Journal of Financial Economics 111 (2014) 589–608

Contents lists available at ScienceDirect

Journal of Financial Economics

journal homepage: www.elsevier.com/locate/jfec

Trading in derivatives when the underlying is scarce ☆

Snehal Banerjee [a,*], Jeremy J. Graveline [b]

[a] Northwestern University, Kellogg School of Management, United States
[b] University of Minnesota, Carlson School of Management, United States

ARTICLE INFO

Article history:
Received 4 March 2013
Received in revised form
14 June 2013
Accepted 15 July 2013
Available online 4 December 2013

JEL classification:
G12
G13

Keywords:
Scarcity
Short-selling
Price distortions
Derivatives
Regulation

ABSTRACT

Regulatory restrictions and market frictions can constrain the aggregate quantity of long and short positions in a security. When these constraints bind, we refer to the security as *scarce*, and its price becomes distorted relative to its value in a frictionless market. We show that an otherwise redundant derivative can reduce the price distortion of the underlying security by relaxing its scarcity. We also show that it is especially important to analyze the underlying and derivative markets jointly when evaluating the impact of regulation, such as short-sales bans and position limits in derivatives, that restricts trade.

1. Introduction

> It is critical that regulatory rulemaking ... is done right, with the proper analysis to ensure that any new rules do not impede the function of the markets they are meant to protect.[1]

Many recent policy changes have focused on restricting trade in derivatives and the securities that underlie them. For example, the Securities and Exchange Commission (SEC) imposed short-selling bans on financial stocks in the fall of 2008, and many European countries imposed similar restrictions for bank stocks during the recent Eurozone crisis. On the derivatives side, the European Union has banned naked credit default swap (CDS) positions on sovereign debt, and the CFTC has proposed position limits on certain commodity derivatives to "diminish, eliminate or prevent" excessive speculation.

By restricting trade in a security, these regulations may distort prices. For instance, a short-sales ban prevents pessimistic short-sellers from trading, but it also prevents long positions from being larger, in aggregate, than the

☆ We are grateful for comments from Darrell Duffie, Nicolae Gârleanu, Brad Jordan, Arvind Krishnamurthy, Xuewen Liu, Moto Yogo, and participants at the Minnesota Junior Finance Conference (2011), the Financial Intermediation Research Society Conference (2012), and the NBER AP Meeting in Palo Alto (2012).

* Corresponding author at: Kellogg School of Management, Northwestern University, 2001 Sheridan Road, Evanston, IL 60208, United States. Tel.: +1 847 491 8335.

E-mail addresses: snehal-banerjee@kellogg.northwestern.edu (S. Banerjee), jeremy@umn.edu (J.J. Graveline).

[1] Timothy Ryan, chief executive officer of Securities Industry and Financial Markets Association (SIFMA), in response to the position limits on commodity derivatives proposed by the Commodity Futures Trading Commission (CFTC) under the mandate of the Dodd-Frank Act (see *Financial Times*, December 5, 2011). In December 2011, SIFMA and the International Swaps and Derivatives Association (ISDA) filed suit against the CFTC over the proposed position limits. In September 2012, less than two weeks before the limits were set to take effect, US district judge Robert Wilkins ruled against the proposed limits, arguing that more analysis was required by the CFTC to assess if the proposals were necessary and appropriate under the law.

0304-405X/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jfineco.2013.11.008

security's outstanding supply. More generally, regulatory restrictions and market frictions, such as transactions costs, search costs, short-sales constraints, and margin requirements, can constrain the aggregate quantity of long and short positions in a security. When these constraints bind, we refer to the security as *scarce*, and the price of the security must adjust to clear the market.

In frictionless markets, a simple derivative security is redundant because it provides exposure to the same source of risk as the underlying asset. However, we show that this redundancy no longer applies when the underlying can be scarce. In fact, the presence of a derivative may affect the price of the underlying itself. Intuitively, derivatives can reduce the scarcity of the underlying by providing a substitute for long and short positions. We characterize how equilibrium prices and trading volume in the underlying asset and its derivative are jointly determined, and provide sufficient conditions under which the presence of a derivative reduces both the scarcity of the underlying and the associated price distortion.

Our model provides a framework to analyze the effect of proposed regulations that restrict trade in the underlying or derivative securities. It is important to note that prices and quantities that are empirically observed in the absence of these proposed trading restrictions cannot be used directly to test the impact of the restrictions. Instead, to evaluate the effects of such policy, one must jointly characterize the equilibrium in the underlying and derivative markets under the counter-factual assumption that the restrictions are in effect. To provide a role for regulatory trading restrictions, we assume that some investors may trade for non-informational motives, and therefore, distort prices relative to the underlying asset's fundamental value. In this framework, we show that even if trade in the derivative is always accompanied by distortionary trading by speculators, restricting trade in the derivative may *increase* the overall price distortion for an underlying asset that is scarce. Moreover, if the underlying asset is scarce and the derivative is not a perfect substitute, then we show that a short-sales ban can actually *lower* the price of the security, instead of raising it. As such, our analysis highlights the importance of accounting for scarcity in the underlying, and jointly modeling the underlying and derivative markets, when evaluating policy decisions.

Our analysis is applicable to any security that may be scarce, including both liquid and illiquid assets. Generally speaking, liquidity captures the ease with which a particular security can be traded, and the literature has focused on the role of various frictions in generating illiquidity.[2] In contrast, the notion of scarcity reflects the aggregate demand for both long and short positions in an asset *relative to* the capacity for such positions that it can support. For example, on-the-run Treasuries are extremely liquid, but the demand for long and short positions in these securities often exceeds the supply that are available to be borrowed in the financing, or repo, market. When this situation occurs, the Treasury is scarce — it is costly to borrow and trades “on special” in the financing market. Off-the-run Treasuries are also extremely liquid, but they are not typically scarce, since the demand for positions is most often concentrated in their on-the-run counterparts. Conversely, while the corporate bond market is much less liquid than the Treasury market, corporate bonds can also be scarce when the demand for positions to hedge or speculate on default risk exceeds the supply of corporate bonds that are available to be traded. Other illiquid assets, such as real-estate and commodities, can also be scarce if there is sufficient demand for positions but they are inherently difficult to borrow and sell short.

The rest of the paper is organized as follows. In the next section, we briefly discuss the related literature. In Section 3, we develop a benchmark model and characterize equilibrium prices and quantities in both markets. We also derive sufficient conditions on preferences and payoff distributions under which the presence of a derivative security reduces the price distortion in the underlying. In Section 4, we develop a general framework that allows us to analyze the implications of scarcity on standard policy changes such as limits on derivative trading and short-selling bans. Section 5 concludes. All proofs are in Appendix A.

2. Related literature

In general, the equilibrium prices of existing securities change when a derivative security, exposed to risks that are not spanned by those securities, is introduced into an economy (e.g., see Detemple and Selden, 1991; Zapatero, 1998; Boyle and Wang, 2001; Bhamra and Uppal, 2009).[3] In contrast to these earlier papers, where derivatives complete the market by allowing investors to trade new sources of risk, we show that the presence of a derivative security can affect the price of the underlying asset, even when both securities span the same risks. In our setup, the derivative makes the market more “complete” by relaxing the constraint on the aggregate capacity for positions in the underlying asset.[4] As such, the mechanism through which derivatives affect the price of the underlying is also distinct from, but complementary to, other channels that have been suggested in the literature, such as reducing transactions costs and search frictions (e.g., Merton, 1989; Gârleanu, 2009).

Our paper is closely related to the large literature that explores the effect of trading frictions on asset prices and, in particular, to earlier models that generate costly

[2] 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, Gârleanu, and Pedersen, 2002; Vayanos and Weill, 2008), and asymmetric information (e.g., Kyle, 1985; Wang, 1993; Gârleanu and Pedersen, 2004). See Amihud, Mendelson, and Pedersen (2005) and Vayanos and Wang (2012) for excellent surveys of the literature on liquidity and asset prices.

[3] However, much of the existing literature on derivatives assumes that the presence of such a derivative does not affect the price of the underlying security — this is described by Hakansson (1979) as “The Catch 22 of Option Pricing.”

[4] Jordan and Kuipers (1997) provide direct empirical evidence of the effect of trading in a derivative on the price of the underlying security in U.S. Treasury markets.

borrowing through frictions such as transaction costs (e.g., Duffie, 1996; Krishnamurthy, 2002), and search costs (e.g., Duffie, Gârleanu, and Pedersen, 2002; Vayanos and Weill, 2008).[5] While our model is more stylized than some of the models in these papers, it offers greater tractability and allows us to characterize how trade in derivative securities affects the cost of borrowing an underlying that may be scarce. Our model also provides an intuitive and flexible framework in which to analyze the effects of regulatory policy that restricts trade in the derivative or the underlying asset.

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), Cuny (1993), and Rahi (1995), and more recent work by Simsek (2011), Kubler and Schmedders (2012), Shen, Yan, and Zhang (2012), and others.[6] In contrast to the general approach in these papers, our model focuses on a particular form of market incompleteness (i.e., the constrained capacity for positions in the underlying), and we specifically derive *how* the price is affected by the presence of a derivative. Moreover, while some of these papers focus on how the introduction of derivatives can increase the distortion in equilibrium allocations (e.g., Simsek, 2011) and prices (e.g., Kubler and Schmedders, 2012), we highlight a channel through which even simple derivatives can help reduce the distortion in the equilibrium price of the underlying. As such, we view our analysis as complementary to these other papers.

Finally, our paper relates to the literature that studies the effect of short-sale constraints on asset prices. Standard intuition suggests that short-sales restrictions constrain pessimists from expressing their views and therefore 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 (e.g., Diamond and Verrecchia, 1987; Bai, Chang, and Wang, 2006; Gallmeyer and Hollifield, 2008). In our model, the mechanism through which short-sale constraints can lower the security's price is analogous to the one in Duffie, Gârleanu, and Pedersen (2002), but we characterize conditions under which the result obtains in the presence of derivatives.

3. Benchmark model

We begin by developing a benchmark model to highlight the effect of derivatives on the scarcity of the underlying asset. Section 3.1 presents the setup of the model and Section 3.2 provides a discussion of our assumptions. In Section 3.3, we present the main analysis, including a characterization of the equilibrium in the underlying and derivative markets. In Section 3.4, we characterize sufficient conditions, for general utility functions and payoff distributions, under which the scarcity of the underlying asset is reduced when investors can trade a derivative.

3.1. Model setup

Securities and payoffs. There are two dates and three securities in the market. The risk-free security is normalized to pay a net return of zero. The risky asset, which we refer to as the underlying, trades at a price P and pays off a normally distributed fundamental value $F \sim N(m,\nu)$ in the next period. The derivative security has a price D, and a payoff of $F+\varepsilon$ in the next period, where $\varepsilon \sim N(0,\delta)$ 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 $R \ge 0$.

Investor beliefs and preferences. There are two groups of investors, indexed by $i \in \{L,S\}$, with constant absolute risk aversion (CARA) utility over next period's wealth, risk-tolerance τ, and common beliefs. An investor from group i is endowed with an exposure ρ_i to the fundamental shock F, where $\rho_S = -\rho_L = \rho$. We denote investor i's position in the underlying asset by x_i, and her position in the derivative by y_i.

We focus on the range of parameters such that L investors are long in the underlying asset and S investors are short. Each short position must be borrowed from a long. However, since the underlying asset is in positive net supply, not every long position can be loaned in equilibrium to a short-seller (i.e., $0 \le -x_S/x_L < 1$). To convey our basic intuition more clearly, we exogenously fix the maximum fraction $0 \le \gamma < 1$ of long positions that are borrowed by (or equivalently, loaned to) shorts. While we take a reduced-form approach in our benchmark model, in Appendices B and C we characterize the conditions under which our results are robust to endogenizing this fraction γ, as discussed below.

Investor i's wealth in the next period is given by

$$W_i = W_{i,0} + \rho_i F + x_i(F - P + \gamma_i R) + y_i(F + \varepsilon - D), \tag{1}$$

where $\gamma_i \le \gamma < 1$ 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 only a fraction γ of long positions can be borrowed by (loaned to) short-sellers. Investor i maximizes expected utility $U_i(W_i)$, given by

$$U(W_i) = -\mathbb{E}\left[\exp\left(-\frac{1}{\tau}W_i\right)\right]. \tag{2}$$

3.2. Discussion of assumptions

In any equilibrium, since the underlying is in positive net supply, not every long position can be borrowed by (or equivalently, loaned to) a short-seller.[7] There are instances when the upper bound, γ, on borrowing and lending is

[5] 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 (2013), which show a positive relation between prices and borrowing fees in bond markets, and D'Avolio (2002), Geczy, Musto, and Reed (2002), and Ofek and Richardson (2003), which document it in equity markets.

