from as low as 25 percent, depending on the security, to 55 percent.[11] While equity prices did fall temporarily in April, there was no instability, and they soon recovered and made new highs. The reason for the differing behavior in 1929 and 1936 is no mystery, of course: margin loans were much less significant in 1936 than in 1929. But the MinEAMASS framework, for the first time, is able to provide a theoretically grounded quantitative argument for why the 1929 market was unstable while the 1936 market was not.

To estimate MinEAMASS, we need to determine the net worth of constrained investors, as well as their leverage ratio and its variance. We then compare this with the total market size in order to evaluate the stability of the market. For 1929, as for most crises, data of sufficient granularity are not available. However, financial statistics and reports from the time allow a rough calculation, which I undertake here.

3.1. Quantity of Margin Credit. Compilations of data from the Federal Reserve indicate that the total value of margin loans outstanding against stocks at the

[11] *Banking and Monetary Statistics*, Table 145.

end of September 1929 was around $18 billion, of which perhaps $14.5 billion was to constrained investors. The remainder of this section explains this estimate in detail.

Margin loans to investors came from two main sources: banks and brokerage firms. Brokerage firms, in turn, funded their share of those loans on the call money market. Data are available on direct bank lending to investors on collateral of securities, as well as on brokers' borrowings on the call market.

Direct bank lending to non-broker-dealers for the purposes of carrying securities was $7.2 billion at the end of September.[12] Brokers' loans on the call market were $8.55 billion[13]. However, this represents broker-dealers' total borrowings on the call market, and includes loans used to finance their own inventories, so only part of this total represents margin loans to their customers. These loans also were not the only source of financing for margin loans; brokers also used their own capital and other customers' credit balances for this purpose. However, according to the Federal Reserve, brokers' borrowings are mainly an indication of margin loans made by brokers.[14]

Beginning in November of 1931, the Fed explicitly collected data on margin loans to customers.[15] These significantly exceeded loans to brokers, typically by about 2.5% of the capitalization of the NYSE in the early 1930's. I thus assume that brokers' loans to customers in September 1929 exceeded brokers' borrowings on the call market by this sum, or about $2.2 billion, although it is recognized that the early 1930's were a different regime than the late 1920's and that this assumption is speculative.

---

[12] *Banking and Monetary Statistics*, Table 19. This important source of credit to securities markets has been left out of all previous analyses of margin borrowing during the 1920's of which I am aware. The notes to the *Banking and Monetary Statistics* make clear that these were also loans extended for the purpose of carrying securities and should thus be included. Beckhart (1932, pp. 155-160) discusses the shifting back and forth of securities credit between bank loans and broker loans.

[13] *ibid.*, Table 142

[14] See *Banking and Monetary Statistics*, p. 435.

[15] *ibid.*, Table 143

Combining the $7.2 billion in bank lending, $8.55 billion in broker loans, and $2.2 billion in margin loans from brokers funded outside the call market gives total margin loans to stock investors (excluding banks) of $18 billion in September 1929.

Compare this to March 1936, when direct bank loans were $211 million and brokers' margin loans to customers were $1.35 billion, for total margin loans of $1.6 billion.

For both of these time frames, the Fed does not break out the data by type of security, meaning that bonds as well as stocks may be included. Because the total value of bonds outstanding in the U.S. totaled at least $70 billion in 1929 (more than half the value of the stock market), an appreciable portion of margin loans could theoretically have been made against bonds. However, the literature appears to contain no discussion of bond purchases on margin in the 1920's. Indeed, with call money rates well above other fixed income benchmark rates in 1929, it would have made little sense to engage in such an investment strategy. I therefore assume all the margin lending was against stocks. Finally, not all margin loans would have been to investors that were near-maximally levered. There were 600,000 margin accounts out of a total of 1.55 million (Galbraith, 1954). If margin accounts had, on average, the same net worth as cash accounts (a doubtful assumption but a useful benchmark), then these margin accounts would have been worth a total of $42 billion, and would, on average, have been levered with about 43 cents of debt for each dollar of equity.

Equity in margin accounts probably varied widely however. Data on this is not available for the 1920's, but it is available for the 1970's and early 1980's,[16] when approximately 70% of margin debt was held by accounts with less than 60% equity in the account, within 10% of the margin requirement. I assume that in light of the sharp tightening in margin requirements from the autumn of 1928 through the summer of 1929, the percentage of constrained accounts in 1929 was probably somewhat larger than this, and therefore that 80% of margin loans were held in accounts that behaved like the speculators in my model.

[16] http://www.nyxdata.com/Data-Products/Facts-and-Figures

This leads to an estimate of margin loans to constrained investors of $14.5 billion in September 1929 and $1.3 billion in March 1936.

3.2. Leverage Ratio of Speculators. Smiley and Keehn (1988) survey evidence on margin requirements prior to the crash, and find that they were generally between 40 and 100 percent, depending on the security. I assume that net worth of constrained investors is divided equally between accounts with 40, 50, and 60 percent margin requirements. This implies an average debt-to-equity ratio of 1.06 with variance 0.12. For 1936, I assume that constrained investors faced the Fed's margin requirement of 55%, leading to a debt-to-equity ratio of 0.82, with zero variance.

3.3. Market Size. The total market capitalization of the New York Stock Exchange was $87 billion at the end of September 1929 (NYSE). For the Curb Exchange[17] I estimate the figure was $22 billion.[18] I follow McGrattan and Prescott (2001) in assuming that total market capitalization in the United States was 1.45 times the value of the NYSE, implying that the total market value of U.S. stocks at the beginning of October 1929 was about $126 billion. Together, these imply a market cap for over-the-counter and regional exchanges of $\$126 - 87 - 22 = \$17$ billion, which is not unreasonable, especially given that all the banks traded in the over-the-counter market to avoid exchange disclosure requirements (Wigmore, 1985). This same method yields a market capitalization of $52 billion for the NYSE and $75 billion for all U.S. stocks in March 1936.

4. Stability Analysis

We are now able to put the components together to complete our stability analysis. We use equation (87) from Chapter 1 to compute MinEAMASS. The results of the analysis are presented below.[19]

[17]Later the American Stock Exchange.
[18]Kuvin (1930) finds a market cap for the Curb Exchange of $17.1 billion in January 1929. Assuming the Curb grew in tandem with the NYSE over the first nine months of 1929 gives a figure of $22 billion at the end of September.
[19]Assumes elasticity equal to one for unconstrained investors.

