Trading in Derivatives when the Underlying is Scarce[*]

Snehal Banerjee[†] Jeremy J. Graveline[‡]

September 2012

Abstract

When market frictions restrict the aggregate quantity of long and short positions in a security, the security becomes "scarce," and its price is distorted relative to its frictionless value. In this case, we show that even simple derivatives, exposed to the same sources of risk as the underlying security, are no longer redundant. Moreover, in equilibrium, trade in derivatives may actually affect the price of the underlying. We characterize equilibrium prices and trading volume in the underlying security and its derivative, and show that improving ease of trade in the derivative can reduce price distortions in the underlying security.

JEL Classification: G12, G13

Keywords: Scarcity, Short-selling, Price distortions, Derivatives

[*] We are grateful for comments from Darrell Duffie, Brad Jordan, Arvind Krishnamurthy, Xuewen Liu, Moto Yogo, and participants at the Minnesota Junior Finance Conference (2011) and the Financial Intermediation Research Society Conference (2012).

[†] Kellogg School of Management, Northwestern University, snehal-banerjee@kellogg.northwestern.edu

[‡] Carlson School of Management, University of Minnesota, jeremy@umn.edu

1 Introduction

In frictionless markets, the aggregate quantity of long and short positions in a security is unrestricted, and so the availability of these positions is perfectly elastic. In reality, frictions prevent long investors and short-sellers from taking arbitrarily large positions. For example, it is unlikely that an outstanding supply of 100 shares of a security can support 10,000 long positions and 9,900 short positions, even if it is extremely liquid. In some cases, investors may find it difficult, or even impossible, to sell securities short — examples include the short-selling restrictions during the financial crisis of 2008 and the recent Eurozone crisis. The theoretical literature has modeled the impact of such frictions on equilibrium prices and quantities. For instance, in Duffie (1996) and Krishnamurthy (2002), long investors are constrained from lending out their entire position to short sellers, while in Duffie, Gârleanu, and Pedersen (2002) and Vayanos and Weill (2008), both long investors and short-sellers face search costs that limit the size of their positions. Moreover, a large number of papers, dating back to Miller (1977), have studied the effect of short-selling restrictions on asset prices. In each of these models, frictions constrain the aggregate quantity of long and short positions in a security. This inelasticity can lead the security to become "scarce" and its price to become distorted relative to its value in a frictionless market.[1]

A derivative security allows long investors and short-sellers to take positions in the same source of fundamental risk as the underlying security. This naturally leads to a question that has largely been unexplored in this literature — how do derivatives affect the scarcity of an underlying security? In frictionless markets, simple derivatives are redundant securities, but we show that this is no longer the case when the underlying security can be scarce. Moreover, we show that the presence of a derivative may affect the price of the underlying security itself. Intuitively, derivatives provide a substitute for long and short positions in the underlying security and therefore make it less scarce. We develop a parsimonious model to capture this intuition, and characterize how equilibrium prices and trading volume in the underlying security and its derivative are jointly determined when the underlying may be scarce.

In our model, agents trade in order to speculate against each other and to hedge their exposure to the risk that drives the fundamental value of a security. For instance, investors

[1] A number of papers document this distortion in the price of scarce securities, in the form of a price premium paid for securities that are costly to borrow (i.e., trade on "special"). These papers include Jordan and Jordan (1997), Krishnamurthy (2002), Goldreich, Hanke, and Nath (2005), and Banerjee and Graveline (2012), which document a positive relation between prices and borrowing fees in bond markets, and Geczy, Musto, and Reed (2002) and Ofek and Richardson (2003), which document it in equity markets.

who wish to bet whether a portfolio of stocks will outperform or underperform the market have an incentive to offset their exposure to the overall stock market. As another example, risk averse market makers have an incentive to hedge risk in their inventories of securities. In the model, agents can perfectly trade the fundamental source of risk with a position in the underlying security, but the aggregate capacity for long and short positions in this security is constrained (i.e., it is not perfectly elastic). If there is no derivative, the security becomes "scarce" when the aggregate demand for long and short positions exceeds its capacity. In this case, the price of security becomes distorted relative to its fundamental value so that markets can clear.[2]

A derivative provides agents with another security with which to take long and short positions, but may offer a noisy exposure to the fundamental risk that investors are concerned about. For example, the cheapest-to-deliver option in many exchange traded derivatives introduces additional noise in their payoff, as does the fact that the secondary market for customized over-the-counter derivatives is often relatively illiquid. In the extreme, the derivative markets for some assets may not even exist (e.g., futures on individual stocks and bonds), and so investors may be forced to use derivatives on related assets. In this setting, agents trade off the higher cost of a position in the underlying security against the noisy exposure that accompanies a position in the derivative. We show that the presence of a derivative security can reduce the distortion in the price of the underlying by reducing its relative scarcity. We also show that the derivative price offers a "cleaner" measure of the frictionless value of the underlying security. Finally, we use the model to explicitly characterize the relative size of the underlying and derivative markets in equilibrium, in terms of investor preferences and beliefs, hedging incentives, and the outstanding supply of the underlying security.

The notion of scarcity is particularly relevant for assets that investors are unable to short-sell, and we explicitly consider this special case of our model in Section 3.4. In particular, our analysis is relevant for assets such as commodities and real-estate, that are inherently difficult to sell short, and for assets that face regulatory or institutional trading restrictions (e.g., financial stocks during the short-selling ban of 2008). Our theory suggests that a restriction on short-selling can make the security more scarce and distort its price, but that access to derivative securities can mitigate these distortions. Moreover, contrary to standard

[2] In this situation, the security becomes costly to borrow. Banerjee and Graveline (2012) show that long investors and short-sellers can pay differently for the security since every short position must be borrowed, but not every long position can be lent to a short-seller. This differential pricing between long investors and short-sellers is what allows the market for the security to clear. Moreover, in equilibrium, the higher cost of borrowing is impounded into a higher price for the security.

intuition (e.g., Miller, 1977), as we characterize in Section 3.4, imposing a short-sales ban may actually *lower* the price of the underlying asset, even when investors have access to derivative securities.

Our main model considers a single period setup in which investors have CARA utility, asset payoffs are normally distributed, and the constraint on borrowing and lending is exogenously fixed — these assumptions are made primarily for tractability and to highlight the intuition for our results. In Section 4 we show that our main results are robust to relaxing these assumptions. In Section 4.1, we characterize weaker sufficient conditions on preferences and payoff distributions under which the presence of a derivative security reduces the price distortion in the underlying. For instance, we show that when demand for both the underlying and derivative is downward sloping, and demand for the underlying (derivative) is more sensitive than the derivative (underlying) to changes in the price of the underlying (derivative, respectively), then the price distortion in the underlying is smaller in the presence of a derivative security. Moreover, we show that with the assumption of CARA utility, our main conclusion holds even if we relax the assumption that payoffs are normally distributed.

In Section 4.2, we endogenize the constraint on borrowing and lending. Specifically, we assume that long investors choose the optimal fraction of their position to lend out, given a non-decreasing, convex (per-unit) cost function. In this case, long investors optimally trade off the marginal cost of lending out more of their position against the marginal benefit of the additional lending fees they collect. As we show, this convex cost of lending introduces an additional source of price distortion in the underlying security; namely, the per unit cost to longs of lending out the security. However, as in the main model, the price distortion in the underlying security decreases as the noise in the derivative payoff decreases.

Finally, in Section 4.3, we explore some dynamic implications in a standard, overlapping generations version of our main model, in which investors are uncertain about whether the security will be scarce in future periods. We show that this forward looking behavior introduces two additional sources of price distortion relative to the static setting. First, the price of the underlying security can be distorted, even if it is not scarce today, because investors expect the security to be scarce in the future. Second, the uncertainty about whether the security will be scarce in the future introduces an additional source of risk, and therefore has an additional effect on the price. Despite these additional sources of price distortion in the dynamic setting, our main result holds — the distortions in the price of the underlying generally decrease as the noise in the derivative security decreases. When investors have access to a perfect derivative (i.e., without any noise), the price distortions disappear completely and, moreover, the price of the underlying security is less volatile.

Our paper provides a framework to analyze a number of recent and current policy decisions in financial markets. For example, as discussed above, our analysis may be useful for understanding the effects of short-selling restrictions such as those enacted during the financial crisis in 2008 and the recent Eurozone crisis. More recently, the European Securities Markets Authority (ESMA) has announced that it will require all revenues, net of operating costs, from securities lending by exchange traded funds to be returned to fund investors. This policy move may lessen the incentives of fund managers to lend securities and thereby reduce the amount of securities that are available to borrow. As a consequence of these policies, these securities can become more scarce, and, all else equal, our model predicts that their prices can become more distorted. Similarly, liquidity coverage ratios in the upcoming Basel III Accord require financial institutions to hold more unencumbered liquid assets, which may also restrict the amount of securities that are available to be borrowed. On the derivatives side, margins on many derivatives contracts have increased, Europe has banned naked CDS positions (buying CDS contracts for speculation rather than as a hedge against an existing long position in an underlying bond), and Dodd-Frank requires position limits on certain commodity derivatives. All of these changes can make derivatives a less attractive substitute for the underlying and therefore may make the underlying more scarce.

The rest of the paper is organized as follows. The next section briefly discusses the related literature. Section 3 develops our benchmark model, and characterizes the equilibrium in this case. To provide more intuition for the results, and to highlight relevant policy implications, this section also considers the special case with a complete ban on short-selling. Section 4 details the extensions of the main model as discussed above. Section 5 concludes.

2 Related Literature

Much of the existing literature on derivatives assumes that the presence of a derivative does not affect the price of the underlying security, even when markets are incomplete, but there are a few notable exceptions. Boyle and Wang (2001) argue that the introduction of a new security into an incomplete market can affect the prices of all the existing traded assets. Bhamra and Uppal (2009) show that a non-redundant derivative can increase the volatility of stock market returns. Zapatero (1998) shows that in a setting with heterogenous beliefs, a security that completes the market affects the volatility of the interest rate. In contrast to these earlier papers, we show that the presence of a derivative security can affect the price of the underlying, even when both securities are exposed to the same source of fundamental risk (so that the derivative security does not complete the market in the traditional sense). In our setup, the derivative makes the market more "complete" by allowing investors to trade larger positions in the fundamental risk, and thereby relaxes the friction that constrains the aggregate trading capacity of the underlying security. Jordan and Kuipers (1997) provide direct empirical evidence of the effect of trading in the derivative on the price of the underlying security in U.S. Treasury markets.

A large literature in finance and economics explores the effect of liquidity on asset prices. Generally speaking, liquidity captures the ease with which a particular security can be traded, and this literature has focussed on the role of various frictions in generating illiquidity. These frictions include transactions costs (e.g., Amihud and Mendelson, 1986; Duffie, 1996; Vayanos, 1998; Krishnamurthy, 2002; Acharya and Pedersen, 2005; Bongaerts, De Jong, and Driessen, 2011), search frictions (e.g., Duffie et al., 2002; Vayanos and Weill, 2008), and asymmetric information (e.g., Kyle, 1985; Wang, 1993; Gârleanu and Pedersen, 2004).[3] Although closely related, there is an important distinction between the notion of liquidity that is explored in these earlier papers and the notion of scarcity that we focus on. While the frictions that make a security illiquid can also make it scarce, the notion of scarcity captures the excess demand for long and short positions in a security *relative* to the trading capacity of the security that may be imposed by these, or other, frictions. For example, recently issued on-the-run Treasuries and more seasoned off-the-run Treasuries are both extremely liquid securities, but typically only on-the-run Treasuries are scarce in that demand for both long and short positions in these securities exceeds the capacity that they can support. Conversely, while corporate bonds are usually considered to be relatively illiquid securities, the large volume of trade in credit default swaps suggests that they may be scarce.

More generally, our paper relates to the literature on financial innovation and security design, including early work by Allen and Gale (1988), Duffie and Jackson (1989), and Cuny (1993), and more recent work by Simsek (2011), Shen et al. (2012) and others.[4] In contrast to the general approach these papers, our model focuses on a particular form of market incompleteness (i.e., the constrained trading capacity of the underlying), and specifically derives *how* the price is affected by the presence of a derivative. Moreover, given our focus, we are also able to characterize the trading volume in these securities.