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

[7] However, there may be significant heterogeneity in the fraction of each long investor's position that they lend to short-sellers. We abstract away from this heterogeneity in order to maintain tractability and more clearly highlight the intuition for our results.

exogenous (e.g., due to short-sales bans or finite market capacity to clear trades). However, one might expect that, in general, γ is determined endogenously. We consider two such scenarios in the appendices.

In Appendix B, we consider a model in which long investors must pay a cost to lend out a fraction γ of their position to short-sellers. In this case, the optimal fraction γ lent out in equilibrium trades off the marginal cost of lending out an additional unit with the marginal benefit of getting the lending fee R from doing so. In Appendix C, we consider an alternate setup in which short-sellers must pay a cost to search for long investors in order to borrow shares and establish a short position. In this case, the optimal search intensity (which, in turn, determines the equilibrium fraction of long positions that are borrowed) sets the marginal cost of searching equal to the increase in a short-seller's expected utility from being able to trade in the underlying security (i.e., from locating a long investor). The analysis in these appendices characterizes conditions under which the results from our benchmark model extend to these settings.

Note that the constraint on borrowing and lending (i.e., γ) can also be interpreted as a collateral constraint faced by investors (e.g., Geanakoplos, 2003; Simsek, 2013). For instance, suppose the initial wealth of L investors reflects their endowment of the risky asset (i.e., $W_0 = QP$ for the aggregate long investor), and the collateral constraint implies that their position in the risky asset can be at most a multiple κ of their initial wealth (i.e., $x_L P \le \kappa W_0$). In this case, the link between the collateral constraint κ and the borrowing constraint γ is given by $\gamma = (\kappa - 1)/\kappa$.[8]

We focus exclusively on a simple derivative that is otherwise redundant if the underlying asset is not scarce. As such, we do not consider more complex derivatives (such as options with nonlinear 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.[9] This assumption allows us to focus exclusively on the role of derivatives in relaxing the scarcity of the underlying.

We assume that the derivative offers a potentially noisy exposure, $F+\varepsilon$, to the fundamental risk, F, so that it may not be a perfect substitute for the underlying asset. 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.[10] 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 the next subsection, we illustrate that when noise in the derivative payoff is present (i.e., $\delta > 0$), it effectively constrains investors' equilibrium positions in the derivative security. In Appendix D, we develop a model that eliminates the noise in the derivative payoff, but instead imposes position limits in the derivative market. Our main result is qualitatively similar in this model — increasing ease of trade in the derivative (i.e., relaxing the limits on derivative positions) reduces the price distortion in the underlying asset by relaxing the effective scarcity that investors face.[11]

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 3.4, 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.

3.3. Market clearing and equilibrium

An 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 Eq. (2) subject to the budget constraint in Eq. (1), 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 market clearing condition for it is given by

$$y_L + y_S = 0. \tag{3}$$

The cash market clearing condition for the underlying asset is given by

$$x_L + x_S = Q, \tag{4}$$

and the financing market clearing condition is given by

$$-x_S \le \gamma x_L. \tag{5}$$

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 short-sellers can borrow at most a fraction γ from longs, Eqs. (4) and (5) imply that the maximum aggregate long position in the underlying is $Q/(1-\gamma)$, and the maximum aggregate short position is $-\gamma Q/(1-\gamma)$.

If borrowing and lending are unconstrained, then the frictionless price of the underlying reflects the (risk-adjusted)

[8] This follows from comparing the constraint on collateral, given by $x_L P \le \kappa QP$, to the constraint on long positions implied by γ, given by $x_L \le Q/(1-\gamma)$.

[9] 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.

[10] For instance, the primary market for over-the-counter interest rate swaps is generally considered to be extremely liquid, but the secondary market is not. As a result, rather than use the secondary market to exit an existing swap position, participants in this market typically take an offsetting position in a new swap (in the primary market) and are, therefore, left with some residual basis risk.

[11] As we discuss in Appendix D, one notable difference with this setup is that when the constraint in the underlying and the position limit in the derivative are both binding, the price of the derivative is bounded, but not determinate without additional assumptions (e.g., bargaining between L and S investors over the gains from trade). This result is because the derivative is in zero net supply and so L and S investors must always pay the same price. In contrast, as we discuss below, a positive borrowing cost R in the underlying asset allows the market to clear even when aggregate positions in it are constrained, since L and S investors pay different *net* prices for the underlying ($P-\gamma R$ and $P-R$, respectively) when $R > 0$.

expected value of its payoff, and is given by

$$P_0 = m - \frac{\nu}{2\tau}Q. \tag{6}$$

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

$$-y_{S,0} = y_{L,0} = 0 \quad \text{and} \tag{7}$$

$$Q - x_{S,0} = x_{L,0} = \tfrac{1}{2}Q + \rho. \tag{8}$$

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

$$D = m - \frac{\nu}{2\tau}Q, \quad P = P_0 + \frac{1+\gamma}{2}R, \quad \text{and} \tag{9a}$$

$$R = \max\left\{0, \frac{\delta}{\delta+\nu}\frac{1}{1-\gamma}\frac{\nu}{\tau}\left(2\rho - \frac{1+\gamma}{1-\gamma}Q\right)\right\}, \tag{9b}$$

and the equilibrium quantities are given by

$$-y_S = y_L = \frac{1-\gamma}{\delta}\frac{\tau}{2}R \quad \textit{and} \quad Q - x_S = x_L = \begin{cases} \dfrac{1}{1-\gamma}Q & \textit{if } R > 0, \\ \dfrac{1}{2}Q + \rho & \textit{if } R = 0. \end{cases} \tag{10}$$

The proof is a special case of Proposition 4 in Section 4 that follows. To gain some intuition for the relation between equilibrium prices and quantities, recall from Eq. (8) that $x_{L,0} = \frac{1}{2}Q + \rho$ and $x_{S,0} = \frac{1}{2}Q - \rho$ are the equilibrium quantities in the underlying asset when its price is P_0 as given by Eq. (6) and the constraint on borrowing and lending does not bind. Market clearing implies that the maximum aggregate long position in equilibrium is $Q/(1-\gamma)$ and the maximum aggregate short position is $-\gamma Q/(1-\gamma)$. Therefore, if

$$x_{L,0} = \frac{1}{2}Q + \rho > \frac{1}{1-\gamma}Q \quad \text{and} \quad x_{S,0} = \frac{1}{2}Q - \rho < -\frac{\gamma}{1-\gamma}Q, \tag{11}$$

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 Eq. (11), the constraint binds if and only if,

$$0 < 2\rho - \frac{1+\gamma}{1-\gamma}Q. \tag{12}$$

The borrowing cost, R (see Eq. (9)), is positive when the constraint binds and Eq. (12) holds. It allows the cash market to clear because longs and shorts pay different *net* prices for the underlying ($P - \gamma R$ and $R - P$ per unit, respectively).

Fig. 1 illustrates the notion of scarcity and its effect on the price of the underlying security. The top panel plots the inverse aggregate demand curves for 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 capacity for aggregate positions beyond the outstanding is to the right of this intersection point (which in the plot is at $\gamma/(1-\gamma) = 2$, so that $\gamma = \frac{2}{3}$), then the security is not scarce, and its price is given by P_0. However, if the capacity for aggregate positions 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. In this case, as the bottom panel of Fig. 1 illustrates, the cash and borrowing markets for the underlying can only clear if the cost to borrow the underlying security becomes positive (i.e., $R > 0$), so that 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. In particular, even though investors have the same beliefs about fundamentals and their hedging demands perfectly offset each other, the equilibrium price for the underlying is higher than P_0 when it is scarce.

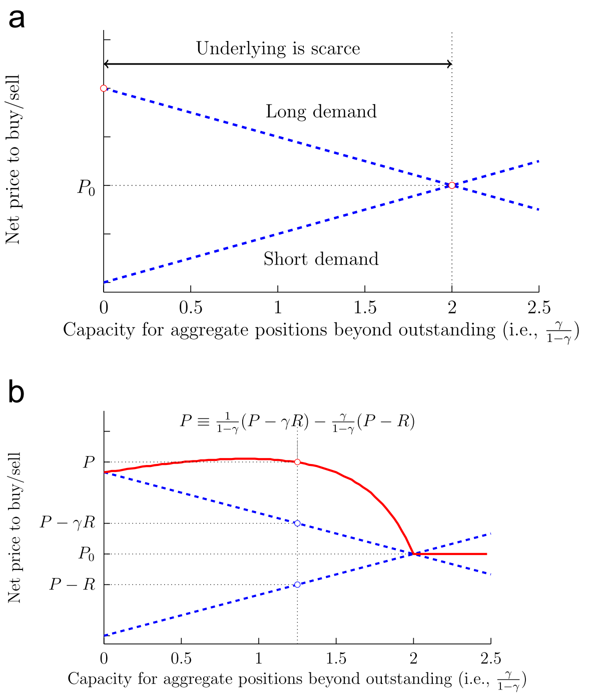

Fig. 1. Scarcity and the distortion in price of the underlying. The top panel plots the frictionless price (horizontal, dotted line), P_0, the (inverse) demand functions for long and short positions in the underlying security (downward sloping and upward sloping dashed lines, respectively), and characterizes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

Positions in the derivative security provide a substitute for positions in the underlying asset, and therefore can

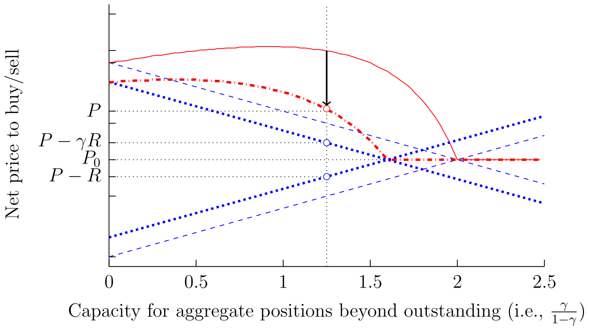

Fig. 2. The effect of the derivative on the price of the underlying. The plot depicts the effect of introducing a derivative on the demand for long and short positions in the underlying, and the resulting effect on the price of the underlying security. The dashed lines reflect the demand for long and 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.

relax the scarcity in the underlying. Investors trade off their demand for positions in the underlying risk against the additional noise ε in the payoff of the derivative. As Fig. 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 net price for the underlying. Therefore, 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. 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 asset. The following proposition summarizes this result.

Proposition 2. The distortion in the price of the underlying asset, $\Delta P \equiv P - P_0$, *decreases with the noise* δ *in the derivative security. Therefore*: (a) *as the noise in the 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 largest if investors do not have access to a derivative* (*or equivalently, the noise in the derivative is infinite*).

Finally, one might expect that the tradeoff between the noise δ and the constraint γ on borrowing and lending would be reflected in both the price *and* the volume of trade in the derivative. However, in our benchmark model, the price D of the derivative is a "cleaner" 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 expectation of F, adjusted for aggregate risk due to the fundamental F. Instead, the tradeoff between γ and δ manifests itself in the equilibrium positions in the derivative. All else equal, derivative positions are smaller when the lending constraint is less binding and when the noise in the derivative is higher (i.e., equilibrium derivative positions are decreasing in γ and δ, respectively).

The scarcity of the underlying is closely related to the lending fee R, which is typically small in the data. However, the prices and quantities that we observe empirically reflect the equilibrium scarcity in the presence of derivatives and in the absence of the proposed regulatory restrictions. Therefore, a small lending fee does not imply that the role of scarcity is unimportant for evaluating regulatory policy. Instead, as we highlighted in the introduction, one must analyze equilibrium prices and quantities under the counterfactual assumptions of imposing the proposed restrictions. We take up this task in Section 4.