	September 1929	April 1936
Margin Borrowings of Constrained Investors ("Speculators")	$14.5 billion	$1.3 billion
Net Worth of Constrained Investors	$13.7 billion	$1.6 billion
Average Debt-to-Equity	1.06	0.82
Variance of Leverage	0.12	0
MinEAMASS	$59 billion	$5.3 billion
Market Capitalization of U.S. Stocks	$126 billion	$75 billion
Instability Ratio	0.47	0.07

TABLE 1

The results of the stability analysis are striking. In 1929, margin borrowings were of the right order of magnitude necessary to destabilize the stock market. By 1936, they were nowhere close. In light of the simplistic behavioral assumptions in the model, quality of data available for the period, and uncertainty around demand elasticity, the results suggest that with proper portfolio data, the instability or near-instability might have been apparent. I have also not attempted to model the additional embedded leverage of investment trusts, which could have raised MinEAMASS in 1929 further.

The events of September and October of 1929 back up the model presented in this paper in qualitative ways as well. Almost since the time of the crash, commentators have noted that there was little news in the autumn of 1929 that would seem capable of sufficiently altering expectations to produce a crash of the magnitude that occurred. The two most widely cited events were regulatory disapproval for a utility stock split and the unearthing of a modest financial fraud in London. Prominent commentators like Irving Fisher that were bullish in early summer were still bullish in early fall; similarly for bearish commentators like Roger Babson. Some modern commentators, such as Rappoport and White (1994) have argued that market participants foresaw the crash months before it actually happened. Economic historians have had difficulty making convincing arguments that changes in expectations were responsible for the crash.

Similarly, while forced liquidations were a major cause of the crash, it is not clear that general credit was, in fact, tightening significantly in the period just prior

to the crash. Despite increases in the Fed's discount rate in the summer, the Fed simultaneously cut the rate at which it would buy bills outright. Interest rates on stock exchange call loans peaked in July. As the Fed urged caution with regard to loans for speculative purposes and non-bank lenders withdrew from the call loan market, banks stepped in to pick up much of the slack.[20] Banks did not experience distress during the crash, and even brokers were not distressed in large numbers. There was thus no general credit crunch during the crash of 1929. This did not come until later.

The shocks that are strong candidates for being triggers of the crash are more consistent with the present theory of instability. We have already discussed the tightening of margin requirements that caused more investors to behave approximately like the model's speculators, mechanistically leveraging to the maximum degree permissible. But there was another large shock in late 1929: a tremendous explosion in new issuance of stocks, $6.6 billion in the twelve months through September 1929, 5% of the value of the equity markets and as much as in 1920-1927 combined. More than $1 billion was issued in September alone, the largest month ever by a factor of two for common stocks and about 1% of the value of the exchange.[21]

Wigmore points out that much of this issuance was initially held by dealers and pools as they aimed to distribute the securities in an orderly manner to the public, but that these institutions began to sell more aggressively to limit their losses as prices fell during September. In the context of the theory presented; this is a supply shock to the quantity of securities that need to be absorbed by retail investors with a downward-sloping demand curve, depressing prices. When markets have significant numbers of constrained investors, as I have proposed they did in 1929, it is the actual issuance, rather than the announcement of it, that causes the price decline.

[20] *Banking and Monetary Statistics*. This has been noted by many other commentators, including Friedman and Schwartz.
[21] *Banking and Monetary Statistics*, Table 137.

Conclusion

Many causes have been proposed for the Great Crash of 1929, but while there has been a great deal of debate in the academic literature about which the relative importance of the various possibilities, there have been virtually no new theories proposed since Irving Fisher's comprehensive assessment in 1930. Here, we have argued that tightening margin requirements in the first nine months of 1929 caused levered investors to become constrained, so that a series of modest negative shocks that buffeted markets in September and October were sufficient to force liquidations and cause a severe crash.

All discussions of the crash of which I am aware give a starring role to the forced liquidations of margin loans, yet none of these have explained why, in a quantitative sense, the stock market was particularly vulnerable to such a crash at the time that it occurred. Here, I have applied the micro-founded measure of financial instability developed in the preceding chapters and shown that this instability ratio shows that the stock markets of 1929 were indeed particularly vulnerable to instability. We have also used this measure to show how the tightening of margin requirements that occurred in 1936 was much less of a threat to the stability of the stock market.

The qualitative evidence on the crash fits the model of chapters one and two as well. The timing of the crash does not appear to have been tied to changes in expectations or fundamentals, but rather to the technical conditions of the market. The contemporary stories of the time are of investors who pyramided leverage upon leverage, and doubled down when their bets paid off. During the crash itself, these investors who wished to buy stocks because their expectations were bullish were forced to sell instead. In short, while good data are difficult to come by, we have seen that the very simple model developed in this paper can explain a good deal of the sudden and uncontrolled nature of the Great Crash of 1929.

Conclusions

This paper has developed a theoretical model of bubbles and financial crises. I have then applied that model in an attempt to better understand the 1998 collapse of hedge fund Long-Term Capital Management as well as the stock market crash of 1929. In particular, we assessed the age-old proposition that excessive leverage is dangerous, and have attempted to provide a rigorous and quantitative formulation of that maxim.

Thus, we have developed a micro-founded measure of financial instability, which I term an *instability ratio*. This measure could allow investors or a macroprudential regulator to understand when the amount and distribution of leverage throughout the financial system is reaching dangerous proportions. By comparison to other measures of financial fragility that have been developed, such as Adrian and Brunnermeier (2008) or Borio and Lowe (2002), my measure stands out in being both micro-founded and stated in sufficiently general form to be operationalizable.

The same model of investor behavior and asset markets that gives rise to my Instability Ratio also produces asset bubbles and sudden crashes, and thus treats financial fragility and asset price bubbles and crashes in a unified framework. Despite the model's simple institutional and behavioral framework, it is able to generate many key stylized facts and dynamics that characterize asset price bubbles, including an accelerating price pattern fueled by credit expansion followed by an endogenous crash or "Minsky Moment" which is accompanied by a sudden evaporation of liquidity on one side of the market.

The model's emphasis on considering stylized versions of real-world institutions has allowed an investigation of policy in a framework that I hope has been intuitive. We have seen how central banks may use interest rate increases or leverage

limitations to pop asset bubbles or prevent their growth, but have also seen that such policy may be counterproductive or only temporarily effective if policy does not cause speculators to reevaluate their enthusiasm.