[3] See Yakov, Mendelson, and Pederson (2005) and Vayanos and Wang (2009) for excellent surveys of the literature on liquidity and asset prices.

[4] See Allen and Gale (1994) and Duffie and Rahi (1995) for comprehensive surveys and discussions of the earlier literature in this area.

Our paper also relates to the literature that studies the effect of short-sale constraints on asset prices. As mentioned earlier, the standard intuition suggests that short-sales constraints increase the price of a security (e.g., Miller, 1977). A number of subsequent papers characterize conditions under which this overpricing result fails to hold. For instance, Diamond and Verrecchia (1987) establish that in a rational expectations model, investors update their valuations to correctly account for the effect of short-sale constraints, and as a result, prices are not biased upwards. Bai, Chang, and Wang (2006) further show that if short-sale constraints prevent informed investors from trading, then prices are lower and more volatile than in the absence of constraints. In a dynamic economy with investors who have heterogeneous beliefs, Gallmeyer and Hollifield (2008) show that a short-sales constraint decreases the price of a security, and increases its volatility, if the optimistic investor's intertemporal elasticity of substitution is less than one. Our model, which we view as complementary to these papers, highlights that short-sale constraints can make a security more scarce, and we characterize conditions under which imposing a short-sale constraint can decrease the price of the security, even in the presence of derivatives.

3 The Model

This section presents the main model of the paper. Section 3.1 presents the setup of the model and Section 3.2 provides a discussion of our assumptions. Section 3.3 presents the main analysis, including a characterization of the equilibrium in the underlying and derivative markets. Finally, Section 3.4 focuses on a special case that is particularly important for recent policy debates — namely, when short-selling is completely banned.

3.1 Model Setup

Assets and Payoffs. There are two dates and three securities in the market. The risk-free asset is normalized to pay a net return of zero. The risky security, which we refer to as the underlying, trades at a price P and pays off a fundamental value F in the next period, which is normally distributed. The derivative security has a price D, and a payoff of $F + \varepsilon$ in the next period, where ε is normally distributed and independent of the fundamental value F. The net supply of the underlying security is given by $Q > 0$, and the derivative is in zero net supply. Short-sellers in the underlying security must borrow it from long investors and pay a borrowing fee of $R \geq 0$. Importantly, there is a limit to how much of the underlying can be lent out (or equivalently, borrowed short). Specifically, longs can lend out, at most, a fraction $0 \leq \gamma < 1$ of their position to shorts; conversely, shorts can borrow, at most, a fraction γ of the position held by longs.

Investor Beliefs and Preferences. Investors have CARA utility over next period wealth. In particular, investor i has risk-tolerance τ_i, and so maximizes expected utility $U_i(W_i)$, given by

$$U_i(W_i) = -\mathbb{E}_i\left[\exp\left\{-\tfrac{1}{\tau_i}W_i\right\}\right]. \tag{1}$$

Investors can have heterogeneous beliefs about the mean of the fundamental F but agree on its variance and the distribution of the noise in the derivative payoff ε. In particular, we denote investor i's beliefs by

$$F \sim N(m_i, \nu) \text{ and } \varepsilon \sim N(0, \delta). \tag{2}$$

Investor i is endowed with an exposure ρ_i to the fundamental shock F, which is realized in the next period. We denote investor i's position in the underlying security by x_i, and her position in the derivative by y_i. Therefore, her wealth in the next period is given by

$$W_i = W_{i,0} + \rho_i F + x_i(F - P + \gamma_i R) + y_i(F + \varepsilon - D), \tag{3}$$

where $\gamma_i \leq \gamma < 1$ if i is long in the underlying (i.e., $x_i > 0$) and $\gamma_i = 1$ if i is short (i.e., $x_i < 0$). The γ_i term captures the fact that short-sellers must borrow each unit they sell at a borrowing fee R, while long investors can only lend up to a fraction γ of their position.

We have two groups of investors indexed by $i \in \{L, S\}$, where $\rho_L < 0 < \rho_S$, and $m_L \geq m_S$. Our analysis focuses on a region of the parameter space where L investors will be long in the underlying security and S investors will be short in equilibrium. In our model, investors trade in order to hedge endowment risk and/or to speculate on their differences in beliefs.

3.2 Discussion of assumptions

For notational simplicity, we assume that the population mass of each type of investor is the same. Since we allow for different risk tolerance (i.e., τ_i) across each group of investors, this assumption is without much loss of generality — increasing the risk tolerance of group i is analogous to increasing the mass of i investors in the economy.

We focus exclusively on a simple derivative that would otherwise be redundant in the absence of scarcity in the underlying. As such, we do not consider more complex derivatives (with non-linear payoffs) that would "complete" the market in a more traditional sense by providing exposure to a risk that investors wish to trade but cannot do so using only the underlying security.[5] This assumption allows us to focus on the role of derivatives in relaxing the scarcity of the underlying. We assume that the derivative offers a potentially noisy exposure, ε, to the fundamental risk, F, so that it may not be a perfect substitute for the underlying security. As we highlighted in the introduction, this noisy exposure to the fundamental is meant to capture market features such as the cheapest-to-deliver option in many exchange traded derivatives or an illiquid secondary market for customized over-the-counter derivatives. For simplicity, we assume that ε is uncorrelated with F. The model can be easily adjusted for the noise ε to be correlated with F by redefining $F + \varepsilon = \alpha F + \eta$, where η is the component of ε that is uncorrelated with F. Moreover, as we show in Section 4.1, if we maintain the assumption that investors exhibit constant absolute risk aversion, our main result is robust to relaxing the assumption that F and ε are normally distributed.

To convey our basic intuition more clearly, we assume that the upper bound γ on the fraction of their positions that longs can lend, or equivalently, shorts can borrow, is exogenously fixed. There are instances when this upper bound is exogenous (e.g., short-sales bans and finite market capacity to clear trades). However, in general, one may expect that the (maximum) fraction that longs lend out is an endogenous decision that depends on transactions costs, search frictions, institutional or regulatory constraints, and the borrowing fee they receive. While the benchmark model takes a reduced form approach, in Section 4.2 we characterize the conditions under which the results are robust to endogenizing this fraction γ (say, by using a non-decreasing and convex cost of lending).

3.3 Market clearing and equilibrium

The equilibrium in this market is defined as the set of prices P, D, and R, and the positions x_i (in the underlying) and y_i (in the derivative) for each group i of investors, such that (i) the positions x_i and y_i maximize the utility of agent i given by equation (1) subject to the budget constraint in equation (3), and (ii) the cash and financing markets for the underlying and the market for the derivative are cleared. The derivative is in zero net supply, so the

[5] For example, if the volatility of an asset is stochastic, then an option on that asset can complete the market (in the traditional sense) by allowing investors to explicitly trade this risk.

market clearing condition for it is given by

$$\sum_i y_i = 0\,. \tag{4}$$

The cash market clearing condition for the underlying is given by

$$\sum_i x_i = Q\,, \tag{5}$$

and the financing market clearing condition for the underlying is given by

$$\underbrace{\gamma\left(\sum_i x_i 1_{\{x_i>0\}}\right)}_{\text{aggregate long positions}} \geq \underbrace{-\left(\sum_i x_i 1_{\{x_i<0\}}\right)}_{\text{aggregrate short positions}}\,. \tag{6}$$

The financing market clearing condition binds with equality when there is a strictly positive fee $R > 0$ to borrow the security, since long investors would like to lend out as much of their position as possible. Moreover, since longs can lend at most a fraction γ of their holdings, equations (5) and (6) imply that the maximum aggregate long position in the underlying is $\frac{1}{1-\gamma}Q$, and the maximum aggregate short position is $-\frac{\gamma}{1-\gamma}Q$.

If borrowing and lending are unconstrained, then the frictionless price of the underlying reflects the (risk-adjusted) expected value of its payoff, and is given by

$$P_0 = \tfrac{\tau_L m_L + \tau_S m_S}{\tau_L + \tau_S} - \tfrac{\nu}{\tau_L + \tau_S}\left(Q + \rho_S + \rho_L\right). \tag{7}$$

Also, there is no cost to borrow the security and the equilibrium quantities are given by

$$-y_{S,0} = y_{L,0} = 0\,, \text{ and} \tag{8}$$

$$Q - x_{S,0} = x_{L,0} = Q + x_0\,, \tag{9}$$

where

$$x_0 = \tfrac{1}{\nu}\tfrac{\tau_L \tau_S}{\tau_L + \tau_S}\left[m_L - \tfrac{\nu}{\tau_L}\left(\rho_L + Q\right) - m_S + \tfrac{\nu}{\tau_S}\rho_S\right]. \tag{10}$$

If the constraint on borrowing and lending does bind, then the price of the underlying is distorted relative to the frictionless price P_0. The following proposition characterizes this distortion, as well as the rest of the equilibrium in this case.

Proposition 1. *Given the economy above with L and S investors, the equilibrium prices are* *given by*

$$D = P_0, \quad P = P_0 + \Delta P, \text{ and } R = \tfrac{\tau_L+\tau_S}{\gamma\tau_L+\tau_S}\Delta P\,, \tag{11}$$

where the price distortion ΔP relative to the friction-less price P_0 in equation (7) is given by

$$\Delta P = \max\left\{0, \tfrac{\delta}{\delta+\nu}\tfrac{\gamma\tau_L+\tau_S}{\tau_L+\tau_S}\tfrac{1}{1-\gamma}\left[m_L - \tfrac{\nu}{\tau_L}\left(\rho_L + \tfrac{1}{1-\gamma}Q\right) - m_S + \tfrac{\nu}{\tau_S}\left(\rho_S - \tfrac{\gamma}{1-\gamma}Q\right)\right]\right\}\,, \tag{12}$$

and the equilibrium quantities are given by

$$-y_S = y_L = \tfrac{1-\gamma}{\delta}\tfrac{\tau_S\tau_L}{\gamma\tau_L+\tau_S}\Delta P\,, \quad \text{and} \tag{13}$$

$$Q - x_S = x_L = \begin{cases} \tfrac{1}{1-\gamma}Q & \text{if } \Delta P > 0, \\ \tfrac{\tau_L}{\tau_L+\tau_S}Q + \tfrac{\tau_L\rho_S-\tau_S\rho_L}{\tau_L+\tau_S} + \tfrac{1}{\nu}\tfrac{\tau_L\tau_S}{\tau_L+\tau_S}\left(m_L - m_S\right) & \text{if } \Delta P = 0. \end{cases} \tag{14}$$

All proofs are in the appendix. To gain some intuition for the relation between equilibrium prices and quantities, recall from equations (9) and (10) that $x_{L,0} = Q + x_0$ and $x_{S,0} = -x_0$ are the equilibrium quantities in the underlying when its price is P_0 as given by equation (7) and borrowing and lending are unconstrained. Market clearing implies that the maximum aggregate long position in equilibrium is $\frac{1}{1-\gamma}Q$ and the maximum aggregate short position is $-\frac{\gamma}{1-\gamma}Q$. Therefore, if

$$x_{L,0} = Q + x_0 > Q + \tfrac{\gamma}{1-\gamma}Q = \tfrac{1}{1-\gamma}Q \quad \text{and} \quad x_{S,0} = -x_0 < -\tfrac{\gamma}{1-\gamma}Q\,, \tag{15}$$

then the constraint on borrowing and lending binds, since the aggregate demand for long and short positions exceeds the capacity that the underlying can support. From equation (15), the constraint binds if and only if,

$$0 < x_0 - \tfrac{\gamma}{1-\gamma}Q = \tfrac{1}{\nu}\tfrac{\tau_L\tau_S}{\tau_L+\tau_S}\left[m_L - \tfrac{\nu}{\tau_L}\left(\rho_L + \tfrac{1}{1-\gamma}Q\right) - m_S + \tfrac{\nu}{\tau_S}\left(\rho_S - \tfrac{\gamma}{1-\gamma}Q\right)\right]\,, \tag{16}$$

which in turn implies a non-zero price distortion (i.e., $\Delta P \neq 0$) in equation (12). A non-zero price distortion ΔP, which implies a non-zero borrowing cost R, allows the cash market to clear because longs and shorts pay different prices for the underlying ($P - \gamma R$ and $R - P$ per unit, respectively). It is important to emphasize that excess trading demand from *either* side of the market (i.e., longs, shorts, or both) can lead to a distortion in the price of the underlying. Condition (16) also highlights that the excess trading demand for the underlying is higher when long investors are more optimistic (i.e., m_L is higher) and short-sellers are more pessimistic (i.e., m_S is lower), and all else equal, these differences lead to a bigger distortion in the price of the underlying.