3.4. General utility functions and payoff distributions

In this subsection, we characterize general sufficient conditions under which the presence of the derivative increases or decreases the price distortion in the underlying. We begin with some general notation. Let $u_i(W_i)$ be agent i's increasing and concave utility function over wealth W_i. As in our benchmark model, we assume that the payoff of the underlying, F, and the noise in the derivative payoff, ε, are independent and that $\mathbb{E}[\varepsilon]=0$. However, we do not impose any additional distributional assumptions. 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,

$$\{x_i(\Pi_i, D), y_i(\Pi_i, D)\} = \arg\max_{x,y} \mathbb{E}_i[u_i(W_i)] \quad \text{where} \tag{13}$$

$$W_i = W_{0,i} + \rho_i F + x(F - \Pi_i) + y(F + \varepsilon - D), \tag{14}$$

and $\rho_L < 0 < \rho_S$. Let P, R, and D denote the equilibrium prices that clear the cash, financing, and derivative markets, respectively. That is,

$$x_L(P-\gamma R, D) + x_S(P-R, D) = Q, \tag{15a}$$

$$\gamma x_L(P-\gamma R, D) + x_S(P-R, D) \geq 0, \tag{15b}$$

and

$$y_L(P-\gamma R, D) + y_S(P-R, D) = 0. \tag{16}$$

When the asset is scarce, Eq. (15b) binds with equality and can be combined with Eq. (15a) to produce

$$x_L(P-\gamma R, D) - Q = \frac{\gamma Q}{1-\gamma} = -x_S(P-R, D). \tag{17}$$

Our main result in Section 3.3 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 is decreasing in the equilibrium derivative positions of the investors, as the following result characterizes.[12]

[12] To highlight the sufficient conditions as generally as possible, we abstract away from the underlying parameter of the payoff distribution or preferences that decreases the equilibrium holdings of the derivative. 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.

Proposition 3. Suppose the underlying security is scarce (i.e., condition (17) holds).

(i) *If for all $i \in \{L,S\}$, we have*

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

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

(ii) *If for all $i \in \{L,S\}$, we have*

$$\frac{\partial y_i}{\partial \Pi_i} - \frac{\partial y_i}{\partial D}\frac{\partial x_i/\partial \Pi_i}{\partial x_i/\partial D} > 0$$

everywhere, then the price of the underlying is lower when there is no derivative available to trade.

In the proof of Proposition 3, we show that the sign of an investor's position in the derivative must be the same as the sign of his or her position in the underlying asset. Therefore, to move towards an equilibrium with no derivative positions, L investors must decrease their position in the derivative and S investors 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

$$P = \frac{1}{1-\gamma}\underbrace{(P-\gamma R)}_{\Pi_L} - \frac{\gamma}{1-\gamma}\underbrace{(P-R)}_{\Pi_S}, \tag{18}$$

a sufficient condition for the price of the underlying to increase (decrease) with smaller derivative positions is that $d\Pi_i/dy_i < 0$ ($d\Pi_i/dy_i > 0$, respectively) for both $i = L$ and $i = S$ investors.

To see why the conditions in Proposition 3 are sufficient for the above conditions, suppose that the net price that agent i pays for the underlying changes by $d\Pi_i$. Since the underlying asset is scarce, the equilibrium positions in it are $x_L = Q/(1-\gamma)$ and $x_S = -\gamma Q/(1-\gamma)$, and must remain unchanged for the market to continue to clear. That is,

$$dx_i = \frac{\partial x_i}{\partial \Pi_i} d\Pi_i + \frac{\partial x_i}{\partial D} dD = 0, \tag{19}$$

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

$$dD = -\frac{\partial x_i/\partial \Pi_i}{\partial x_i/\partial D} d\Pi_i. \tag{20}$$

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

$$dy_i = \frac{\partial y_i}{\partial \Pi_i} d\Pi_i + \frac{\partial y_i}{\partial D} dD = \left(\frac{\partial y_i}{\partial \Pi_i} - \frac{\partial y_i}{\partial D}\frac{\partial x_i/\partial \Pi_i}{\partial x_i/\partial D}\right) d\Pi_i, \tag{21}$$

which implies the sign of $d\Pi_i/dy_i$ is characterized by the conditions in Proposition 3. The following result provides intuition for the conditions in Proposition 3.

Corollary 1. Assume that the underlying security is scarce and the demand curves for the underlying and the derivative are downward sloping (i.e., $\partial x_i/\partial \Pi_i < 0$ and $\partial y_i/\partial D < 0$).

1. *Suppose that, for all $i \in \{L,S\}$, either of the two following conditions always holds:*

$$\left|\frac{\partial y_i}{\partial \Pi_i}\right| < \left|\frac{\partial x_i}{\partial \Pi_i}\right| \text{ and } \left|\frac{\partial x_i}{\partial D}\right| < \left|\frac{\partial y_i}{\partial D}\right| \text{ or} \tag{22a}$$

$$\left|\frac{\partial x_i}{\partial D}\right| < \left|\frac{\partial x_i}{\partial \Pi_i}\right| \text{ and } \left|\frac{\partial y_i}{\partial \Pi_i}\right| < \left|\frac{\partial y_i}{\partial D}\right|. \tag{22b}$$

Then

(a) *The price of the underlying is higher in the absence of a derivative if the substitution effect* always *dominates the wealth effect for both securities (i.e., $\partial x_i/\partial D > 0$ and $\partial y_i/\partial \Pi_i > 0$).*

(b) *The price of the underlying is lower in the absence of a derivative if the wealth effect* always *dominates the substitution effect for both securities (i.e., $\partial x_i/\partial D < 0$ and $\partial y_i/\partial \Pi_i < 0$).*

Suppose that, for all $i \in \{L,S\}$, either of the two following conditions always holds:

2. $$\left|\frac{\partial x_i}{\partial \Pi_i}\right| < \left|\frac{\partial y_i}{\partial \Pi_i}\right| \text{ and } \left|\frac{\partial y_i}{\partial D}\right| < \left|\frac{\partial x_i}{\partial D}\right| \text{ or} \tag{23a}$$

$$\left|\frac{\partial x_i}{\partial \Pi_i}\right| < \left|\frac{\partial x_i}{\partial D}\right| \text{ and } \left|\frac{\partial y_i}{\partial D}\right| < \left|\frac{\partial y_i}{\partial \Pi_i}\right|. \tag{23b}$$

Then

(a) *The price of the underlying is lower in the absence of a derivative if the substitution effect always dominates the wealth effect for both securities (i.e., $\partial x_i/\partial D > 0$ and $\partial y_i/\partial \Pi_i > 0$).*

(b) *The price of the underlying is higher in the absence of a derivative if the wealth effect always dominates the substitution effect for both securities (i.e., $\partial x_i/\partial D < 0$ and $\partial y_i/\partial \Pi_i < 0$).*

The sufficient conditions in Corollary 1 depend on (i) the relative sensitivity of the optimal demand for each security to the price of the securities (as described in Eqs. (22) and (23)) and (ii) whether the substitution effect dominates the wealth (or income) effect for both securities. The first part of condition (22) requires that, for a given change in the price of a security, the optimal demand for that security changes by more than the optimal demand for the other security. Alternatively, the second part of condition (22) requires that, for the same change in price of both the underlying and the derivative, the optimal demand for a security is more sensitive to its own price. Under either of these intuitive conditions,[13] the price of the underlying is higher in the absence of a derivative if the substitution effect always dominates the wealth effect, but lower if the wealth effect always dominates the substitution effect. It seems counterintuitive to assume that the wealth effect *always* dominates the substitution effect, therefore, the sufficient conditions for 1(a) above seem most natural.

It is difficult to derive more primitive sufficient conditions because the partial derivatives depend on both the assumptions of preferences and the payoff distributions. However, the conditions in part (i) of Proposition 3 are always satisfied when investors exhibit constant absolute risk aversion. Therefore, the results from our main model

[13] In contrast, condition (23) seems less intuitive since it requires that the optimal demand for a security is *always* less sensitive to its own price.

survive even when we the relax the assumption that payoffs are normally distributed.

Corollary 2. If both agents have preferences with constant absolute risk aversion (*CARA*), *then for* $i \in \{L, S\}$, *we always have*

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

Therefore, if both groups of agents have CARA preferences, then the price of the underlying is higher when there is no derivative available to trade. This result holds for any distributions of the fundamental payoff F and noise ε *in the derivative payoff* (*with finite first and second moments*).

4. The general model

In this section, we generalize our benchmark model by allowing for agents with heterogeneous beliefs and trade by speculators. The first subsection presents the setup of the model and Section 4.2 characterizes the equilibrium prices and quantities. In the following subsections, we characterize the equilibrium in special cases of the general models. In Section 4.3, we assume that L and S investors can only trade in the underlying, but speculators can trade both the underlying and derivative securities. In Section 4.4, we analyze the equilibrium when speculators can only trade in the derivative security. This case is useful to consider the effects of restricting trade in derivatives when there are speculative investors who can distort the underlying price through their trades in the derivatives market. In Section 4.5, we consider the effects of introducing long-only investors (e.g., mutual funds) in the underlying on equilibrium prices for the underlying and derivative securities. Finally, in Section 4.6, we analyze the special case in which short-sales of the underlying security are completely banned and characterize the conditions under which imposing such a ban can lead to a *decrease* in the price of the underlying, even in the presence of a derivative.

4.1. Model setup

The assumptions about security payoffs and prices are as in our benchmark model of Section 3. We refer to the L and S investors as *hedgers*, since they use the underlying and derivative securities to hedge shocks to their endowments (i.e., ρ_i). In addition, we assume there exist *speculators*, denoted by B, who have no endowment shocks (i.e., $\rho_B = 0$) but instead *bet* on the realization of F. An investor of type i has CARA utility with risk-tolerance τ_i. While investors are assumed to agree on the distribution of ε, they may have different beliefs about F. In particular, while the objective (or “true”) distribution of F is given by $F \sim N(m, \nu)$ as before, we assume that investor i's beliefs about F are given by

$$F \sim N(m_i, \nu), \tag{24}$$

where $m_L \geq m_S$. Again, we focus on the parameter space where L investors have long positions and S investors have short positions in equilibrium.

Whether or not investors have a distortionary effect on prices depends on whether or not they are assumed to be trading on fundamental information. For instance, if differences in expectations arise due to differences in information, then equilibrium prices aggregate information and the primary source of price distortion in our setup is the scarcity of the underlying security.[14] However, more generally, these differences in beliefs may represent other, non-informational motives for trade that are often interpreted as distortionary, and therefore may generate a role for regulation.

Implicit in the discussion of regulatory policy that restricts trade in a security is the notion that some subset of this trade leads to undesirable distortions in prices. For instance, a short-sales ban is often considered when regulators believe that short-sellers are overly pessimistic and drive asset prices below their fundamental value. Similarly, regulators may propose restrictions on derivative positions in commodities if they feel that speculators with beliefs that are disconnected from fundamentals may distort prices. Given a stand on investors' information and trading motives, our framework allows us to characterize the effect of regulatory restrictions when the underlying security is scarce. For instance, in Section 4.4, we assume that speculators have distorted beliefs, so that, in the absence of scarcity in the underlying (e.g., if $\gamma = 1$), restricting trade in the derivative security would reduce the price distortion in the underlying. Similarly, in Section 4.6, we assume that S investors have overly pessimistic beliefs and evaluate the impact of a short-sales ban. To emphasize, it is not our objective to argue that short-sellers or speculators are distortionary. Instead, we analyze the effects of commonly proposed regulatory restrictions under these assumptions when the underlying security may be scarce.

It is also worth noting that, rather than analyze the welfare implications of regulatory policy, we instead restrict attention to price distortions relative to a frictionless benchmark. Since investors in our model have heterogeneous beliefs about the distribution of fundamentals, it is unclear how one defines a welfare criterion — see Brunnermeier, Simsek, and Xiong (2012) and Gilboa, Samuelson, and Schmeidler (2012) for recent approaches. In the context of our model, a frictionless benchmark price is a less ambiguous objective.

4.2. Equilibrium characterization

The market clearing conditions for the underlying asset are given by

$$x_L + x_S + x_B = Q \quad \text{and} \quad \gamma x_L + \gamma_B x_B + x_S \geq 0, \tag{25}$$

where $\gamma_B = 1$ if $x_B < 0$ and $\gamma_B = \gamma$ if $x_B > 0$. The market clearing condition for the derivative becomes

$$y_L + y_S + y_B = 0. \tag{26}$$

[14] From Proposition 4 (below), we have

$$P - \frac{\gamma\tau_L + \tau_S}{\tau_L + \tau_S} R = \frac{\tau_L m_L + \tau_S m_S}{\tau_L + \tau_S} - \frac{\nu}{\tau_L + \tau_S}(Q + \rho_L + \rho_S - y_B).$$

Therefore, if investors can observe both R and P, then they can infer a linear combination of expectations and aggregate supply as in standard noisy rational expectations models (e.g., Grossman and Stiglitz, 1980). It may be possible to study a very stylized model with asymmetric information within our framework, but we leave this for future work.

The following proposition characterizes the equilibrium prices and positions in the underlying security and the derivative.

