CONTENTS

6

Herding in Finance, Stock Market Crashes, Frenzies, and Bank Runs

The last chapter illustrated herding and informational cascades in a general context. This chapter shows that herding can also arise in financial markets and describes how herding behavior can be used to explain interesting empirical observations in finance. For example, herding can result in stock market crashes and frenzies in auctions. The stock market might still be rising prior to a crash if bad news is hidden and not reflected in the price. A triggering event can reveal this hidden news and lead to a stock market crash. Crashes and frenzies in auctions are described in greater detail in Section 6.1.3.

Another example is the use of investigative herding models to show that traders have a strong incentive to gather the same short-run information. Trading based only on short-run information guarantees that the information is reflected in the price early enough before traders unwind their acquired positions. Section 6.2 illustrates the different reasons why traders might want to unwind their positions early and highlights the limits of arbitrage. It also throws new light on Keynes' comparison of the stock market with a beauty contest.

This short-run focus of investors not only affects the stock price but can also potentially affect corporate decision making. In Section 6.3 we cover two models which show that if investors focus on the short-run, and if corporate managers care about the stock market value, then corporate decision making also becomes short-sighted.

Finally, bank run models are closely linked to herding models. Seminal bank run papers are presented in Section 6.4. While the early papers did not appeal to herding models directly, this connection is explicitly drawn in the more recent research on bank runs. Insights from the bank run literature can also help us get a better understanding of international financial crises. For example, the financial crisis in Southeast Asia in the late 1990s is often viewed as a big bank run.

6.1. Stock Market Crashes

A stock market crash is a significant drop in asset prices. A crash often occurs even when there is no major news event. After each stock market crash, the popular literature has rushed to find a culprit. The introduction of stop loss orders combined with margin calls and forced sales caused by the decline in value of assets that served as collateral were considered to be possible causes for the crash of 1929. Early writings after the stock market crash of 1987 attributed the crash exclusively to dynamic portfolio insurance trading. A dynamic portfolio trading strategy, also called program trading, allows investors to replicate the payoff of derivatives. This strategy was often used to synthesize a call option payoff structure which provides an insurance against downward movements of the stock price. In order to dynamically replicate a call option payoff, one has to buy stocks when the price increases and sell shares when the price declines. Stop loss orders, sales triggered by the fall of value of collateral, and dynamic trading strategies were obvious candidates to blame for the 1929 and 1987 crashes, respectively, since they did not obey the law of demand and were thus believed to destabilize the market. Day traders who trade over the internet are the most likely candidates to be blamed for the next stock market crash.

Pointing fingers is easy, but more explicit theoretical models are required to fully understand the mechanism via which a stock market crash occurs. A good understanding of these mechanisms may provide some indication of how crashes can be avoided in the future. The challenge is to explain sharp price drops triggered by relatively unimportant news events. Theoretical models which explain crashes can be grouped into four categories:

(1) liquidity shortage models;
(2) multiple equilibria and sunspot models;
(3) bursting bubble models; and
(4) lumpy information aggregation models.

Each of these class of models can explain crashes even when all agents act rationally. However, they differ in their prediction of the price path after the stock market crash. Depending on the model, the crash can be a correction and the stock market can remain low for a substantial amount of time or it can immediately bounce back.

The first class of models argues that the decline in prices can be due to a temporary *liquidity shortage*. The market dries up when nobody is willing to buy stocks at a certain point in time. This can be due to unexpected selling pressure by program traders. These sales might be mistakenly interpreted as sales driven by bad news. This leads to a large price decline. In this setting, asymmetric information about the trading motive is crucial for generating a stock market crash. The model by Grossman (1988) described in the next section illustrates the informational difference between traded securities and dynamic trading strategies that replicate the payoff of derivatives. Crashes which are purely driven by liquidity shortage are of a temporary nature. In other words, if the price drop was caused by liquidity problems, one would expect a fast recovery of the stock market.

The second class of models shows that large price drops that cannot be attributed to significant news events related to the fundamental value of an asset may be triggered by *sunspots*. A sunspot is an extrinsic event, that is, a public announcement which contains no information about the underlying economy. Nevertheless, sunspots can affect the economic outcome since agents use them as a coordination device and, thus, they influence agents' beliefs. The economy might have multiple equilibria and the appearance of a sunspot might indicate a shift from the high asset price equilibrium to an equilibrium with lower prices. This leads to a large change in the fundamental value of the asset. This area of research was discussed earlier in Section 2.3 and will only be partly touched upon in this section. Note that all movement between multiple equilibria need not be associated with sunspots. Gennotte and Leland (1990) provide an example of a crash that arises even in the absence of sunspots. In their model there are multiple equilibria for a range of parameter values. The price drop in Gennotte and Leland (1990) is not caused by a sunspot. As the parameter values change slightly, the high-price equilibrium vanishes and the economy jumps discontinuously to the low-price equilibrium. This model will be described in detail in the next section.

The third class of models attributes crashes to *bursting bubbles*. In contrast to models with multiple equilibria or sunspot models, a crash which is caused by a bursting bubble may occur even when the fundamental value of the asset does not change. In this setting, there is an excessive asset price increase prior to the crash. The asset price exceeds its fundamental value and this is mutually known by all market participants, yet it is not common knowledge among them. Each trader thinks that the other traders do not know that the asset is overpriced. Therefore, each trader believes that he can sell the risky asset at a higher – even more unrealistic – price to somebody else. At one point the bubble has to burst and the prices plummet. A crash due to a bursting bubble is a correction and one would not expect prices to rebound after the crash. Although bursting bubbles provide a very plausible explanation for crashes, bubbles are hard to explain in theoretical models without introducing asymmetric information or boundedly rational behavior. The possibility of bubbles under asymmetric information is the focus of Section 2.3 of this survey and is therefore not discussed again in this section.

A sharp price drop in theoretical models can also occur even when no bubble exists. That is, it is not mutual knowledge that the asset price is too high. Often traders do not know that the asset is overpriced, but an additional price observation combined with the knowledge of the past price path makes them suddenly aware of the mispricing. Models involving this *lumpy information aggregation* are closely related to herding models. The economy might be in a partial informational cascade until the cascade is shattered by a small event. This event triggers an information revelation combined with a significant price drop. Section 6.1.2 illustrates the close link between herding models with exogenous sequencing and sequential trading models. Frenzies in descending multi-unit Dutch auctions – as covered in Section 6.1.3 – are closely related to herding outcomes in models with endogenous sequencing. The difference between these trading models and pure herding models is that herding is not only due to informational externalities. In most settings, the predecessor's action causes both an informational externality as well as a payoff externality. A stock market crash caused by lumpy informational aggregation is often preceded by a steady increase in prices. The crash itself corrects this mispricing and, hence, one does not expect a fast recovery of the stock market.

The formal analysis of crashes that follows can be conducted using different model setups. We first look at competitive REE models before we examine sequential trade models. We illustrate how temporary liquidity shortage, dynamic portfolio insurance, and lumpy information revelation by prices can explain crashes. The discussion of these models sheds light on the important role of asymmetric information in understanding stock market crashes.

6.1.1. Crashes in Competitive REE Models

In a competitive REE model, many traders simultaneously submit orders. They take prices as given and can trade any quantity of shares in each trading round. In this setting, crashes can occur because of temporary liquidity shortage, multiple equilibria due to portfolio insurance trading, and sudden information revelation by prices. We begin by looking at Grossman's (1988) model where program trading can lead to temporary liquidity shortage.

Temporary Liquidity Shortage and Portfolio Insurance Trading

Grossman (1988) was written before the stock market crash in October 1987. In his model poor information about hedging demand leads to a large price decline. The original focus of the paper was to highlight the informational difference between traded options and synthesized options. Its main conclusion is that derivative securities are not redundant, even when their payoffs can be replicated with dynamic trading strategies. This is because the price of a traded derivative reveals information, whereas a synthesized option does not.[1] In a world where investors have asymmetric information about the volatility of the underlying stock price, the price of a traded option provides valuable information about the underlying asset's future volatility. The equilibrium price path and the volatility of a risky asset are driven by news announcements about its liquidation value as well as by investors' risk aversion.

In Grossman (1988) there are three periods, $t = 1, 2, 3$. There are public announcements about the value of the stock in period $t = 2$ and in $t = 3$. After the second announcement in $t = 3$, every investor knows the final liquidation value of the stock. Each public announcement can be either good or bad, that is $S_t^{\text{public}} \in \{\text{g}, \text{b}\}$, where $t = 1, 2$. Consequently, the price in $t = 3$ can take on one of four values: $P_{3\text{bb}}$, if both signals are bad; $P_{3\text{bg}}$, if the public announcement in $t = 2$ is good but the one in $t = 3$ is bad, $P_{3\text{gb}}$, or $P_{3\text{gg}}$. The price in $t = 2$, $P_{2\text{g}}$ or $P_{2\text{b}}$, depends on the investors' risk aversion. In this model, there is a fraction f of investors whose risk aversion increases significantly as their wealth declines. These investors are only willing to hold a risky asset as long as their wealth does not fall below a certain threshold. As the price of the stock declines due to a bad news announcement in $t = 2$, and with it the value of their portfolio, investors become much more risk averse and less willing to hold risky stocks. They would only be willing to hold the stock in their portfolio if the expected rate of return, $(P_3 - P_2)/P_2$, is much higher. This can only be achieved if the price in $t = 2$ drops drastically. Given their risk aversion, these traders want to insure themselves against this price decline in advance. Thus, they would like to

[1] Section 2.2.2 discusses the informational difference between traded securities and trading strategies at a more abstract level.

hold a position which exhibits a call option feature. To achieve this they can either buy additional put options in $t = 1$ or alternatively they can employ a dynamic hedging strategy which replicates the call option payoff structure. This dynamic trading strategy requires the investor to sell stocks when the price is falling in $t = 2$ and buy stocks when it is rising. These sales lead to an even larger price decline. The larger the fraction f of investors with decreasing risk aversion, the larger the number of traders who either follow this dynamic trading strategy or buy a put option. Thus, the volatility of the stock price in $t = 2$ increases as f increases.

To counteract this large price decline, there are also less risk averse market timers who are willing to bear part of this risk and provide liquidity at a much lower expected rate of return. These market timers can only provide liquidity to the extent that they have not committed their funds in other investment projects in $t = 1$. Market timers have to decide in $t = 1$ how much capital M to set aside to profitably smooth out temporary price movements. The amount of capital M that market timers put aside in $t = 1$ depends on their expectations about market volatility, that is, on the expected fraction f of risk averse investors who might insure themselves with dynamic hedging strategies or by buying put options.

Grossman (1988) compares three scenarios:

1. If the extent of adoption of dynamic hedging strategies f is known to everybody in $t = 1$, then market timers reserve funds in $t = 1$ for market interventions in $t = 2$. They will do so as long as this intervention is more profitable than using these funds for other purposes. Their activity stabilizes the market and reduces the price volatility in $t = 2$.

2. If the extent of dynamic hedging strategies f is not known in $t = 1$, but put options are traded in $t = 1$, the price of the put option reveals the expected volatility in $t = 2$. The price of the put option in $t = 1$ might even fully reveal f. It provides the market timers with valuable information about how much money M to put aside. Market timers stabilize the market as in the case where f is directly observable. Note that it is only required that a liquid option market exists which reveals information about the volatility of the underlying stock. Intermediaries who write put options can hedge their position with dynamic trading strategies.

3. If the extent of hedging strategies f in the market is not known and not revealed by an option price, the market timers face uncertainty about the profitability of their price smoothing activity in $t = 2$. If they underestimate the degree of dynamic hedging activity, they do not

have enough funds in $t = 2$ to exploit the high price volatility. This makes the prices much more volatile and might explain stock market crashes. After a slightly negative news announcement in $t = 2$, the price drops dramatically since all dynamic hedgers become much more risk averse and sell their stocks. Market timers also do not have enough funds in reserve to exploit this cheap buying opportunity. The market only bounces back later when the market timers can free up money from other investment projects and provide liquidity. In Grossman (1988) the market price bounces back in $t = 3$ as all uncertainty is resolved in that period.

Note that as long as the put option price reveals f, the put option payoff can be replicated with dynamic trading strategies. However, if all traders switch to dynamic hedging strategies, the option market breaks down and thus f is not revealed to the traders. In this case the volatility of the underlying stock is not known. This makes an exact replication of the option payoff impossible.

Large price movements in Grossman (1988) are due to a lack of liquidity provision by market timers, who underestimate the extent of sales due to portfolio insurance trading. In this model, traders do not try to infer any information about the value of the underlying stock from its price. It is arguable whether dynamic hedging demand alone can trigger a price drop of over 20 percent as experienced in October 1987. Portfolio insurance trading covered only \$60–90 million in assets, which represents only 2–3 percent of the outstanding equity market in the US. Although sales by portfolio insurers were considerable, they did not exceed more than 15 percent of total trading volume. Contrary to the experience of recent shocks, Grossman's model also predicts that the price would rebound immediately after the temporary liquidity shortage is overcome. Therefore, this model might better capture the "almost crash" caused by the Long Term Capital Management (LTCM) crisis during the fall of 1998 than the more long-lived crash of 1987.

Multiple Equilibria in a Static REE

While the stock market crashes in Grossman (1988) because market timers who have not put enough money aside cannot submit orders after a price drop due to sales by program trades, in Gennotte and Leland (1990) the market crashes because some other market participants incorrectly interpret this price drop as a bad signal about the fundamental value of the stock. In the latter model, traders hold asymmetric information about the value of the stock and, thus, the price of the underlying stock is also a signal about its fundamental liquidation value. Consequently, even these other market participants start selling their shares. Combining asymmetric information about the fundamental value of the stock with uncertainty about the extent of dynamic hedging strategies can lead to a larger decline in price in $t = 2$. The reason is that the traders wrongly attribute the price drop to a low fundamental value rather than to liquidity shortage. They might think that many other traders are selling because they received bad information about the fundamental value of the stock, while actually many sell orders are triggered by portfolio insurance trading.

Gennotte and Leland (1990) employ a static model even though stock market crashes or price changes occur over time. As the parameters change over time, the price equilibrium changes. The repetition of a static model can often be considered as a sufficient representation of a dynamic setting. Thus, comparative static results with respect to some parameters in a static model can be viewed as dynamic changes over time. A stock market crash – defined as a large price movement triggered by a small news announcement – occurs if a small change in the underlying information parameter causes a discontinuous drop in the equilibrium price.

The authors model this discontinuity in a static REE limit order model à la Hellwig (1980) with two different kinds of informed traders:[2]

(1) (value-)informed traders, who each receives an idiosyncratic individual signal $S^i = v + \epsilon^i$ about the liquidation value $v \sim \mathcal{N}(\mu_v, \sigma_v^2)$;
(2) (supply-)informed traders, who know better whether the limit order book is due to informed trading or uninformed noise trading.