The model also contains insights relevant to the debate over proper prudential regulation. Most importantly, I show that financial fragility is related not to the financial system's *net* derivative exposure (which is always zero) but rather to investors' *gross* notional exposure. This suggests that position sizes should be monitored carefully. Further, I show that despite the attention that has been paid to procyclical margin policy as a cause of bubbles and crises, the endogenous boom and bust pattern of an asset bubble can occur even with constant margin requirements.

In the second half of the paper, we undertook a historical examination of the collapse of Long-Term Capital Management and the 1929 stock market using the model of chapters one and two as a framework for analysis.

While data were necessarily incomplete, I have brought together a number of sources to compile what I believe is the most complete portfolio information available for LTCM at the time of its collapse. Using these data, I have shown that LTCM alone was large and levered enough that a disorderly default would have been a cause for concern in some important markets, despite its peak net worth of less than $5 billion.

My investigation of the Great Crash of '29 has shown how margin borrowing, which had been a standard feature of stock market investing for decades, took on characteristics that exposed the stock market to a sudden break downward and drying up of liquidity. In particular, I have argued that the market suffered from a series of modest negative shocks in the autumn of 1929 that, together with a tightening of margin requirements that caused investors to become collateral-constrained, gave rise to a moment of instability that triggered the crash. I have drawn on many accounts of the crash to undertake my investigation, although incredibly, an in-depth academic investigation of the crash itself has never, to my knowledge, been written.

Throughout this paper, I have taken a modeling approach that is complementary to the existing literature. It explicitly relies on the results of others—such as Grossman and Vila (1992) and Liu and Longstaff (2004), as well as the historical record—to determine the behavior of investors, rather than deriving this behavior from first principles. This has allowed me to place my focus squarely on the macro-level implications of those behaviors and to derive what I hope are interesting dynamics in a relatively accessible framework. We employed this framework to develop a micro-founded, operationalizable measure to determine when excessive leverage becomes a systemic threat; we used it to show how collateral-constrained optimistic investors behave in a way that can cause an accelerating asset bubble that endogenously bursts; we explored how policy-makers could respond using interest rates and leverage limits; and we showed in a quantitative manner the impact that leverage employed by investors during historical episodes in 1929 and 1998 had on market stability.

APPENDIX A

12 Bubble Simulations

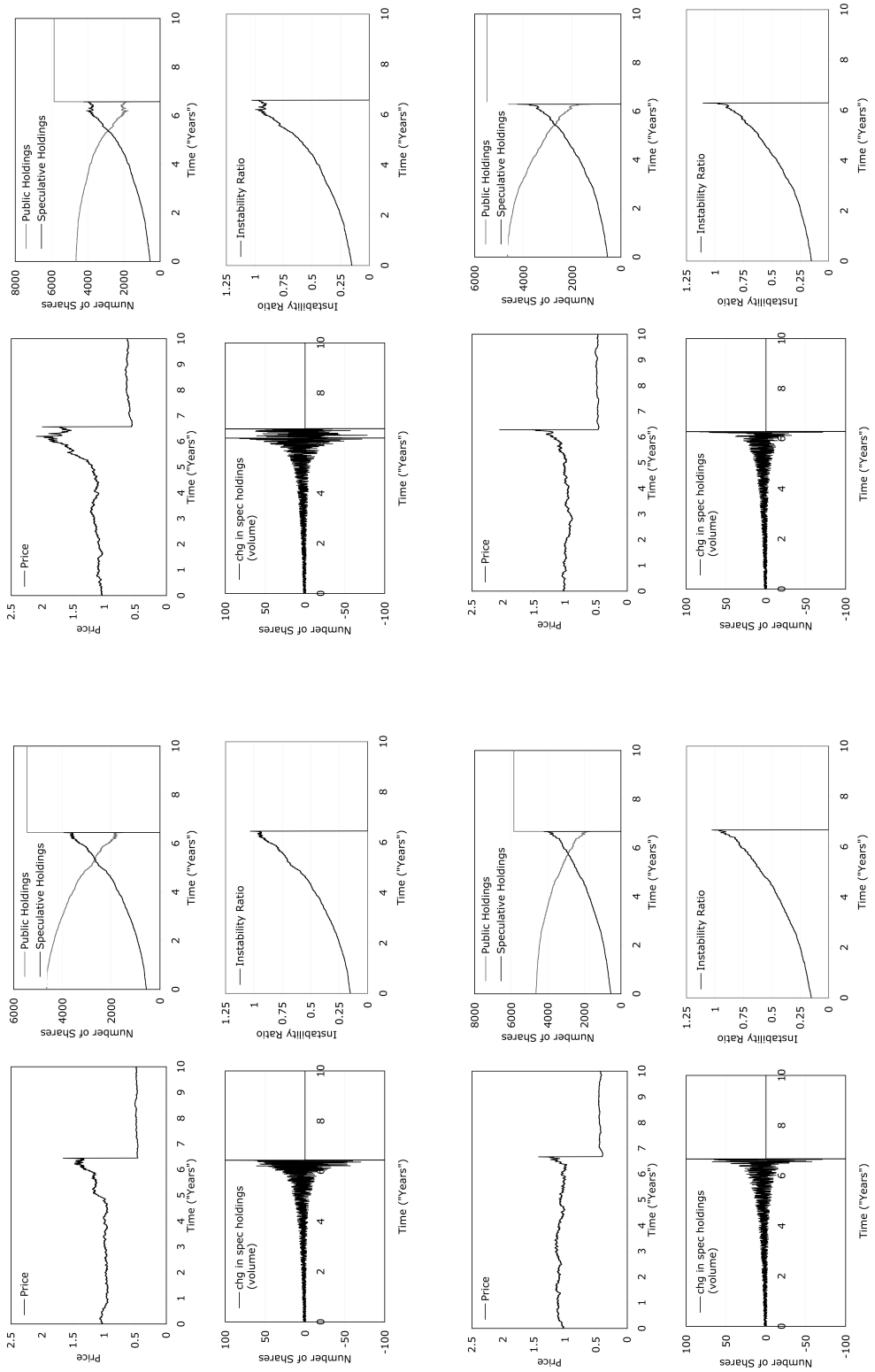

FIGURE 1. Simulations 1-4.