Proposition 4. Given the economy above with L, S, and B investors, the equilibrium prices are given by

$$D = P_0 + \frac{\nu}{\tau_L+\tau_S}(x_B+y_B) + \frac{\delta}{\tau_L+\tau_S} y_B, \tag{27}$$

$$R = \max\left\{0, \frac{\delta}{\delta+\nu}\left[R_0 + \frac{1}{1-\gamma}\left(\frac{\nu}{\tau_L}x_B^+ - \frac{\nu}{\tau_S}x_B^-\right)\right]\right\}, \tag{28}$$

$$P = P_0 + \frac{\nu}{\tau_L+\tau_S}(x_B+y_B) + \frac{\gamma\tau_L+\tau_S}{\tau_L+\tau_S}R, \tag{29}$$

where $x_B^- = x_B 1_{\{x_B<0\}}$, $x_B^+ = x_B 1_{\{x_B>0\}}$,

$$P_0 \equiv \frac{\tau_L m_L + \tau_S m_S}{\tau_L+\tau_S} - \frac{\nu}{\tau_L+\tau_S}(Q+\rho_L+\rho_S) \quad \text{and} \tag{30}$$

$$R_0 \equiv \frac{1}{1-\gamma}\left[\frac{\nu}{\tau_S}\left(-\frac{\gamma}{1-\gamma}Q+\rho_S\right) - m_S + m_L - \frac{\nu}{\tau_L}\left(\frac{1}{1-\gamma}Q+\rho_L\right)\right]. \tag{31}$$

The equilibrium quantities $\{x_i, y_i\}_{i\in\{L,S,B\}}$ *are characterized in the proof.*

The characterization of the equilibrium in the above result highlights the incremental effect of the presence of B investors. In particular, note that P_0 is the frictionless price of the underlying asset if it is traded by L and S investors only, once we allow for heterogeneity in beliefs, risk-tolerances, and endowment shocks. Similarly, in the absence of B investors, the borrowing rate for the underlying is given by $\max\{0, R_0\}$ if L and S investors cannot trade the derivative, and by $\max\{0, R_0\delta/(\delta+\nu)\}$ if they can trade the derivative.

Note that when B investors participate in the underlying (i.e., $x_B \neq 0$) and the underlying is scarce, a larger demand from speculators increases the borrowing rate R, *irrespective* of whether they are long or short. This result is because, all else equal, their demand for positions in the underlying increases its scarcity. Moreover, due to a risk-sharing effect, a larger long (short) position from B investors increases (decreases, respectively) the price P.

It is also interesting to note that B investors do not always acquire a position in the underlying security. Specifically, as we characterize in the proof, when the underlying is scarce (i.e., $R>0$), and the investors' beliefs m_B are close to the hedgers' risk-adjusted valuation of the asset (i.e., P_0), B investors optimally choose not to participate in the underlying market. However, even in this case, the speculator can affect the price of the underlying. Since B investors trade in the derivative security, the net exposure of L and S investors in the derivative need not be zero. This net aggregate exposure to the fundamental shock F leads to a risk adjustment in the price of the underlying security. For instance, if B investors are net long in the derivative (i.e., $y_B>0$), then L and S investors must be net short, which implies that the total exposure of the hedgers to the fundamental shock is lower — as a result, the risk-premium component of P is smaller, and therefore P is higher. Similarly, if B investors are net short in the derivative (i.e., $y_B<0$), hedgers must be net long, which leads to a bigger risk-premium adjustment and hence a lower price for the underlying.[15] Moreover, even though speculators do not affect the borrowing rate R if they do not trade in the underlying security (i.e., when $x_B=0$), they do affect the net cost of being long and short in the underlying security (i.e., $P-\gamma R$ and $R-P$, respectively) through their effect on P. In particular, if speculators have an aggregate long position in the derivative, then, all else equal, the price P of the underlying is higher, which increases the net cost of a long position (i.e., $P-\gamma R$), but decreases the net cost of a short position (i.e., $R-P$).

While the general model is tractable and offers a great deal of flexibility, the equilibrium comparative statics results are not transparent given the large number of parameters. In the following subsections, we consider special cases of the above model which are more parsimonious, but highlight the important consequences of our analysis for regulatory policy, and the intuition for these results.

4.3. When L and S investors only trade in the underlying

There may be investors who are prohibited from accessing derivative markets, while speculators (or arbitrageurs) are able to able to trade in both the underlying and derivative markets.[16] This setting can be analyzed as a special case of our model. Specifically, suppose that L and S investors have correct beliefs (i.e., $m = m_L = m_S$), while speculators have incorrect beliefs (i.e., $m_B \neq m$), but only speculators can trade in the derivative market. The following corollary characterizes the equilibrium prices in this case, and the effect of speculators on the price of the underlying.

Corollary 3. Suppose that L and S investors have correct beliefs (i.e., $m_L = m_S = m$*) but can only trade in the underlying market (i.e.,* $y_L = y_S = 0$*), while speculators have incorrect beliefs (i.e.,* $m_B \neq m$*) and can trade both the underlying and derivative securities. Then, if the underlying security is scarce, the cost of borrowing R and the price of the underlying P are always higher in the presence of speculators than in their absence, irrespective of whether they have long or short positions in equilibrium.*

Since L and S investors do not trade the derivative, market clearing implies that the net derivative position across all speculators must be zero (i.e., $y_B=0$), and that the borrowing rate R is given by the expression in (28) with $\delta\to\infty$, or equivalently, $\delta/(\delta+\nu)=1$.[17] In this case, any distortionary effects of speculation are directly through the price and

[15] A casual argument proposed for why speculators in the derivative security should not affect the price of the underlying is based on the observation that derivatives are in zero net supply. For instance, if speculators are optimistic and so have an aggregate long exposure, other investors must be willing to take the other side, and so the buying pressure should not affect the price of the underlying security. However, what this casual argument fails to recognize is that the aggregate short exposure of these other traders (the hedgers in our model) in the derivative market will affect their positions in the underlying security, which in turn, affects the price of the underlying.

[16] We are grateful to the referee for highlighting this situation.

[17] When there is only one group of speculators, they do not trade in the derivative security. However, as the proof of Proposition 4 establishes, even when there are multiple groups of speculators, the net derivative position across them all must be zero (i.e., $\sum_{Bi} y_{Bi} = y_B = 0$), and the conclusions of Corollary 3 follow.

borrowing rate for the underlying security. If the underlying is scarce, an increase in speculative trading in either direction (i.e., long or short) increases the borrowing rate R. In addition to this effect, speculators increase the price of the underlying if they are long, and decrease it if they are short. However, the first effect dominates and the overall effect of speculators on the price of the underlying is given by

$$P-\left(P_0+\frac{\gamma\tau_L+\tau_S}{\tau_L+\tau_S}R_0\right)=\frac{1}{1-\gamma}\frac{\nu}{\tau_L}x_B^+-\frac{\gamma}{1-\gamma}\frac{\nu}{\tau_S}x_B^-, \tag{32}$$

which implies that both optimistic and pessimistic speculators (i.e., $x_B>0$ and $x_B<0$, respectively) always increase the price of the underlying security. Intuitively, since L and S investors do not access the derivative markets, the presence of the derivative does not reduce the scarcity of the underlying. Therefore, any trading by speculators leads to a price distortion in the underlying. As we discuss in the next subsection, this result is in contrast to the effect of speculators when they can only access the derivatives market but hedgers can trade both securities.

4.4. *The effect of speculation only in the derivative market*

In this subsection, we consider the case where speculators can only trade in the derivative security, and hedgers are exposed to their distortionary trades through the derivatives market. This scenario is often cited by regulators as justification for limiting speculative trade in derivative securities. For instance, on November 15, 2011, the European Parliament adopted regulation that restricts investors from entering into uncovered, or “naked,” CDS on sovereign debt after November 2012, in an effort to curb speculation against a country's default.[18] Similarly, under the mandate of the Dodd-Frank Act, the CFTC proposed position limits on commodity derivatives so as: “(i) To diminish, eliminate, or prevent excessive speculation as described under this section; (ii) To deter and prevent market manipulation, squeezes, and corners; (iii) To ensure sufficient market liquidity for bona fide hedgers; and (iv) To ensure that the price discovery function of the underlying market is not disrupted.”[19]

To analyze this situation, we assume that the L and S investors in our model have correct beliefs (i.e., $m=m_L=m_S$), while speculators have incorrect beliefs (i.e., $m_B\neq m$) and can only trade the derivative security (i.e., $x_B=0$). To simplify the analysis, we also assume that all investors have the same risk-tolerance (i.e., $\tau_i=\tau$), and that the endowment shocks of L and S investors perfectly offset each other (i.e., $\rho_S=-\rho_L=\rho$). The frictionless benchmark in this setup is when there are no constraints on borrowing and lending (i.e., $\gamma=1$), and no speculators (i.e., $\tau_B=0$). In this case, the price of the underlying security simplifies to[20]

$$P_0=m-\frac{\nu}{2\tau}Q. \tag{33}$$

Relative to this frictionless benchmark, if the underlying is scarce, the price distortion when investors have access to the derivative is given by

$$\Delta P=\frac{1+\gamma}{2}R+\frac{\nu}{2\tau}y_B=\frac{1+\gamma}{2}\frac{\delta}{\nu+\delta}R_0+\frac{\nu}{2\tau}y_B, \tag{34}$$

and the distortion when investors do not have access to the derivative is given by

$$\Delta P=\frac{1+\gamma}{2}R_0. \tag{35}$$

Therefore, despite the distortionary effects of speculative traders, the presence of a derivative may reduce the overall price distortion in the underlying security, through its effect on scarcity. The following result characterizes sufficient conditions for this result.

Proposition 5. Suppose that L and S investors have correct beliefs (i.e., $m_L=m_S=m$), while speculators have incorrect beliefs (i.e., $m_B\neq m$) and can only trade in the derivatives market (i.e., $x_B=0$). Moreover, suppose that $\tau_i=\tau$ for all i and $\rho_S=-\rho_L=\rho$.

(i) *If*

$$\left|\frac{2}{3}(m_B-m)+\frac{\nu}{3\tau}Q\right|<(1+\gamma)R_0,$$

then the price distortion in the underlying (relative to the frictionless price) is lower in the presence of the derivative with speculators than in the absence of derivatives.

(ii) *If*

$$\left|\frac{2}{3}(m_B-m)+\frac{\nu}{3\tau}Q\right|>\frac{\nu+2\delta}{\nu}(1+\gamma)R_0,$$

then the price distortion in the underlying (relative to the frictionless price) is higher in the presence of the derivative with speculators than in the absence of derivatives.

(iii) *If speculators are long (i.e., $y_B>0$), or sufficiently short (i.e.,*

$$y_B<-(1+\gamma)\frac{\delta}{\nu+\delta}\frac{\tau}{\nu}R_0,$$

then the price distortion in the underlying (relative to the frictionless price) is increasing in the aggregate position of speculators (i.e., $|y_B|$). Otherwise, if

$$-(1+\gamma)\frac{\delta}{\nu+\delta}\frac{\tau}{\nu}R_0<y_B<0,$$

then the price distortion is increasing in the aggregate position of speculators.

The proof of the above result follows from the expressions for the price distortion in the presence and absence of derivatives with speculators. For parts (i) and (ii), note

[18] See http://ec.europa.eu/internal_market/securities/short_selling_en.htm.

[19] See http://www.cftc.gov/LawRegulation/DoddFrankAct/Rulemakings/DF_26_PosLimits/index.htm.

[20] Alternatively, one can set the frictionless benchmark as the case in which there are no constraints on borrowing and lending (i.e., for $\gamma=1$), no noise in the derivative (i.e., $\delta=0$), and speculators have correct beliefs (i.e., $m_B=m$). In this case, the price of the underlying security is $P_0=m-Q\nu/(3\tau)$. As is apparent, the results in Proposition 5 remain unchanged for this benchmark.

that the price distortion with the derivative is lower when

$$-\frac{1+\gamma}{2}\frac{\nu+2\delta}{\nu+\delta}R_0 < \frac{\nu}{2\tau}y_B < \frac{1+\gamma}{2}\frac{\nu}{\nu+\delta}R_0, \tag{36}$$

and higher when either

$$\frac{\nu}{2\tau}y_B > \frac{1+\gamma}{2}\frac{\nu}{\nu+\delta}R_0 \quad \text{or} \quad \frac{\nu}{2\tau}y_B < -\frac{1+\gamma}{2}\frac{\nu+2\delta}{\nu+\delta}R_0, \tag{37}$$

and that the optimal derivative position for the speculators is given by

$$y_B = \frac{\tau}{\delta+\nu}(m_B-D) = \frac{2}{3}\frac{\tau}{\delta+\nu}(m_B-m)+\frac{1}{3}\frac{\nu}{\delta+\nu}Q. \tag{38}$$

Intuitively, the presence of the derivative reduces the price distortion in the underlying when the security is scarce (i.e., R_0 is large), the outstanding supply Q is small, and the disagreement between speculators and hedgers is not large (i.e., $|m_B-m|$ is small). Notably, whether the presence of the derivative reduces price distortion is unaffected by the noise in the payoff (i.e., δ), although the magnitude of the change in the price distortion is obviously affected. Finally, part (iii) emphasizes that even in the presence of a derivative, completely restricting trade by speculators may lead to a larger price distortion in the underlying than allowing for some trade by (short) speculators.