Supply-informed traders can infer more information from the equilibrium price P_1. The aggregate supply in the limit order book is given by the normally distributed random variable $u = \bar{u} + u_S + u_L$. That is, u is divided into the part $\bar{u}$ which is known to everybody, u_S which is only known to the supply-informed traders, and the liquidity supply u_L which is not known to anybody. The individual value-informed trader's demand is, as usual, given by

$$x^i = \frac{E[v|S^i, P_1] - P_1}{\rho \operatorname{Var}[v|S^i, P_1]}.$$

[2] To facilitate comparison across papers, I have adjusted the notation to $S^i = p'_i$, $v = p$, $\mu_v = \bar{p}$, $\sigma_v^2 = \Sigma$, $p_1 = p_0$, $u = m$, $u_L = L$, $u_S = S$.

Similarly, the supply-informed trader's demand is given by

$$x^j = \frac{E[v|u_S, P_1] - P_1}{\rho \operatorname{Var}[v|u_S, P_1]}.$$

In addition to the informed traders' demand, there is an exogenous demand from portfolio traders who use dynamic trading strategies. Their demand $\pi(P_1)$ rises as the price increases and declines as the price falls.

As long as $\pi(P_1)$ is linear and common knowledge, the equilibrium price $P_1 = f(v - \mu_v - k_1 u_L - k_2 u_S)$ is a linear function with constants k_1 and k_2. In this linear case, the price P_1 is normally distributed. For nonlinear hedging demands $\pi(P_1)$, the argument of the price function, $f^{-1}(P_1)$, is still normally distributed and, therefore, the standard technique for deriving conditional expectations for normally distributed random variables can still be used. Discontinuity in $f(\cdot)$ makes "crashes" possible, that is, a small change in the argument of $f(\cdot)$ leads to a large price shift. $f(\cdot)$ is linear and continuous in the absence of any program trading, $\pi(P_1) = 0$. This rules out crashes.

Nevertheless, even for $\pi(P_1) = 0$ an increase in the supply can lead to a large price shift. Gennotte and Leland (1990) derive elasticities measuring the percentage change in the price relative to the percentage change in supply. This price elasticity depends crucially on how well a supply shift can be observed. The price change is small if the change in supply is common knowledge, that is, the supply change is caused by a shift in $\bar{u}$. If the supply shift is only observed by supply-informed traders, the price change is still moderate. This occurs because price-informed and supply-informed traders take on a big part of this additional supply even if the fraction of informed traders is small. Supply-informed traders know that the additional excess supply does not result from different price signals while price-informed traders can partially infer this from their signal. If, on the other hand, the additional supply is not observable to anybody, a small increase in the liquidity supply u_L can have a large impact on the price. In this case, traders are reluctant to counteract the increase in liquidity supply u_L by buying stocks since they cannot rule out the possibility that the low price is due to bad information that other traders might have received. Regardless of whether the supply shift is known to everyone, someone, or no one, the equilibrium price is still a linear continuous function of the fundamentals and thus no crash occurs.

By adding program trading demand, the price P_1 becomes even more volatile since $\pi(P_1)$ is an increasing function. Dynamic hedgers buy stocks when the price increases and sell stocks when the price declines. This violates the law of demand. As long as $\pi(P_1)$ is linear, $P_1 = f(\cdot)$ is continuous and linear. Crashes only occur when the program trading is large enough to cause a discontinuous price correspondence $f(\cdot)$. The discontinuity stems from the nonlinearity of program trading $\pi(P_1)$ and the lack of knowledge of the amount of program trading $\pi(P_1)$. Crashes are much more likely and prices are more volatile if some investors underestimate the supply due to program trading. Gennotte and Leland (1990) illustrate their point by means of an example of a put-replicating hedging strategy (synthetic put). In this example, the excess demand curve is downward sloping as long as all traders or at least the supply informed traders know the level of program trading demand. In the case where hedging demand is totally unobserved, the demand curve looks like an "inverted S." There are multiple equilibria for a certain range of aggregate supply.[3] The aggregate supply can be depicted as a vertical line. Thus as the aggregate supply shifts, the equilibrium with the high asset price vanishes and the asset price discontinuously falls to a lower equilibrium level. This is illustrated in Figure 6.1.

Gennotte and Leland's (1990) explanation of a stock market crash provides a different answer to the question of whether the market will bounce back after the crash. In contrast to Grossman (1988), the price can remain at this lower level even when the supply returns to its old

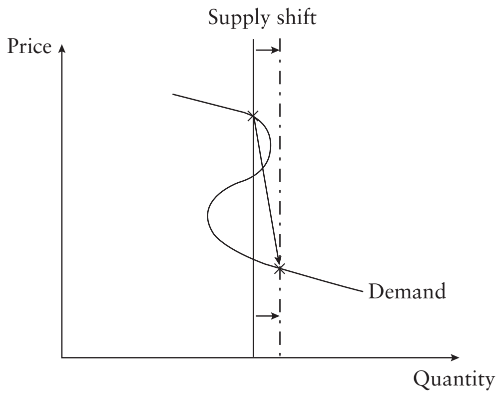

Figure 6.1. *Price crash in a multiple equilibrium setting*

[3] In this range, crashes can also be generated by sunspots. A different realization of the sunspot might induce traders to coordinate in the low-price equilibrium instead of the high-price equilibrium.

level. The economy stays in a different equilibrium with a lower asset price.

The reason why uninformed portfolio trading has a larger impact in Gennotte and Leland (1990) than in Grossman (1988) is that it affects other investors' trading activities as well. Asymmetric information about the asset's fundamental value is a crucial element of the former model. Program trading can lead to an "inverted S"-shaped excess demand curve. As a consequence, there are multiple equilibria in a certain range of parameters and the price drops discontinuously as the underlying parameter values of the economy change only slightly. It is, however, questionable whether this discontinuity in the static setup would also arise in a fully fledged dynamic model. In a dynamic model, traders would take into account the fact that a possible small parameter change can lead to a large price drop. Therefore, traders would already start selling their shares before the critical parameter values are reached. This behavior might smooth out the transition and the dynamic equilibrium will not necessarily exhibit the same discontinuity.

Delayed Sudden Information Revelation in a Dynamic REE

Romer (1993) illustrates a drastic price drop in a dynamic two-period model. In this model, a crash can occur in the second period since the price in the second trading round leads to a sudden revelation of information. It is assumed that traders do not know the other traders' signal quality. The price in the first trading round cannot reveal both the average signal about the value of the stock as well as precision of the signals, that is higher-order uncertainty. In the second trading round, a small commonly known supply shift leads to a different price which partially reveals higher-order information. This can lead to large price shocks and stock market crashes.

In Romer (1993) each investor receives one of three possible signals about the liquidation value of the single risky asset, $v \sim \mathcal{N}(\mu_v, \sigma_v^2)$:[4]

$$S^j = v + \epsilon_{S^j},$$

where $\epsilon_{S^2} = \epsilon_{S^1} + \delta^2$, $\epsilon_{S^3} = \epsilon_{S^2} + \delta^3$ and ϵ_{S^1}, δ^2, δ^3 are independently distributed with mean of zero and variance $\sigma^2_{\epsilon_{S^1}}$, $\sigma^2_{\delta^2}$, $\sigma^2_{\delta^3}$, respectively. Thus, S^j is a sufficient statistic for S^{j+1}. There are two equally likely states of the world for the signal distribution. *Either* half of the traders receive

[4] The notation in the original article is: $v = \alpha, S_j = s_j, \mu_v = \mu, \sigma_v^2 = V_\alpha, u_1 = Q, \mu_{u_1} = \bar{Q}, \sigma^2_{u_1} = V_Q$.

signal S^1 and the other half receive signal S^2 *or* half of the traders receive signal S^2 and the other half receive signal S^3. It is obvious that traders who receive signal S^1 (or S^3) can infer the relevant signal distribution since each investor knows the precision of his own signal. Only traders who receive signal S^2 do not know whether the other half of the traders have received the more precise signal S^1 or the less precise signal S^3. As usual, the random supply in period 1 is given by the independently distributed random variable $u_1 \sim \mathcal{N}(\mu_{u_1}, \sigma^2_{u_1})$.[5]

The stock holdings in equilibrium of S^1-traders, $x^1(S^1)$, can be directly derived using the projection theorem. S^1-traders do not make any inference from the price since they know that their information is sufficient for any other signal. Traders with S^3-signals face a more complex problem. They know the signal distribution precisely but they also know that they have the worst information. In addition to their signal S^3, they try to infer signal S^2 from the price P_1. The equilibrium price in $t = 1$, P_1, is determined by $x^2(S^2, P_1) + x^3(S^3, P_1) = u_1$ (assuming a unit mass of each type of investor). Since an S^3 trader knows $x^2(\cdot)$, $x^3(\cdot)$, and the joint distribution of S^2, S^3, and u_1, he can derive the distribution of S^2 conditional on S^3 and P_1. Since $x^2(S^2, P_1)$ is not linear in S^2, $x^3(S^3, P_1)$ is also nonlinear. S^2-investors do not know the signal precision of the other traders. Therefore, the $\mathrm{Var}[P_1|S^2]$ depends on the higher-order information, that is, on whether the other half of the traders are S^1- or S^3-investors. S^2-traders use P_1 to predict more precisely the true signal distribution, that is, to predict the information quality of other traders. If they observe an extreme price P_1, then it is more likely that other investors received signal S^3. On the other hand, if P_1 is close to the expected price given their own signal S^2, then it is more likely that the others are S^1-traders. S^2-investors' demand functions $x^2(S^2, P_1)$ are not linear in P_1 since P_1 changes not only the expectations about v, but also its variance. This nonlinearity forces Romer (1993) to restrict his analysis to a numerical example. His simulation shows that S^2-investors' demand functions are very responsive to price changes.

Romer's (1993) key insight is that a small shift in aggregate supply in period $t = 2$ induces a price change which allows the S^2-investors to infer the precision of the other traders' signals. A small supply change leads to the revelation of "old" information which has a significant impact

[5] Even without the random supply term u_1, the REE is not (strong form) informationally efficient since a single price cannot reveal two facts, the signal and the signal's quality. The structure is similar to the partially revealing REE analysis in Ausubel (1990). However, if there is no noisy supply, the no-trade theorem applies.

on prices. Note that in contrast to Grundy and McNichols (1989), discussed in Section 4.1.2, the supply shift in period $t = 2$ is common knowledge among all traders. An uncertain supply shift would prevent S^2-investors from learning the type of the other investors with certainty. Romer (1993) uses this insight to explain the October 1987 market meltdown.

In his model the stock market crash in $t = 2$ is a price correction. The revelation of information through P_2 makes investors aware of the early mispricing. Therefore, in contrast to Grossman (1988) but in line with Gennotte and Leland (1990), this model does not predict any rebounding of the price after the stock price.

In Section II of his paper, Romer (1993) develops an alternative model to explain stock market crashes. In this model, informed traders trade at most once. They can trade immediately if they pay a fee. Else, they can save the fee and but then their trade will be executed at a random time or not at all. This model is closer in spirit to the sequential trade models that are covered in the next section.

Modeling crashes within a dynamic REE setup gets complex very quickly. Even the analysis in Romers' (1993) two-period REE setup is restricted to numerical simulations. One needs models which cover a longer time horizon to really understand the dynamics of stock market crashes. The more simplistic sequential trade models provide one possible framework for a dynamic analysis.

6.1.2. *Crashes in Sequential Trade Models*

Sequential trade models are more tractable and, thus, allow us to focus on the dynamic aspects of crashes. The literature based on sequential trade models also analyzes the role of portfolio insurance trading and stresses the importance of asymmetric information to explain crashes.

The economic insights of the herding literature provide a basis for understanding stock market crashes. An informational cascade or a partial informational cascade can arise in trading models. If the market is in a partial cascade, the actions of predecessors need not lead to a price change for a long time. Eventually, a fragile partial cascade might burst and cause a significant price change. This is in contrast to a full information cascade which never bursts. Using Lee's (1998) terminology, an informational avalanche occurs when a partial cascade bursts.

Sequential trade models à la Glosten and Milgrom (1985) and herding models à la Bikhchandani, Hirshleifer, and Welch (1992) share some common features:

1. Traders can only buy or sell a fixed number of shares. Their action space is, therefore, discrete.
2. Agents also trade one after the other.

This replicates an exogenous sequencing model where the timing of agents' trade is exogeneously specified. In descending Dutch auctions, traders can decide when to trade and thus they are closely related in spirit to herding models with endogenous sequencing. The latter class of models is discussed in the next section.

Portfolio Insurance Trading in Sequential Trade Models

As in Gennotte and Leland (1990), Jacklin, Kleidon, and Pfleiderer (1992) attribute the stock market crash in 1987 to imperfect information aggregation caused by an underestimation of the extent of dynamic portfolio insurance trading. The authors reach this conclusion after introducing dynamic program trading strategies in the sequential trade model of Glosten and Milgrom (1985). The market maker sets a competitive bid and ask price at the beginning of each trading round. Given this price schedule, a single trader has the opportunity to buy or sell a fixed number x of shares or to not trade at all. The probability that an informed trader trades in this period is μ. This trader knows the final liquidation value of the stock $v \in \{v_L, v_M, v_H\}$. An informed trader buys (sells) the stock when its value stock v is higher (lower) than the ask (bid) price and does not trade at all if v is between the bid and ask price. An uninformed trader trades in this period with probability $(1 - \mu)$. Uninformed traders are either dynamic hedgers or liquidity traders. The fraction of dynamic hedgers θ is not known and can be either θ_H or θ_L. The strategy of dynamic hedgers is exogeneously modeled in a very stylized manner and exhibits some similarity to herding behavior. Dynamic hedgers either buy or sell shares. They buy shares for two reasons: to start a new dynamic hedging strategy or to continue with an existing strategy. In the latter case, they buy shares if the trading (in)activity in the previous trading round increases their judgment about the value of the stock. In addition, dynamic hedgers sell shares with some probability. They always buy or sell shares and are never inactive in the market. This distinguishes them from informed traders and liquidity traders. Liquidity traders buy or sell x shares with the same probability r or do not trade at all with the remaining probability $1 - 2r$.

The authors illustrate the price path by means of a numerical simulation. One can rule out a stock market crash following a significant price rise as long as the fraction θ is known to the market makers. However, the price might rise sharply if the market maker underestimates the degree of dynamic portfolio trading. The market maker mistakenly interprets buy orders from dynamic hedgers as informed traders with positive information. This leads to a sharp price increase. After many trading rounds, the fact that he observes only few "no trade outcomes" makes him suspicious that the earlier order might have come from dynamic hedgers. He updates his posterior about θ and significantly corrects the price. This leads to a stock market crash. Since the crash is a price correction, one does not expect the price to bounce back.

Jacklin, Kleidon, and Pfleiderer (1992) focus solely on dynamic trading strategies and make no reference to the herding literature. However, rational hedging also generates similar behavior. The articles described next explicitly draw the connection between the herding literature and trading games and, hence, provide deeper insights.