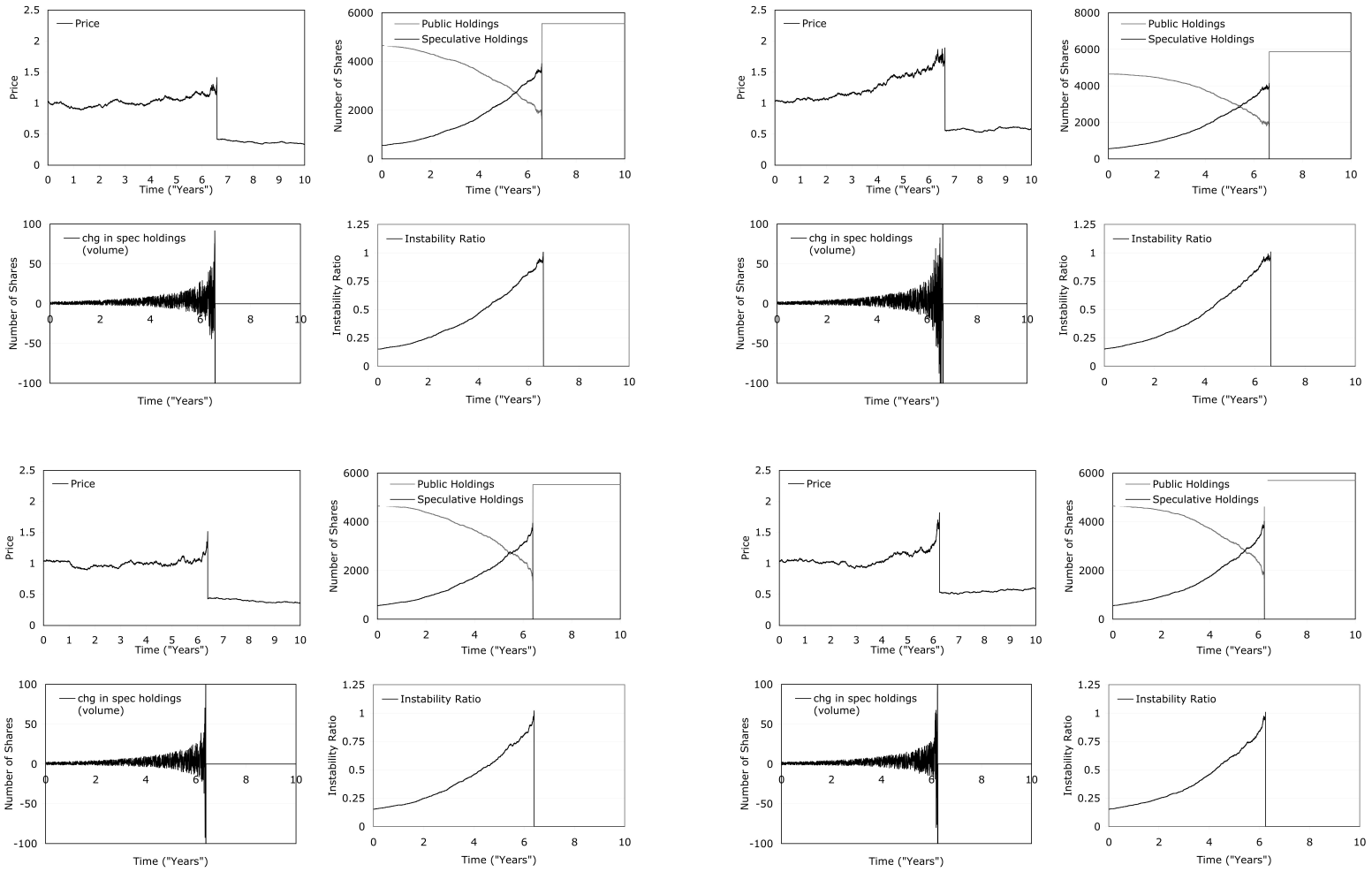

FIGURE 2. Simulations 5-8.