This result highlights the tradeoff that regulators face when restricting trade in derivatives. While allowing for unrestricted trade in derivatives in the presence of aggressive speculators may increase the price distortion in the underlying, excessive trading restrictions can make the underlying more scarce and, thereby, distort prices.

4.5. The effect of speculation only in the underlying

As documented by Kaplan, Moskowitz, and Sensoy (2013), Rizova (2011), Evans, Ferreira, and Porras Prado (2012), and others, a large part of the equity lending market has recently been dominated by long-only investors who lend stocks (e.g., mutual funds). An open question in this literature is whether the lending activity of such buy-and-hold investors distorts the price of the underlying security.[21] We can analyze this situation as a special case of our model in which B investors are restricted to have only long positions in the underlying security and are unable to trade in the derivative.[22] In particular, suppose that L and S investors have correct beliefs (i.e., $m=m_L=m_S$), and B investors can only trade in the underlying market (i.e., $y_B=0$). The following corollary characterizes the equilibrium prices in this case, and the effect of long-only (or short-only) investors on the price of the underlying and the derivative.

Corollary 4. Suppose that L and S investors have correct beliefs (i.e., $m_L=m_S=m$), and speculators can only trade in the underlying security (i.e., $y_B=0$). Then:

(i) *In the presence of long-only speculators (i.e., $x_B>0$), the price of the derivative D, the borrowing cost R, and the price of the underlying P are all higher than in their absence.*

(ii) *In the presence of short-only speculators (i.e., $x_B<0$), the price of the derivative D is always lower and the borrowing cost R is always higher than in their absence. The price of the underlying is higher in their presence if $\tau_S(1-\gamma)/(\tau_S+\gamma\tau_L)<\delta/(\delta+\nu)$, and lower if $\tau_S(1-\gamma)/(\tau_S+\gamma\tau_L)>\delta/(\delta+\nu)$.*

As discussed before, when the underlying security is scarce, the effect of speculative trading in the underlying is two-fold. First, higher demand from either side leads to an increase in the borrowing rate since it increases the scarcity in the underlying. Second, due to risk-sharing in the underlying, a larger long position from speculators increases the price of the underlying, while a larger short position decreases the price. The overall effect on the underlying price then depends on the relative magnitudes of the two effects, and their signs, as is characterized by the expression:

$$P-\left(P_0+\frac{\gamma\tau_L+\tau_S}{\tau_L+\tau_S}\frac{\delta}{\nu+\delta}R_0\right) = \frac{\nu}{\tau_L+\tau_S}x_B+\frac{\gamma\tau_L+\tau_S}{\tau_L+\tau_S}\frac{\delta}{\delta+\nu}\times\frac{1}{1-\gamma}\left(\frac{\nu}{\tau_L}x_B^+-\frac{\nu}{\tau_S}x_B^-\right). \tag{39}$$

In the case of long-only speculators, the two effects reinforce each other, and as such, *unambiguously* increase the price of the underlying. However, for short-only speculators in the underlying, the overall effect depends on

$$P-\left(P_0+\frac{\gamma\tau_L+\tau_S}{\tau_L+\tau_S}\frac{\delta}{\nu+\delta}R_0\right) = \left(1-\frac{\gamma\tau_L+\tau_S}{\tau_S(1-\gamma)}\frac{\delta}{\delta+\nu}\right)\frac{\nu}{\tau_L+\tau_S}x_B. \tag{40}$$

In particular, note that when the derivative is a perfect substitute (i.e., $\delta=0$), the price of the underlying decreases with shorting demand from B investors, while in the absence of the derivative (i.e., $\delta=\infty$), the price of the underlying increases (as in Section 4.3).

Finally, it is important to note that even though B investors do not trade the derivative security, they affect the price of the derivative through their trading in the underlying since the two securities are substitutes. If B investors have a net long position in the underlying, they increase the price of the underlying. All else equal, this makes the derivative a more attractive substitute for long investors, which drives up its price. Similarly, when B investors have a net short position in the underlying, this drives down the price of the derivative. Therefore, our model generates an interesting prediction: increased short-selling by investors in the underlying always *decreases* the derivative price, but *increases* the price of the underlying when the derivative payoff is sufficiently noisy.

4.6. The effect of short-sales bans

Regulations that restrict short-selling are often justified as a means to limit the effect of pessimistic investors who would otherwise distort prices below their fundamental

[21] We thank the referee for pointing out this case.

[22] While we interpret B investors as speculators in the earlier subsections, we do not think of mutual funds as speculators. However, it is notationally convenient to denote the mutual funds as B investors, so that we can appeal to the general results in Section 4.2.

value. For instance, in September 2008, the SEC temporarily banned short-selling in financial stocks "to combat market manipulation that threatens investors and capital markets."[23] More recently, a number of EU member countries imposed short-sales bans for bank stocks to "restrict the benefits that can be achieved by spreading false rumors."[24]

To capture this scenario, we assume that there are no speculators in the market (i.e., $\tau_B=0$), that L investors have correct beliefs about fundamentals (i.e., $m_L=m$), and that S investors are pessimistic (i.e., $m_S<m$). In this case, while short-sellers are pessimistic, the following proposition shows that for some parameter values, banning short-sales (i.e., setting $\gamma=0$) actually reduces the price of the underlying asset, even in the presence of derivatives.

Corollary 5. Suppose there are no speculators (i.e., $\tau_B=0$), L investors have correct beliefs (i.e., $m_L=m$), and S investors are pessimistic (i.e., $m_S<m$).

(i) *If short-selling is allowed, then the underlying price is given by*

$$P=\frac{\tau_L m+\tau_S m_S}{\tau_L+\tau_S}-\frac{\nu}{\tau_L+\tau_S}(Q+\rho_S+\rho_L)+\frac{\gamma\tau_L+\tau_S}{\tau_L+\tau_S}R, \quad \textit{where} \tag{41}$$

$$R=\max\left\{0,\frac{\frac{\nu}{\tau_S}\left(\rho_S-\frac{\gamma}{1-\gamma}Q\right)-\frac{\nu}{\tau_L}\left(\frac{1}{1-\gamma}Q+\rho_L\right)+m-m_S}{(1-\gamma)\left(1+\frac{\nu}{\delta}\right)}\right\}. \tag{42}$$

(ii) *If short-selling is not allowed, then the underlying price is given by*

$$P_{ns}=\frac{\tau_L m+\tau_S m_S}{\tau_L+\tau_S}-\frac{\nu}{\tau_L+\tau_S}(Q+\rho_S+\rho_L)+\frac{\tau_S}{\tau_L+\tau_S}R_{ns} \tag{43}$$

where

$$R_{ns}=\max\left\{0,\frac{\frac{\nu}{\tau_S}\rho_S-\frac{\nu}{\tau_L}(Q+\rho_L)+m-m_S}{\left(1+\frac{\nu}{\delta}\right)}\right\}. \tag{44}$$

The price of the derivative in either case is given by

$$D=\frac{\tau_L m+\tau_S m_S}{\tau_L+\tau_S}-\frac{\nu}{\tau_L+\tau_S}(Q+\rho_S+\rho_L). \tag{45}$$

Even though short-selling is not allowed, it is instructive to decompose the underlying price into a frictionless component and a shadow cost of borrowing the security (i.e., R_{ns}). For instance, in the special case when trading in derivatives is not allowed (i.e., $\delta\to\infty$ which implies $\delta/(\delta+\nu)=1$) and short-selling is not allowed (i.e., $\gamma=0$), the shadow cost of borrowing the security is given by

$$R_{ns}=\frac{\nu}{\tau_S}\rho_S-\frac{\nu}{\tau_L}(Q+\rho_L)+m-m_S, \tag{46}$$

and the price of the underlying reduces to the familiar expression

$$P_{ns}=\frac{\tau_L m+\tau_S m_S}{\tau_L+\tau_S}-\frac{\nu}{\tau_L+\tau_S}(Q+\rho_S+\rho_L)+\frac{\tau_S}{\tau_L+\tau_S}R_{ns} \tag{47}$$

$$P_{ns}=m-\frac{\nu}{\tau_L}(Q+\rho_L). \tag{48}$$

That is, if L investors must hold the underlying security and cannot trade the derivative with the S investors, then the price of the underlying security is completely determined by the beliefs and preferences of L investors. If short-selling is not allowed, but S investors are allowed to trade the derivative security, then their beliefs affect the value of the underlying security. In fact, lower noise in the derivative payoff (i.e., lower δ), decreases the price P_{ns} when short-sales are not allowed, since S investors are able to express their beliefs more strongly. In the limit, if there is no noise in the derivative (i.e., $\delta=0$), then imposing the short-sales constraint has no effect on the price of the underlying. In contrast, regardless of the noise in the derivative payoff, imposing a short-selling ban has no effect on the price of the derivative security itself. However, positions in the derivative security *do* change in response to the ban — as in our benchmark model in Section 3, the optimal demand for the derivative security depends on the (shadow) cost of borrowing R, which depends on whether the short-selling ban is in effect.

Finally, note that the derivative of P with respect to γ at $\gamma=0$ is given by

$$\left.\frac{\partial P}{\partial\gamma}\right|_{\gamma=0}=\frac{\delta}{\delta+\nu}\left(m-m_s+\nu\left[\frac{\rho_S}{\tau_S}-\frac{1}{\tau_L}(\rho_L+2Q)\right]\right), \tag{49}$$

which is positive when Q is small. Thus, if the underlying security is scarce relative to its outstanding supply, its price may be lower when short-sales are banned (i.e., $\gamma=0$). Intuitively, when the underlying is scarce but short-selling is not allowed, the shadow cost of borrowing the underlying is not zero (i.e., $R_{ns}\neq 0$), but none of this borrowing cost is reflected in the price P_{ns}. In this case, relaxing the short-selling ban increases the price of the underlying since it allows L investors to lend some of their positions at R_{ns}, which increases the price they are willing to pay for the underlying asset. The above result generalizes a similar result in Duffie, Gârleanu, and Pedersen (2002) to a setting in which investors can trade derivatives.

To summarize, we have the following results concerning the effect of a short-sales ban on an underlying asset that may be scarce, when investors have access to a derivative.

Proposition 6. Suppose that there are no speculators (i.e., $\tau_B=0$), L investors have correct beliefs (i.e., $m_L=m$), S investors are pessimistic (i.e., $m_S<m$). Then:

(i) *if the aggregate supply of the underlying is low enough, imposing a short-sales ban can lower the price of the underlying security, even when L and S can trade the derivative,*
(ii) *the price of the underlying security decreases as the noise in the derivative decreases (i.e., δ decreases), even in the presence of a short-sales ban,*
(iii) *imposing a short-sales ban has no effect on the price of the underlying if investors can trade a derivative with no noise (i.e., $\delta=0$),*

[23] See http://www.sec.gov/news/press/2008/2008-211.htm.

[24] See http://www.ft.com/intl/cms/s/0/9a55839a-c42d-11e0-ad9a-00144feabdc0.html.

(iv) *imposing a short-sales ban has the largest effect on the price of the underlying if investors cannot trade a derivative* (*i.e.*, $\delta \to \infty$, *or equivalently*, $\delta/(\delta+\nu)=1$), *and*

(v) *imposing a short-sales ban has no effect on the derivative price but increases the size of equilibrium derivative positions*.

As these results highlight, regulation that bans short-selling to mitigate the distortionary effects of trading by pessimistic investors may be ineffective if investors can trade derivatives, and can actually decrease the underlying price if the security is sufficiently scarce. These results are in contrast to the over-pricing effect of short-sale constraints in many standard models (e.g., Miller, 1977) that is often used to motivate such regulation.

While beyond the scope of our model, in settings where the noise in the derivative payoff is endogenously affected by the constraint on borrowing and lending in the underlying (i.e., if δ is affected by γ), parts (i) and (ii) of the above proposition may be particularly relevant. For instance, consider a setting in which market makers in the derivative market must hedge their positions by trading in the underlying security. In this case, one might expect that restricting short-sales (i.e., decreasing γ) makes it more difficult for market makers to hedge their derivative exposures and therefore increases the noise δ in the derivative payoff.[25] As a result, in this setting, (i) and (ii) predict opposite effects on the underlying price. In particular, the first part of the result implies that a decrease in γ might lead to lower P, while the second part implies that an increase in δ leads to a higher P. In an alternative setting, suppose that investors face search frictions in trading both the underlying and derivative securities. In this case, increasing scarcity in the underlying by restricting short-sales might lead to increased liquidity and lower search costs in the derivative security, as more investors coordinate their trading in the latter security. As a result, the two effects could reinforce each other a decrease in γ and a decrease in δ might both lead to a decrease in P.