Herding and Crashes in Sequential Trade Models

Avery and Zemsky (1998) illustrate a sequential trade model with an information structure similar to the herding model in Bikhchandani, Hirshleifer, and Welch (1992). A fraction μ of the traders are informed while $(1-\mu)$ are uninformed liquidity traders. Liquidity traders buy, sell, or stay inactive with equal probability. Each informed trader receives a noisy individual signal about the value of the stock $v \in \{0, 1\}$. The signal is correct with probability $q > \frac{1}{2}$. In a sequential trade model, the predecessor's action not only causes a *positive informational externality* as in Bikhchandani, Hirshleifer, and Welch (1992), but also a *negative payoff externality*. The price changes since the market maker also learns from the predecessor's trade. Hence, he adjusts the bid and ask schedule accordingly. This changes the payoff structure for all successors. Avery and Zemsky (1998) show that the price adjusts in such a way that it offsets the incentive to herd. This is the case because the market maker and the insiders learn at the same rate from past trading rounds. Therefore, herding will not occur given pure *value uncertainty*. In general, as long as the signals are monotonic, the herding incentives are offset by the market maker's price adjustment. Consequently, a (full) informational cascade does not arise.

Indeed, informational cascades can be ruled out even for information structures which lead to herding behavior since the authors assume that there is always a minimal amount of "useful" information. Hence, the price converges to the true asset value and the price process exhibits no "excess volatility," regardless of the assumed signal structure, due to the price process' martingale property. This implies that large mispricings followed by a stock market crash occur only with a very low probability.

Avery and Zemsky (1998) also explicitly analyze some nonmonotonic signal structures. As in Easley and O'Hara (1992), they introduce higher-order uncertainty via *event uncertainty*. Insiders receive either a perfect signal that no new information has arrived, that is, the value of stock remains $v = \frac{1}{2}$, or a noisy signal which reports the correct liquidation value $v \in \{0, 1\}$ with probability q. Viewed differently, all insiders receive either a totally useless signal whose precision is $q' = 1/2$ (no information event) or all insiders receive possibly different signals but with the same precision $q' = q \in (1/2, 1]$. The precision, q', is known to the insiders, but not to the market maker. In other words, the market maker does not know whether an information event occurred or not. This asymmetry in higher-order information between insiders and the market maker allows insiders to learn more from the price process (trading sequence) than the market maker. Since the market maker sets the price, the price adjustment is slower. Bikhchandani, Hirshleifer, and Welch (1992) can be viewed as an extreme case where prices are essentially "fixed." Slow price adjustment reduces the payoff externalities which could offset the information externality. Consequently, traders might herd in equilibrium. However, no informational cascade arises since the market maker can gather information about the occurrence of an information event. Surprisingly, herding increases the market maker's awareness of information events and does not distort the asset price. Therefore, herding in a setting with only "event uncertainty" cannot explain large mispricings or stock market crashes.

A more complex information structure is needed to simulate crashes. Avery and Zemsky (1998) consider a setting with two types of informed traders in order to explain large mispricings. One group of traders receives their signals with low precision q_{L}, whereas the other receives them with high precision $q_{\mathrm{H}} = 1$, that is, they receive a perfect signal. The proportion of insiders with the perfect signal is either high or low and it is not known to the market maker. The authors call this information structure *composition uncertainty*. This information structure makes it difficult for the market maker to differentiate between a market composed of well-informed traders following their perfect signal from one with poorly informed traders who herd. In both situations a whole chain of informed traders follows the same trade. If the prior probability is very low that poorly informed traders are operating in the market, a chain of buy orders make the market maker think that a large fraction of traders is perfectly informed. Thus, he increases the price. If the unlikely event occurs in which only poorly informed traders herd, the asset price may exceed its liquidation value v. The market maker can infer only after many trading rounds that the uninformed traders have herded. In that case, the asset price crashes. Avery and Zemsky (1998) refer to this event as a bubble even though it is not a bubble in the sense described in Section 2.3. Bubbles only occur if traders mutually know that the price is too high yet they still hold or buy the asset. This is the case since they think that they can unwind the position at an even higher price before the liquidation value is paid out. Bubbles in a sequential trading setting à la Glosten and Milgrom (1985) can never occur since this setting does not allow agents to trade a second time. That is, traders cannot unwind their acquired position. All traders have to hold the asset until the liquidation value is paid out.

Gervais (1997) is similar to Avery and Zemsky (1998). However, it shows that uncertain information precision can lead to full informational cascades where the insider's information precision never gets fully revealed. Thus, the bid–ask spread does not reflect the true precision. In Gervais (1997) all agents receive a signal with the same precision, $q_{\mathrm{H}} > q_{\mathrm{L}}$, $q_{\mathrm{L}} > \frac{1}{2}$, or $q_{\mathrm{no}} = \frac{1}{2}$. If the signal precision is $q_{\mathrm{no}} = \frac{1}{2}$, the signal is useless, that is, no information event occurs. In contrast to Avery and Zemsky (1998), the signals do not refer to the liquidation value of the asset, v, directly, but only to a certain aspect v_t of v. More formally, the trader who can trade in trading round t receives a noisy signal S_t about the component v_t. There is only one signal for each component v_t, which takes on a value $1/T$ or $-1/T$ with equal probability of $\frac{1}{2}$. The final liquidation value of the asset is then given by $v = \sum_{t=1}^{T} v_t$. As in Glosten and Milgrom (1985), the risk neutral market maker sets competitive quotes. If the bid–ask spread is high, insiders trade only if their signal precision is high. The trade/no-trade sequence allows the market maker to update his beliefs about the quality of the insider's signals. He can also update his beliefs about the true asset value v. Therefore, the competitive spread has to decrease over time. Note that the trading/quote history is more informative for insiders because they already know the precision of the signal. When the competitive bid–ask spread decreases below a certain level, insiders will engage in trading independent of the precision of their signal. This prevents the competitive market maker from learning more about the signals' precision, that is, the economy ends up in a cascade state with respect to the precision of the insider's signals.

In Madrigal and Scheinkman (1997) the market maker does not set a competitive bid–ask spread. Instead, he sets the bid and ask prices which maximize his profit. The price function in this one-period model displays a discontinuity in the order flow. As in Gennotte and Leland (1990), this discontinuity can be viewed as a price crash since an arbitrarily small change in the market variables leads to a large price shock.

Crashes due to Information Avalanches

Lee's (1998) model departs in many respects from the Glosten–Milgrom setting. It is still the case that in each period only a single trader receives a signal about the liquidation value $v \in \{0, 1\}$. However, in Lee (1998) the trader can decide when to trade and he can also trade more than once.[6] In particular, traders have the possibility of unwinding their position in later trading rounds. This model is, therefore, much closer in spirit to herding models with endogenous sequencing. The trades are also not restricted to a certain number of shares. However, when agents want to trade they have to pay a one-time fixed transaction fee c to open an account with a broker. There are no liquidity traders or dynamic hedgers in this model; there are only risk averse informed traders. Traders are assumed to be price takers. Prior to each trading round the market maker sets a *single* price at which all orders in this trading round will be executed. This is in contrast to the earlier models where the market maker sets a whole price schedule, or at least a bid and an ask price. The market maker's single price $p_t = E[v|\{x_t^i\}_{i \in \mathbb{I}}]$ is based on all observed individual orders in all the *previous* trading rounds. The market maker loses money on average since he cannot charge a bid-ask spread even though informed traders are better informed than he is. This "odd" assumption simplifies matters and is necessary to induce informed traders to trade. Otherwise the no-trade (speculation) theorem of Milgrom and Stokey (1982) would apply in a setting without liquidity traders.

Each informed trader receives one of N possible signals $S_n \in \{S_1, \ldots, S_N\} =: \mathcal{S}$ which differ in their precision. The signals satisfy the monotone likelihood property and are ranked accordingly. The market maker can observe each individual order and since there are no liquidity traders he can fully infer the information of the informed trader. However, by assumption the market maker can only adjust the price for the next trading round. The price in the next trading period then fully reflects the informed trader's signal and, thus, the informed trader has no informational advantage after his trade is completed. Due to the market maker's risk neutrality, no risk premium is paid and, hence, the

[6] The notation departs from that in the original article: $v = Y$, $x_t^i = z_t^i$, $S_n = \theta_n$, $\mathcal{S} = \Theta$.

risk averse insider is unwilling to hold his risky position. He will unwind his entire position immediately in the next trading round. This trading strategy of 'acquiring and unwinding in the next round' would guarantee informed traders a certain capital gain. Consequently, it would be optimal for the informed traders to trade an infinite number of stocks in the first place. In order to avoid this, Lee (1998) assumes that in each period the liquidation value v might become common knowledge with a certain probability γ. This makes the capital gains random and, thus, restrains the trading activity of the risk averse informed traders. In short, the model setup is such that the informed agents trade at most twice. After they buy the asset they unwind their position immediately in the next period. Therefore, the trader's decision is de facto to wait or to trade now and unwind the position in the next trading round. This makes the "endogenous reduced action space" of the trading game *discrete*.

As trading goes on and the price converges (maybe wrongly) to the value $v = 0$ or $v = 1$, the price impact of an individual signal and thus the capital gains for informed traders become smaller and smaller. It is possible that the expected capital gains are so small that it is not worthwhile for the informed trader to pay the transaction costs c. This is especially the case for traders with less precise signals. Consequently, all traders with less precise signals $S_n \in \hat{\mathcal{S}} \subset \mathcal{S}$ will opt for a "wait and see strategy." That is, all traders with signals $S_n \in \hat{\mathcal{S}}$ herd by not trading. In Lee's words, the economy is in a *partial informational cascade*. When agents do not trade based on their information, this information is not revealed and, hence, the market accumulates a lot of hidden information which is not reflected in the current stock price. An extreme signal can shatter this partial informational cascade, as shown in Gale (1996) in Section 5.2.2. A trader with an extreme signal might trade when his signal strongly indicates that the price has converged to the wrong state. This single investor's trade not only induces some successors to trade but might also enlighten traders who received their signal earlier and did not trade so far. It might now be worthwhile for them to pay the transaction costs c and to trade based on their information. These traders are now eager to trade immediately in the same trading round as long as the market maker has committed himself to the same price. Consequently, there will be an avalanche of orders and all the hidden information will be revealed. In other words, an *informational avalanche* in the form of a stock market crash occurs. The subsequent price after the stock market crash is likely to be closer to the true liquidation value. The analysis in Lee (1998) also shows that the whole price process will eventually end up in a total informational cascade, that is, where no signal can break up the cascade.

Information avalanches in Lee (1998) hinge on the assumption that the market maker cannot adjust his quoted price within a trading round even when there are many individual orders coming in. Since the market maker is forced to execute a large order flow at a price that is much too high, he is the biggest loser in the event that a crash occurs. It would be interesting to determine the extent to which this assumption can be relaxed without eliminating the occurrence of informational avalanches. As in almost all models discussed so far, there is no reason why a crash has to be a price decline. It can also be a sharp price increase. This is a general criticism of almost all models given the empirical observations that one mostly observes sharp price declines.

In contrast to the standard sequential trade models, Lee's (1998) analysis has the nice feature that traders can choose endogenously when to trade and what amount to trade. In the standard auction theory covered in the next sections, bidders can also choose the timing of their bid. However, their quantity is fixed to a unit demand.

6.1.3. Crashes and Frenzies in Auctions and War of Attrition Games

While in the standard sequential trade models à la Glosten and Milgrom (1985) the order of trades is exogenous, auctions with ascending or descending bidding allow bidders to decide when to bid or stop bidding. Thus, these models correspond more to endogenous sequencing herding models. In contrast to pure informational herding models but like sequential trade models, the bidders' decisions cause both an information externality as well as a payoff externality. The information externality might even relate to the payoff externality. This is the case when the predecessor holds private information and his action affects the payoff structure of the successors. For example, when a bidder quits in a standard ascending auction (Japanese version), he reveals to the remaining bidders that there is one less competitor. This is a positive payoff externality for the remaining bidders. In addition, he reveals a signal about the common value of the good.

This section only covers the small part of the auction literature that focuses on crashes.[7] Due to its central role in this literature, let us first

[7] The auction literature was initiated by Vickrey (1961). There are several excellent overview articles that describe this literature. We refer the interested reader to Klemperer (1999, 2000), Matthews (1995), McAfee and McMillan (1987), and Milgrom (1989).

discuss the revenue equivalence theorem developed by Myerson (1981) and Riley and Samuelson (1981).

The Revenue Equivalence Theorem

The revenue equivalence theorem (RET) is the most important theorem in auction theory. It states that under certain conditions any auction mechanism that (1) assigns the good to the bidder with the highest signal and (2) grants the bidder with the lowest feasible signal a zero surplus, leads to the same expected revenue to the seller. This equivalence holds for a fixed number of risk neutral bidders and if the signals are independently drawn from a common, strictly increasing, atomless distribution, for example on $[\underline{V}, \bar{V}]$ It applies to a pure private value auction. It also extends to a pure common value auction provided the individual signals S^i are independent and the common value is a function of them, that is, $v = f(S^1, \ldots, S^I)$.

Let us outline the intuitive reasoning for this result. Without loss of generality, we choose signals $S^i = v^i$ such that they coincide with the unconditional value of the asset for bidder i. Suppose the expected payoff for a bidder with private signal v^i is $U^i(v^i)$. If the v^i-bidder tries to mimic a bidder with a signal $v^i + \Delta v$, his payoff would be the payoff of a $(v^i + \Delta v)$-bidder with the difference that, in the case that he wins the object, he values it Δv less than the $(v^i + \Delta v)$-bidder. He would receive the object with probability $P(v^i + \Delta)$ if he mimics the $(v^i + \Delta v)$-bidder. In any mechanism the bidder should have no incentive to mimic somebody else, that is, $U(v^i) \geq U(v^i + \Delta v) - \Delta v \, \mathrm{Pr}(v^i + \Delta v)$. Similarly, the $(v^i + \Delta v)$-bidder should not want to mimic the v^i-bidder, that is, $U(v^i + \Delta v) \geq U(v^i) + \Delta v \, \mathrm{Pr}(v^i)$. Combining both inequalities leads to

$$Pr(v^i) \leq \frac{U^i(v^i + \Delta v) - U^i(v^i)}{\Delta v} \leq \mathrm{Pr}(v^i + \Delta v).$$

For very small deviations $\Delta v \to 0$, this reduces to

$$\frac{\mathrm{d}U^i}{\mathrm{d}v} = \mathrm{Pr}(v^i).$$

Integrating this expression leads to the following expected payoff function.

$$U^i(v^i) = U^i(\underline{V}) + \int_{x=\underline{\mathrm{V}}}^{v^i} \mathrm{Pr}(x) \, \mathrm{d}x.$$

This payoff function determines the expected payoff for any v^i-type bidder. The no mimic conditions are satisfied as long as the bidder's payoff function is convex, that is, the probability of winning the object increases in v^i.

The risk neutral bidder's expected payoff $U(v^i)$ is given by his expected value of the object $E[v|v^i] = v^i$ times the probability of receiving the object, minus his expected transfer payment, T, in short, by $v^i \mathrm{Pr}^i(v^i) - T$. Two different auction mechanisms lead to the same payoff for a v^i-bidder if the bidder with the lowest signals receives the same payoff $U^i(\underline{V})$ in both auction mechanisms. If in addition the probability of winning is the same, then the expected transfer payoff for any type of bidder is the same in both auctions and so is the expected revenue for the seller.

The revenue equivalence theorem is extremely useful and powerful. Instead of analyzing the more complicated actual auction mechanism, one can restrict the analysis to simpler mechanisms by appealing to the revenue equivalence theorem.