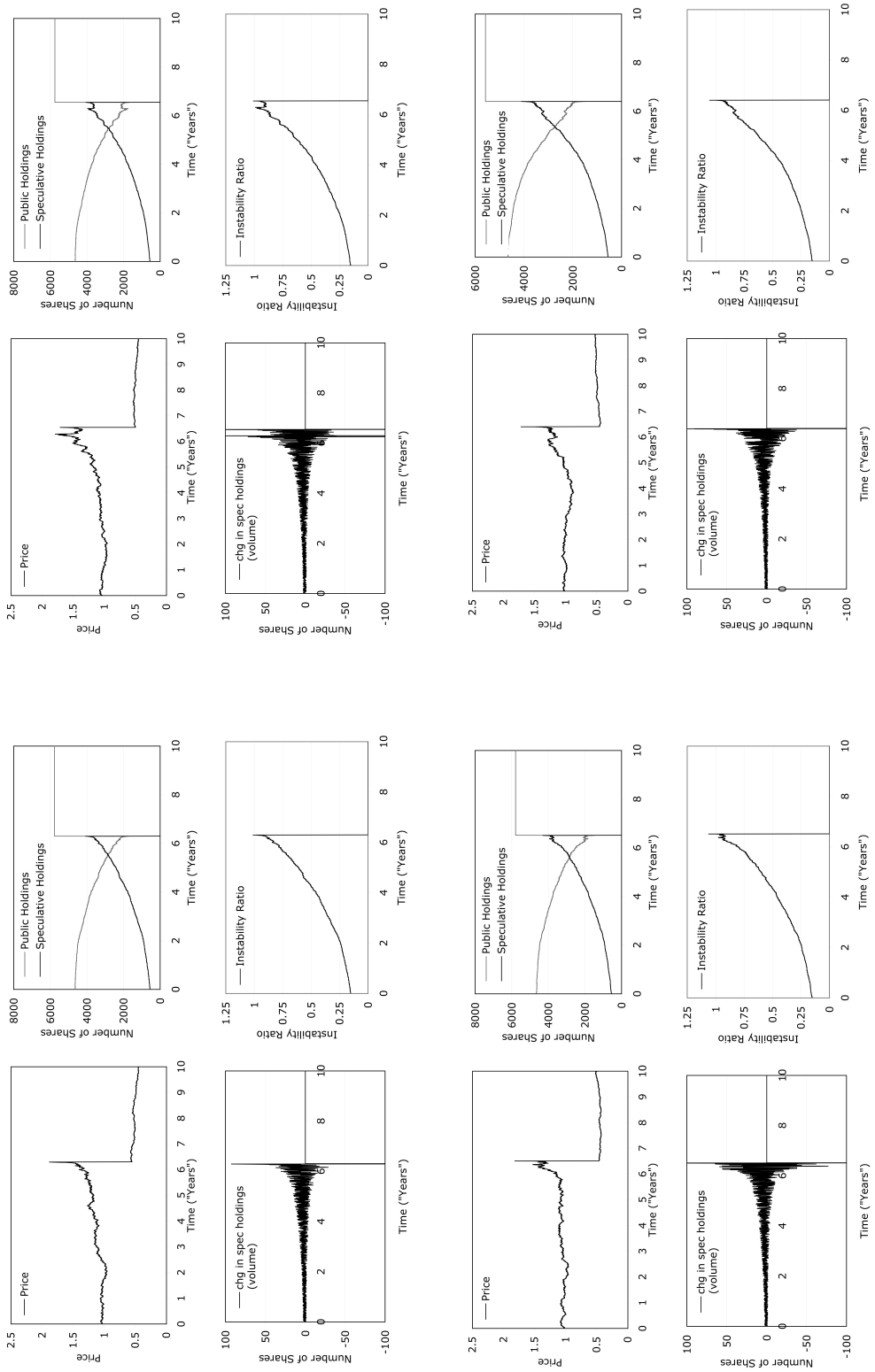

FIGURE 3. Simulations 9-12.

References

Abreu, Dilip and Markus Brunnermeier, "Bubbles and Crashes," *Econometrica*, 2003, *71* (1), 173–204.

Acemoglu, Daron, Simon Johnson, and James Robinson, "The Rise of Europe: Atlantic Trade, Institutional Change and Economic Growth," *American Economic Review*, 2005, *95* (3), 546–579.

Adrian, Tobias and Hyun Song Shin, "Liquidity and Leverage," Staff Report 328, Federal Reserve Bank of New York 2008.

_____ **and Markus K. Brunnermeier**, "CoVaR," Staff Report 348, Federal Reserve Bank of New York 2008.

Agate, Elizabeth, *Woodlands, a Practical Handbook*, Doncaster: BTCV, 2002. Available online at http://handbooks.btcv.org.uk/handbooks/content/section/3767.

Allen, F., S. Morris, and A. Postlewaite, "Finite Bubbles with Short Sale Constraints and Asymmetric Information," *Journal of Economic Theory*, 1993, *61* (2), 206 – 229.

Allen, Franklin and Douglas Gale, "Bubbles and Crises," *Economic Journal*, January 2000, *110* (460), 236–55.

_____ **and Douglas M. Gale**, "An Introduction to Financial Crises," Working Paper 07-20, Wharton Financial Institutions 2007.

Allen, Frederick Lewis, *Only Yesterday: an Informal History of the 1920's*, New York: Perennial Classics, 2000, c1931.

Allen, Robert C., "Economic structure and agricultural productivity in Europe, 1300-1800," *European Review of Economic History*, 2000, *3*, 1–25.

_____, 2001. Wage and price indexes for various European cities and countries. Available at http://www.nuff.ox.ac.uk/General/Members/homepage.aspx?nuffid=ALLERC00.

_____, "Progress and Poverty in Early Modern Europe," *Economic History Review*, 2003, *LVI* (3), 403–443.

_____, "English and Welsh Agriculture, 1300-1850: Outputs, Inputs and Income," January 2005. Working paper. Available at http://www.nuffield.ox.ac.uk/users/allen/unpublished/AllenE&W.pdf.

_____, "Engel's Pause: A Pessimist's Guide to the Industrial Revolution," *Oxford University Department of Economics Discussion Paper*, 2007, (315). Available at http://www.nuff.ox.ac.uk/users/Allen/unpublished/solow-11.pdf.

_____, *The British Industrial Revolution in Global Perspective*, Cambridge: Cambridge University Press, 2009.

Anonymous, "Mosler's Moral: Just Small Enough to Fail," *Institutional Investor*, October 1998.

Antras, Pol and Hans-Joachim Voth, "Productivity Growth during the English Industrial Revolution: A Dual Approach," *UPF Economics and Business Working Papers*, October 2000, (495). Available at http://www.econ.upf.edu/docs/papers/downloads/495.pdf.

Ashraf, Quamrul and Oded Galor, "Dynamics and Stagnation in the Malthusian Epoch: Theory and Evidence," 2008. Brown University Working Paper, available at http://www.brown.edu/Departments/Economics/Papers/2008/2008-14_paper.pdf.

Beattie, Mollie, Charles Thompson, and Lynn Levine, *Working with Your Woodland: A Landowner's Guide*, Hanover, N.H.: University Press of New England, 1993.

Beckhart, Benjamin Haggott, *The New York Money Market, Vol. III*, Columbia University Press, 1932.

Bernanke, Ben, "Asset-Price 'Bubbles' and Monetary Policy," October 15, 2002. Remarks before the New York Chapter of the National Association for Business Economics, New York, NY. Available at http://www.federalreserve.gov/BoardDocs/Speeches/2002/20021015/default.htm#f5.

_____, "Monetary Policy and the Housing Bubble," January 3, 2010. Remarks before the American Economic Association, Atlanta, GA. Available at http://www.bis.org/review/r100113a.pdf.

_____ **and Mark Gertler**, "Agency Costs, Net Worth, and Business Fluctuations," *American Economic Review*, 1989, *79* (1), 14–31.

Bierman, Harold, Jr., *The Great Myths of 1929 and the Lessons to be Learned*, New York: Greenwood Press, 1991.

Blanchard, Olivier and Stanley Fischer, *Lectures on Macroeconomics*, Cambridge, MA: MIT Press, 1990.

Blanchard, Olivier J. and Mark W. Watson, "Bubbles, Rational Expectations and Financial Markets," NBER Working Papers 0945, National Bureau of Economic Research, Inc June 1983.

Board of Governors of the Federal Reserve System, *Banking and Monetary Statistics 1914-1941* 1943. Available at http://fraser.stlouisfed.org/publications/bms/.

Borio, Claudio and Philip Lowe, "Asset Prices, Financial and Monetary Stability: Exploring the Nexus," *BIS Working Papers*, 2002, (114).

Broadberry, Stephen, Bruce Campbell, Alexander Klein, Mark Overton, and Bas van Leeuwen, "British Economic Growth, 1270-1870," July 14, 2010.

_____, _____, **and Bas van Leeuwen**, "The Sectoral Distribution of the Labour Force and Labour Productivity in Britain, 1381-1851," March 17, 2011.

Brunnermeier, Markus and L. Pedersen, "Predatory Trading," *Journal of Finance*, August 2005, *LX* (4).