5. Conclusions

When the demand for long and short positions exceeds an asset's capacity to support such positions, it becomes *scarce* and its price is distorted relative to its value in a frictionless market. In this case, we show that even simple derivatives are no longer redundant since the presence of a derivative can alleviate the scarcity in the underlying asset, and, therefore, reduce the distortion in its price. Finally, we show that accounting for scarcity is important in analyzing the effects of regulatory policy, such as short-selling bans and derivative position limits, that restrict trade in an underlying asset or its derivative.

To highlight the role of derivatives in relaxing the scarcity of the underlying asset, we focus exclusively on a simple (linear) derivative that is otherwise redundant. Our model predicts that, all else equal, assets with simple derivatives that are close substitutes should have lower borrowing fees and therefore smaller price distortions. A number of empirical papers that study how the introduction of options impacts the underlying asset find evidence consistent with our model's predictions for simple derivatives. For example, Sorescu (2000) finds that the price of the underlying stock tends to fall when (nonlinear) options on the stock are introduced. Danielsen and Sorescu (2001) provide empirical evidence that is consistent with the notion that the introduction of options mitigates short-sale constraints. However, one must be cautious in interpreting these results as conclusive evidence for our model. While options may relax the scarcity of the underlying asset, they can also affect the equilibrium by allowing investors to trade new sources of risk (such as the volatility of the asset).[26]

In testing the empirical predictions of our model, it is also important to account for a number of features of the market that we have abstracted from in order to more clearly highlight the role of derivatives in relaxing the scarcity of the underlying. For instance, we do not consider the effect of financial intermediaries who make markets in the derivative securities. If market makers need to hedge their derivative positions with positions in the underlying, then increased trade in derivatives can lead to an increase in the scarcity of the underlying. One must also keep in mind that the introduction of derivatives is endogenous — it may be more likely that derivatives are introduced on underlying securities that are scarce. In particular, it is important to control for this endogeneity in any cross-sectional tests of our model's predictions.

Finally, 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, may also be important for understanding the relative size of derivative markets across various assets. For instance, although U.S. equity indexes are extremely liquid, they are often accompanied by very large futures markets, which may be driven by the fact that it can be difficult to simultaneously short-sell all of the components of an index.[27] 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). Our model may also be useful for understanding why even extremely liquid securities like U.S. Treasuries may be accompanied by extremely large futures markets.[28] As our model highlights, the size of the derivative market may be driven, not by the constraint on borrowing and lending in the underlying per se, but by the demand for both long and short positions *relative* to the trading capacity of the underlying that results from this constraint.

[25] In the extreme, investors may be unwilling or unable to make markets in the derivative security when the underlying is sufficiently scarce (e.g., single name futures contracts), and instead only offer derivatives on correlated securities (e.g., broader index futures).

[26] For examples, see Detemple and Selden (1991), Zapatero (1998), Boyle and Wang (2001), and Bhamra and Uppal (2009).

[27] For example, in October 2011, the average daily trading volume in Standard and Poor's 500 (S&P 500) futures on the Chicago Mercantile Exchange (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.

[28] 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.

Appendix A. Proofs of main results

Proof of Proposition 3. 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.,

$$dx_i=\frac{\partial x_i}{\partial \Pi_i}d\Pi_i+\frac{\partial x_i}{\partial D}dD=0, \tag{50}$$

which implies the price of the derivative must change by

$$dD=-\frac{\partial x_i/\partial \Pi_i}{\partial x_i/\partial D}d\Pi_i. \tag{51}$$

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

$$dy_i=\frac{\partial y_i}{\partial \Pi_i}d\Pi_i+\frac{\partial y_i}{\partial D}dD=\left(\frac{\partial y_i}{\partial \Pi_i}-\frac{\partial y_i}{\partial D}\frac{\partial x_i/\partial \Pi_i}{\partial x_i/\partial D}\right)d\Pi_i. \tag{52}$$

Hence,

$$\frac{\partial y_i}{\partial \Pi_i}-\frac{\partial y_i}{\partial D}\frac{\partial x_i/\partial \Pi_i}{\partial x_i/\partial D}<0 \tag{53}$$

is a sufficient condition for $dy_i/d\Pi_i<0$, and vice versa, which, combined with the observation that $y_L=-y_S$ and $P=\Pi_L/(1-\gamma)-\Pi_S\gamma/(1-\gamma)$, gives us the result. □

Proof of Corollary 2. Dropping the subscript i, the first order conditions of the optimal portfolio problem in Eq. (13) are

$$0=\mathbb{E}[\{F-\Pi\}u'(W_0+\rho F+x[F-\Pi]+y[F+\varepsilon-D])], \tag{54a}$$

$$0=\mathbb{E}[\{F+\varepsilon-D\}u'(W_0+\rho F+x[F-\Pi]+y[F+\varepsilon-D])]. \tag{54b}$$

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

$$\frac{\partial x}{\partial \Pi}=\frac{(A+2B+C)\mathbb{E}[u'+x(F-\Pi)u'']-(A+B)x\mathbb{E}[(F+\varepsilon-D)u'']}{AC-B^2}, \tag{55a}$$

$$\frac{\partial y}{\partial \Pi}=\frac{Ax\mathbb{E}[(F+\varepsilon-D)u'']-(A+B)\mathbb{E}[u'+x(F-\Pi)u'']}{AC-B^2}, \tag{55b}$$

$$\frac{\partial x}{\partial D}=\frac{(A+2B+C)y\mathbb{E}[(F-\Pi)u'']-(A+B)\mathbb{E}[u'+y(F+\varepsilon-D)u'']}{AC-B^2}, \tag{55c}$$

$$\frac{\partial y}{\partial D}=\frac{A\mathbb{E}[u'+y(F+\varepsilon-D)u'']-(A+B)y\mathbb{E}[(F-\Pi)u'']}{AC-B^2}, \tag{55d}$$

where

$$A=\mathbb{E}[(F-\Pi)^2u''], \tag{56a}$$

$$B=\mathbb{E}[(F-\Pi)(\varepsilon-D+\Pi)u''], \tag{56b}$$

$$C=\mathbb{E}[(\varepsilon-D+\Pi)^2u'']. \tag{56c}$$

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

$$\mathbb{E}[(F-\Pi)u'']=-k\underbrace{\mathbb{E}[(F-\Pi)u']}_{0\text{ by Eq. (54a)}}=0, \tag{57}$$

$$\mathbb{E}[(F+\varepsilon-D)u'']=-k\underbrace{\mathbb{E}[(F+\varepsilon-D)u']}_{0\text{ by Eq. (54b)}}=0. \tag{58}$$

Also,

$$\mathbb{E}[(F-\Pi)(\varepsilon-D+\Pi)u''] \tag{59a}$$

$$=\mathbb{E}\left[\left(F-\mathbb{E}\left[F\frac{u'}{\mathbb{E}[u']}\right]\right)\left(\varepsilon-\mathbb{E}\left[\varepsilon\frac{u'}{\mathbb{E}[u']}\right]\right)u''\right]\quad\text{by Eq. (54)} \tag{59b}$$

$$=-k\mathbb{E}[u']\mathbb{E}\left[\left(F-\mathbb{E}\left[F\frac{u'}{\mathbb{E}[u']}\right]\right)\left(\varepsilon-\mathbb{E}\left[\varepsilon\frac{u'}{\mathbb{E}[u']}\right]\right)\frac{u'}{\mathbb{E}[u']}\right] \tag{59c}$$

$$=-k\mathbb{E}[u']\underbrace{\mathbb{E}\left[\left(F-\mathbb{E}\left[F\frac{u'}{\mathbb{E}[u']}\right]\right)\frac{u'}{\mathbb{E}[u']}\right]}_{0}\underbrace{\mathbb{E}\left[\left(\varepsilon-\mathbb{E}\left[\varepsilon\frac{u'}{\mathbb{E}[u']}\right]\right)\frac{u'}{\mathbb{E}[u']}\right]}_{0} \tag{59d}$$

$$=0. \tag{59e}$$

To move from Eq. (59c) to Eq. (59d), note that $u'/\mathbb{E}[u']$ defines a change of probability measure under which F and ε are independent (since they are independent under the original measure).[29] Combining these results, we have

$$\frac{\partial x}{\partial \Pi}=\frac{(A+C)\mathbb{E}[u']}{AC}=\frac{\mathbb{E}[(F-\Pi)^2u'']\mathbb{E}[u']+\mathbb{E}[(\varepsilon-D+\Pi)^2u'']\mathbb{E}[u']}{\mathbb{E}[(F-\Pi)^2u'']\mathbb{E}[(\varepsilon-D+\Pi)^2u'']}, \tag{60a}$$

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

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

so that

$$\mathbb{E}\left[\left(F-\mathbb{E}\left[F\frac{u'}{\mathbb{E}[u']}\right]\right)\left(\varepsilon-\mathbb{E}\left[\varepsilon\frac{u'}{\mathbb{E}[u']}\right]\right)\frac{u'}{\mathbb{E}[u']}\right]$$
$$=\mathbb{E}\left[(F-\mathbb{E}[Fg(F)h(\varepsilon)])(\varepsilon-\mathbb{E}[\varepsilon g(F)h(\varepsilon)])g(F)h(\varepsilon)\right]$$

$$=\underbrace{\mathbb{E}[(F-\mathbb{E}[Fg(F)])g(F)]}_{0}\underbrace{\mathbb{E}[(\varepsilon-\mathbb{E}[\varepsilon h(\varepsilon)])h(\varepsilon)]}_{0}.$$

.

$$\frac{\partial y}{\partial \Pi} = \frac{\partial x}{\partial D} = -\frac{\partial y}{\partial D} = -\frac{A\mathbb{E}[u']}{AC} = -\frac{\mathbb{E}[(F-\Pi)^2 u'']\mathbb{E}[u']}{\mathbb{E}[(F-\Pi)^2 u'']\mathbb{E}[(\varepsilon - D+\Pi)^2 u'']}, \tag{60b}$$

which establishes the result. □

Proof of Proposition 4. We consider a more general specification than in the proposition, in which there are two types of speculators, indexed by $i \in \{BO, BP\}$ (for optimists and pessimists, respectively). The first-order conditions for investor $i \in \{L, S, BO, BP\}$ imply that

$$\tau_i(m_i - P + \gamma_i R) = \nu(x_i + y_i + \rho_i) \quad \text{and} \tag{61}$$

$$\tau_i(m_i - D) = \nu(x_i + y_i + \rho_i) + \delta y_i. \tag{62}$$

Let $P_0 = (\tau_L m_L + \tau_S m_S)/(\tau_L + \tau_S) - (\nu/(\tau_L+\tau_S))(Q + \rho_L + \rho_S)$, and let $x_B = x_{BO} + x_{BP}$, and $y_B = y_{BO} + y_{BP}$. The market clearing condition for the derivative implies that

$$\tau_L m_L + \tau_S m_S - (\tau_L + \tau_S)D = \nu(Q + \rho_L + \rho_S - (x_B + y_B)) - \delta y_B \tag{63}$$

$$\Rightarrow D = P_0 + \frac{\nu}{\tau_L + \tau_S}(x_B + y_B) + \frac{\delta}{\tau_L + \tau_S} y_B \tag{64}$$

and market clearing in the cash-market implies

$$P = P_0 + \frac{\nu}{\tau_L + \tau_S}(x_B + y_B) + \frac{\gamma \tau_L + \tau_S}{\tau_L + \tau_S} R. \tag{65}$$

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

$$R(1-\gamma) = \frac{\nu}{\tau_S}(x_S + \rho_S + y_S) - m_S - \frac{\nu}{\tau_L}(x_L + \rho_L + y_L) + m_L \tag{66}$$

and since $(\nu/\tau_S)y_S - (\nu/\tau_L)y_L = (\nu/\delta)(\gamma - 1)R$, we have

$$R = \max\left\{0, \frac{\frac{\nu}{\tau_S}(x_S + \rho_S) - m_S + m_L - \frac{\nu}{\tau_L}(x_L + \rho_L)}{(1-\gamma)\left(1 + \frac{\nu}{\delta}\right)}\right\}. \tag{67}$$