Frenzies and Crashes

Bulow and Klemperer's (1994) auction article emphasizes frenzies and crashes within a multi-unit Dutch auction. As in the real option literature, a potential buyer has to decide whether to buy now or later, rather than now or never.[8] Bulow and Klemperer (1994) consider a private value setting, wherein each of $K + L$ potential buyers' *private* value for one good v^i is independently drawn from a distribution $F(v^i)$ which is strictly increasing and atomless on $[\underline{V}, \bar{V}]$. A seller offers K identical units of a good for sale to $K + L$ potential buyers. The seller can commit himself to a specific selling procedure. Hence, the seller receives the whole social surplus except the information rent, which goes to the bidders. Crashes and frenzies arise in any selling mechanism and are derived using the revenue equivalence theorem.

For concreteness, Bulow and Klemperer (1994) illustrate crashes and frenzies in a multi-unit Dutch auction. The seller starts at a high price and lowers it continuously until a purchase occurs. Then, the seller asks the remaining bidders whether somebody has changed his mind and would like to buy at this price too. If this is the case, he sells the goods to them and if some additional goods remain he asks the remaining bidders again

[8] The trade-off in the real option literature is that by delaying the purchase, the investor incurs waiting costs but gains the opportunity to learn something about the common value of the product.

whether their willingness to pay has changed. The authors define multiple sales at a single price as a *frenzy*. If nobody changes his mind, that is, if nobody else is willing to buy at this price, the seller continues to lower the price. If too many bidders have changed their mind and want to buy at this price, he runs a new Dutch auction among these bidders with the remaining goods. This might lead to higher prices. Therefore, the bidder is faced with a trade-off. On the one hand, if he waits, the price may be lower, but on the other hand waiting also increases the likelihood of a frenzy which could lead to a higher price or to the possibility that he walks home empty handed. In general, somebody else's purchase generates a negative externality for the remaining bidders since the number of remaining goods diminishes and with it the probability of receiving a good at this price decreases. Nevertheless, the option to wait changes the buyers "willingness to pay" in comparison to a setting where the seller commits to a single take-it-or-leave-it price.

Since the revenue equivalence theorem applies in this multi-unit setting, each bidder's expected payment must be the same for any auction design. In particular, the willingness to pay $\omega(v^i)$ for a bidder with private value v^i equals the expected price a bidder would pay in a standard multi-unit English auction. For k remaining goods and $k + l$ remaining bidders, each bidders "willingness to pay" is equal to his expectation of the $(k + 1)$st highest value out of the $k + l$ remaining values, provided this value is lower than his own valuation v. For bidders with high v^i, the willingness to pay is almost the same. To illustrate this, consider the bidder with the highest possible value, $\bar{v}$. He knows for sure that his valuation is the highest. Therefore, his estimate of the $(k + 1)$ highest valuation is the k highest of the other $(k + l - 1)$ bidders. This estimate decreases only slightly for bidders with lower v^i's as long as they are pretty sure that they are among the k bidders with the highest values. In other words, the WTP $\omega(v)$ is very flat, especially for high private values, v, compared to the standard demand curve – which represents the buyers' willingness to accept a take-it-or-leave-it final offer. Figure 6.2, which is taken from Bulow and Klemperer (1994), illustrates this point for a uniform distribution.

The WTP for the remaining bidders changes when one of the other bidders buys one of the goods. A purchase reduces the number of remaining goods available for the rest of the bidders. This increases the price each remaining bidder expects to pay (negative payoff externality) and thus shifts all other bidders' WTP functions upwards. Therefore, when the seller offers more sales at the same selling price, bidders with close enough v values might change their mind and come forward to buy at

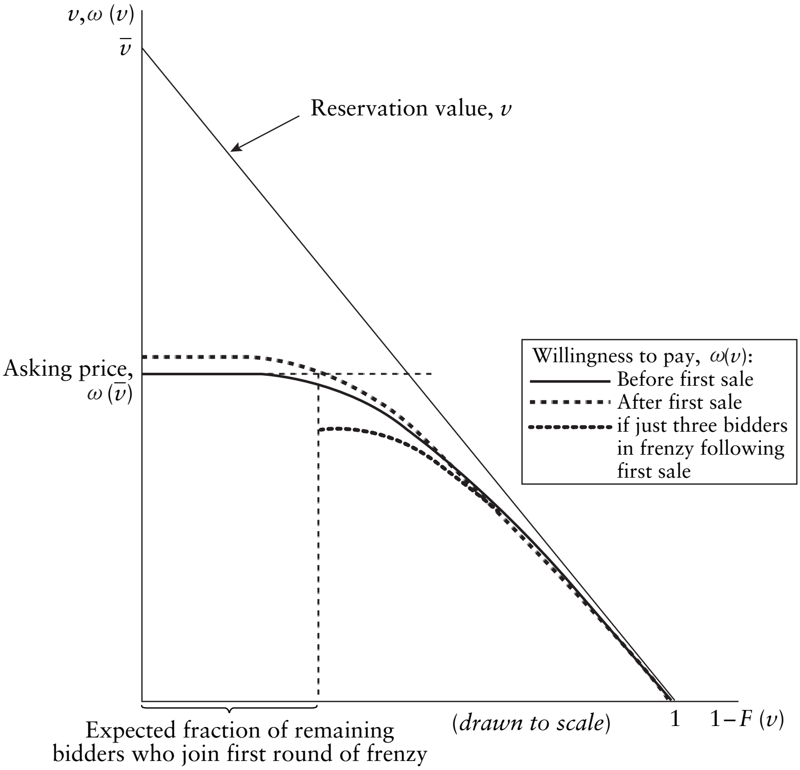

Figure 6.2. *Frenzy in an auction*

the same price. Since the WTP function $\omega(1 - F(v^i))$ is very flat, it is very likely that many bidders will come forward in the second round of sale at the same price. That is a frenzy might emerge. More specifically one of the following three scenarios can occur. (1) More bidders than expected change their mind but the demand can be satisfied. In this case the *frenzy* feeds itself since the bidders who did not buy in the second round might change their mind after observing that so many bidders have decided to buy in the second round. The seller offers the good at the same price in a third round and so on. (2) Too many bidders come forward and the seller cannot satisfy the demand. The seller then initiates a new descending multi-unit Dutch auction among these bidders by starting at the original starting price. (3) Although all WTPs increased no bidder or less than the expected number of bidders are willing to participate in the second round. In this case a "crash" occurs where it becomes common knowledge among all bidders that no purchase will occur until the price has fallen to a strictly lower level. The seller goes on lowering the price and one observes a discrete price jump. It needs to be stressed that these effects will be even stronger in a common value environment. With common values, the purchase is an additional signal for the remaining bidders that the value of the good is high. Thus, the remaining bidders' WTP increases even further.

War of Attrition Game

A *war of attrition* (*chicken*) *game* is like a sealed bid *all pay auction*. In an all pay auction, each bidder pays his own bid independently of whether he wins the object or not. In a war of attrition game the player who suffers the longest, that is, pays the most, wins the prize. Each player's strategy is like a bidding strategy, which specifies the point at which to stop suffering. The only difference between a war of attrition game and an all pay auction is that in the war of attrition game the player's payment does not go to the seller of an object, but is socially wasted. In a setting with independent private values of the object, the expected surplus of each player is the same in both games as it is in any auction, as long as the assumptions for the RET apply. This allows us to switch to the mechanism of the second price auction which is much easier to analyze.

Bulow and Klemperer (1999) use the RET to analyze a generalized war of attrition game. In Bulow and Klemperer (1999) each of $K + L$ player can win one of K objects. The player pays one unit per period as long as he stays in the race. After quitting, his costs reduce to $c \leq 1$ (possibly to zero) until L players ("losers") quit, that is, the game ends. If the players pay no costs after dropping out, that is, $c = 0$, $L-1$ players quit immediately and the remaining $K+1$ players play a standard multi-unit war of attrition game analyzed in Fudenberg and Tirole (1986). This result is derived by means of the RET. The expected total suffering can be calculated using a simpler standard $K + 1$ price auction because of the RET. We know from the $K+1$ price auction that the expected total payment, or suffering in this case, of all $K + L$ players coincides with expected $K + 1$ highest evaluation. After $L - 1$ lowest types drop out, the expected total amount of suffering of the remaining $K+1$ players is still the same. This can only be the case if the $L-1$ lowest types drop out immediately. Bulow and Klemperer (1999) also analyze the case where each player has to suffer until the game ends, independently of whether he drops out early or not, that is, $c = 1$. In this case the drop out strategy is independent of the number of players and of other players' drop out behavior. The optimal drop out strategies for intermediate cases of $0 < c < 1$ are also characterized by Bulow and Klemperer (1999).

6.2. Keynes' Beauty Contest, Investigative Herding, and Limits of Arbitrage

In reality, traders do not receive all information for free. They have to decide whether and which information to gather prior to trading. This affects their trading behavior as well as the stock price movements. Models in this section illustrate that traders have an incentive to gather the same information and ignore long-run information.

In his famous book *The General Theory of Employment, Interest and Money*, Keynes (1936) compared the stock market with a *beauty contest*. Participating judges – rather than focusing on the relative beauty of the contestants – try to second-guess the opinion of other judges. It seems that they would rather choose the winner than the most beautiful girl. Similarly in stock markets, investors' search effort is not focused on fundamentals but on finding out the information that other traders will trade on in the near future. Their intention is to trade on information right before somebody else trades on the same information. In Keynes' words, "skilled investment today is to '*beat the gun*' … ." This section argues that this is a rational thing to do, in particular if the investor – for whatever reason – intends to unwind the acquired position early.

In a setting where traders have to decide which information to collect, the value of a piece of information for the trader might depend on the other traders' actions. New information allows traders to update their estimate about the value of assets. Hence, in their view assets might become mispriced. Yet, private information only provides investors with a profitable trading opportunity if (1) they can acquire a position without immediately revealing their private information, *and* (2) they are able to unload their acquired position at a price which reflects their private information. In other words, as long as they acquire the position, the asset has to remain mispriced. However, when investors liquidate the required position, the price has to incorporate their information. Traders cannot exploit their knowledge if they are forced to liquidate before the asset is priced correctly. The mispricing might become even worse in the medium term. In this case, forced early liquidation leads to trading losses.

An asset is mispriced if its price does not coincide with the equilibrium price (*absolute* asset pricing). The exploitation of a mispriced asset is often referred to as *arbitrage*. Arbitrage – in the strict theoretical sense – refers only to mispricing *relative* to other assets. It involves no risk since one buys and sells assets such that future payoff streams exactly offset each other and a positive current payoff remains. In an incomplete market setting, there are often insufficient traded securities that exactly offset future payoff streams of an asset. The asset is, thus, not redundant. Therefore, mispricing of (nonredundant) assets need not lead to arbitrage opportunities in the strict sense. Practitioners often call all trading strategies which exploit mispricing "risky arbitrage trading." In contrast to riskless arbitrage trading, these trading strategies exploit mispricing even though the future payoff streams cannot be offset. Some models in this section adopt this broader definition of arbitrage, thereby essentially covering any information based trading.

Whether or not an asset's mispricing is corrected before the trader has to liquidate his position depends on whether the same information spreads to other traders. This new information is only fully reflected in the asset price when other market participants also base their trading activity on it. Brennan (1990) noted the strong interdependence of individual information acquisition decisions. In a market with many investors the value of information about a certain (latent) asset may be very small if this asset pays a low dividend and no other investor acquires the same information. If, on the other hand, many investors collect this information, the share price adjusts and rewards those traders who gathered this information first. Coordinating information collection activities can, therefore, be mutually beneficial.

There are various reasons why investors unwind their position early before the final liquidation value of the stock is known. The following sections discuss three different reasons why investors might want to unwind their position early. Traders try to unwind their position early because of:

(1) short-livedness;
(2) risk aversion in an incomplete market setting;
(3) portfolio delegation in a principal–agent setting.

If the traders liquidate their position early then they care more about future price developments than about the true fundamental value of the stock. Consequently, traders *prefer short-run information* to long-run information. They might even ignore long-run information. The future development of the asset price also depends on the information other traders gather. This explains why traders have an incentive to gather the same information, that is, why they herd in information acquisition. *Investigative herding* is the focus of Froot, Scharfstein, and Stein (1992).

6.2.1. Unwinding due to Short Horizons

Short-Livedness and Myopia

Short-lived agents convert their stocks and other savings into consumption latest in their last period of life. Agents with *short horizons* may live longer but they think only a few periods ahead. Their current behavior is often similar to that of short-lived individuals. In addition, myopic people's behavior is *dynamically inconsistent*. In the current period, myopic investors ignore some future payoffs. However, they will value them in some future period. The marginal rate of *temporal* substitution between consumption in two periods changes dramatically over time. Myopia is, therefore, an extreme form of hyperbolic discounting and has to be attributed to boundedly rational behavior. Myopia alters an agent's trading strategy since they do not take into account how their current trading affects their future optimal trading. However, the backwards induction argument still applies, which rules out major alterations of the price process in a setting with exogenous information acquisition. Nevertheless, there might exist additional equilibria if risk averse agents are short-lived or myopic. In these equilibria asset prices are very volatile and traders demand a higher risk premium since short-lived agents care only about the next period's price and dividend. Spiegel (1998) illustrates this in an overlapping generations (OLG) model. Similarly in DeLong, Shleifer, Summers and Waldmann (1990) short-livedness combined with risk aversion prevents arbitrageurs from driving prices back to their riskless fundamental value.The risk averse arbitrageurs care only about next period's price which is risky due to the random demand of noise traders.

Introducing Endogenous Information Acquisition

In models with endogenous information acquisition, short-livedness can also have a large impact on the price process. Brennan (1990) noticed the interdependence of agents' information acquisition decisions for low dividend paying (latent) assets. He formalizes this argument with an overlapping generations (OLG) model where agents only live for three periods. A short life span might force traders to liquidate their position before the information is reflected in the price. This is the case if the other traders did not gather the same information.

In Froot, Scharfstein, and Stein (1992) traders are also forced to unwind their acquired position in period $t = 3$ even though their information might be only reflected in the price in $t = 4$. Consequently, all traders worry only about the short-run price development since they have to unwind their position early. They can only profit from their information if it is subsequently reflected in the price. Since this is only the case if enough traders observe the same information, each trader's optimal information acquisition depends on the others' information acquisition. The resulting positive information spillover explains why traders care more about the information of others than about the fundamentals. Froot, Scharfstein, and Stein's (1992) analysis focuses on investigative herding. Herding in information acquisition would not occur in the stock market if agents only cared about the final liquidation value. In that case, information spillovers would be negative and thus, it would be better to have information that others do not have. Consequently, investors would try to collect information related to different events.