Figure 1 illustrates the notion of scarcity and its effect on the price of the underlying security. The top panel plots the inverse demand curves for aggregate long and short positions in the underlying. Since L investors are willing to hold larger long positions when the net price they pay for the security is lower, the aggregate demand curve for long positions is downward sloping. Similarly, the aggregate demand curve for short positions is upward sloping since S investors are willing to hold larger short positions only if the net price of the underlying is higher. The aggregate demand curves intersect at the price P_0, which is the frictionless price for the underlying security. If the maximum available fraction that can be borrowed or lent (i.e., γ) is to the right of this intersection point (which in the plot is at $\gamma = 0.8$), then the security is not scarce, and the price is given by P_0. However, if the maximum fraction γ is constrained to the left of the intersection, then the demand for long and short positions cannot clear at the same net price for long investors and short-sellers, and the security is scarce.[6] In this case, as the bottom panel of Figure 1 illustrates, the cash and borrowing markets for the underlying can only clear if the cost to borrow the underlying security becomes non-zero (i.e., $R > 0$), and long investors and short-sellers pay different *net* costs for the underlying security (i.e., $P - \gamma R$ and $P - R$, respectively). As a result, the equilibrium price P of the underlying security (plotted as the solid curve in the bottom panel) is distorted relative to the frictionless price P_0.

Positions in the derivative security provide a substitute for positions in the underlying, and therefore can relax the scarcity in the underlying. Investors trade off their excess demand for positions in the underlying risk against the additional noise ε in the payoff of the derivative. As Figure 2 illustrates, when investors can trade in a derivative security, the aggregate demand for long and short positions in the underlying are smaller at each price. As a result, in the presence of the derivative, the underlying security is scarce for a smaller range of γ (the intersection of the demand curves for long and short positions shifts left), and the price distortion in the underlying security is smaller. Therefore, as the following result summarizes, in contrast to a frictionless market where the derivative is a redundant security, the presence of the derivative in this case can affect the price of the underlying.

Corollary 1. *The distortion in the price of the underlying security ΔP increases with the noise δ in the derivative security. This implies that: (a) as the noise in the derivative*

[6] A special case, which we consider in greater detail in Section 3.4, is $\gamma = 0$, where no short-sales are allowed.

Figure 1. Scarcity and the distortion in price of the underlying
The top panel plots the frictionless price (horizontal, dotted line), P_0, the (inverse) demand functions for long and short positions in the underlying security (downward sloping and upward sloping dashed lines, respectively), and characterizes the range of γ for which the underlying security is scarce. In addition, the bottom panel plots the equilibrium price (solid, non-linear curve), P.

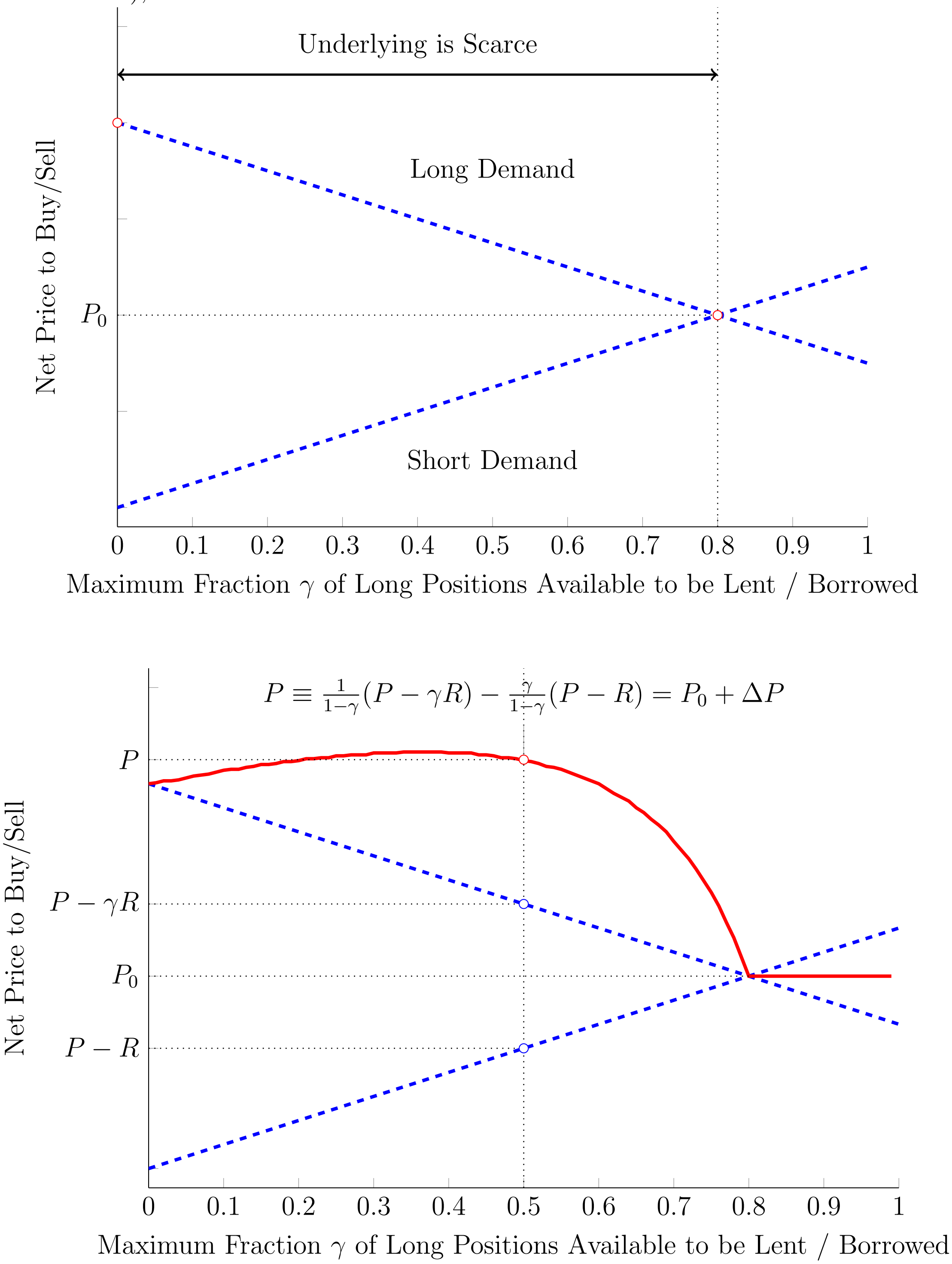
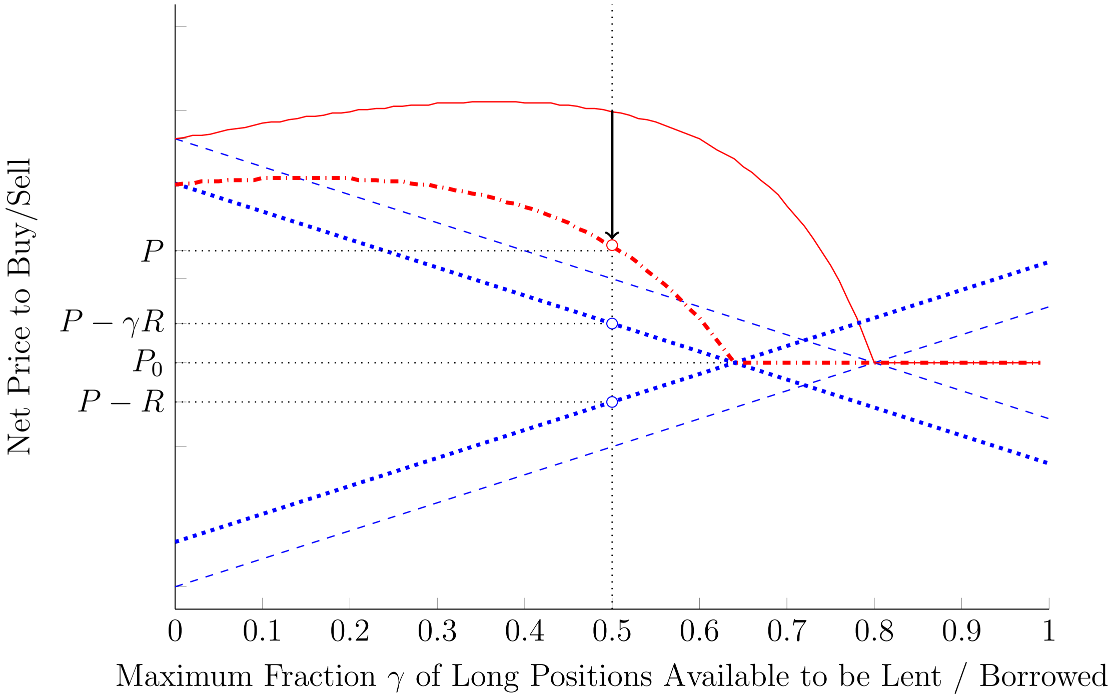


Figure 2. The effect of the derivative on the price of the underlying

The plot depicts the effect of introducing a derivative on the demand for long and short positions in the underlying, and the resulting effect on the price of the underlying security. The dashed lines reflect the demand for long and short positions when there is no derivative, while the dotted lines reflect the demand for long and short positions in the presence of the derivative. Similarly, the solid line reflects the equilibrium price of the underlying security when there is no derivative, while the dot-dashed line reflects the underlying price in the presence of a derivative.



(i.e., δ) becomes arbitrarily small, the price distortion in the underlying disappears, i.e., $\lim_{\delta \to 0} \Delta P = 0$*, and (b) the price distortion in the underlying is the largest if investors do not have access to a derivative (or equivalently, the noise in the derivative is infinite).*

The above result has important implications for the recent debate on derivatives trading and excessive speculation. Note that when the underlying is scarce, the the price distortion ΔP is increasing in the difference in beliefs (i.e., $m_L - m_S$) all else equal. One can interpret this component of the price distortion as the effect of speculative behavior by investors. As the above result highlights, our model predicts that allowing investors to trade in the derivative *decreases* this speculative distortion in the price of the underlying. This prediction is in contrast to recent arguments for restricting trade in derivative markets, which claim that allowing for unrestricted trade in derivatives markets leads to excessive speculative behavior and distortions in the price of the underlying security. Our model suggests that the opposite may be true. When there is excess demand for trading the underlying (i.e., condition (16) holds), but investors cannot trade derivative securities, the price of the underlying security must move to clear markets. This leads to price distortions relative to the market's risk-adjusted expectation of fundamentals.[7]

Finally, note that when there is trade in the derivative, one might expect that both the price and the volume traded in the derivative would reflect the tradeoff between the noise δ and the constraint γ on borrowing and lending. However, in our model, the price D of the derivative is a "clean" measure of the (risk-adjusted) expected value of the underlying security, in that it does not depend on δ or γ. Specifically, the derivative price D is the market's risk-tolerance weighted average expectation of F, adjusted for aggregate risk due to the fundamental F, and is unaffected the lending constraint γ in the underlying.[8] Since the derivative is in zero net supply, it allows long investors and short-sellers to take arbitrarily large positions (as long as they offset each other), and so it can never be scarce. As such, the price is unaffected by position constraints and reflects the marginal valuations of the investors in the market. Instead, the tradeoff between γ and δ manifests itself in the equilibrium trading positions in the derivative. Straightforward calculations show that the size of equilibrium positions in the derivative are decreasing in γ and δ. That is, all else equal, derivative positions are smaller when the lending constraint is less binding and when

[7] As we shall see in the dynamic setting of Section 4.3, improving the ease of trade in derivatives can also lower the price volatility of the underlying.