Brunnermeier, Markus K., "Bubbles," in Steven N. Durlauf and Lawrence E. Blume, eds., *The New Palgrave Dictionary of Economics*, Basingstoke: Palgrave Macmillan, 2008.

_____ **and Lasse Pedersen**, "Market Liquidity and Funding Liquidity," *Review of Financial Studies*, 2009, *22* (6), 2201–2238.

Buxton, Neil K., *The Economic Development of the British Coal Industry*, London: Batsford Academic, 1978.

Chimney Sweep, Inc., "Sweep's Library: Firewood Comparison Charts." http://www.chimneysweeponline.com/howood.htm, accessed 8/31/09.

Chowdhry, Bhagwan and Vikram K. Nanda, "Leverage and Market Stability: The Role of Margin Rules and Price Limits," *Journal of Business*, April 1998, *71* (2).

Church, Roy, *The History of the British Coal Industry, 1830-1913: Victorian Pre-emininence*, Vol. 3, Oxford: Clarendon Press, 1986.

Clark, Gregory, "The Long March of History: Farm Laborers' Wages in England 1208-1850," 2001. Working paper, UC-Davis. Available at http://www.econ.ucdavis.edu/faculty/gclark/papers/long_march_of_history.pdf.

_____, "Land Rental Values and the Agrarian Economy: England and Wales, 1500-1914," *European Review of Economic History*, 2002, *6*, 281–308.

_____, "The Price History of English Agriculture, 1209-1914," October 2003. Working paper, UC Davis. Available at http://www.econ.ucdavis.edu/faculty/gclark/papers/Agprice.pdf.

_____, "The Condition of the Working Class in England, 1209-2004," *Journal of Political Economy*, 2005, *113* (6), 1307–1340.

_____ **and David Jacks**, "Coal and the Industrial Revolution, 1700-1869," *European Review of Economic History*, March 2007, *11* (1), 39–72.

_____ **and Gillian Hamilton**, "Survival of the Richest: The Malthusian Mechanism in Pre-Industrial England," *Journal of Economic History*, September 2006, *66* (3), 707–736.

Committee on Banking and Currency, *Stock Exchange Practices*, United States Senate, 72nd Congress, 1934. Available at http://fraser.stlouisfed.org/publications/sensep/.

Committee on the Global Financial System, *A Review of Financial Market Events in Autumn 1999*, Basel: Bank for International Settlements, 1999.

Coy, Peter and Suzanne Woolley, "Failed Wizards of Wall Street," *Business Week*, September 21, 1998. Available at http://www.businessweek.com/1998/38/b3596001.htm.

Crafts, N. F. R., *British Economic Growth During the Industrial Revolution*, Oxford: Clarendon Press, 1985.

_____, "The New Economic History and the Industrial Revolution," in Peter Mathias and John A. Davis, eds., *The First Industrial Revolutions*, Oxford: Basil Blackwell, 1989, pp. 25–43.

_____, "Exogenous or Endogenous Growth? The Industrial Revolution Reconsidered," *The Journal of Economic History*, December 1995, *55* (4), 745–772.

_____, "Solow and Growth Accounting: a Perspective from Quantitative Economic History," March 2008. Working paper, University of Warwick. Available at http://www2.warwick.ac.uk/fac/soc/economics/staff/faculty/crafts_/working_papers/solowga.pdf.

_____ **and C. Knick Harley**, "Output Growth and the British Industrial Revolution: a Restatement of the Crafts-Harley View," *Economic History Review*, 1992, *XLV* (4), 703–730.

_____ **and** _____, "Precocious British Industrialization: A General Equilibrium Perspective," in Leandro Prados de la Escosura, ed., *Exceptionalism and Industrialisation*, Cambridge: Cambridge University Press, 2002, pp. 86–110. Available at http://www.lse.ac.uk/collections/economicHistory/pdf/wp6702.pdf.

_____ **and Terence C. Mills**, "From Malthus to Solow: How Did the Malthusian Economy Really Evolve?," January 2007. Working paper, University of Warwick. Available at http://www2.warwick.ac.uk/fac/soc/economics/staff/faculty/crafts_/working_papers/from_malthus_to_solow.pdf.

Crockford, K. J. and P. S. Savill, "Preliminary Yield Tables for Oak Coppice," *Forestry*, 1991, *64* (1), 29–50.

Davis, Ralph, "English Foreign Trade, 1700-1774," *The Economic History Review*, 1962, *15* (2), 185–303.

_____, *The Industrial Revolution and British Overseas Trade*, Atlantic Highlands, N.J.: Humanities Press, 1979.

De Long, J. Bradford and Andrei Shleifer, "The Stock Market Bubble of 1929: Evidence from Closed-end Mutual Funds," *The Journal of Economic History*, 1991, *51* (3), 675–700.

_____, _____, **Lawrence H. Summers, and Robert J. Waldmann**, "Noise Trader Risk in Financial Markets," *Journal of Political Economy*, August 1990, *98* (4), 703–38.

de Vries, Jan and Ad Van der Woude, *The First Modern Economy*, Cambridge University Press, 1997.

Deane, Phyllis and W. A. Cole, *British Economic Growth 1688-1959*, Cambridge University Press, 1967.

DeWald, Scott, Scott Josiah, and Becky Erdkamp, "Heating with Wood: Producing, Harvesting and Processing Firewood," *NebGuide*, 2005. Available at www.ianrpubs.unl.edu/epublic/live/g1554/build/g1554.pdf.

Diamond, Douglas W. and Philip H. Dybvig, "Bank Runs, Deposit Insurance and Liquidity," *Journal of Political Economy*, 1983, *91* (3).

_____ **and Robert E. Verrecchia**, "Information Aggregation in a Noisy Rational Expectations Economy," *Journal of Financial Economics*, 1980, (9), 221–235.

Dunbar, Nicholas, *Inventing Money: The Story of Long-Term Capital Management and the Legends Behind It*, Chichester: Wiley, 2000.

Farmer, J. Doyne and John Geanakoplos, "The Virtues and Vices and Equilibrium and the Future of Financial Economics," *Complexity*, Jan/Feb 2009, *14* (3), 11–38.

Feinstein, Charles H., "National Statistics," in Charles H. Feinstein and Sidney Pollard, eds., *Studies in Capital Formation in the United Kingdom, 1750-1920*, Oxford: Clarendon Press, 1988, pp. 258–471.

Fisher, Irving, *The Stock Market Crash and After*, New York: The MacMillan Company, 1930.

Flinn, Michael W., *The History of the British Coal Industry, 1700-1830: The Industrial Revolution*, Vol. 2, Oxford: Clarendon Press, 1984.