In Froot, Scharfstein, and Stein (1992) each individual can only collect one piece of information. Each trader has to decide whether to receive a signal about event A or event B. The trading game in Froot , Scharfstein, and Stein (1992) is based on Kyle (1985). The asset's liquidation value is given by the sum of two components, δ^A and δ^B,

$$v = \delta^A + \delta^B,$$

where $\delta^A \sim \mathcal{N}(0, \sigma^2_{\delta^A})$ refers to event A and $\delta^B \sim \mathcal{N}(0, \sigma^2_{\delta^B})$ refers to the independent event B. Each trader can decide whether to observe either δ^A or δ^B, but not both. After observing δ^A or δ^B a trader submits a market order to the market maker at $t = 1$. Half of the submitted market orders are executed at $t = 1$ and the second half at $t = 2$. The period in which an order is processed is random. Liquidity traders submit market orders of aggregate random size u_1 in each period $t = 1$ and u_2 in $t = 2$. As in Kyle (1985) the risk neutral market maker sets a competitive price in each period based on the observed total net order flow. Thus, the price only partially reveals the information collected by the informed traders. Traders acquire their position either in $t = 1$ at a price P_1 or in $t = 2$ at a price P_2, depending on when their order, which was submitted in $t = 1$, will be executed. At $t = 3$ all traders, that is, insiders and liquidity traders, unwind their position and by assumption the risk neutral market maker takes on all risky positions.

The fundamental value $v = \delta^A + \delta^B$ is publicly announced either in period $t = 3$ before the insiders have to unwind their position or in period $t = 4$ after they unwind their portfolio. With probability α, v is known in $t = 3$ and with probability $(1 - \alpha)$ it is known in $t = 4$. If the fundamentals are known to everybody in $t = 3$, the acquired positions are unloaded at a price $P_3 = \delta^A + \delta^B$. If the public announcement occurs only in $t = 4$, the price does not change in period $t = 3$, that is, $P_3 = P_2$ and the traders unwind their position at the price P_2. In this case the expected profit per share for an insider is $\frac{1}{2}(P_2 - P_1)$. The $\frac{1}{2}$ results from the fact that the insider's order is only processed early with a probability $\frac{1}{2}$. Only then does the insider receive the shares at a price of P_1, which he can later sell in $t = 3$ for $P_3 = P_2$. The trader makes no profit if his order is executed late and the fundamentals are only announced in $t = 4$. With probability α, the fundamentals δ^A and δ^B are already announced at $t = 3$, that is, $P_3 = v$. In this case a trader who submitted an order at $t = 1$ and buys a share for P_1 or P_2 with equal probability, sells it at $t = 3$ for $P_3 = v$. His expected profit in this case is given by $v - \frac{1}{2}[P_1 + P_2]$. Thus, the overall expected profit per share for an informed trader is

$$E\left\{\alpha\left[v - \frac{P_1 + P_2}{2}\right] + (1 - \alpha)\left[\frac{P_2 - P_1}{2}\right]\right\}.$$

In both cases the profit is determined by P_3, the price at which the informed trader unwinds his position; $P_3 = v$ with probability α. Thus, δ^A and δ^B are equally important, with probability α. With probability $(1-\alpha)$, $P_3 = P_2$. Since P_2 depends on the information set of all informed traders, each insider cares about the information that the other traders are collecting.

For illustrative reasons let us consider the polar case $\alpha = 0$, that is δ^A and δ^B are only publicly announced in $t = 4$. If all other investors collect information δ^A, then information δ^B is worthless in this case since δ^B will only enter into the price in $t = 4$. In period $t = 4$ investors will have already unwound their positions. Consequently, all investors will herd to gather information δ^A and nobody will collect information δ^B.[9] Thus, the short horizons of traders creates positive informational spillovers which lead to herding in information acquisition.

In an even more extreme scenario, if all investors herd on some noise term ζ, which is totally unrelated to the fundamental value $v = \delta^A + \delta^B$, a rational investor is (weakly) better off if he also collects information ζ rather than information about fundamentals alone. If $\alpha = 0$ and all other investors are searching for ζ, the fundamentals δ^A and δ^B are only reflected in P_4. The price at which the traders have to close their position, P_3 $(= P_2)$ might depend on the "sunspot" ζ, given their strategies.

[9] In a (Nash) equilibrium the information that other traders are collecting is mutual knowledge.

For the more general case of $\alpha > 0$, where δ^A and δ^B might already be announced in $t = 3$, herding in information might still occur. This is still the case if α is sufficiently small. In contrast if $\alpha = 1$, each trader prefers to collect information about events that are not the main focus of the other traders' information gathering effort. In short, individuals' search efforts are "strategic substitutes" if $\alpha = 1$, and "strategic complements" if $\alpha = 0$.

The above reasoning can also be analyzed in a multiperiod overlapping generations (OLG) framework. A new generation of short-sighted traders enters the market in each period. Inefficient herding still occurs in the following OLG setting. Generation t speculators can study one of k pieces of information. At the end of period t, one of these pieces will be randomly drawn and publicly announced. In the following period $t + 1$, a new piece of information can be studied. Thus, each trader in each generation can study one of k pieces of information. For each generation it pays off to have accidentally studied the information that gets publicly announced at the end of the period. Since this only occurs with a probability $1/k$ it is more worthwhile to collect information which is also studied by other traders and thus is reflected in the price for sure. In short, herding in information acquisition may also occur in this OLG setup.

Arbitrage Chains

Dow and Gorton's (1994) "arbitrage chains" model stresses that the value of exploiting a certain piece of information depends on the likelihood that another insider will receive the same information in the next trading round and drive the price closer to its fundamental. Only then can the insider, who lives for two periods, unwind his position at a profit. If there is no agent who trades on the same information in the next trading round then the trader would have been better off by investing in a bond since he would have saved transaction costs, c. In contrast to Froot, Scharfstein, and Stein (1992), Dow and Gorton (1994) consider an infinite horizon economy $t = -\infty, \ldots, \infty$ with overlapping generations (OLG). Each agent lives only for two periods. All young people receive a fixed endowment W. Consumption takes place only in the agents' second period of life and thus agents try to save. Agents can save by buying a bond with riskless return of r or a stock which pays a dividend of either 1 or 0 in each period. The dividend payments are serially uncorrelated and a dividend of 1 is paid with (prior) probability π. Another differentiating feature of Dow and Gorton's model from Froot, Scharfstein, and Stein (1992) is that the information acquisition process is assumed to be exogenous. The insider cannot decide *which* information to gather. A young trader receives a perfect signal about the dividend payments in t^{div} with a certain probability $\gamma_t = \gamma_{t+1}\epsilon = \epsilon^{(t^{\text{div}}-t)}$.[10] This probability converges smoothly to 1 as t approaches t^{div}. In addition to the young informed trader, uninformed hedgers might also be present in certain periods. They are active in the market place with probability $\frac{1}{2}$. These traders have a strong incentive to trade for hedging reasons as their wage of 0 or 1 in the next period is perfectly negatively correlated with the dividend payments. In short, each generation consists either of a single informed arbitrageur and/or uninformed hedgers or nobody.

The price setting is similar to Kyle (1985). A single competitive market maker sets the price after observing the order flow. In contrast to Kyle (1985), he also observes each individual order. He can deduce the orders from the old generations since they unwind their earlier trades. This unwinding keeps the market maker's inventory from growing ever larger. Although the market maker can observe each individual order he does not know whether the orders from the young generation is due to hedging needs or informed trading. Young uninformed hedgers try to hedge their wage income risk by buying x_t stocks. The informed trader might also buy x_t stocks. Given the market maker's beliefs, the informed trader can only hide behind hedgers if he submits a buy order of the same size x_t. Any other order size would reveal to the market maker that he trades for informational reasons. An informed young trader will only buy the stock if he receives a positive signal about the dividend payment in the near future t^{div}. If the dividend payment in t^{div} is more than K periods away, he will ignore his signal. The market maker knows that the insider might get a signal about the dividend payment (in t^{div}). Prior to $t^{\text{div}} - K$, the market maker always sets the price p_t equal to π/r since nobody is trading for informational reasons. The stock price is equal to the average dividend payment π in perpetuity, discounted at the rate r. However, an insider might be trading in periods closer to t^{div} and thus the market maker adjusts the price according to the observed order flow from the young generation. If the order flow is $2x$, the market maker knows for sure that an informed trader submitted a buy order and thus the dividend payments in t^{div} will be 1. Therefore, he adjusts the price to $p_t = \pi/r + (1+r)^{-(t^{\text{div}}-t)}(1-\pi)$. If nobody submitted an order, the insider might have received bad news or no news. Therefore, the market maker will lower the stock price. If he observes a single buy order x

[10] For consistency with the rest of the chapter, we replace the original notation δ_t with γ_t and T with t^{div}.

then he does not know whether an informed trader or a hedger submitted it. The aggregate order flow of x might stem from the insider if he received a good signal and no hedger was active in the market place, or it might stem from the hedgers. In the latter case, the arbitrageur has received either no signal or a bad signal. Dow and Gorton's (1994) model specification is such that the market maker's belief β about the dividend payment in t^{div} is not affected in this case. The market maker adjusts the price only slightly to reflect the fact that the expected dividend payment in t^{div} of β is now one period closer and thus requires less discounting.

Given this pricing rule, the insider's profit is highest in the case where only he transacts with the market maker when he buys the stock and one period later when he sells his stock; the new generation's order flow is $2x$, as this fully reveals the private information to the market maker. Dow and Gorton (1994) show that the optimal trading strategy for an insider is to ignore any long-run information that refers to dividend payments which are more than K periods in the future. This is the result of two effects: (1) As long as t^{div} is in the distant future, it is very unlikely that the information will be reflected in the next period's price. Therefore, it is not worthwhile to pay the (round trip) transaction costs c. (2) The second effect is due to discounting. If the information refers to a positive dividend payment ($= 1$) in the distant future, its present value and thus the present capital gains will be smaller. Given that transaction costs c have to be paid immediately, short-run information is more valuable. Both effects together make it optimal for an insider to ignore any information concerning dividend payments not within a K periods' reach. In other words, an insider only trades on short-run information. A whole chain of insiders might emerge who trade on their information in this window of K periods prior to t^{div}.

In OLG models, bubbles are possible if long-run information is ignored. Consider a situation where all traders in one generation – except the market maker – know that the asset is mispriced. They might not trade on this information if the probability is low that the next generation's young traders will have the same information and also not trade on it.

Dow and Gorton (1994) depart from the standard models in two ways. (1) They introduce trading costs c and (2) they assume exogenously short livedness/horizons. But even when all traders have long horizons, transaction costs alone make very long-run information worthless. This is due to the discounting effect described above. Transaction costs cause a short-term bias in the kind of information that is incorporated in asset prices. Traders' short horizons multiply this bias. To see this, even when there is an informed insider in each trading round, that is, $\gamma = 1$, the profits of short-sighted agents are only half that of the long-horizon decision maker. The reason is that, with probability $\frac{1}{2}$, no hedger will arrive in the next period. In this case, the market maker cannot infer the insider's information even in the next period and, thus, the "unwinding" price will not fully reflect the insider's information. As the probability that an insider trades in the next trading round γ decreases, so do the expected capital gains for myopic traders. The smaller the probability γ, the higher the potential capital gain has to be in order to make up for the transaction costs c.

Dow and Gorton's OLG model can be easily extended to a setting with endogenous information acquisition. Obviously, traders will be unwilling to purchase long-run information. Herding in information acquisition might occur if traders have to choose between different short-run information referring to the same dividend payment at t^{div}, for example, between an imprecise signal $S^A_{t,T}$ and an imprecise signal $S^B_{t,T}$. On the other hand, traders with long horizons would not herd. Agents are, however, endogenously myopic if they have to pay a "cost of carry" in each period instead of the one-time transaction cost c.

6.2.2. *Unwinding due to Risk Aversion in Incomplete Markets Settings*

The short livedness assumed in Froot, Scharfstein, and Stein (1992) induce the traders to unwind their position early. In Hirshleifer, Subrahmanyam, and Titman (1994) and Holden and Subrahmanyam (1996) informed traders have long horizons but they want to unwind their position for risk-sharing purposes after their information is revealed. This implicitly makes them partly myopic, that is, they care about both the intermediate price and the fundamental value.

Hirshleifer, Subrahmanyam, and Titman (1994) show that herding in information acquisition occurs under certain parameter values in their competitive REE model. After they have decided which information to collect, a continuum of competitive risk averse traders receive their signal accidentally early or late. Before focusing on the information acquisition decision, Hirshleifer, Subrahmanyam, and Titman (1994) derive interesting results pertaining to the investors' trading pattern. For the time being, let us consider the case where all risk averse investors search for the *same* information δ about the liquidation value v of a single risky asset. Let

$$v = \bar{v} + \delta + \epsilon,$$

where $\bar{v}$ is known and $\delta \sim \mathcal{N}(0, \sigma_\delta^2)$ and $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$ are independently distributed. Some investors, whose mass is M, receive information δ accidentally early, that is, already in $t = 1$, whereas the others, whose mass is $(N - M)$ are informed later. Both groups of traders receive the same information δ, but at different times. All traders maximize CARA utility functions of the final wealth W_3, that is, $U = -\exp(-\rho W_3)$. The demand for the risky asset by the early-informed is denoted by $x_t^e(\delta, \cdot)$, whereas that by the late-informed is $x_t^l(\cdot, \cdot)$. The aggregate demand of liquidity traders is modeled by the random variables $u_1 \sim \mathcal{N}(0, \sigma_{u_1}^2)$ in $t = 1$ and $\Delta u_2 \sim \mathcal{N}(0, \sigma_{\Delta u_2}^2)$ in $t = 2$.[11] Finally, there is also a group of risk neutral competitive market makers (such as scalpers and floor brokers) who observe the limit order book, that is, the noisy aggregate demand schedules, but not the information δ. The noisy aggregate demand function is $X_1(\cdot) = Mx_1^e(\delta, \cdot) + (N - M)x_1^l(\cdot) + u_1$ in $t = 1$ and $X_2(\cdot) = Mx_2^e(\delta, \cdot) + (N - M)x_2^l(\delta, \cdot) + u_1 + \Delta u_2$ in $t = 2$. Given risk neutrality and competitiveness of the market makers, the market makers set a semi-strong efficient price with respect to their information sets, that is, $P_1 = E[v \mid X_1(\cdot)]$ and $P_2 = E[v \mid X_1(\cdot), X_2(\cdot)]$.

In equilibrium, investors conjecture the following linear price relations:

$$P_2 = \bar{v} + a\delta + bu_1 + c\Delta u_2$$
$$P_1 = \bar{v} + e\delta + fu_1.$$

The equilibrium is derived by backward induction. At $t = 2$ both groups of investors, early and late informed, know δ and, therefore, their stock holding is as usual

$$x_2^e(\delta, P_2) = x_2^l(\delta, P_2) = \frac{\bar{v} + \delta - P_2}{\rho\sigma_\epsilon^2}.$$

[11] All demand functions are expressed in stock holdings, therefore the additional demand in $t = 2$ is given by $\Delta u_2 := u_2 - u_1$.