[8] As we show in Section 4.3, in a dynamic model, if the payoff of the derivative depends on the future price of the underlying security, and the price of the underlying depends on the lending constraint γ, then we would expect D to mechanically depend on γ. However, even in this case, the derivative price can be interpreted as "clean" since it is not affected by the market friction in the underlying directly, but only via a mechanical relationship through its payoff.

the noise in the derivative is higher. As expected, derivative positions are also increasing in the dispersion in beliefs across investors (i.e., $m_L - m_S$).

3.4 Short-sales ban

An important special case of the model is when short-selling is not allowed at all, i.e., $\gamma = 0$. Since S investors cannot hold a short position, L investors must hold the entire supply of the underlying security, i.e., $x_L = Q$.[9] When investors do not have access to the derivative security (or equivalently, if $\delta = \infty$), then the equilibrium price of the underlying is given by

$$P_{ns} = m_L - \tfrac{\nu}{\tau_L}\left(Q + \rho_L\right) = P_0 + \Delta P_{ns}, \tag{17}$$

where the price distortion is given by the expression in equation (12), but reflects the fact that no short-selling and no derivative trading is allowed (i.e., $\gamma = 0$ and $\delta = \infty$):

$$\Delta P_{ns} = \tfrac{\tau_S}{\tau_S+\tau_L}\left(\tfrac{\nu}{\tau_S}\rho_S - m_S + m_L - \tfrac{\nu}{\tau_L}\left(Q+\rho_L\right)\right). \tag{18}$$

On the other hand, when investors are able to trade the derivative security, market clearing in the derivative implies that

$$D = \tfrac{\tau_L m_L + \tau_S m_S}{\tau_S+\tau_L} - \tfrac{\nu}{\tau_S+\tau_L}\left(Q + \rho_S + \rho_L\right) = P_0, \tag{19}$$

and the price of the underlying is given by

$$P = P_0 + \Delta P = \tfrac{\nu}{\delta+\nu} P_0 + \tfrac{\delta}{\delta+\nu} P_{ns}, \tag{20}$$

where the price distortion, ΔP , reflects the fact that, while short-selling is not allowed, investors can trade in the derivative:

$$\Delta P = \tfrac{\delta}{\delta+\nu}\tfrac{\tau_S}{\tau_S+\tau_L}\left(\tfrac{\nu}{\tau_S}\rho_S - m_S + m_L - \tfrac{\nu}{\tau_L}\left(Q+\rho_L\right)\right) = \tfrac{\delta}{\delta+\nu}\Delta P_{ns}. \tag{21}$$

As in the case when short-sales are allowed (but constrained), the price of the underlying security is less distorted relative to the market's risk-adjusted expectation of F, and in the limit, as the noise in the derivative becomes arbitrarily small (i.e., $\delta \to 0$), the price distortion vanishes so that $\lim_{\delta\to 0} P = P_0$.

[9] Some tedious algebra shows that with $\gamma = 0$, if condition (16) is not satisfied, then S investors hold a long position. In this case, the short-sales ban is not binding.

Finally, in the presence of the derivative security, note that the difference between the price distortion in the underlying when a fraction γ of the long positions can be lent (i.e., ΔP), and the price distortion when short-sales are not allowed (i.e., ΔP_{ns}), is given by

$$\Delta P - \Delta P_{ns} = \tfrac{\delta}{\delta+\nu}\tfrac{\gamma}{1-\gamma}\left(m_L - \tfrac{\nu}{\tau_L}\rho_L - m_S + \tfrac{\nu}{\tau_S}\rho_S\right) - \tfrac{\gamma\nu}{(1-\gamma)^2}Q\left(\tfrac{\gamma}{\tau_s} + \tfrac{2-\gamma}{\tau_L}\right). \tag{22}$$

The above difference is positive if the supply Q of the underlying security is small relative to the other parameters in expression (22). As a result, if the security is extremely scarce relative to its outstanding supply, its price may be lower when short-sales are banned (i.e., $\gamma = 0$). This result is in contrast to the over-pricing effect of short-sale constraints in many standard models (e.g., Miller, 1970), and may have important policy implications for regulating short-selling in securities.

4 Extensions

The benchmark model in Section 3 is stylized in order to maintain tractability and highlight the intuition for our results. In this section, we explore the robustness of our main results when we relax some of our earlier assumptions. In Section 4.1, we characterize sufficient conditions under which our main result obtains for general utility functions and payoff distributions. In Section 4.2, we describe the effects of endogenizing the constraint on borrowing and lending for equilibrium prices and quantities. Finally, in Section 4.3, we consider a dynamic version of the main model which allows us to study the effects of uncertainty about future scarcity in a parsimonious and tractable manner.

4.1 General utility functions and payoff distributions

We begin with some general notation. Let $u_i(W_i)$ be agent i's increasing and concave utility function over wealth W. The payoff next period on the underlying security is F and the payoff on the derivative is $F+\varepsilon$, were ε and F are independent and $\mathbb{E}[\varepsilon] = 0$. Let $x_i(\Pi_i, D)$ and $y_i(\Pi_i, D)$ be agent i's optimal demand for the underlying and derivative respectively, where D is the price for the derivative and $\Pi_i = P - \gamma_i R$ is the *net* price that agent i pays for the underlying. That is,

$$\{x_i(\Pi_i, D), y_i(\Pi_i, D)\} = \arg\max_{x,y} \mathbb{E}_i[u_i(W_i)], \text{ where} \tag{23}$$

$$W_i = W_{0,i} + \rho_i F + x(F - \Pi_i) + y(F + \varepsilon - D). \tag{24}$$

Let P, R, and D denote the equilibrium prices that clear the cash, financing, and derivative markets respectively. That is

$$x_L(P - \gamma R, D) + x_S(P - R, D) = Q, \tag{25a}$$

$$\gamma x_L(P - \gamma R, D) + x_S(P - R, D) \geq 0, \tag{25b}$$

and

$$y_L(P - \gamma R, D) + y_S(P - R, D) = 0. \tag{26}$$

When the security is scarce, equation (25b) binds with equality and can be combined with equation (25a) to produce

$$x_L(P - \gamma R, D) - Q = \frac{\gamma Q}{1 - \gamma} = -x_S(P - R, D). \tag{27}$$

Our main result in Section 3 is that the price of the underlying security is higher (more distorted) when there is no derivative. In the more general current setting, it is sufficient to show that the equilibrium price of the underlying decreases in the equilibrium derivative positions of the investors, as the following result characterizes.[10]

Proposition 2. *If the underlying security is scarce (i.e., condition (27) holds) and, for both groups of investors $i \in \{L, S\}$, we have*

$$\frac{\partial y_i}{\partial \Pi_i} - \frac{\partial y_i}{\partial D}\frac{\partial x_i/\partial \Pi_i}{\partial x_i/\partial D} < 0, \tag{28}$$

then the price of the underlying is higher when there is no derivative available to trade.

In the proof of Proposition 2, we show that L investors have positive positions in the derivative and S investors have negative positions. Therefore, to move towards an equilib-

[10] It is important to note that we abstract away from the underlying parameter of the payoff distribution or preferences that decreases the equilibrium holdings of the derivative. This is done to highlight the sufficient conditions as generally as possible. However, what we have in mind is a change in a fundamental parameter (e.g., the variance of the noise in the derivative payoff) that affects investors' positions in the derivative, and our goal in this section is to characterize the effect of this change in derivative positions on the price of the underlying security.

rium with no derivative positions, long investors must decrease their position in the derivative and short-sellers must increase their position (i.e., they must take a smaller short position in the derivative). Since the price of the underlying security can be expressed as

$$P = \frac{1}{1-\gamma}\underbrace{(P-\gamma R)}_{\Pi_L} - \frac{\gamma}{1-\gamma}\underbrace{(P-R)}_{\Pi_S}, \tag{29}$$

a sufficient condition for the price of the underlying to increase with smaller derivative positions for L and S investors is

$$\frac{d\Pi_L}{dy_L} < 0 \quad \text{and} \quad \frac{d\Pi_S}{dy_S} < 0\,. \tag{30}$$

To see why equation (28) characterizes the relevant sufficient condition in equation (30), suppose that the net price that agent i pays for the underlying changes by $d\Pi_i$. Since the underlying security is scarce, the equilibrium positions in the underlying security are $x_L = \frac{1}{1-\gamma}Q$ and $x_S = -\frac{\gamma}{1-\gamma}Q$, and must remain unchanged for the market to continue to clear. That is,

$$dx_i = \frac{\partial x_i}{\partial \Pi_i} d\Pi_i + \frac{\partial x_i}{\partial D} dD = 0, \tag{31}$$

which, in turn, implies that the equilibrium price of the derivative must change by

$$dD = -\frac{\partial x_i/\partial \Pi_i}{\partial x_i/\partial D} d\Pi_i. \tag{32}$$

Along this equilibrium path, the change in agent i's optimal position in the derivative is

$$dy_i = \frac{\partial y_i}{\partial \Pi_i} d\Pi_i + \frac{\partial y_i}{\partial D} dD = \left(\frac{\partial y_i}{\partial \Pi_i} - \frac{\partial y_i}{\partial D}\frac{\partial x_i/\partial \Pi_i}{\partial x_i/\partial D}\right) d\Pi_i, \tag{33}$$

and condition (28) follows.

The following result provides intuitive sufficient conditions for equation (28) (and hence, Proposition 2).

Corollary 2. *The price of the underlying is higher in the absence of a derivative if either*

(i) $0 < \frac{\partial y_i}{\partial \Pi_i} < -\frac{\partial x_i}{\partial \Pi_i}$ *and* $0 < \frac{\partial x_i}{\partial D} \leq -\frac{\partial y_i}{\partial D}\,, \quad i \in \{L, S\}$, *or*

(ii) $0 < \frac{\partial x_i}{\partial D} < -\frac{\partial x_i}{\partial \Pi_i}$ *and* $0 < \frac{\partial y_i}{\partial \Pi_i} \leq -\frac{\partial y_i}{\partial D}\,, \quad i \in \{L, S\}\,.$

In both cases, the demand curves for the underlying and the derivative are downward sloping (i.e., $\partial x_i/\partial \Pi_i < 0$ and $\partial y_i/\partial D < 0$), and the underlying and derivative are substitutes for each other (i.e., $\partial x_i/\partial D > 0$ and $\partial y_i/\partial \Pi_i > 0$). The first sufficient condition implies that demand for the underlying (derivative) is more sensitive than the demand for the derivative (underlying) is to price changes in the underlying (derivative, respectively). The second sufficient condition implies that the demand for the underlying (derivative) is more sensitive to price changes in the underlying (derivative) than to price changes in the derivative (underlying, respectively).

The following result establishes that Proposition 2 is always satisfied when investors exhibit constant absolute risk aversion, and so the results from our main model survive even when we the relax assumption that payoffs are normally distributed.

Corollary 3. *If both agents have preferences with constant absolute risk aversion (CARA), then condition (28) holds. This implies that the price of the underlying is higher when there is no derivative available to trade for any distribution of the fundamental payoff F and noise ε in the derivative payoff (with finite first and second moments).*

When investors have CARA utility, there are no wealth (income) effects and so the demand curve for each security is downward sloping. As we show in the proof, the asset payoffs are positively correlated (under risk-neutral probabilities) and so they are substitutes. Finally, since there is additional noise in the payoff to the derivative, demand for it is less sensitive to changes in the price of the underlying than is demand for the underlying is (i.e., $\frac{\partial y_i}{\partial \Pi_i} < -\frac{\partial x_i}{\partial \Pi_i}$).

4.2 Endogenizing the fraction γ lent by longs

In our main model of Section 3, we exogenously fix the maximum fraction γ of their position that long investors can lend out. While this assumption is made for tractability, in this subsection, we show that our results are qualitatively similar when the fraction γ lent by longs is determined endogenously in equilibrium. In particular, if long investors face a cost $c(\gamma)$ to lend out a fraction γ of their position, then their wealth is given by

$$W_L = W_{L,0} + \rho_L F + x_L (F - P + \gamma R - c(\gamma)) + y_L (F + \varepsilon - D). \tag{34}$$

We show that in this case, the equilibrium is characterized by the following proposition.