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

$$x_S = -\frac{\gamma}{1-\gamma} Q - x_B^- \quad \text{and} \quad x_L = \frac{1}{1-\gamma} Q - x_B^+, \tag{68}$$

where

$$x_B^- = (x_{BO} 1_{\{x_{BO} < 0\}} + x_{BP} 1_{\{x_{BP} < 0\}}) \quad \text{and} \tag{69}$$

$$x_B^+ = (x_{BO} 1_{\{x_{BO} > 0\}} + x_{BP} 1_{\{x_{BP} > 0\}}). \tag{70}$$

This implies

$$R = \max\left\{0, \frac{\delta}{\delta + \nu}\left(R_0 + \frac{1}{1-\gamma}\left(\frac{\nu}{\tau_L} x_B^+ - \frac{\nu}{\tau_S} x_B^-\right)\right)\right\}, \tag{71}$$

where

$$R_0 \equiv \frac{1}{1-\gamma}\left(\frac{\nu}{\tau_S}\left(-\frac{\gamma}{1-\gamma} Q + \rho_S\right) - m_S + m_L - \frac{\nu}{\tau_L}\left(\frac{1}{1-\gamma} Q + \rho_L\right)\right).$$

The equilibrium quantities can be computed by plugging in the expressions for the price into the first-order conditions for each type of investor. Since $y_{Bi} = (\tau_{Bi}/\delta)(P - \gamma_{Bi} R - D)$ and $x_{Bi} + y_{Bi} = (\tau_{Bi}/\nu)(m_{Bi} - P + \gamma_{Bi} R)$, we have for $j \neq i$:

$$y_{Bi} = \frac{\tau_{Bi}}{\delta}\left(\frac{\gamma \tau_L + \tau_S}{\tau_L + \tau_S} R - \frac{\delta}{\tau_L + \tau_S}(y_{BO} + y_{BP}) - \gamma_{Bi} R\right) \tag{72}$$

$$\Rightarrow \left(1 + \frac{\tau_{Bi}}{\tau_L + \tau_S}\right) y_{Bi} = \frac{\tau_{Bi}}{\delta}\left(\frac{\gamma \tau_L + \tau_S}{\tau_L + \tau_S} - \gamma_{Bi}\right) R - \frac{\tau_{Bi}}{\tau_L + \tau_S} y_{Bj}. \tag{73}$$

But this implies

$$\left(1 + \frac{\tau_{Bi}}{\tau_L + \tau_S}\right) x_{Bi} = \frac{\tau_{Bi}}{\nu}(m_{Bi} - P_0) - \frac{\tau_{Bi}}{\tau_L + \tau_S} x_{Bj} + \left(\frac{\tau_{Bi}}{\nu} + \frac{\tau_{Bi}}{\delta}\right)\left(\gamma_{Bi} - \frac{\gamma \tau_L + \tau_S}{\tau_L + \tau_S}\right) R. \tag{74}$$

When $R = 0$, the above reduces to

$$\left(1 + \frac{\tau_{Bi}}{\tau_L + \tau_S}\right) y_{Bi} = -\frac{\tau_{Bi}}{\tau_L + \tau_S} y_{Bj} \Rightarrow y_{Bi} = -\frac{\tau_{Bi}}{\tau_L + \tau_S} y_B \Rightarrow y_{Bi} = 0 \tag{75}$$

and this implies

$$x_{Bi} = \frac{\tau_{Bi}}{\nu}\left[\frac{(\tau_L + \tau_S + \tau_{BO} + \tau_{BP}) m_{Bi} - (\tau_{BO} m_{BO} + \tau_{BP} m_{BP}) - (\tau_L + \tau_S) P_0}{\tau_L + \tau_S + \tau_{BO} + \tau_{BP}}\right]. \tag{76}$$

When $R > 0$, then

$$\begin{aligned} x_{Bi}\left(1 + \frac{\tau_{Bi}}{\tau_L + \tau_S}\right) - \frac{\tau_{Bi}}{\nu}\left(\gamma_{Bi} - \frac{\gamma \tau_L + \tau_S}{\tau_L + \tau_S}\right)\frac{1}{1-\gamma}\left(\frac{\nu}{\tau_L} x_{Bi}^+ - \frac{\nu}{\tau_S} x_{Bi}^-\right) \\ = \frac{\tau_{Bi}}{\nu}(m_{Bi} - P_0) - \frac{\tau_{Bi}}{\tau_L + \tau_S} x_{Bj} \\ + \frac{\tau_{Bi}}{\nu}\left(\gamma_{Bi} - \frac{\gamma \tau_L + \tau_S}{\tau_L + \tau_S}\right)\left(R_0 + \frac{1}{1-\gamma}\frac{\nu}{\tau_L} x_{Bj}^+ - \frac{1}{1-\gamma}\frac{\nu}{\tau_S} x_{Bj}^-\right). \end{aligned} \tag{77}$$

This implies that

(i) If $x_B > 0$, then $\gamma_B = \gamma$, and so

$$x_{Bi} = \frac{\frac{\tau_{Bi}}{\nu}(m_{Bi} - P_0) - \frac{\tau_{Bi}}{\nu}\frac{\tau_S}{\tau_S + \tau_L}\left((1-\gamma)R_0 + \frac{\nu}{\tau_L} x_{Bj}^+ - \frac{\nu}{\tau_S} x_{Bj}^-\right) - \frac{\tau_{Bi}}{\tau_L + \tau_S} x_{Bj}}{1 + \frac{\tau_{Bi}}{\tau_L + \tau_S} + \frac{\tau_{Bi}}{\tau_L}\frac{\tau_S}{\tau_L + \tau_S}} > 0, \tag{78}$$

and so $m_{Bi} - P_0 > (\tau_S/(\tau_S + \tau_L))(1-\gamma)R_0 + (\nu/\tau_L)x_{Bj}^+$ is sufficient.

(ii) If $x_B < 0$, then $\gamma_B = 1$ and so

$$x_{Bi} = \frac{\frac{\tau_{Bi}}{\nu}(m_{Bi} - P_0) + \frac{\tau_{Bi}}{\nu}\frac{\tau_L}{\tau_L + \tau_S}\left((1-\gamma)R_0 + \frac{\nu}{\tau_L} x_{Bj}^+ - \frac{\nu}{\tau_S} x_{Bj}^-\right) - \frac{\tau_{Bi}}{\tau_L + \tau_S} x_{Bj}}{1 + \frac{\tau_{Bi}}{\tau_L + \tau_S} + \frac{\tau_{Bi}}{\tau_S}\frac{\tau_L}{\tau_L + \tau_S}} < 0 \tag{79}$$

and so $m_{Bi}-P_0 < -(\tau_L/(\tau_L+\tau_S))(1-\gamma)R_0+(\nu/\tau_S)x_{Bj}^-$ is sufficient.

(iii) Finally, if

$$-\frac{\tau_L(1-\gamma)}{\tau_L+\tau_S}R_0+\frac{\nu}{\tau_S}x_{Bj}^- < m_{Bi}-P_0 < \frac{\tau_S(1-\gamma)}{\tau_S+\tau_L}R_0+\frac{\nu}{\tau_L}x_{Bj}^+,$$

then $x_B=0$ and $y_B=\tau_B(m_B–D)/(\delta+\nu)$.

The expressions in the statement of the proposition follow from the above conditions and noting there is only one group of speculators (i.e., $x_B=x_{Bi}$ and $x_{Bj}=0$ for $j\neq i$). □

Appendix B. Costly lending by long investors

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

$$W_L=W_{L,0}+\rho_L F+x_L(F-P+\gamma R-c(\gamma))+y_L(F+\varepsilon-D). \tag{80}$$

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

Proposition 7. 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

$$D=P_0,\quad P=P_0+\Delta P\quad \text{and}\quad R=\frac{\tau_L+\tau_S}{\gamma^*\tau_L+\tau_S}\left(\Delta P+\frac{\tau_L}{\tau_L+\tau_S}c(\gamma^*)\right), \tag{81}$$

where the price distortion ΔP relative to the frictionless price P_0 in Eq. (6) is given by

$$\Delta P=\frac{\gamma^*\tau_L+\tau_S}{(\tau_L+\tau_S)(1-\gamma^*)}\max\left\{0,\frac{\delta}{\delta+\nu}\left[\begin{array}{c} m_L-\frac{\nu}{\tau_L}\left(\rho_L+\frac{1}{1-\gamma^*}Q\right)\\ -m_S+\frac{\nu}{\tau_S}\left(\rho_S-\frac{\gamma^*}{1-\gamma^*}Q\right)\end{array}\right]-c(\gamma^*)\right\} -\frac{\tau_L c(\gamma^*)}{\tau_L+\tau_S} \tag{82}$$

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

$$-y_S=y_L=\frac{1}{\delta}\left(1-\gamma^*+\frac{1}{R}c(\gamma^*)\right)\frac{\tau_L\tau_S}{\gamma^*\tau_L+\tau_S}\left(\Delta P+\frac{\tau_L}{\tau_L+\tau_S}c(\gamma^*)\right), \tag{83}$$

$$Q-x_S=x_L=\begin{cases}\dfrac{1}{1-\gamma^*}Q & \text{if } R>0,\\[2ex] \dfrac{\tau_L(Q+\rho_S+\rho_L)}{\tau_L+\tau_S}-\rho_L & \text{if } R=0.\end{cases} \tag{84}$$

Proof. The first-order conditions for investor L imply that

$$c_\gamma=R, \tag{85}$$

$$y_L\delta=\tau_L(P-\gamma R+c-D), \tag{86}$$

$$(x_L+y_L+\rho_L)\nu=\tau_L(m_L-P+\gamma R-c), \tag{87}$$

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

$$y_S\delta=\tau_S(P-R-D), \tag{88}$$

$$(x_S+y_S+\rho_S)\nu=\tau_S(m_S-P+R). \tag{89}$$

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 $c_\gamma(\gamma^*)=R(\gamma^*)$, and let c denote the cost at that optimal fraction (i.e., $c=c(\gamma^*)$). The market clearing conditions for the derivative and underlying cash-market imply

$$D=P-\frac{\tau_S+\gamma\tau_L}{\tau_S+\tau_L}R+\frac{\tau_L}{\tau_S+\tau_L}c, \tag{90}$$

$$P=\frac{\tau_L m_L+\tau_S m_S}{\tau_L+\tau_S}+\frac{\tau_S+\gamma\tau_L}{\tau_S+\tau_L}R-\frac{\tau_L}{\tau_S+\tau_L}c-\frac{\nu}{\tau_L+\tau_S}(Q+\rho_L+\rho_S), \tag{91}$$

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

$$y_S=-\left(1-\gamma+\frac{c}{R}\right)\frac{\tau_S\tau_L}{\delta(\tau_S+\tau_L)}R=-y_L. \tag{92}$$

Finally, note that

$$m_S-P+R-(m_L-P+\gamma R-c)=\frac{\nu}{\tau_S}(x_S+y_S+\rho_S)-\frac{\nu}{\tau_L}(x_L+y_L+\rho_L), \tag{93}$$

and

$$\frac{\delta}{\tau_S}y_S-\frac{\delta}{\tau_L}y_L=P-R-D-(P-\gamma R+c-D), \tag{94}$$

which imply

$$R(1-\gamma)+c=m_L-m_S+\frac{\nu}{\tau_S}(x_S+y_S+\rho_S)-\frac{\nu}{\tau_L}(x_L+y_L+\rho_L), \tag{95}$$

$$\Rightarrow R=\frac{m_L-m_S+\frac{\nu}{\tau_S}(x_S+\rho_S)-\frac{\nu}{\tau_L}(x_L+\rho_L)}{(1-\gamma)\left(1+\frac{\nu}{\delta}\right)}-\frac{c}{1-\gamma}. \tag{96}$$

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

$$c_\gamma(\gamma)=\frac{m_L-\frac{\nu}{\tau_L}\left(\frac{1}{1-\gamma}Q+\rho_L\right)-m_S+\frac{\nu}{\tau_S}\left(\rho_S-\frac{\gamma}{1-\gamma}Q\right)}{(1-\gamma)\left(1+\frac{\nu}{\delta}\right)}-\frac{c(\gamma)}{1-\gamma}, \tag{97}$$

or equivalently,

$$c(\gamma)+(1-\gamma)c'(\gamma)=\frac{m_L-\frac{\nu}{\tau_L}\rho_L-m_S+\frac{\nu}{\tau_S}\rho_S-\left(\frac{\nu}{\tau_S}\gamma+\frac{\nu}{\tau_L}\right)\frac{1}{1-\gamma}Q}{\left(1+\frac{\nu}{\delta}\right)}. \tag{98}$$

Since $c(\cdot)\geq 0$, $c'(\cdot)\geq 0$, and $c''(\cdot)\geq 0$, the left hand side (LHS) is non-negative and increasing in γ. On the other hand, the right hand side (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

$$c(0)+c'(0) < \frac{m_L - \frac{\nu}{\tau_L}\rho_L - m_S + \frac{\nu}{\tau_S}\rho_S - \frac{\nu}{\tau_L}Q}{\left(1+\frac{\nu}{\delta}\right)}, \tag{99}$$

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. □

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 $-\tau_L c(\gamma^*)/(\tau_L+\tau_S)$, 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 (12), the underlying security is scarce only when

$$\frac{\delta}{\delta+\nu}\left(m_L - m_S + \frac{\nu}{\tau_S}\left(\rho_S - \frac{\gamma^*}{1-\gamma^*}Q\right) - \frac{\nu}{\tau_L}\left(\frac{1}{1-\gamma^*}Q+\rho_L\right)\right) - c(\gamma^*) > 0. \tag{100}$$

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 (102) 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 $R(\gamma) = c_\gamma(\gamma)$.[30] As before, there is 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 8. 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).