At $t = 1$ only the group of early-informed investors knows δ. Their stock holding is

$$x_1^e(\delta, P_1) = \frac{E[P_2|\mathcal{F}_1^e] - P_1}{\rho}\left[\frac{1}{\text{Var}[P_2|\mathcal{F}_1^e]} + \frac{1}{\sigma_\epsilon^2}\right] + \frac{\bar{v} + \delta - E[P_2|\mathcal{F}_1^e]}{\rho\sigma_\epsilon^2}.$$

The demand of early-informed traders consists of two components. The first term captures the speculative demand due to an expected price change. The second term is the expected final stock holding which the early-informed traders try to acquire at the "on average" better price P_1. Investors who receive their signal only at $t = 2$ demand nothing at $t = 1$, that is, $x_1^l = 0$. This is due to the fact that they do not have superior information as compared to the market makers in $t = 1$. Since the market makers are risk neutral (1) no risk premium is offered and (2) the expected P_2 is unbiased. In other words, risk averse late-informed traders cannot hedge their period 2 demands already at $t = 1$.

There are five equilibrium configurations for the coefficients of the price relations in this economy. No trading occurs in the fully revealing equilibrium. In addition, there are two equilibria where prices do not move, that is, $P_1 = P_2$. Hirshleifer, Subrahmanyam, and Titman (1994) focus on the remaining two equilibria in which trading occurs and the price is not the same in both periods. In these equilibria, both price changes $(P_1 - P_0)$ and $(P_2 - P_1)$, are positively correlated with δ. On average P_2 reveals more about δ than P_1. This is due to the fact that the market makers' information set, which determines the price, improves when two noisy aggregate demand curves are observed. Both aggregate demand curves depend on information δ. Since Δu_2 is independent of u_1, the correlation between u_1 and u_2 eases the inference of δ from both demand curves. However, the price changes, $(P_1 - P_0)$ and $(P_2 - P_1)$, themselves are uncorrelated and thus prices follow a martingale process given the market makers' filtration.

The trading behavior of the early-informed investors exhibits speculative features. They take on large positions in $t = 1$ and "on average" partially unwind their position in $t = 2$ at a more favorable price P_2. More precisely, their trading in $t = 1$, x_2^e, is positively correlated with the price change $(P_2 - P_1)$ in $t = 2$. However, their trading in $t = 2$ is negatively correlated with this price change. Therefore, these investors partially unwind their position and realize capital gains "on average." The intuition for this result is as follows. No risk premium is paid since the market makers' are risk neutral. Thus, risk averse traders would be unwilling to take on any risky stock position in the absence of any informational advantage. Early-informed investors have an informational advantage since they receive the signal δ in $t = 1$ and, hence, they are willing to take on some risk. Their informational advantage, together with the existence of noise traders, compensates them for taking on the risk represented by the random variable ϵ. However, the informational advantage of early-informed traders with respect to the late-informed traders vanishes in $t = 2$ for two reasons. First, late-informed traders receive the same signal δ. Thus, early-informed traders share the risk with late-informed traders in $t = 2$, that is, $\mathrm{Cov}(x_2^l, x_2^e) > 0$. Second, the informational advantage of the early-informed traders with respect to the market makers shrinks as well, since market makers can observe an additional limit order book at $t = 2$. This limit order book carries information for the market makers, especially since the stock holding of the noise traders is correlated in both periods. This allows the market makers to get a better idea about δ and, thus, P_2 should be "on average" closer to $\bar{v} + \delta$ than P_1. In period two, both these effects cause early-informed traders to partially unwind the position they built up in the previous period. The unwinding behavior of early-informed traders in this sequential information arrival model also stimulates trading volume.

The fact that early-informed traders on average unwind their position in $t = 2$ is in sharp contrast to models based on Kyle (1985). In these models the risk neutral insider tries to buy the stocks in small pieces in order to hide behind noise trading, that is, his stock holding over time is positively correlated. However, Brunnermeier (1998) shows in a Kyle (1985) setting with a more general information structure that speculative trading by a risk neutral insider can also arise for strategic reasons. This is in contrast to Hirshleifer, Subrahmanyam, and Titman (1994) where speculative trading is only due to investors' risk aversion.

Having analyzed the trading stage, Hirshleifer, Subrahmanyam, and Titman (1994) show that herding can occur in the information acquisition stage. At the time when they decide which information to collect, traders do not know whether they will find the information early or late. The authors derive expressions for utility levels of the early-informed and late-informed individuals. The authors then provide a numerical example in which the ex-ante utility before knowing when one receives the information is increasing in the total mass of informed traders. If this is the case, it is worthwhile for traders to concentrate on the same informational aspects, that is, gather information about the same stocks. In other words, traders will herd in information acquisition.

Whether the ex-ante utility of a higher mass of informed traders really increases depends on the parameters, especially on σ_ϵ^2. There are three main effects: (1) Increasing the mass of informed traders leads to more late-informed traders. This makes it easier for early-informed traders to unwind larger positions in $t = 2$. There are more traders in $t = 2$ that are willing to share the risk resulting from ϵ. (2) This, however, is disadvantageous for the late-informed traders since there is tougher competition among them and the extent of noise trading does not change. (3) Increasing the mass of informed traders also increases the number of early-informed traders. This decreases the utility of both early-informed and late-informed traders. In order to obtain herding, the former effect has to outweigh the latter two. This requires that σ_ϵ^2 is sufficiently high. The authors try to extend their analysis by introducing some boundedly rational elements. This extension lies outside the scope of the current literature survey.

Less Valuable Long-term Information due to Unexpected Intermediate Price Moves

In Hirshleifer, Subrahmanyam, and Titman (1994) all traders search for the *same* piece of information which they randomly receive earlier or later. In contrast, in Holden and Subrahmanyam (1996) traders can decide whether to search for *short-term* information or for *long-term information*. They choose between two signals which are reflected in value at different points in time. Holden and Subrahmanyam (1996) show that under certain conditions all risk averse traders focus exclusively on the short-term signal. Trading based on long-term information has the disadvantage that unexpected price changes can occur before the collected long-term information is fully reflected in the price.

The liquidation payoff of a single risky asset in their model is given by

$$v = \bar{v} + \delta^{\text{short}} + \eta + \delta^{\text{long}} + \epsilon,$$

where δ^{short}, η, δ^{long}, and ϵ are mutually independent normally distributed and $\bar{v}$ is normalized to zero without loss of generality. Traders who acquire short-term information observe δ^{short} at $t = 1$. At $t = 2$, δ^{short} as well as η becomes publicly known and thus they are fully reflected in the price P_2. No trader receives a signal about η in $t = 1$. δ^{long} and ϵ are made public in $t = 3$. Consequently, they are only fully incorporated in the price P_3 in $t = 3$. The long-run information signal reveals δ^{long} to the informed trader in $t = 1$. Note that the markets are incomplete since the components of v cannot be traded directly. This assumption is essential for the analysis.

A competitive REE model is employed as in Hirshleifer, Subrahmanyam, and Titman (1994). A mass of M long-term informed traders and a mass of $N = 1 - M$ short-term traders submit limit orders to the limit order book. The aggregate order size of the liquidity traders is random and is given by u_1 in $t = 1$ and Δu_2 in $t = 2$. A group of risk neutral market makers observes only the publicly available information and the noisy aggregate demand schedule, that is, the limit order book. Like in Hirshleifer, Subrahmanyam, and Titman (1994) the market makers act competitively and they are risk neutral. Hence, their information sets determine the prices.

Analyzing the equilibrium backwards, the mass of short-term traders, N, and of long-term traders M, is kept fixed at the second stage and is endogenized at the first stage. Backward induction is also applied within the trading subgame for deriving the optimal stock holdings of informed risk averse traders. At $t = 2$, the stock holding demand is standard for the long-term informed traders,

$$x_2^l = \frac{\delta^{\text{long}} + \delta^{\text{short}} + \eta - P_2}{\rho\sigma_\epsilon^2}$$

and for the short-term informed traders,

$$x_2^s = \frac{E[\delta^{\text{long}}|\mathcal{F}_2^s] + \delta^{\text{short}} + \eta - P_2}{\rho[\sigma_\epsilon^2 + \text{Var}[\delta^{\text{long}}|\mathcal{F}_2^s]]} = 0.$$

$x_2^s = 0$, since the market makers have the same information set as the short-term-informed traders and, therefore, the numerator in the above equation is zero. In economic terms, it would not make a lot of sense for risk averse short-term investors to hold risky stocks if the risk neutral market makers have the same information. Since x_2^s is zero, x_1^s is the same as in a myopic setting:

$$x_1^s = \frac{E[P_2|\mathcal{F}_1^s] - P_1}{\rho\,\text{Var}[P_2 \mid \mathcal{F}_1^s]}.$$

Short-term informed traders try to exploit the expected price change $(P_2 - P_1)$ and they close their position at $t = 2$. Long-term traders' stock holding at $t = 1$ is

$$x_1^l = \frac{E[P_2|\mathcal{F}_1^l] - P_1}{\rho\mathcal{S}_1} + \varrho E[x_2^l|\mathcal{F}_1^l],$$

where $\mathcal{S}_1$ and ϱ are nonstochastic quantities.

Holden and Subrahmanyam (1996) derive the REE only for a special case and continue their analysis with numerical simulations. In equilibrium, long-term traders reduce their period 1 demand if the variance of η is very high. η's realization is announced at $t = 2$. Early-informed traders do not want to expose themselves to the announcement risk generated by η (which is reflected in P_2). They engage in heavier trading after a large part of the uncertainty about the asset's value is resolved.

Holden and Subrahmanyam (1996) endogenize M and, thus, $N = 1-M$. The equilibrium mass M can be derived by comparing the ex-ante utilities of short-term informed traders with the utility of long-term informed traders. They show that for certain cases the ex-ante utility from collecting short-term information is higher for $M \in [0,1]$ than the utility from gathering the long-term signal. Thus, all traders search for the short-term signal in equilibrium. This is the case if the traders are sufficiently risk averse and σ_ϵ^2 is substantially high. Intuitively, short-term informed investors can only make use of their information from the price change $(P_2 - P_1)$ provided there are noise traders in $t = 1$ distorting P_1. Since η makes P_2 risky, high variance in η reduces their aggressiveness. Long-term informed traders can exploit their information from both price changes, $(P_2 - P_1)$ and $(P_3 - P_2)$. As described above, high variance of η makes long-term informed agents delay their purchase. Therefore, they are more active at $t = 2$ and they exploit $(P_3 - P_2)$ to a greater degree. If the variance of ϵ is very high, that is, speculating at $t = 2$ is very risky, long-term informed traders are very cautious at $t = 2$. Thus, they cannot make as much money out of their information as short-term informed traders can.

Holden and Subrahmanyam (1996) further show that as the degree of liquidity trading increases, both types of information are more valuable. Short-term investors profit more from higher variance in noise trading, at least for the case where it is the same in both periods.

The authors also address the question of whether long-term information can be made more valuable by making it short-term. In other words, is it profitable for long-term informed investors to disclose their information already in $t = 2$? The impact of early credible disclosures is discussed in the last section of their paper.

6.2.3. *Unwinding due to Principal–Agent Problems*

A wealth constrained trader who has discovered a profitable trading strategy might have to borrow money in order to trade on his superior information. However, the lending party might fear that the trader could default on loan repayment. The trader might be overconfident and his trading strategy might not be as profitable as he claims. In order to reassure the lender, the trader has to signal in the early stages that his trading strategy is paying off. If this is (accidentally) not the case, the lender will withdraw his money and the trader will be forced to liquidate his position early. Consequently, the trader will care a lot about short-term price movements.

Portfolio delegation leads to a similar principal–agent problem. It leads to a principal–agent relationship between the individual investor and the fund manager. Many individual investors delegate their portfolio management to fund managers. The share of investments undertaken by institutional investors is steadily increasing. Pension funds, mutual funds, as well as hedge funds are becoming predominant players in both the stock market and foreign exchange market. These professional traders conduct the bulk of informed trading.

It is very hard for an individual investor to find out whether a certain fund manager is really able to make extra profits. Bhattacharya and Pfleiderer (1985) show that optimal incentive contracts for the remuneration of fund managers might alleviate this problem by screening good from bad managers. Nevertheless, a linear remuneration contract is often the optimal one and full screening is not possible. Portfolio delegation might also induce managers to "churn bubbles" as shown in Allen and Gorton (1993).

The threat of early withdrawal of their funds is a much more powerful device for individual investors than is designing the optimal ex-ante remuneration contract. The fund manager might then be forced to liquidate part of his acquired position. The power of early withdrawal of funds changes the fund managers' incentives dramatically. Shleifer and Vishny (1990, 1997) show that it limits traders' ability to exploit arbitrage opportunities and thus has a profound impact on the assets' price process. Paradoxically, a good manager is most likely to be forced to liquidate his position when it is most profitable to extend the arbitrage opportunity.

Limits of Arbitrage

In Shleifer and Vishny (1997) only liquidity traders and fund managers are active in the stock market. Individual investors do not trade directly. They entrust their money F_1 to a fund manager who trades on their behalf. The fund manager's ability to pick the right stocks is not known to the investors. Good fund managers have found a riskless arbitrage opportunity. They know the fundamental value v of the stock with certainty. Bad fund managers have no additional information and just want to gamble with others people's money. Investors cannot screen the good managers from the bad ones, by assumption. There are two trading rounds, $t = 1$ and $t = 2$. In period $t = 3$ the true value of the stock v is common knowledge and the price adjusts accordingly to $P_3 = v$. The price in $t = 2$, P_2, in this limit order model is determined by the aggregate demand from fund managers and liquidity traders. The fund manager faces a liquidation risk in $t = 2$. Individual investors can withdraw their funds conditional on P_2.

Shleifer and Vishny (1997) focus on the case where (1) investors have entrusted their money to a "good" fund manager, and (2) the asset price goes even further down in $t = 2$ even though the asset was already undervalued in $t = 1$, that is, $P_1 < v$. This is due to sell orders submitted by the uninformed liquidity traders in $t = 2$. In the eyes of the individual investors, the additional price drop can be the result of three factors: (1) a random error term, or (2) sell orders by liquidity traders, or (3) sell orders by other informed traders in the case that the true value of the stock is lower. If the latter case were true, then the fund manager would have made the wrong decision and most probably he has no extraordinary skills to find arbitrage opportunities. Given that the individual investors can only observe the price process, it is rational for them to conclude that they probably gave their money to a bad fund manager. Consequently, they will withdraw some of their money. Shleifer and Vishny (1997) assume in their reduced form model that the fund size in $t = 2$ is $F_2 = F_1 - aD_1(1 - P_2/P_1)$, where D_1 is the amount of money the fund manager invested in the stock. The higher the coefficient a is, the more sensitive are individual investors to past performance. If the price does not change, the money in the fund remains constant. If the price increases, even more investors provide money to the fund, that is, $F_2 > F_1$. But in the case where the arbitrage opportunity becomes even more profitable, that is, when $P_2 < P_1$, investors withdraw money for fear of having entrusted their money to a bad fund manager. If the fund manager fully exploited the arbitrage opportunity, that is, he invested the whole fund into the stock, $D_1 = F_1$, he is forced to unwind part of his position although he is sure that the price will come back in $t = 3$. He incurs a loss by unwinding his position at an even lower price. Knowing that the investors will withdraw some money if the price goes down in $t = 2$, the fund manager will invest only part D_1 of the fund F_1 in the undervalued asset in $t = 1$. In general, the fund manager does not fully exploit the arbitrage opportunity. He will only invest the whole fund F_1

if the mispricing is very large thus making it very unlikely that the price will go down further.