Proposition 3. *Suppose that L investors pay a per-unit cost $c(\gamma)$ in order to lend out a fraction γ of their portfolio, where $c(\gamma)$ is non-negative, non-decreasing, convex, $c(0) = 0$*

and $c'(0) = 0$. Then, the equilibrium prices are given by

$$D = P_0, \quad P = P_0 + \Delta P, \text{ and } R = \tfrac{\tau_L+\tau_S}{\gamma^*\tau_L+\tau_S}\left(\Delta P + \tfrac{\tau_L}{\tau_L+\tau_S}c(\gamma^*)\right), \tag{35}$$

where the price distortion ΔP relative to the friction-less price P_0 in equation (7) is given by

$$\begin{aligned}\Delta P = & \tfrac{\gamma^*\tau_L+\tau_S}{\tau_L+\tau_S}\tfrac{1}{1-\gamma^*}\max\left\{0, \tfrac{\delta}{\delta+\nu}\left[m_L - \tfrac{\nu}{\tau_L}\left(\rho_L + \tfrac{1}{1-\gamma^*}Q\right) - m_S + \tfrac{\nu}{\tau_S}\left(\rho_S - \tfrac{\gamma^*}{1-\gamma^*}Q\right)\right] - c(\gamma^*)\right\} \\ & - \tfrac{\tau_L c(\gamma^*)}{\tau_L+\tau_S},\end{aligned} \tag{36}$$

and the optimal fraction γ^* lent out is characterized by $R = c_\gamma(\gamma^*)$. The equilibrium positions are given by

$$-y_S = y_L = \tfrac{1}{\delta}\left(1-\gamma^* + \tfrac{1}{R}c(\gamma^*)\right)\tfrac{\tau_L\tau_S}{\gamma^*\tau_L+\tau_S}\left(\Delta P + \tfrac{\tau_L}{\tau_L+\tau_S}c(\gamma^*)\right), \text{ and} \tag{37}$$

$$Q - x_S = x_L = \begin{cases} \tfrac{1}{1-\gamma^*}Q & \text{if } R > 0, \\ \tfrac{\tau_L(Q+\rho_S+\rho_L)}{\tau_L+\tau_S} - \rho_L & \text{if } R = 0. \end{cases} \tag{38}$$

The conditions on the cost function $c(\cdot)$ ensure that the equilibrium fraction lent out by long investors, γ^*, is between zero and one. Comparing the above result to Proposition 1, an endogenous fraction γ leads to a number of differences in equilibrium prices and quantities. First, note that there is always a price distortion, given by $-\frac{\tau_L}{\tau_L+\tau_S}c(\gamma^*)$, relative to the frictionless price P_0. This price distortion reflects the (risk-tolerance weighted) cost that long investors incur to lend out the equilibrium fraction γ^* of the underlying security. Second, in comparison to condition (16), the underlying security is scarce only when

$$\tfrac{\delta}{\delta+\nu}\left(m_L - m_S + \tfrac{\nu}{\tau_S}\left(\rho_S - \tfrac{\gamma^*}{1-\gamma^*}Q\right) - \tfrac{\nu}{\tau_L}\left(\tfrac{1}{1-\gamma^*}Q + \rho_L\right)\right) - c(\gamma^*) > 0. \tag{39}$$

Since the long investor pays $P + c - \gamma R$ per unit position in the underlying, while the short pays $R - P$, longs and shorts pay different prices even when the cost of borrowing (i.e., R) is zero. The security is scarce only when, for a zero lending fee (i.e., $R = 0$), the aggregate demand for long and short positions do not clear the market.

When there is excess demand for the underlying security (i.e., condition (39) holds), both prices (i.e., R and P) and quantities (i.e., γ) adjust in equilibrium. The degree to which each adjusts depends on the specific cost function $c(\gamma)$, and the equilibrium quantity γ^* is pinned down by the L investors' indifference condition $R(\gamma) = c_\gamma(\gamma)$.[11] As before, there is

[11] Note that this is the indifference condition for the L investors when they take the lending fee R as given.

trade in the derivative security only when the underlying is scarce.

The next result establishes that even though the borrowing constraint is endogenous in this setting, the distortion in the price of the underlying increases in the noise δ in the derivative, and is highest in the absence of the derivative.

Proposition 4. *Suppose that L investors pay a per-unit cost $c(\gamma)$ in order to lend out a fraction γ of their portfolio, where $c(\gamma)$ is non-negative, non-decreasing, convex, $c(0)=0$ and $c'(0)=0$. Then the price distortion ΔP in the underlying security increases with the variance δ of the noise in the derivative security. This implies that, all else equal, the price distortion in the underlying is largest when investors do not have access to a derivative (or equivalently, the noise in the derivative becomes arbitrarily large).*

4.3 Dynamic Considerations

In this subsection, we analyze a simple dynamic extension of the main model in Section 3. We assume that there are overlapping generations of L and S investors who have CARA utility over next period's wealth. This assumption allows us to maintain tractability while introducing an important feature of dynamic models, namely, that an investor's optimal demand depends not only on her beliefs about fundamentals, but also on her beliefs about future prices.

Specifically, suppose that there are three assets: a risk-free asset, the underlying risky security in aggregate supply Q, and the derivative security. The risk-free asset pays a net risk-free rate of $r>0$. The underlying risky security pays dividends F_t at date t, has a price P_t, and a borrowing cost R_t. The derivative security at date t has a price D_t and pays off $P_{t+1}+F_{t+1}+\varepsilon_{t+1}$ in the next period. Finally, investor i is endowed with an exposure $\rho_{i,t}$ to the payoff of the underlying security at date $t+1$. Denoting investor i's position in the underlying and the derivative by $x_{i,t}$ and $y_{i,t}$, respectively, investor i's wealth evolves as

$$W_{i,t+1}=\begin{array}{c} W_{i,t}(1+r)+x_{i,t}\left(P_{t+1}+F_{t+1}-(1+r)\left(P_t-\gamma_i R_t\right)\right) \\ +y_{i,t}\left(P_{t+1}+F_{t+1}+\varepsilon_{t+1}-(1+r)D_t\right)+\rho_{i,t}\left(P_{t+1}+F_{t+1}\right) \end{array}, \tag{40}$$

where $\gamma_i \le \gamma < 1$ if i is long and $\gamma_i = 1$ if i is short. Investor i's beliefs about dividends and the error term are given by

$$F_t \sim N\left(m_{i,}\nu\right), \text{ and } \varepsilon \sim N\left(0,\delta\right), \tag{41}$$

respectively. For notational simplicity, and without loss of generality, we assume that investors have homogenous beliefs about dividends F_{t+1} (i.e, $m_L = m_S = m$).[12] We assume that $\rho_{S,t} = -\rho_{L,t} = \rho_t$ at each date, where $\rho_t \in \{0, \rho\}$ with probability $\{1-\pi, \pi\}$, respectively. This specification allows us to parsimoniously capture the possibility that the underlying security will be scarce in some periods but not in others. Finally, we assume that F_t, ε_t and ρ_t are independent across each other and over time.

As before, the market clearing conditions are given by equations (4) through (6). If there are no constraints on borrowing and lending (i.e., $\gamma = 1$), then the frictionless price of the underlying security in the dynamic model is given by the present value of the discounted stream of dividends, and can be expressed as

$$P_0 = \tfrac{1}{r}\left(m - \tfrac{\nu}{\tau_L+\tau_S} Q\right) . \tag{42}$$

However, as in the main model, the presence of the constraint on borrowing and lending implies that the price of the underlying is distorted relative to this frictionless price, P_0. The following proposition characterizes the stationary equilibrium of the dynamic model.

Proposition 5. *Suppose that $\rho > \frac{1}{1-\gamma}\frac{\gamma\tau_L+\tau_S}{\tau_L+\tau_S}Q$, so that the underlying security is scarce when $\rho_t = \rho$. Then, the stationary equilibrium is characterized by the following prices:*

$$P_t = P_0 + \Delta P_t, \quad D_t = P_t - \tfrac{\gamma\tau_L+\tau_S}{\tau_L+\tau_S} R_t, \quad R_t = \begin{cases} 0 & \text{if } \rho_t = 0 \\ R & \text{if } \rho_t = \rho \end{cases}, \tag{43}$$

and the distortion in the price of the underlying relative to the frictionless price P_0 in expression (42) is given by

$$\Delta P_t = \tfrac{\gamma\tau_L+\tau_S}{\tau_L+\tau_S} R_t + \tfrac{1}{r}\left(\tfrac{\gamma\tau_L+\tau_S}{\tau_L+\tau_S}\pi R - \tfrac{1}{\tau_L+\tau_S}\left(V-\nu\right)Q\right), \tag{44}$$

where $V = var\left(P_{t+1}+F_{t+1}\right) = \nu + \pi\left(1-\pi\right)\left(\frac{\gamma\tau_L+\tau_S}{\tau_L+\tau_S}\right)^2 R^2$, and R solves the cubic equation

$$R = \frac{\delta V}{\delta + V}\frac{\rho\left(\frac{1}{\tau_S}+\frac{1}{\tau_L}\right) - \frac{1}{1-\gamma}\left(\frac{\gamma}{\tau_S}+\frac{1}{\tau_L}\right)Q}{\left(1+r\right)\left(1-\gamma\right)}. \tag{45}$$

[12] Note that in order to recover heterogeneity in beliefs, one could replace ρ_i in the restricted model by $\rho_i - \frac{\tau_i}{\nu}m_i$, and all the results would follow immediately.

The equilibrium quantities are given by

$$-y_{S,t} = y_{L,t} = \tfrac{1-\gamma}{\delta}\left(1+r\right)\tfrac{\tau_L \tau_S}{\tau_L+\tau_S} R_t, \ \textit{and } Q - x_{S,t} = x_{L,t} = \begin{cases} \tfrac{\tau_L}{\tau_L+\tau_S} Q & \textit{if } \rho_t = 0 \\ \tfrac{1}{1-\gamma} Q & \textit{if } \rho_t = \rho \end{cases}. \tag{46}$$

Moreover, in the limit, as the noise in the derivative payoff goes to zero, the distortion in the price of the underlying also goes to zero, i.e., $\lim_{\delta \to \infty} \Delta P_t = 0$, *and consequently, the volatility in the price of the underlying is lower.*

As in the static model from Section 3, the underlying security is scarce when the aggregate demand for long and short positions in the underlying exceeds its capacity, given γ and Q. Not surprisingly, the condition for $R_t > 0$, given by

$$\rho\left(\tfrac{1}{\tau_L} + \tfrac{1}{\tau_S}\right) - \tfrac{1}{1-\gamma}\left(\tfrac{1}{\tau_L} + \tfrac{\gamma}{\tau_S}\right) Q > 0, \tag{47}$$

is exactly the condition for scarcity in equation (16), when $m_L = m_S$ and $\rho_S = -\rho_L = \rho$.

However, unlike the static case, the price of the underlying can be distorted even if the underlying security is not scarce in the current period (i.e., $\Delta P_t \neq 0$ even when $R_t = 0$). This result follows from the fact that, in a dynamic model, the current price may be distorted not only if the underlying security is scarce in the current period, but also if investors expect it to be scarce in a future period. In particular, the price distortion ΔP_t has an intuitive decomposition:

$$\Delta P_t = \underbrace{\tfrac{\gamma\tau_L+\tau_S}{\tau_L+\tau_S} R_t}_{\text{current scarcity}} + \tfrac{1}{r}\left(\underbrace{\tfrac{\gamma\tau_L+\tau_S}{\tau_L+\tau_S}\pi R}_{\text{expected scarcity}} - \underbrace{\tfrac{1}{\tau_L+\tau_S}\left(V-\nu\right)Q}_{\text{uncertainty about scarcity}}\right). \tag{48}$$

The first component of ΔP_t reflects whether or not the underlying security is scarce in the current period, and is common to both the dynamic and static models. The second component reflects the distortion in the current price because investors expect the underlying to be scarce in the next period with probability π, which affects their willingness to pay for the security in the current period. Finally, the third component reflects the effect that investors' uncertainty about future scarcity has on the price. Since the underlying will be scarce in some states of the world but not in others, investors face additional uncertainty about the future price, which affects their valuation in the current period.[13] Since V is

[13] Note that the third term is driven by the variance in next period's price. If investors expect that the underlying will be scarce next period with probability zero or one (i.e., $\pi = 0$ or $\pi = 1$, respectively), then $V = \nu$, and the last component of the price distortion is zero.

increasing in R, the second and third components act in opposite directions, which makes it difficult to determine whether the overall distortion in prices (i.e., ΔP_t), is positive or negative. However, as in the static model, reducing the noise in the payoff of the derivative security helps reduce the distortion in the price of the underlying, and in the limit, eliminates it completely. Moreover, since the variation in ΔP_t is an additional source of volatility, eliminating it reduces the volatility in the underlying security.