Fostel, Ana and John Geanakoplos, "Leverage Cycles and the Anxious Economy," *American Economic Review*, 2008, *98* (4), 1211–1244.

Friedman, Milton and Anna Jacobson Schwartz, *A Monetary History of the United States, 1867-1960*, Princeton University Press, 1963.

Galbraith, John Kenneth, *The Great Crash: 1929*, New York: Houghton Mifflin, 1997, c1954.

Galor, Oded and David N. Weil, "Population, Technology, and Growth: From Malthusian Stagnation to the Demographic Transition and Beyond," *The American Economic Review*, September 2000, *90* (4), 806–828.

Geanakoplos, John, "Liquidity, Default, and Crashes," *Cowles Foundation for Research in Economics Working Papers*, 2003, (1074).

Gennotte, Gerard and Hayne Leland, "Market Liquidity, Hedging and Crashes," *American Economic Review*, 1990, *80* (5), 999–1021.

Glover, Thomas J., *Pocket Ref*, Littleton, Colorado: Sequoia Publishing, Inc., 2003.

Greenspan, Alan, "Testimony before the House of Representatives Committee on Financial Services," October 1, 1998. Available at http://financialservices.house.gov/banking/10198gre.shtml.

Grossman, Sanford J. and Jean-Luc Vila, "Dynamic Trading with Leverage Constraints," *Journal of Financial and Quantitative Analysis*, June 1992, *27* (2), 151–168.

Hamilton, Rick, "Forest Soils and Site Index," *Woodland Owner Notes*, 1995, *7*. Available at http://www.ces.ncsu.edu/forestry/pdf/WON/won07.pdf.

Hammersley, G., "The Charcoal Iron Industry and Its Fuel, 1540-1750," *The Economic History Review*, 1973, *26* (4), 593–613.

Hansen, Gary D. and Edward C. Prescott, "Malthus to Solow," *American Economic Review*, 2002, *92*, 1205–1217.

Harley, C. Knick, "Foreign Trade: Comparative Advantage and Performance," in Roderick Floud and Donald McCloskey, eds., *The Economic History of Britain Since 1700*, Vol. 1, Cambridge University Press, 1994, pp. 300–331.

_____, "Reassessing the Industrial Revolution: A Macro View," in Joel Mokyr, ed., *The British Industrial Revolution: An Economic Perspective*, Boulder, Colorado: Westview Press, 1999, pp. 160–205.

_____ **and N. F. R. Crafts**, "Simulating the Two Views of the British Industrial Revolution," *The Journal of Economic History*, September 2000, *60* (3), 819–841.

Hatcher, John, *Plague, Population and the English Economy, 1348-1530*, Basingstoke: Macmillan, 1977.

_____, *The History of the British Coal Industry, Before 1700: Towards the Age of Coal*, Vol. 1, Oxford: Clarendon Press, 1993.

Hull, John C., *Options, Futures and Derivatives*, Upper Saddle River, NJ: Prentice Hall, 2006.

Iso, Takatoshi and Tokuo Iwaisako, "Explaining Asset Bubbles in Japan," *NBER Working Papers*, 1995, (5358).

Kamada, Koichiro and Kentaro Nasu, "How Can Leverage Regulations Work for the Stabilization of Financial Systems," Working Paper 10-E-2, Bank of Japan March 2010.

Kindleberger, Charles P., *The World in Depression, 1929-1939*, London: Allen Lane, 1973.

_____, *Manias, Panics and Crashes: A History of Financial Crises*, New York: John Wiley, 1996.

Kiyotaki, Nobuhiro and John Moore, "Credit Cycles," *Journal of Political Economy*, April 1997, *105* (2), 211–248.

Koch, Christof, *Biophysics of Computation*, Oxford University Press, 1999.

Koo, Richard C., *The Holy Grail of Macroeconomics: Lessons from Japan's Great Recession*, Singapore: John Wiley & Sons, 2008.

Kremer, Michael, "Population Growth and Technological Change: One Million B.C. to 1990," *The Quarterly Journal of Economics*, August 1993, *108* (3), 681–716.

Krugman, Paul, "Balance Sheets, the Transfer Problem and Financial Crises," *International Tax and Public Finance*, 1999, *6*, 459–472.

Kuvin, Leonard, "Stock Price Indexes of the New York Curb Exchange," *Journal of the American Statistical Association*, 1930, *25* (169), pp. 51–62.

Lee, Ronald D., "Population in Pre-Industrial England: An Econometric Analysis," *Quarterly Journal of Economics*, 1973, *87* (4), 581–607.

———, "Accidental and Systematic Change in Population History: Homeostasis in a Stochastic Setting," *Explorations in Economic History*, 1993, *30*, 1–30.

——— **and Michael Anderson**, "Malthus in State Space," *Journal of Population Economics*, 2002, *15*, 195–220.

Lindert, Peter H., "English Population, Wages and Prices: 1541-1913," *Journal of Interdisciplinary History*, 1985, *115* (4), 609–634.

——— **and Jeffrey G. Williamson**, "Revising England's Social Tables 1688-1812," *Explorations in Economic History*, 1982, *19*, 385–408.

Liu, Jun and Francis A. Longstaff, "Losing Money on Arbitrage: Optimal Dynamic Portfolio Choice in Markets with Arbitrage Opportunities," *Review of Financial Studies*, Autumn 2004, *17* (3), 611–641.

Lowenstein, Roger, *When Genius Failed: The Rise and Fall of Long Term Capital Management*, New York: Random House, 2000.

Lucas, Robert E., "The Industrial Revolution: Past and Future," *Federal Reserve Bank of Minneapolis 2003 Annual Report Essay*, May 2004. Available at http://www.minneapolisfed.org/publications_papers/pub_display.cfm?id=3333.

MacKenzie, Donald, "Long-Term Capital Management and the Sociology of Arbitrage," *Economy and Society*, August 2003, *32* (3), 349–380.

Maddison, Angus, *The World Economy: A Millennial Perspective*, OECD, 2001.

Mahoney, Paul G., "The Stock Pools and the Securities Exchange Act," *Journal of Financial Economics*, 1999.

Mathias, Peter, "The Industrial Revolution: Concept and Reality," in Peter Mathias and John A. Davis, eds., *The First Industrial Revolutions*, Oxford: Basil Blackwell, 1989, pp. 1–24.

McDonough, William, "Testimony before the House of Representatives Committee on Financial Services," October 1, 1998. Available at http://financialservices.house.gov/banking/10198don.shtml.