This shows that even pure long-run arbitrage opportunities are risky since investors might withdraw their money early. Fund managers face the risk of interim liquidation. Pure arbitrage opportunities are very rare in reality and traders mostly discover expected arbitrage opportunities. Therefore, risky arbitrage is not only risky for fund managers because they cannot exactly replicate the payoff stream but also because they face an "early liquidation risk."

The consequence is that fund managers search for less risky arbitrage opportunities. In order to minimize the "early liquidation risk," they can either concentrate their research efforts on short-run information which will be made public very soon, or on information which is the focus of sufficiently many other arbitrageurs. This makes it more likely that information is reflected in the price soon. Professional arbitrage is concentrated in a few markets like in the bond market and foreign exchange market but is hardly ever present in the stock market. This is the same "arbitrage chain" argument which is formalized by Dow and Gorton (1994). Given that fund managers focus only on short-run arbitrage opportunities, long-run assets, whose positive dividend payoffs will be in the far future, are more mispriced in equilibrium. No fund manager will exploit long-run arbitrage opportunities out of fear that he has to liquidate the position early when individual investors withdraw their funds. Put differently, long-run arbitrage opportunities must provide much higher returns than short-run arbitrage opportunities in order to compensate for the additional liquidation risk. This might also explain why stock market returns – contrary to what the capital asset pricing model (CAPM) suggests – do not only depend on systematic risk but also on idiosyncratic risk. The risk of wrong intermediate price movements makes arbitrage trading less attractive and thus must lead to higher returns.

Induced Collection of Short-Run Information

Gümbel (1999) explicitly models the principal–agent relationship and its implication in the stock market. He shows that the individual investors actually prefer that fund managers primarily search for *short-term information* and exploit short-term arbitrage opportunities. This allows the investors to quickly infer the manager's ability and to lay off an unable manager.

In Gümbel (1999) the risk neutral investor delegates his investment decision to a risk neutral fund manager whose ability to choose the right trading strategy is unknown. There are two underlying risky assets in this economy which pay a dividend of either 0 or 1 in each period $t \in 0, 1, 2, \ldots, \infty$. Let us assume for illustrative purposes that the individual dividend payments are securitized and are traded. In addition, there are traded bonds whose fixed return is r. The fund manager can gather either *short-term* information *or long-term information* without cost. He always receives one noisy signal {up, down} for each of the two stocks. Short-term information provides two noisy signals about the dividend payments of both assets in the next period $t + 1$, whereas long-term information provides two noisy signals about the dividend payments of both assets in $t + 2$.

There is a pool of potential fund managers, who invest on behalf of the investor. A fraction γ has high ability and the rest is of low ability. In contrast to Shleifer and Vishny (1997) neither the principal nor the fund managers know their type and both learn the manager's type at the same speed. Fund managers receive one signal for each of the two stocks. Each signal's realization is either "up" or "down." Bad fund manager's signals are always correct for one stock and incorrect for the other one. Either the signal for stock A is correct and the one for stock B is incorrect or vice versa with equal probability. Good fund managers' signals have the same structure with probability $(1 - \nu)$. However, with probability $\nu(\mu)$, their short-term (long-term) signals $S_t^{\text{short},j}(S_t^{\text{long},j})$ are correct for both assets $j \in 1, 2$. The trading game for each asset is a binary version of Kyle (1985). Liquidity traders in both markets as well as informed fund managers submit market orders to the market makers. The market makers only observe the aggregate order flows X_t^j of the asset j and set informationally efficient prices P_t^j. The liquidity trader submits a random order of fixed size $-x$ or $+x$ with equal probability. Whether the fund manager submits a buy or sell order depends on his signal. As long as the probability that he is of high type is sufficiently high, he will submit a buy (sell) order if he gets a positive (negative) signal. In order to disguise his order behind the liquidity traders' orders, his order size is also either $-x$ or x. The market maker could immediately identify any other order size as an order originating from the fund manager. The aggregate order flow is thus $-2x$, 0, $+2x$. If the aggregate order flow is $-2x$ or $2x$, the market maker can perfectly infer the fund manager's information. The market maker cannot figure out whether the manager submitted a buy or sell order only if the aggregate order flow is zero. Only in this case does the fund manager make a nonzero trading profit. This feature of the model simplifies the analysis.

The fund manager's trading activity depends on whether the fund manager has collected long-term information or short-term information. In the case that the manager is induced to collect *long-run information* about the dividend payments in $t+2$, he submits an order in t. This order will be executed at the price P_t, which the market maker sets based on the observed aggregate order flow in t. The market maker receives a private signal signal $S_{t+1}^{mm,j} \in \{\text{up}, \text{down}\}$ after he has executed the order at the price P_t. This signal predicts the correct d_{t+2} with probability $q \in [\frac{1}{2}, 1]$. The fund manager has the opportunity to unwind his acquired position prior to trading in $t+1$. Unwinding perfectly reveals his signal to the market maker. Consequently, the "unwinding price" is determined by the fund manager's information together with the market maker's signal $S_{t+1}^{mm,j}$. The manager is indifferent between unwinding and holding the asset until it pays the dividend in $t+2$. This is because the competitive risk neutral market maker sets the (semi-strong) informationally efficient price and the manager has the same information as the market maker about d_{t+1}. That is, they expect d_{t+1} to be zero. De facto, a fund manager with long-run information trades an asset in t whose "unwinding value" prior to trading in $t+1$ is

$$\frac{1}{1+r}E[d_{t+2}^j | S_{t+1}^{mm,j}, q, S_t^{\text{long},j}].$$

The informational advantage for the manager with respect to the market maker in t results from his knowledge of $S_t^{\text{long},j}$. Note since d_{t+2} is only paid out in $t+2$, the unwinding value has to be discounted by one period. $S_{t+1}^{mm,j}$ generates an additional noise term for the fund manager's "unwinding price" and thus does not affect the manager's expected profit.

Managers who gather *short-term information* trade an asset in t whose value in $t+1$ is d_{t+1}^j. The fund manager's best estimate in t is $E[d_{t+1}^j | S_t^{\text{short},j}]$. The manager's informational advantage is, however, smaller since the market maker also holds some information about d_{t+1} prior to trading in period t. This is because (1) the market maker received a private signal $S_t^{mm,j}$ about d_{t+1}^j, and (2) he might have learned something from other fund managers who unwind the long-term position that they acquired by observing the signal about d_{t+2} in t.[12]

In summary, long-run information is advantageous for the manager since the market maker does not know it yet, i.e. he has not observed

[12] This will not occur in equilibrium since the fund manager will gather short-term information in equilibrium.

the signal $S_{t+1}^{mm,j}$ yet. On the other hand, short-run information of good fund managers is assumed to be more precise, that is, $\nu > \mu$, and one trades an asset whose dividend of 0 or 1 is paid out in $t+1$ rather than $t+2$. This reduces the loss from discounting. Proposition 1 of the paper shows that a high type manager trades more profitably with long-run information if $\mu > \nu(1+r)4q(1-q)$.

The decision whether to gather short-run or long-run information not only affects the direct trading profits, but also affects how quickly one learns the manager's ability. Short-run information not only has the advantage that it is more precise since $\nu > \mu$ but it also provides the principal a better update about the manager's ability already in $t+1$. This again influences the employment decision of the principal, that is, when to fire the manager and hire a new agent from the pool of potential managers. If the manager traded in the right direction for both assets, he is of high quality with probability one, since a bad manager always trades in the wrong direction for at least one asset. If he has traded in the wrong direction for one asset, it is more likely that he is a bad manager.[13] If one of the manager's two first trades is wrong, it is better for the principal to replace him with a new manager from the pool.

If the manager collects long-run information, the principal's ability to evaluate the agent in $t+1$ by observing his unwinding decision depends on the quality q of the market maker's signal, $S_{t+1}^{mm,j}$. Let us consider the two polar cases $q = \frac{1}{2}$ and $q = 1$. If $q = \frac{1}{2}$ the market maker's signal is worthless. Since $S_{t+1}^{mm,j}$ has no informational content, the market maker only learns the fund manager's signal if he unwinds it prior to trading in $t+1$. He cannot evaluate whether the manager received a correct long-term signal or not. If $q = 1$ the market maker receives a perfect signal about the dividend payment in $t+2$. Hence, he can infer whether the manager received a correct long-term signal or not. If he has received such a signal, then he is for sure of high ability; if not, it might still be the case that he received bad information because he was unlucky. Note that since $\nu > \mu$, it is more likely that a good manager who gathers long-run information is unluckier than one who gathers short-run information. Nevertheless, trading in the wrong direction makes it more likely that he is a bad manager and thus the principal fires him and hires a new manager from the pool. Note that a higher q makes long-run information more attractive for two reasons: (1) it allows a quicker evaluation of the manager's ability, and (2) it makes the short-run information

[13] Note that if the manager himself knows that he is of low ability, his trades would always contradict one of his signals.

less valuable since the market maker already knows part of the private information that the fund manager will collect.

The paper assumes that the decision to collect long-run versus short-run information is contractible and thus is decided by the principal. The main result of this paper is that for certain parameter values, learning about the manager's ability induces the principal to search for short-term information even though long-term information would be more valuable. Short-run information allows the principal to dismiss bad managers early. Focusing exclusively on short-run information leads to long-run mispricing.

6.3. Firms' Short-Termism

Mispricing of assets is not very harmful if it does not affect the real decision making within firms. This section illustrates that short-sightedness of investors leads to short-termism in firms' investment decisions.

Shleifer and Vishny (1990) argue convincingly that managers care about the stock price of their company. Corporate managers' remunerations are very closely linked to the stock price via stock options. They risk being fired because of a possible take-over if the company's equity is underpriced. Corporate managers have a vital interest that their investment decisions are reflected correctly in the stock price. Investors' focus on short horizons leads to systematically less accurate pricing of long-term assets, for example, stocks of firms whose investment projects only lead to positive return in the far future. Corporate managers who are averse to mispricing, therefore, focus on short-term projects.

In Brandenburger and Polak (1996) managers ignore their superior information and follow the opinion of the market. The market can observe the corporate manager's action and try to infer the manager's superior information, which is then reflected in the stock price. Since the manager cares about the short-run stock price, he has an incentive to manipulate his action and thus the market's inference. The result is that the corporate manager does not follow his superior information in equilibrium.

In the first part of Brandenburger and Polak (1996), a single risk neutral manager has to choose between action L(left) and R(right). The payoff of his action depends on the state of the world. In state λ, action L pays off \$1 and action R pays nothing. In state ρ, the payoff structure is exactly the opposite. Action L's payoff is 0 and action R pays off 1. The true state is ρ with prior probability $\pi > \frac{1}{2}$. The prior distribution might reflect public (short-run) information which is known to the whole market. The manager receives an additional signal $S^i \in \{l, r\}$ which tells him the true state with precision $q > \pi$, that is, $q = \Pr(l|\lambda) = \Pr(r|\rho)$. Since the signal is more precise than the prior, a manager who maximizes the long-run value of his company should follow his signal. However, it takes a while until the true payoffs are realized and reflected in the stock price. In the meantime, the market tries to infer the manager's signal and updates the short-run market price. If the manager could truthfully announce his signal to the market, he would always follow his signal and the market price would adjust accordingly. The trading game is such that the price reflects the posterior probability of the market. Note that the market price would be higher if the manager received signal r instead of signal l. This is due to the biased prior $\pi > \frac{1}{2}$.

In Brandenburger and Polak (1996) the manager cannot truthfully announce his signal. The market participants try to infer the signal from the manager's observed action R or L. However, there exists no pure strategy equilibrium in which the manager would follow his signal. If such an equilibrium existed, then the market participants would believe that the manager's strategy is to always follow his signal. Therefore, they would think that they can perfectly infer the manager's signal from his action. Consequently, they would update the stock price accordingly. The stock price after observing action R would be higher than that after observing L. This occurs because of the bias in the prior $\pi > \frac{1}{2}$. Since the manager cares about the current stock price, he has an incentive to deviate from the strategy that always chooses action R. Always choosing R is indeed the best BNE in pure strategies. The manager ignores his signal completely and – since in equilibrium the market participants know this – the stock price reflects the fact that the manager's action is always R. Even though the stock price is informationally efficient, the manager's decisions are clearly (allocatively) inefficient.

There are, however, mixed strategy equilibria in which the manager at least partly uses his information. The manager ignores part of his information since he sometimes chooses R even though he has received signal l. In the mixed strategy equilibria, the market participants know which strategy the manager applies but they cannot fully infer his signal. Mixed strategies can, therefore, be thought of as "garblings" of signals.[14] Traders can partly infer the manager's signal. The mixing probabilities have to be such that the market participants' posteriors

[14] Note the similarity to Crawford and Sobel (1982). In Crawford and Sobel (1982) the sender of the message cares about the receivers' opinion since it affects his action. In

that the manager has chosen the right action are the same. In other words, the market conjecture is such that the short-run stock price does not depend on the action of the manager. Consequently, the manager is indifferent between both actions in equilibrium and has no incentive to deviate from his mixed strategy. The bias in the unbalanced prior $\pi > \frac{1}{2}$, which drives the nonexistence result of informative pure strategy equilibria, has to be counterbalanced by the mixed strategy. Observing an action L has to be a stronger indication of signal l than observing R is for signal r. The stronger the bias, the more mixing is necessary and thus the higher the loss of information. The key is actually not the skewness or bias of the prior but the fact that the two decisions yield unequal posteriors about the expected profit of the firm.

In the second part of the paper, a dynamic model is introduced. Many firms receive a signal about the state λ or ρ and have to sequentially choose action L or R. Informational cascades like in the herding model à la Bikhchandani, Hirshleifer, and Welch (1992) arise. One might suspect that by applying mixed strategies the information aggregation problem due to herding might be alleviated. On the contrary, Brandenburger and Polak (1996) show that with share price maximization, equilibrium choices are strictly less efficient than under herding behavior. The successors can infer less information from their predecessor's decision but it is still optimal for them to herd on the inferred information and to disregard their own private signal.

There are numerous other papers dealing with short-termism of firms induced by the stock market. Grant, King, and Polak (1996) provide a good survey of this literature.

6.4. Bank Runs and Financial Crisis

Bank runs and bank panics are special forms of herding behavior. A bank run occurs when the deposit holders of a bank suddenly withdraw their money. If a run on a single bank spreads over to other banks, it can cause a panic in the whole banking system. Strong spillover effects can lead to contagion where many banks get into solvency problems.

This section focuses solely on the herding aspect of bank runs and thus ignores a large part of the banking literature. Interested readers are directed to Freixas and Rochet (1997) for a comprehensive coverage of

Brandenburger and Polak (1996) the sender cares about the action and thus the market participants' opinion because it affects the short-run stock price.

the banking literature. Although withdrawals by deposit holders occur sequentially in reality, the literature typically models bank runs as a simultaneous move game. An exception is Chen (1999) who explicitly models a bank run in a sequential setting.

Banks as Liquidity Insurance Providers

One role of banks is to transform illiquid technologies into liquid payoffs, and also to provide liquidity insurance. Diamond and Dybvig's (1983) seminal paper illustrates this role of banks and builds on initial insights presented in Bryant (1980). In their model, banks offer demand deposits to match the agents' liquidity needs with projects' maturities. However, these demand deposits open up the possibility of bank runs.