5 Conclusions

Even the most liquid securities are subject to frictions that constrain the capacity of the security to support arbitrarily large long and short positions. When the demand for long and short positions exceeds a security's capacity the security becomes "scarce" and its price is distorted relative to its risk-adjusted fundamental value. We show that the presence of a derivative security that allows investors to take arbitrarily large positions can alleviate this relative scarcity of the underlying, and therefore reduce the distortion in its price. We characterize the relative size of the underlying and derivative markets in terms of investor preferences and the outstanding supply of the underlying security. Finally, our model provides a simple benchmark in which one can evaluate the effects of policy changes such as short-selling bans and restrictions on derivative positions.

The notion that simple derivatives can "complete" the market by allowing long investors and short-sellers to take larger aggregate positions in the same source of risk as the underlying, is potentially important in understanding the relative size of derivative markets across various assets. For instance, although U.S. equity indices are extremely liquid, they are often accompanied by very large futures markets,[14] which may be driven by the fact that it can be difficult to simultaneously short-sell all of the components of an index. Similarly, the fact that many corporate bonds are often difficult to borrow (and short-sell) may be an important driver of the recent surge in the size of the corporate CDS markets (e.g., Gupta and Sundaram, 2011). Importantly, our model may be useful in understanding why even extremely liquid securities like U.S. Treasuries may be accompanied by extremely large futures markets.[15] As our model highlights, the size of the derivative market is driven, not by

[14] For example, in October 2011, the average daily trading volume in S&P 500 futures on the CME was about $270B notional, while trade in all stocks on the NYSE, NASDAQ, and SPDRs (an exchange traded fund that mimics the S&P 500) was about half that.

[15] In October 2011, the average daily trading volume in all Treasury securities was around $500B, while trade in four Treasury futures on the Chicago Mercantile Exchange was about $220B notional.

the constraint on borrowing and lending in the underlying per se, but by the excess trading demand *relative* to the trading capacity of the underlying that is imposed by this constraint.

References

Acharya, V. and Pedersen, L., 2005, Asset pricing with liquidity risk, *Journal of Financial Economics*, 77(2), pp. 375–410.

Allen, F. and Gale, D., 1988, Optimal security design, *Review of Financial Studies*, 1(3), pp. 229–263.

Allen, F. and Gale, D., 1994, *Financial innovation and risk sharing.* The MIT Press.

Amihud, Y. and Mendelson, H., 1986, Asset pricing and the bid-ask spread, *Journal of financial Economics*, 17(2), pp. 223–249.

Bai, Y., Chang, E., and Wang, J., 2006, *Asset prices under short-sale constraints.* Ph.D. thesis, University of Hong Kong.

Banerjee, S. and Graveline, J. J., 2012, The cost of short-selling liquid securities, *Journal of Finance.*

Bhamra, H. S. and Uppal, R., 2009, The effect of introducing a non-redundant derivative on the volatility of stock-market returns when agents differ in risk aversion, *Review of Financial Studies*, 22(6), pp. 2303–2330.

Bongaerts, D., De Jong, F., and Driessen, J., 2011, Derivative pricing with liquidity risk: Theory and evidence from the credit default swap market, *The Journal of Finance*, 66(1), pp. 203–240.

Boyle, P. and Wang, T., 2001, Pricing of new securities in an incomplete market: the catch 22 of no-arbitrage pricing, *Mathematical Finance*, 11(3), pp. 267–284.

Cuny, C. J., 1993, The role of liquidity in futures market innovations, *Review of Financial Studies*, 6(1), pp. 57–78.

Diamond, D. W. and Verrecchia, R. E., 1987, Constraints on short-selling and asset price adjustment to private information, *Journal of Financial Economics*, 18(2), pp. 277 – 311.

Duffie, D., 1996, Special repo rates, *Journal of Finance*, 51(2), pp. 493–526.

Duffie, D., Gârleanu, N., and Pedersen, L., 2002, Securities lending, shorting, and pricing, *Journal of Financial Economics*, 66(2-3), pp. 307–339.

Duffie, D. and Jackson, M., 1989, Optimal innovation of futures contracts, *Review of Financial Studies*, 2(3), pp. 275–296.

Duffie, D. and Rahi, R., 1995, Financial market innovation and security design: An introduction, *Journal of Economic Theory*, 65(1), pp. 1 – 42.

Gallmeyer, M. and Hollifield, B., 2008, An examination of heterogeneous beliefs with a short-sale constraint in a dynamic economy, *Review of Finance*, 12(2), pp. 323–364.

Gârleanu, N. and Pedersen, L., 2004, Adverse selection and the required return, *Review of Financial Studies*, 17(3), pp. 643–665.

Geczy, C., Musto, D., and Reed, A., 2002, Stocks are special too: An analysis of the equity lending market, *Journal of Financial Economics*, 66(2-3), pp. 241–269.

Goldreich, D., Hanke, B., and Nath, P., 2005, The price of future liquidity: Time-varying liquidity in the us treasury market, *Review of Finance*, 9(1), pp. 1–32.

Gupta, S. and Sundaram, R. K., 2011, Cds credit-event auctions. Working Paper.

Jordan, B. and Jordan, S., 1997, Special repo rates: An empirical analysis, *Journal of Finance*, pp. 2051–2072.

Jordan, B. and Kuipers, D., 1997, Negative option values are possible: The impact of treasury bond futures on the cash us treasury market, *Journal of Financial Economics*, 46(1), pp. 67–102.

Krishnamurthy, A., 2002, The bond/old-bond spread, *Journal of Financial Economics*, 66(2-3), pp. 463–506.

Kyle, A. S., 1985, Continuous auctions and insider trading, *Econometrica*, 53(6), pp. 1315—1335.

Miller, E. M., 1977, Risk, uncertainty, and divergence of opinion, *The Journal of Finance*, 32(4), pp. 1151–1168.

Ofek, E. and Richardson, M., 2003, Dot-com mania: The rise and fall of internet stock prices, *Journal of Finance*, 58, pp. 1113–1138.

Shen, J., Yan, H., and Zhang, J., 2012, Collateral-motivated financial innovation.

Simsek, A., 2011, Speculation and risk sharing with new financial assets. Technical report, National Bureau of Economic Research.

Vayanos, D., 1998, Transaction costs and asset prices: A dynamic equilibrium model, *Review of Financial Studies*, 11(1), pp. 1–58.

Vayanos, D. and Wang, J., 2009, Theories of liquidity, *Foundations and Trends in Finance.*

Vayanos, D. and Weill, P.-O., 2008, A search-based theory of the on-the-run phenomenon, *Journal of Finance*, 63(3), pp. 1361–1398.

Wang, J., 1993, A model of intertemporal asset prices under asymmetric information, *The Review of Economic Studies*, 60(2), pp. 249–282.

Yakov, A., Mendelson, H., and Pederson, L., 2005, Liquidity and asset prices, *Foundation and Trends in Finance*, 1(4), pp. 1–96.

Zapatero, F., 1998, Effects of financial innovations on market volatility when beliefs are heterogeneous, *Journal of Economic Dynamics and Control*, 22(4), pp. 597–626.

Appendix

Proof of Proposition 1. The first order conditions for investor i implies that

$$\begin{aligned} \tau_i \left(m_i - P + \gamma_i R\right) &= \nu \left(x_i + y_i + \rho_i\right) \,, \quad \text{and} &(49)\\ \tau_i \left(m - D\right) &= \nu \left(x_i + y_i + \rho_i\right) + \delta y_i \,, &(50)\end{aligned}$$

which implies

$$y_i \delta = \tau_i \left(P - \gamma_i R - D\right) \,. \tag{51}$$

The market clearing condition for the derivative implies that

$$\begin{aligned} D &= P - \tfrac{\tau_S + \tau_L \gamma}{\tau_S + \tau_L} R \,, &(52)\\ y_L &= \tfrac{\tau_L}{\delta} \left(P - D - \gamma R\right) = \left(1 - \gamma\right) \tfrac{\tau_S \tau_L}{\tau_S + \tau_L} \tfrac{R}{\delta} \,, \quad \text{and} &(53)\\ y_S &= \tfrac{\tau_S}{\delta} \left(P - D - R\right) = -\left(1 - \gamma\right) \tfrac{\tau_S \tau_L}{\tau_S + \tau_L} \tfrac{R}{\delta} \,. &(54)\end{aligned}$$

The cash-market clearing condition for the underlying implies that

$$P = \tfrac{\tau_L m_L + \tau_S m_S}{\tau_S + \tau_L} + \tfrac{\tau_S + \tau_L \gamma}{\tau_S + \tau_L} R - \tfrac{\nu}{\tau_S + \tau_L} \left(Q + \rho_S + \rho_L\right) \,. \tag{55}$$

Finally, comparing the risk-return tradeoffs for L and S implies that

$$\begin{aligned} \frac{\nu}{\tau_L} \left(x_L + \rho_L + y_L\right) &= m_L - P + \gamma R \,, \quad \text{and} &(56)\\ \frac{\nu}{\tau_S} \left(x_S + \rho_S + y_S\right) &= m_S - P + R \,, &(57)\end{aligned}$$

which, in turn, implies,

$$R = \max \left\{ 0, \frac{\frac{\nu}{\tau_S} \left(x_S + \rho_S\right) - m_S + m_L - \frac{\nu}{\tau_L} \left(x_L + \rho_L\right)}{\left(1 - \gamma\right) \left(1 + \frac{\nu}{\delta}\right)} \right\} \,. \tag{58}$$

Note that when $R = 0$, we have

$$\tfrac{\nu}{\tau_L} \left(x_L + \rho_L\right) - m_L = \tfrac{\nu}{\tau_S} \left(x_S + \rho_S\right) - m_S, \tag{59}$$

and combining with market clearing implies that

$$\begin{aligned} x_L &= \tfrac{\tau_L}{\tau_L + \tau_S} Q - \tfrac{\tau_S \rho_L - \tau_L \rho_S}{\tau_L + \tau_S} + \tfrac{1}{\nu} \tfrac{\tau_L \tau_S}{\tau_L + \tau_S} \left(m_L - m_S\right) \,, &(60)\\ x_S &= \tfrac{\tau_S}{\tau_L + \tau_S} Q + \tfrac{\tau_S \rho_L - \tau_L \rho_S}{\tau_L + \tau_S} - \tfrac{1}{\nu} \tfrac{\tau_L \tau_S}{\tau_L + \tau_S} \left(m_L - m_S\right) \,. &(61)\end{aligned}$$

When $R > 0$, the lending constraint binds since longs want to lend as much as they can,

$$x_S = -\tfrac{\gamma}{1 - \gamma} Q \quad \text{and} \quad x_L = \tfrac{1}{1 - \gamma} Q \,, \tag{62}$$

so that

$$R = \frac{\frac{\nu}{\tau_S} \left(-\frac{\gamma}{1 - \gamma} Q + \rho_S\right) - m_S + m_L - \frac{\nu}{\tau_L} \left(\frac{1}{1 - \gamma} Q + \rho_L\right)}{\left(1 - \gamma\right) \left(1 + \frac{\nu}{\delta}\right)} \,. \tag{63}$$

This completes the characterization of equilibrium prices and quantities. ■

Proof of Proposition 2. We begin by making the following observations:

1. In any equilibrium, unless $\varepsilon = 0$, $P - R < D < P - \gamma R$. If $D > P - \gamma R$, then long investors would

short the derivative and so would shorts, so markets cannot clear. If $D < P - R$, then short sellers would be long the derivative and so would longs, so markets cannot clear.