McGrattan, Ellen R. and Edward C. Prescott, "The 1929 Stock Market: Irving Fisher Was Right," *International Economic Review*, 2004, *45*, 991–1009.

Mendels, Franklin F., "Proto-Industrialization: The First Phase of the Industrialization Process," *The Journal of Economic History*, 1972, *32* (1), 241–261.

Minsky, Hyman P., *Stabilizing an Unstable Economy*, Yale University Press, 1986.

Mishkin, Frederic S., "The Transmission Mechanism and the Role of Asset Prices in Monetary Policy," *NBER Working Papers*, 2001, (8617).

Perrold, Andre, "Long-Term Capital Management L.P.," *Harvard Business School Case Studies 200-007/8/9/10*, 1999.

Phelps Brown, Henry and Sheila V. Hopkins, *A Perspective of Wages and Prices*, London: Meuthen, 1981.

Pomeranz, Kenneth, *The Great Divergence: China, Europe, and the Making of the Modern World Economy*, Princeton University Press, 2000.

President's Working Group on Financial Markets, *Hedge Funds, Leverage and the Lessons of Long-Term Capital Management*, U.S. Treasury, 1999.

Rappoport, Peter and Eugene N. White, "Was the Crash of 1929 Expected?," *American Economic Review*, 1994, *84* (1), pp. 271–281.

Romer, Christina D, "The Great Crash and the Onset of the Great Depression," *The Quarterly Journal of Economics*, August 1990, *105* (3), 597–624.

Rosenfeld, Eric, "Long-Term Capital Management: 10 Years Later," February 19, 2009. Presentation at Sloane School of Business. Available at http://techtv.mit.edu/videos/2450-eric-rosenfeld-15437-presentation-21909.

Rostow, Walter W., *The Stages of Economic Growth*, Cambridge University Press, 1960. Third edition published 1990.

Sachs, Jeffrey et. al., "Ending Africa's Poverty Trap," *Brookings Papers on Economic Activity*, 2004, *1*, 117–240.

Sato, Ryuzo and Yoshio Niho, "Population Growth and the Development of a Dual Economy," *Oxford Economic Papers*, 1971, *23* (3), 418–436.

Scheinkman, Jose A. and Wei Xiong, "Overconfidence and Speculative Bubbles," *Journal of Political Economy*, 2003, *111* (6), 1183–1220.

Scholes, Myron, "The Near-Crash of 1998," *AEA Papers and Proceedings*, 2000, *90* (2), 17–21.

Shleifer, Andrei and Robert W. Vishny, "The Limits of Arbitrage," *Journal of Finance*, March 1997, *LII* (1), 35–55.

Smiley, Gene and Richard H. Keehn, "Margin Purchases, Brokers' Loans and the Bull Market of the Twenties," *Business and Economic History*, 1988, *17*.

Smith, Adam, *The Wealth of Nations*, London: Penguin Books, 1999. Originally published 1776.

Sobel, Robert, *The Big Board: A History of the New York Stock Market*, New York: Free Press, 1965.

Solow, Robert M., "A Contribution to the Theory of Economic Growth," *The Quarterly Journal of Economics*, February 1956, *70* (1), 65–94.

Taylor, John B., "The Financial Crisis and Policy Responses: An Empirical Analysis of What Went Wrong," 2008. Available at www.stanford.edu/ johntayl/FCPR.pdf.

Temin, Peter, *Lessons from the Great Depression*, Cambridge, MA: MIT Press, 1989.

Thurner, Stefan, J. Doyne Farmer, and John Geanakoplos, "Leverage Causes Fat Tails and Clustered Volatility," Quantitative Finance Papers 0908.1555, arXiv.org August 2009.

U.S. Department of Energy, "Bioenergy Conversion Factors." http://bioenergy.ornl.gov/papers/misc/energy_conv.html, accessed 8/31/09.

Van Zanden, Jan Luiten, "Economic Growth in the Golden Age," *Economic and Social History in the Netherlands*, 1992, *4*, 5–25. Available at http://docserver.repository.knaw.nl/18569_291_zanden.pdf.

_____, "The Dutch Economy in the Very Long Run—Growth in Production, Energy Consumption and Capital in Holland (1500-1805) and the Netherlands (1805-1910)," in A. Szirmai, B. Van Ark, and D. Pilat, eds., *Explaining Economic Growth: Essays in Honour of Angus Maddison*, Amsterdam: Elsevier Science Publishers B.V., 1993, pp. 267–283.

_____, "Simulating Early Modern Economic Growth," 2004. IISH working paper, available at http://www.iisg.nl/research/jvz-simulating.pdf.

_____, "Cobb-Douglas in Pre-Modern Europe," May 2005. IISH working paper, available at www.iisg.nl/research/jvz-cobbdouglas.pdf.

Walker, Roger, "Mass, Weight, Density or Specific Gravity of Wood." http://www.simetric.co.uk/si_wood.htm, accessed 8/31/09.

Warde, Paul, *Energy Consumption in England and Wales, 1560-2000*, Naples: Consiglio Nazionale delle Richerche, 2007.

Weir, David R., "Life Under Pressure: France and England, 1670-1870," *Journal of Economic History*, 1984, *44* (1), 27–47.

Weisdorf, Jacob L., "From Domestic Manufacture to Industrial Revolution: Long-Run Growth and Agricultural Development," *Oxford Economic Papers*, 2006, *58* (2), 264–287.

White, Eugene N., "The Stock Market Boom and Crash of 1929 Revisited," *Journal of Economic Perspectives*, April 1990.

Wigmore, Barrie A., *The Crash and its Aftermath: A History of Securities Markets in the United States, 1929-1933*, Westport, CT: Greenwood Press, 1985.

Wrigley, E. A., *People, Cities and Wealth*, Oxford: Basil Blackwell, 1987.

———, "The Transition to an Advanced Organic Economy," *Economic History Review*, 2006, *LIX* (3), 435–480.

——— **and R. S. Schofield**, *The Population History of England 1541-1871*, Cambridge: Cambridge University Press, 1981.

Xiong, Wei, "Convergence trading with wealth effects: an amplification mechanism in financial markets," *Journal of Financial Economics*, November 2001, *62* (2), 247–292.

Yuan, Kathy, "Asymmetric Price Movements and Borrowing Constraints: A Rational Expectations Equilibrium Model of Crises, Contagion, and Confusion," *Journal of Finance*, 2005, *LX* (1), 379–411.