In Diamond and Dybvig (1983) there are two technologies in which money can be invested for future consumption: an illiquid technology and a storage technology. The illiquid technology is a long-run investment project that requires one unit of investment. It can be liquidated early in $t = 1$ at a salvage value of $L \leq 1$.[15] If one carries on with the project until $t = 2$, the project pays off a fixed gross return of $R > 1$. In addition to the productive long-run investment project, agents also have access to a costless storage technology. Agents can devote a fraction of their endowment to the illiquid investment project and store the rest in the costless storage technology. The savings opportunities are summarized in Table 6.1.

There is a continuum of ex-ante identical agents who have an endowment of one unit each. Each agent faces a preference shock prior to $t = 1$. Depending on this shock, each agent consumes either in $t = 1$ or in $t = 2$. They are either "early diers," who consume in $t = 1$ or

Table 6.1.

Investment projects	$t = 0$	$t = 1$	$t = 2$
Risky investment project			
(a) continuation	-1	0	$R > 1$
(b) early liquidation	-1	$L \leq 1$	0
Storage technology			
(a) from $t = 0$ to $t = 1$	-1	$+1$	
(b) from $t = 1$ to $t = 2$		-1	$+1$

[15] Diamond and Dybvig (1983) restrict their analysis to $L = 1$. To illustrate the utility improving role of asset markets, we consider the more general case of $L \leq 1$.

"late diers," who consume in $t = 2$. In other words, early diers derive utility $U^1(c_1)$ only from consumption in $t = 1$, whereas late diers derive utility $U^2(c_2)$ only from consumption in $t = 2$. Since the agents do not know ex-ante whether they will die early or not, they would like to insure themselves against their uncertain liquidity needs.

Without markets or financial intermediaries each agent would invest x in the long-run investment project and store the rest $(1 - x)$. Early diers who liquidate their project consume $c_1 = xL + (1 - x) \in [L, 1]$, while late diers consume $c_2 = xR + (1 - x) \in [1, R]$. The ex-ante utility of each agent is given by $qU(c_1) + (1 - q)U(c_2)$, where q denotes the probability of dying early. This utility can be improved if trading of assets is allowed in $t = 1$.

Financial markets allow agents to sell their stake in the long-run investment project in $t = 1$. In this case, the higher consumption levels $c_1 = 1$ and $c_2 = R$ can be achieved even if $L < 1$ as long as a fraction $(1-q)$ is invested in the illiquid asset on aggregate. Instead of liquidating the long-run asset in $t = 1$, early diers can sell their asset to the late diers in exchange for c_1-consumption at a price of $P = 1$. Note that the price of the asset in $t = 1$ has to be 1 in order to ensure that agents are indifferent between storage and investing in the investment project in $t = 0$.

However, the consumption pattern of $c_1 = 1$ for early diers and $c_2 = R$ for late diers is typically not ex-ante optimal since it does not provide an optimal insurance against the ex-ante risk that one can be either an early or late dier. Ex-ante optimal consumption levels must satisfy

$$\frac{\partial U}{\partial c_1}(\cdot) = R\frac{\partial U}{\partial c_2}(\cdot).$$

The allocation $(c_1 = 1, c_2 = R)$ is ex-ante optimal only for special utility functions. Within the class of HARA utility functions, this allocation is only ex-ante optimal for the log-utility function. For utility functions with a relative risk aversion coefficient, γ, larger than unity,

$$\frac{\partial U}{\partial c_1}(1) > R\frac{\partial U}{\partial c_2}(R).$$

Thus, a contract which offers $c_1 = 1$, and $c_2 = R$ is not ex-ante optimal. In other words, given $\gamma > 1$, a feasible contract $c_1^* > 1$ and $c_2^* < R$ which satisfies

$$\frac{\partial U}{\partial c_1}(c_1^*) = R\frac{\partial U}{\partial c_2}(c_2^*)$$

is ex-ante preferred to $c_1 = 1$ and $c_2 = R$.

A bank can commit itself to perform this transfer of resources from c_2 to c_1. Competitive banks offer deposit contracts (c_1^*, c_2^*) which maximize the agents' ex-ante utility. Free entry in the banking sector and the absence of aggregate risk ensures this. In equilibrium, the bank makes zero profit, invests x^* into the investment project, and stores the rest $(1-x^*)$. The stored reserves are enough to satisfy the early diers demand in $t = 1$, that is, $qc_1^* = (1 - x^*)$, while the rest is paid out to the late diers in $t = 2$. Thus, $(1 - q)c_2^* = Rx^*$.

In Diamond and Dybvig (1983) the bank can observe neither the consumer type nor his private storage activity from $t = 1$ to $t = 2$. Therefore, the bank has to provide the right incentives such that late diers do not withdraw their money early and store it for later consumption in $t = 2$. As long as only early diers withdraw their demand deposit c_1 from the bank in $t = 1$, the bank is prepared for this money outflow and does not need to liquidate the long-run asset. In this case, no late dier has an incentive to withdraw his money early and hence deposit contracts are optimal.

Bank Runs as a Sunspot Phenomenon

However, if other late diers start withdrawing money early, then the bank does not have enough reserves and is forced to liquidate its long-run projects. For each additional late dier who withdraws c_1^* units from the bank, the bank has to liquidate more than one unit. The bank promised a payment of $c_1^* > 1$, which was optimal given the deposit holder's relative risk aversion coefficient $\gamma > 1$. If the salvage value L is strictly smaller than 1, the bank has to liquidate even a larger fraction of the long-run investment project. This reduces the possible payments in $t = 2$ and thus the incentive for late diers not to withdraw their money early. Diamond and Dybvig (1983) assume that the bank must honor a *sequential service constraint*. Depositors reach the teller one after the other and the bank honors its contracts until it runs out of money. The sequential service constraint gives depositors the incentive to withdraw their money as early as possible if they think that late diers will also withdraw their demand deposits early in $t = 1$ and make the bank insolvent. This payoff externality triggers the herding behavior. The authors assume the sequential service constraint even though they formally employ a simultaneous move game. In short, there also exists a bank run equilibrium in which all agents immediately withdraw their deposits in $t = 1$ and the bank is forced to liquidate its assets. In the bank run case deposit contracts are not necessarily optimal. Whether the Pareto inferior bank run equilibrium arises or the full insurance equilibrium arises might depend on sunspots. Sunspots, as explained in Section 2.3, are commonly observed extrinsic random variables which serve as a coordination device.

Suspension of convertibility eliminates the bank run equilibrium as long as the fraction of early diers q is deterministic. If the bank commits itself to serve only the first q customers who show up to withdraw their demand deposits, no assets need be liquidated and the bank has enough money to pay c_2^*. Consequently, no late dier has an incentive to withdraw any money in $t = 1$ in the first place. In short, the anticipation of suspension of convertibility prevents bank runs.

If the fraction of early diers q is random, the suspension of convertibility does not prevent bank runs since the bank does not know when to stop paying out money in $t = 1$.[16] On the other hand, a governmental *deposit insurance* financed by an inflation tax can eliminate the bank run equilibrium even for a random q. If the deposit guarantee of c_1^* is nominal, an inflation tax that depends on early withdrawals can reduce the real value of the demand deposit. This provides the late diers with the necessary incentive not to withdraw their money early.

Jacklin (1987) shows that agents can achieve the same optimal consumption level (c_1^*, c_2^*) with dividend paying equity contracts instead of bank deposits. Furthermore, dividend paying equity contracts eliminate the Pareto inferior bank run equilibrium. However, the optimal consumption level cannot be achieved with equity contracts in a more general setting with smooth preferences where both types of agents consume in both periods.

Possibility of Information-Induced Bank Runs in a Unique Equilibrium

Jacklin and Bhattacharya (1988) compare demand deposits with equity contracts. In their model bank runs are not due to sunspots, but changes in the fundamental variables. The payoff of the long-run investment project $\tilde{R}$ is random in Jacklin and Bhattacharya (1988) and some traders receive information about $\tilde{R}$ prior to their withdrawal. In contrast to

[16] The randomness of q also affects the bank's investment decision x. In Diamond and Dybvig (1983) this has no impact since $L = 1$ and thus investing in $t = 0$ and liquidating in $t = 1$ provides the same return as storage.

Table 6.2.

Investment projects	$t = 0$	$t = 1$	$t = 2$
Illiquid risky project	-1	$L = 0$	$R = \begin{cases} R_H & \text{Pr } 1 - \theta \\ R_L & \text{Pr } \theta \end{cases}$
Storage technology			
(a) from $t = 0$ to $t = 1$	-1	$+1$	
(b) from $t = 1$ to $t = 2$		-1	$+1$

Diamond and Dybvig (1983), there is only one unique equilibrium. In this equilibrium bank runs occur in some states of the world.[17]

Another distinction between Diamond and Dybvig (1983) and Jacklin and Bhattacharya (1988) is that the latter authors assume smooth preferences $U^i(c_1, c_2) = u(c_1^i) + \beta^i u(c_2^i)$. Hence, agents want to consume a positive amount in both periods. Impatient agents put more weight on consumption in $t = 1$ and patient agents put more weight on consumption in $t = 2$, that is, $1 \geq \beta^2 > \beta^1 > 0$. Smooth preferences rule out the possibility that the optimal consumption profile can be implemented with dividend paying equity on a bank instead of demandable deposits.

The payoff structure in Jacklin and Bhattacharya (1988) is summarized in Table 6.2. The payoff structure differs from the one in Diamond and Dybvig (1983) in two ways. First, the salvage value of the illiquid investment project, L, is zero in $t = 1$. Second, the final payoff of the illiquid project R in $t = 2$ is random. The probability of a high return R_H is $(1 - \theta)$ and the probability of a low return R_L is θ. In the latter case, the bank can pay at most a fraction R_L/R_H of the maximum payment in $t = 2$. Agents learn their time preference β in $t = 1$. That is, they discover how strongly they prefer to consume the bulk of their endowment in $t = 1$ instead of in $t = 2$. A fixed fraction α of the more patient "late consumers" also receive a signal about the payoff of the illiquid project. This signal allows the informed late consumers to update their prior θ to $\hat{\theta}$.

Nonpatient consumers with low β^1 always withdraw a large fraction of their deposits from the bank in $t = 1$. Uninformed patient consumers keep their deposits with the bank, while informed patient consumers withdraw their money early if the posterior of the bad event R_L, $\hat{\theta}$, is above the threshold level $\bar{\theta}$. Jacklin and Bhattacharya (1988) show that the bank run threshold level $\bar{\theta}$ decreases as the variance of R increases.

[17] In this respect, their model is similar to Postlewaite and Vives (1987) who develop an alternative setup with a unique equilibrium over a range of parameter values.

Chari and Jagannathan (1988) analyze information induced bank runs where uninformed late consumers infer information from the aggregate withdrawal rate. In their setup, all agents are risk neutral with a utility function $U^i(c_1, c_2) = c_1 + \beta^i c_2$. Type 1 agents are early consumers and their β^1 is close to zero. Type 2 agents with high β^2 are late consumers. Risk neutrality eliminates the bank's role as a liquidity insurer. The fraction $q \in \{0, q_1, q_2\}$ of impatient early consumers is random in Chari and Jagannathan (1988). As in Jacklin and Bhattacharya (1988), a fraction α of late consumers receive a signal about the random return of the illiquid investment project $R \in \{R_\mathrm{L}, R_\mathrm{H}\}$. However, this fraction is also random with $\alpha \in \{0, \bar{\alpha}\}$. In short, in Chari and Jagannathan (1988) the fraction of impatient consumers q, the return R, and the fraction α of informed late consumers is random. In contrast to Diamond and Dybvig (1983), the authors do not assume the sequential service constraint. In their model all deposit holders arrive simultaneously and there is a *pro rata* allocation of the funds. If short-term funds are not sufficient, the bank can prematurely liquidate the long-run project. As long as the total aggregate withdrawals do not exceed some threshold $\bar{K}$ the salvage value of the long-run investment project is $L = 1$. Otherwise, premature liquidation is costly, that is, $L < 1$.

A large withdrawal of deposits can be (1) due to a large fraction of impatient consumers, that is a high realization of q, or (2) due to the fact that informed patient consumers received a bad signal about R. Since uninformed patient consumers cannot distinguish between both forms of shocks, they base their decision solely on aggregate withdrawals. Uninformed patient consumers might misinterpret large withdrawals due to a high q as being caused by a bad signal received by informed late consumers. This induces them to withdraw their funds and forces banks to liquidate their investment projects. Wrong inference by the uninformed deposit holders can lead to bank runs even when $R = R_H$. The liquidation costs erode the bank's assets and the possible payouts in $t = 2$. In Chari and Jagannathan (1988), the early withdrawal by deposit holders causes an information externality and a payoff externality. The early withdrawal sends a signal to the uninformed deposit holders that the return of the long-run asset is probably low (information externality) and also forces the bank to conduct costly liquidation (payoff externality).[18]

[18] In Gorton (1985) a bank can stop a bank run if $R = R_\mathrm{H}$. By paying a verification cost, it is able to credibly communicate the true return R_H and suspend convertibility.

Potential bank runs can also serve as a discipline device for bank managers to make the right investment decisions. Calomiris and Kahn (1991) focus on this aspect in a model with endogenous information acquisition by the deposit holders. Their analysis explains why demand deposit contracts are the dominant form of savings.

Financial Crisis

A single bank run can easily spill over to other banks. A bank panic involves runs on many banks and might lead to a collapse of the whole banking system. Bhattacharya and Gale (1987) provide a model illustrating bank panics in a setting that focuses on the role of the interbank loan market. Chen's (1999) paper illustrates contagious runs on multiple banks in a herding model where deposit holders can decide sequentially. The analysis highlights the crucial role of information externalities and payoff externalities. The latter is due to the sequential servicing constraint.

In a broader context, all these problems arise from short-run financing of long-run high-yield investment opportunities. A fund manager who invests on behalf of individual investors also faces the same problem. As discussed in Section 6.2.3, the fear of early withdrawal of funds makes him reluctant to exploit profitable long-run arbitrage opportunities.

The discrepancy of maturities between investment projects and their short-term financing might explain the scope of the financial crisis in Southeast Asia at the end of the 1990s. Bad news about the lack of an efficient corporate governance structure might have justified a certain correction. However, it triggered a significant outflow of funds from these countries due to herding behavior, as in a bank run. This resulted in a plummeting of share prices and large-scale currency devaluations, thereby forcing these countries to also liquidate useful long-run investment projects.

Radelet and Sachs (1998) contrast this reasoning with other possible causes of the recent Asian crises. Each cause leads to different predictions of the price path and requires different remedies. No measures should be taken if the crash is just a price correction, for example, the bursting of a bubble. On the other hand, if the crisis is due to herding behavior as in bank runs, capital controls are a useful device to avoid the Pareto inferior bank-run equilibrium. Policy makers who are able to differentiate between these different causes can develop the right remedies to reduce the impact of future crises and minimize the social hardship faced by large fractions of the population.

INDEX