2. Given the above, $y^L > 0$ and $y^S < 0$. If $y^L < 0$, then the long could do better by selling some of the underlying instead since $D < P - \gamma R < P$. Similarly, if $y^S > 0$, then short-sellers would be better off reducing their short position a little, since $P - R < D$.

3. If the constraints are binding before the derivative, they will be binding after the derivative, unless $\varepsilon = 0$. If not, $R = 0$, which implies $P = D$. But since $\varepsilon \neq 0$, no one trades the derivative, and so we have a contradiction.

These observations imply that generically, L investors have positive positions in the derivative, S investors have negative positions in the derivative, and the rest of the argument follows the text of the paper. Specifically, the equilibrium position in the underlying does not change for either investor i.e.,

$$dx_i = \frac{\partial x_i}{\partial \Pi_i} d\Pi_i + \frac{\partial x_i}{\partial D} dD = 0, \tag{64}$$

which implies the price of the derivative must change by

$$dD = -\frac{\partial x_i / \partial \Pi_i}{\partial x_i / \partial D} d\Pi_i \, , \tag{65}$$

Along this path, the change in agent i's optimal position in the derivative is

$$dy_i = \frac{\partial y_i}{\partial \Pi_i} d\Pi_i + \frac{\partial y_i}{\partial D} dD = \left(\frac{\partial y_i}{\partial \Pi_i} - \frac{\partial y_i}{\partial D} \frac{\partial x_i / \partial \Pi_i}{\partial x_i / \partial D} \right) d\Pi_i \, . \tag{66}$$

Hence,

$$\frac{\partial y_i}{\partial \Pi_i} - \frac{\partial y_i}{\partial D} \frac{\partial x_i / \partial \Pi_i}{\partial x_i / \partial D} < 0 \tag{67}$$

is a sufficient condition for $\frac{dy_i}{d\Pi_i} < 0$, which, combined with the observation that $y_L = -y_S$ and $P = \frac{1}{1-\gamma}\Pi_L - \frac{\gamma}{1-\gamma}\Pi_S$, gives us the result.■

Proof of Corollary 3. Dropping the subscript i, the first order conditions of the optimal portfolio problem in equation (23) are

$$0 = \mathbb{E}\left[\{F - \Pi\} u'\left(W_0 + \rho F + x\left[F - \Pi\right] + y\left[F + \varepsilon - D\right]\right)\right] \, , \tag{68a}$$
$$0 = \mathbb{E}\left[\{F + \varepsilon - D\} u'\left(W_0 + \rho F + x\left[F - \Pi\right] + y\left[F + \varepsilon - D\right]\right)\right] \, . \tag{68b}$$

Differentiating each of these equations with respect to Π and D yields

$$\frac{\partial x}{\partial \Pi} = \frac{(A + 2B + C)\, \mathbb{E}\left[u' + x\left(F - \Pi\right) u''\right] - (A + B)\, x \mathbb{E}\left[\left(F + \varepsilon - D\right) u''\right]}{AC - B^2} \, , \tag{69a}$$
$$\frac{\partial y}{\partial \Pi} = \frac{A x \mathbb{E}\left[\left(F + \varepsilon - D\right) u''\right] - (A + B)\, \mathbb{E}\left[u' + x\left(F - \Pi\right) u''\right]}{AC - B^2} \, , \tag{69b}$$
$$\frac{\partial x}{\partial D} = \frac{(A + 2B + C)\, y \mathbb{E}\left[\left(F - \Pi\right) u''\right] - (A + B)\, \mathbb{E}\left[u' + y\left(F + \varepsilon - D\right) u''\right]}{AC - B^2} \, , \tag{69c}$$
$$\frac{\partial y}{\partial D} = \frac{A \mathbb{E}\left[u' + y\left(F + \varepsilon - D\right) u''\right] - (A + B)\, y \mathbb{E}\left[\left(F - \Pi\right) u''\right]}{AC - B^2} \, , \tag{69d}$$

where

$$\begin{aligned} A &= \mathbb{E}\left[\left(F-\Pi\right)^2 u''\right] \,, \\ B &= \mathbb{E}\left[\left(F-\Pi\right)\left(\varepsilon - D + \Pi\right) u''\right] \,, \\ C &= \mathbb{E}\left[\left(\varepsilon - D + \Pi\right)^2 u''\right] \,. \end{aligned}$$

For an investor with CARA utility we have $u'' = -k\,u'$ for some constant $k > 0$, so that

$$\mathbb{E}\left[\left(F-\Pi\right) u''\right] = -k \underbrace{\mathbb{E}\left[\left(F-\Pi\right) u'\right]}_{0 \text{ by Eq. (68a)}} = 0 \,, \tag{71}$$

$$\mathbb{E}\left[\left(F+\varepsilon-D\right) u''\right] = -k \underbrace{\mathbb{E}\left[\left(F+\varepsilon-D\right) u'\right]}_{0 \text{ by Eq. (68b)}} = 0 \,. \tag{72}$$

Also,

$$\begin{aligned} \mathbb{E}\left[\left(F-\Pi\right)\left(\varepsilon-D+\Pi\right)u''\right] &= \mathbb{E}\left[\left(F - \mathbb{E}\left[F\frac{u'}{\mathbb{E}\left[u'\right]}\right]\right)\left(\varepsilon - \mathbb{E}\left[\varepsilon\frac{u'}{\mathbb{E}\left[u'\right]}\right]\right)u''\right] \,, \quad \text{by Eq. (68)} \,, && \text{(73a)} \\ &= -k\mathbb{E}\left[u'\right]\mathbb{E}\left[\left(F - \mathbb{E}\left[F\frac{u'}{\mathbb{E}\left[u'\right]}\right]\right)\left(\varepsilon - \mathbb{E}\left[\varepsilon\frac{u'}{\mathbb{E}\left[u'\right]}\right]\right)\frac{u'}{\mathbb{E}\left[u'\right]}\right] \,, && \text{(73b)} \\ &= -k\mathbb{E}\left[u'\right]\underbrace{\mathbb{E}\left[\left(F - \mathbb{E}\left[F\frac{u'}{\mathbb{E}\left[u'\right]}\right]\right)\frac{u'}{\mathbb{E}\left[u'\right]}\right]}_{0}\underbrace{\mathbb{E}\left[\left(\varepsilon - \mathbb{E}\left[\varepsilon\frac{u'}{\mathbb{E}\left[u'\right]}\right]\right)\frac{u'}{\mathbb{E}\left[u'\right]}\right]}_{0} , && \text{(73c)} \\ &= 0 && \text{(73d)} \end{aligned}$$

To move from equation (73b) to equation (73c) note that $u'/\mathbb{E}\left[u'\right]$ defines a change of probability measure under which F and ε are independent (since they are independent under the original measure).[16] Combining these results, we have

$$\frac{\partial x}{\partial \Pi} = \frac{\left(A+C\right)\mathbb{E}\left[u'\right]}{AC} = \frac{\mathbb{E}\left[\left(F-\Pi\right)^2 u''\right]\mathbb{E}\left[u'\right] + \mathbb{E}\left[\left(\varepsilon-D+\Pi\right)^2 u''\right]\mathbb{E}\left[u'\right]}{\mathbb{E}\left[\left(F-\Pi\right)^2 u''\right]\mathbb{E}\left[\left(\varepsilon-D+\Pi\right)^2 u''\right]} \,, \tag{76a}$$

$$\frac{\partial y}{\partial \Pi} = \frac{\partial x}{\partial D} = -\frac{\partial y}{\partial D} = -\frac{A\mathbb{E}\left[u'\right]}{AC} = -\frac{\mathbb{E}\left[\left(F-\Pi\right)^2 u''\right]\mathbb{E}\left[u'\right]}{\mathbb{E}\left[\left(F-\Pi\right)^2 u''\right]\mathbb{E}\left[\left(\varepsilon-D+\Pi\right)^2 u''\right]} \,, \tag{76b}$$

which establishes the result. ■

[16] More specifically, since $u'\left(W\right) = e^{-kW}$ and therefore,

$$\frac{u'}{\mathbb{E}\left[u'\right]} = g\left(F\right)h\left(\varepsilon\right) \,, \quad \text{where} \quad \mathbb{E}\left[g\left(F\right)\right] = 1 = \mathbb{E}\left[h\left(\varepsilon\right)\right] \,, \tag{74}$$

so that

$$\begin{aligned} &\mathbb{E}\left[\left(F - \mathbb{E}\left[F\frac{u'}{\mathbb{E}\left[u'\right]}\right]\right)\left(\varepsilon - \mathbb{E}\left[\varepsilon\frac{u'}{\mathbb{E}\left[u'\right]}\right]\right)\frac{u'}{\mathbb{E}\left[u'\right]}\right] \\ &= \mathbb{E}\left[\left(F - \mathbb{E}\left[Fg\left(F\right)h\left(\varepsilon\right)\right]\right)\left(\varepsilon - \mathbb{E}\left[\varepsilon g\left(F\right)h\left(\varepsilon\right)\right]\right)g\left(F\right)h\left(\varepsilon\right)\right] \,, && \text{(75a)} \\ &= \underbrace{\mathbb{E}\left[\left(F - \mathbb{E}\left[Fg\left(F\right)\right]\right)g\left(F\right)\right]}_{0}\underbrace{\mathbb{E}\left[\left(\varepsilon - \mathbb{E}\left[\varepsilon h\left(\varepsilon\right)\right]\right)h\left(\varepsilon\right)\right]}_{0} \,. && \text{(75b)} \end{aligned}$$

Proof of Proposition 3. The first order conditions for investor L imply that:

$$
\begin{aligned}
c_\gamma &= R\,, && (77)\\
y_L\delta &= \tau_L\left(P-\gamma R+c-D\right)\,, && (78)\\
\left(x_L+y_L+\rho_L\right)\nu &= \tau_L\left(m_L-P+\gamma R-c\right)\,, && (79)
\end{aligned}
$$

while the first order conditions for the S investors are given by:

$$
\begin{aligned}
y_S\delta &= \tau_S\left(P-R-D\right)\,, && (80)\\
\left(x_S+y_S+\rho_S\right)\nu &= \tau_S\left(m_S-P+R\right)\,. && (81)
\end{aligned}
$$

Suppose that an interior optimal fraction γ^* exists (we shall verify this below). For notational simplicity, let γ denote the optimal fraction γ^* that is characterized by $c_\gamma\left(\gamma^*\right)=R\left(\gamma^*\right)$, and let c denote the cost at that optimal fraction (i.e., $c=c\left(\gamma^*\right)$). The market clearing conditions for the derivative and underlying cash-market imply

$$
\begin{aligned}
D &= P-\tfrac{\tau_S+\gamma\tau_L}{\tau_S+\tau_L}R+\tfrac{\tau_L}{\tau_S+\tau_L}c\,, && (82)\\
P &= \tfrac{\tau_L m_L+\tau_S m_S}{\tau_L+\tau_S}+\tfrac{\tau_S+\gamma\tau_L}{\tau_S+\tau_L}R-\tfrac{\tau_L}{\tau_S+\tau_L}c-\tfrac{\nu}{\tau_L+\tau_S}\left(Q+\rho_L+\rho_S\right)\,, && (83)
\end{aligned}
$$

which in turn imply that the equilibrium positions in the derivative are given by

$$y_S=-\left(1-\gamma+\tfrac{c}{R}\right)\tfrac{\tau_S\tau_L}{\delta(\tau_S+\tau_L)}R=-y_L. \tag{84}$$

Finally, note that

$$
\begin{aligned}
m_S-P+R-\left(m_L-P+\gamma R-c\right) &= \tfrac{\nu}{\tau_S}\left(x_S+y_S+\rho_S\right)-\tfrac{\nu}{\tau_L}\left(x_L+y_L+\rho_L\right), && (85)\\
\tfrac{\delta}{\tau_S}y_S-\tfrac{\delta}{\tau_L}y_L &= P-R-D-\left(P-\gamma R+c-D\right), && (86)
\end{aligned}
$$

which imply

$$
\begin{aligned}
R\left(1-\gamma\right)+c &= m_L-m_S+\tfrac{\nu}{\tau_S}\left(x_S+y_S+\rho_S\right)-\tfrac{\nu}{\tau_L}\left(x_L+y_L+\rho_L\right)\,, && (87)\\
&= m_L-m_S+\tfrac{\nu}{\tau_S}\left(x_S+\rho_S\right)-\tfrac{\nu}{\tau_L}\left(x_L+\rho_L\right)-\tfrac{\nu}{\delta}\left(\left(1-\gamma\right)R+c\right)\,, && (88)\\
\Rightarrow R &= \frac{m_L-m_S+\frac{\nu}{\tau_S}\left(x_S+\rho_S\right)-\frac{\nu}{\tau_L}\left(x_L+\rho_L\right)}{\left(1-\gamma\right)\left(1+\frac{\nu}{\delta}\right)}-\frac{c}{1-\gamma}\,. && (89)
\end{aligned}
$$

Finally, to establish the existence of the optimal fraction γ^*, note that it must solve the equation

$$c_\gamma\left(\gamma\right)=\frac{m_L-\frac{\nu}{\tau_L}\left(\frac{1}{1-\gamma}Q+\rho_L\right)-m_S+\frac{\nu}{\tau_S}\left(\rho_S-\frac{\gamma}{1-\gamma}Q\right)}{\left(1-\gamma\right)\left(1+\frac{\nu}{\delta}\right)}-\frac{c\left(\gamma\right)}{1-\gamma}\,, \tag{90}$$

or equivalently,

$$c\left(\gamma\right)+\left(1-\gamma\right)c'\left(\gamma\right)=\frac{m_L-\frac{\nu}{\tau_L}\rho_L-m_S+\frac{\nu}{\tau_S}\rho_S-\left(\frac{\nu}{\tau_S}\gamma+\frac{\nu}{\tau_L}\right)\frac{1}{1-\gamma}Q}{\left(1+\frac{\nu}{\delta}\right)}\,. \tag{91}$$

Since $c\left(\cdot\right)\geq 0$, $c'\left(\cdot\right)\geq 0$, and $c''\left(\cdot\right)\geq 0$, the LHS is non-negative and increasing in γ. On the other hand, the RHS is positive for $\gamma=0$, decreasing in γ, and strictly negative for $\gamma=1$. Therefore, if the LHS is less than the RHS at $\gamma=0$, i.e., if

$$c\left(0\right)+c'\left(0\right)<\frac{m_L-\frac{\nu}{\tau_L}\rho_L-m_S+\frac{\nu}{\tau_S}\rho_S-\frac{\nu}{\tau_L}Q}{\left(1+\frac{\nu}{\delta}\right)}\,, \tag{92}$$

then there is a solution $\gamma^* \in [0,1]$ to the above equation. In particular, if $c(0) = 0$ and $c'(0) = 0$, then the condition is satisfied, which completes the proof. ■

Proof of Proposition 4. As the proof of Proposition 3 establishes, the assumptions on the cost function ensure that there is an optimal γ^* in equilibrium between zero and one. Moreover, recall that the cost function is non-decreasing and convex (i.e., $c_\gamma \geq 0$ and $c_{\gamma\gamma} \geq 0$). Recall that the cost of borrowing can be expressed as $R = R^0 - \frac{c(\gamma^*)}{1-\gamma^*}$, where R^0 is given by

$$R^0 = \frac{m_L - m_S + \frac{\nu}{\tau_S}\left(\rho_S - \frac{\gamma^*}{1-\gamma^*}Q\right) - \frac{\nu}{\tau_L}\left(\frac{1}{1-\gamma^*}Q + \rho_L\right)}{\left(1-\gamma^*\right)\left(1+\frac{\nu}{\delta}\right)} \tag{93}$$

This implies that

$$R_\delta = \frac{\partial R}{\partial \delta} = \frac{\nu}{\delta\left(\nu+\delta\right)}R^0, \text{ and } \quad R_\gamma = \frac{\partial R}{\partial \gamma} = -\frac{Q\left(\frac{1}{\tau_S}+\frac{1}{\tau_L}\right)}{\left(1-\gamma\right)^3\left(\frac{1}{\nu}+\frac{1}{\delta}\right)} + \tfrac{1}{1-\gamma}R^0 - \left(\tfrac{1}{1-\gamma}c_\gamma + \tfrac{1}{(1-\gamma)^2}c\right). \tag{94}$$

Finally, since $R = c'(\gamma^*)$ in equilibrium, we have that $\frac{dR}{d\delta} = R_\delta + R_\gamma \frac{d\gamma}{d\delta}$ must be equal to $\frac{dc_\gamma}{d\delta} = c_{\gamma\gamma}\frac{d\gamma}{d\delta}$, which implies

$$\frac{d\gamma}{d\delta} = \frac{R_\delta}{c_{\gamma\gamma} - R_\gamma} = \frac{\frac{\nu}{\delta(\nu+\delta)}R^0}{c_{\gamma\gamma} - \left(-\frac{Q\left(\frac{1}{\tau_S}+\frac{1}{\tau_L}\right)}{(1-\gamma)^3\left(\frac{1}{\nu}+\frac{1}{\delta}\right)} + \frac{1}{1-\gamma}R^0 - \left(\frac{1}{1-\gamma}c_\gamma + \frac{1}{(1-\gamma)^2}c\right)\right)} = \frac{\frac{\nu}{\delta(\nu+\delta)}R^0}{c_{\gamma\gamma} + \frac{Q\left(\frac{1}{\tau_S}+\frac{1}{\tau_L}\right)}{(1-\gamma)^3\left(\frac{1}{\nu}+\frac{1}{\delta}\right)}} > 0$$

Since $R = c_\gamma(\gamma^*)$ and $c_{\gamma\gamma} > 0$, we have that for $\gamma < \gamma^*$, $c_\gamma(\gamma) < c_\gamma(\gamma^*)$, and so

$$\Delta P = \tfrac{\tau_L\gamma^* + \tau_S}{\tau_L+\tau_S}R - \tfrac{\tau_L}{\tau_L+\tau_S}c = \tfrac{\tau_S}{\tau_L+\tau_S}R + \tfrac{\tau_L}{\tau_L+\tau_S}\left(\gamma^* R - c\right) \geq 0. \tag{95}$$

Moreover, change in the price distortion due to a change in δ can be expressed as:

$$\frac{d\Delta P}{d\delta} = -\frac{\tau_L}{\tau_L+\tau_S}c_\gamma\frac{d\gamma}{d\delta} + \frac{\tau_L}{\tau_L+\tau_S}R\frac{d\gamma}{d\delta} + \frac{\tau_L\gamma+\tau_S}{\tau_L+\tau_S}\frac{dR}{d\delta}, \tag{96}$$

$$= \frac{\tau_L\gamma+\tau_S}{\tau_L+\tau_S}\frac{dR}{d\delta} = \frac{\tau_L\gamma+\tau_S}{\tau_L+\tau_S}\frac{dc_\gamma}{d\delta} = \frac{\tau_L\gamma+\tau_S}{\tau_L+\tau_S}c_{\gamma\gamma}\frac{d\gamma}{d\delta} > 0. \tag{97}$$

Hence, the distortion in price is increasing in the noise δ. ■

Proof of Proposition 5. We conjecture that $R_t = R$ when $\rho_t = \rho$, and zero otherwise. Let $V = \text{var}\left(F_{t+1} + P_{t+1}\right)$. Then, the first order conditions for investor i are given by

$$\tau_i\left(\mathbb{E}\left[P_{t+1} + F_{t+1}\right] - (1+r)\left(P_t - \gamma_i R_t\right)\right) = \left(x_{i,t} + y_{i,t} + \rho_{i,t}\right)V \tag{98}$$

$$\tau_i\left(\mathbb{E}\left[P_{t+1} + F_{t+1}\right] - (1+r)D_t\right) = \left(x_{i,t} + y_{i,t} + \rho_{i,t}\right)V + y_{i,t}\delta \tag{99}$$

$$\Rightarrow \tau_i(1+r)\left(P_t - \gamma_i R_t - D_t\right) = \delta y_{i,t} \tag{100}$$

Market clearing in the derivative implies

$$D_t = P_t - \frac{\gamma\tau_L + \tau_S}{\tau_L + \tau_S}R_t \tag{101}$$

and market clearing in the underlying implies

$$P_t = \tfrac{1}{1+r}\left(\mathbb{E}\left[P_{t+1} + F_{t+1}\right] - \tfrac{1}{\tau_L+\tau_S}VQ\right) + \tfrac{\gamma\tau_L+\tau_S}{\tau_L+\tau_S}R_t \tag{102}$$

$$\Rightarrow \mathbb{E}\left[P_t\right] = \tfrac{1}{1+r}\left(\mathbb{E}\left[P_{t+1} + F_{t+1}\right] - \tfrac{1}{\tau_L+\tau_S}VQ\right) + \tfrac{\gamma\tau_L+\tau_S}{\tau_L+\tau_S}\pi R \tag{103}$$

In a stationary equilibrium, since all random variables are i.i.d., we have $\mathbb{E}\left[P_t\right] = \mathbb{E}\left[P_{t+1}\right]$ and so

$$P_t = \tfrac{1}{r}\left(\mathbb{E}\left[F_{t+1}\right] - \tfrac{1}{\tau_L+\tau_S}VQ + \tfrac{\gamma\tau_L+\tau_S}{\tau_L+\tau_S}\pi R\right) + \tfrac{\gamma\tau_L+\tau_S}{\tau_L+\tau_S}R_t, \tag{104}$$

and so the distortion is

$$\Delta P = \tfrac{1}{r}\left(\tfrac{\gamma\tau_L+\tau_S}{\tau_L+\tau_S}\pi R - \tfrac{1}{\tau_L+\tau_S}\left(V-\nu\right)Q\right) + \tfrac{\gamma\tau_L+\tau_S}{\tau_L+\tau_S}R_t. \tag{105}$$

Finally, note that

$$\left(1+r\right)\left(1-\gamma\right)R_t = \frac{V}{\tau_S}\left(x_{S,t} + y_{S,t} + \rho_{S,t}\right) - \frac{V}{\tau_L}\left(x_{Lt} + y_{L,t} + \rho_{L,t}\right) \tag{106}$$

$$\frac{\delta}{\tau_S}y_S - \frac{\delta}{\tau_L}y_L = -\left(1+r\right)\left(1-\gamma\right)R_t \tag{107}$$

which in turn imply:

$$R_t = \frac{\frac{V}{\tau_S}\left(x_{S,t} + \rho_{S,t}\right) - \frac{V}{\tau_L}\left(x_{L,t} + \rho_{L,t}\right)}{\left(1+r\right)\left(1-\gamma\right)\left(1+\frac{V}{\delta}\right)}, \tag{108}$$

and note that $R_t > 0$ only if $\rho_t = \rho$, and $\rho > \frac{1}{1-\gamma}\frac{\gamma\tau_L+\tau_S}{\tau_L+\tau_S}Q$ (since $x_{L,t} = \frac{1}{1-\gamma}Q$ and $x_{S,t} = -\frac{\gamma}{1-\gamma}Q$ when $R_t > 0$). This implies that our conjecture is verified: $R_t = R$ when $\rho_t = \rho$ and is zero otherwise, and that R is characterized by the equation

$$R = \frac{\delta V}{\delta + V}\frac{\rho\left(\frac{1}{\tau_S} + \frac{1}{\tau_L}\right) - \frac{1}{1-\gamma}\left(\frac{\gamma}{\tau_S} + \frac{1}{\tau_L}\right)Q}{\left(1+r\right)\left(1-\gamma\right)}, \tag{109}$$

where

$$V = \operatorname{var}\left(F_{t+1} + P_{t+1}\right) = \nu + \pi\left(1-\pi\right)\left(\tfrac{\gamma\tau_L+\tau_S}{\tau_L+\tau_S}\right)^2 R^2. \tag{110}$$

Note that while the price distortion ΔP may be positive or negative, depending on the parameters, as $\delta \to 0$, the first order conditions imply that $R_t \to 0$, and this in turn, implies that $\lim_{\delta\to 0}\Delta P_t = 0$. ■