Proof. As the proof of Proposition 7 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 $R = R^0 - c(\gamma^*)/(1-\gamma^*)$, where R^0 is given by

$$R^0 = \frac{m_L - m_S + \frac{\nu}{\tau_S}\left(\rho_S - \frac{\gamma^*}{1-\gamma^*}Q\right) - \frac{\nu}{\tau_L}\left(\frac{1}{1-\gamma^*}Q+\rho_L\right)}{(1-\gamma^*)\left(1+\frac{\nu}{\delta}\right)}. \tag{101}$$

This implies that

$$R_\delta = \frac{\partial R}{\partial \delta} = \frac{\nu}{\delta(\nu+\delta)}R^0 \quad \text{and} \tag{102}$$

$$R_\gamma = \frac{\partial R}{\partial \gamma} = -\frac{Q\left(\frac{1}{\tau_S}+\frac{1}{\tau_L}\right)}{(1-\gamma)^3\left(\frac{1}{\nu}+\frac{1}{\delta}\right)} + \frac{1}{1-\gamma}R^0 - \left(\frac{1}{1-\gamma}c_\gamma + \frac{1}{(1-\gamma)^2}c\right). \tag{103}$$

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

$$\frac{d\gamma}{d\delta} = \frac{R_\delta}{c_{\gamma\gamma}-R_\gamma} = \frac{\frac{\nu}{\delta(\nu+\delta)}R^0}{c_{\gamma\gamma} + \frac{Q\left(\frac{1}{\tau_S}+\frac{1}{\tau_L}\right)}{(1-\gamma)^3\left(\frac{1}{\nu}+\frac{1}{\delta}\right)}} > 0. \tag{104}$$

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

$$\Delta P = \frac{\tau_L\gamma^* + \tau_S}{\tau_L+\tau_S}R - \frac{\tau_L}{\tau_L+\tau_S}c = \frac{\tau_S}{\tau_L+\tau_S}R + \frac{\tau_L}{\tau_L+\tau_S}(\gamma^* R - c) \geq 0. \tag{105}$$

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

$$\frac{d\Delta P}{d\delta} = -\frac{\tau_L}{\tau_L+\tau_S}c_\gamma\frac{d\gamma}{d\delta} + \frac{\tau_L}{\tau_L+\tau_S}R\frac{d\gamma}{d\delta} + \frac{\tau_L\gamma+\tau_S}{\tau_L+\tau_S}\frac{dR}{d\delta}, \tag{106}$$

$$\frac{d\Delta P}{d\delta} = \frac{\tau_L\gamma+\tau_S}{\tau_L+\tau_S}\frac{dR}{d\delta} = \frac{\tau_L\gamma+\tau_S}{\tau_L+\tau_S}\frac{dc_\gamma}{d\delta} = \frac{\tau_L\gamma+\tau_S}{\tau_L+\tau_S}c_{\gamma\gamma}\frac{d\gamma}{d\delta} > 0. \tag{107}$$

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

Appendix C. Costly search by short sellers

In this appendix, we develop a search model in which the equilibrium fraction of the outstanding supply of the underlying asset that is borrowed/lent is determined endogenously. Specifically, suppose short-sellers pay a utility cost $c(\lambda)$ to search, which results in successfully meeting a long investor with probability λ. Conditional on meeting a long investor, they submit demand $\{x_{S,1}, y_{S,1}\}$. On the other hand, if they do not meet a long investor, they submit demand $\{x_{S,0} \equiv 0, y_{S,0}\}$. As a result, short-sellers solve the following problem:

$$\max_{\lambda, x_{S,0}, x_{S,1}, y_{S,0}, y_{S,1}} \begin{array}{c} \lambda\mathbb{E}[u(W_0 + x_{S,1}(F-(P-R)) + y_{S,1}(F+\varepsilon-D) + \rho_S F)] \\ +(1-\lambda)\mathbb{E}[u(W_0 + x_{S,0}(F-(P-R)) + y_{S,0}(F+\varepsilon-D) + \rho_S F)] \\ -c(\lambda). \end{array} \tag{108}$$

Similarly, if a long investor meets a short-seller (which happens with probability λ), he submits a demand

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

$\{x_{L,1}, y_{L,1}\}$. On the other hand, if they do not meet a short-seller, they submit demand $\{x_{L,0}, y_{L,0}\}$. As a result, short-sellers solve the following problem:

$$\max_{x_{L,0},x_{L,1},y_{L,0},y_{L,1}} \begin{array}{l} \lambda\mathbb{E}[u(W_0+x_{L,1}(F-(P-R))+y_{L,1}(F+\varepsilon-D)+\rho_L F)] \\ +(1-\lambda)\mathbb{E}[u(W_0+x_{L,0}(F-P)+y_{L,0}(F+\varepsilon-D)+\rho_L F)]. \end{array} \tag{109}$$

The market clearing conditions are given by the following:

$$\lambda(x_{L,1}+x_{S,1})+(1-\lambda)(x_{S,0}+x_{L,0})=Q, \tag{110}$$

$$\lambda x_{S,1}+\lambda x_{L,1}\geq 0, \tag{111}$$

$$\lambda(y_{L,1}+y_{S,1})+(1-\lambda)(y_{S,0}+y_{L,0})=0, \tag{112}$$

$$x_{S,0}=0. \tag{113}$$

To simplify notation, we shall restrict attention to the parameter assumptions from Section 3 of the benchmark model. Specifically, suppose $\rho_S=-\rho_L=\rho$ and suppose the risk-tolerance of both groups is given by τ. In this case, the following proposition characterizes the equilibrium.

Proposition 9. Suppose that S investors pay a per-unit cost $c(\lambda)$ in order to search for L investors with probability λ, where $c(\lambda)$ is non-negative, non-decreasing, and strictly convex. Then, the equilibrium prices are given by

$$D=m-\frac{\nu}{2\tau}Q, \quad R=\max\left\{0, \frac{\delta}{\delta+\nu}\frac{\nu}{\tau}\left(\rho-\frac{1}{1-\lambda}Q\right)\right\} \quad \textit{and} \tag{114}$$

$$P=D+R-\frac{\nu}{2\tau}\left(\frac{\delta}{\delta+\nu}(x_{L,1}+x_{S,1}-Q)\right) \tag{115}$$

where

$$x_{L,1}=\begin{cases}\rho & \textit{if } R>0\\ \dfrac{1}{1+\lambda}(Q+2\lambda\rho) & \textit{if } R=0,\end{cases} \tag{116}$$

$$x_{S,1}=\begin{cases}-\rho & \textit{if } R>0\\ -\dfrac{1}{1+\lambda}(2\rho-Q) & \textit{if } R=0,\end{cases} \tag{117}$$

and λ is the unique solution to

$$c'(\lambda)=\begin{array}{l}\mathbb{E}[u(W_0+x_{S,1}(F-(P-R))+y_{S,1}(F+\varepsilon-D)+\rho_S F)] \\ \qquad -\mathbb{E}[u(W_0+y_{S,0}(F+\varepsilon-D)+\rho_S F)]\end{array}. \tag{118}$$

Proof. The expressions for the prices follow from computing the optimal demand for the underlying and the derivative in each scenario and applying the market clearing conditions. In particular, one can show that:

- If the underlying is scarce, then we have

$$y_{L,1}=y_{S,1}=\frac{\nu}{\nu+\delta}\frac{1}{2}Q, \tag{119}$$

$$y_{L,0}=\frac{\nu}{\nu+\delta}\frac{1}{2}\left(2\rho-\frac{1+\lambda}{1-\lambda}Q\right), \quad y_{S,0}=-\frac{\nu}{\nu+\delta}\left(\rho-\frac{1}{2}Q\right), \tag{120}$$

$$x_{L,1}=-x_{S,1}=\rho, \quad x_{L,0}=\frac{1}{1-\lambda}Q. \tag{121}$$

- If the underlying is not scarce, i.e., $R=0$

$$y_{L,1}=y_{S,1}=y_{L,0}=\frac{1}{2}\frac{\nu}{\nu+\delta}\frac{1-\lambda}{1+\lambda}(2\rho-Q), \tag{122}$$

$$y_{S,0}=-\frac{1}{2}\frac{\nu}{\nu+\delta}(2\rho-Q), \tag{123}$$

$$x_{L,1}=x_{L,0}=\frac{1}{1+\lambda}(Q+2\lambda\rho), \quad x_{S,1}=-\frac{1}{1+\lambda}(2\rho-Q). \tag{124}$$

Finally, the optimal λ is pinned down by the first-order condition for λ in the S investors' optimization problem, given by

$$c'(\lambda)=DEU \quad \text{where } DEU\equiv EU_1-EU_0, \tag{125}$$

$$EU_1\equiv\mathbb{E}[u(W_0+x_{S,1}(F-(P-R))+y_{S,1}(F+\varepsilon-D)+\rho_S F)], \tag{126}$$

$$EU_0\equiv\mathbb{E}[u(W_0+y_{S,0}(F+\varepsilon-D)+\rho_S F)]. \tag{127}$$

Note that

$$EU_0=-e^{\{-(1/\tau)(W_0+m\rho+(1/8\tau)(\nu/(\nu+\delta))(Q\nu(Q-4\rho)-4\delta\rho^2))\}} \tag{128}$$

whether or not the underlying is scarce. On the other hand,

$$EU_1=-e^{\{-(1/\tau)(W_0+m\rho+(1/8\tau)(\nu/(\nu+\delta))Q\nu(Q-4\rho))\}} \tag{129}$$

if the underlying is scarce, and

$$EU_1=-e^{\{-(1/\tau)(W_0+m\rho+(1/8\tau)(\nu/(\nu+\delta))(4\delta(Q-2\rho)^2/(1+\lambda)^2+Q\nu(Q-4\rho)-4\delta\rho^2))\}} \tag{130}$$

if the underlying is not scarce. As such, *DEU* only depends on λ through EU_1 when the underlying is scarce, and moreover, *DEU* is non-increasing in λ. Therefore, if $c(\lambda)$ is strictly convex, i.e., $c'(\lambda)$ is increasing, then there is a unique solution λ^* to Eq. (125), which characterizes the equilibrium search probability λ. □

The above result establishes that many of the results from our benchmark model in Section 3 continue to hold in this setting, in which the equilibrium fraction of the outstanding supply of the underlying asset that is lent/borrowed is determined as a function of the endogenous search probability λ. For instance, note that: (i) the price of the derivative does not depend on the search probability λ or the noise in the derivative payoff δ, and (ii) the repo rate R and the price distortion $\Delta P=P-P_0$ are decreasing in λ. Moreover, as in the benchmark model of the paper, for a fixed search probability, increasing the noise in the derivative δ makes the underlying more scarce and, therefore, increases the price distortion.

However, when short-sellers can endogenously choose their search probability (which determines the scarcity of the underlying) in response to the noise in the derivative payoff, then increasing the noise in the derivative may *decrease* the scarcity and price distortion in the underlying. Intuitively, the equilibrium search probability λ is determined by the short-seller's first-order condition which sets the marginal cost of searching $c'(\lambda)$ equal to the increase in her expected utility from being able to trade in the underlying security (i.e., from locating a long investor). As such, more noise in the derivative increases the relative benefit (in expected utility) for each short-seller from meeting a long investor, which incentivizes her to search more aggressively in equilibrium. However, by searching more aggressively, short-sellers reduce the scarcity in the underlying, and therefore reduce the price distortion.

The following proposition characterizes the conditions under which an increase in δ leads to an increase in the price distortion, and shows that this is always the case when $\delta=0$, i.e., when the derivative security is initially frictionless.

Proposition 10. Suppose that S investors pay a per-unit cost $c(\lambda)$ in order to search for L investors with probability λ, where $c(\lambda)$ is non-negative, non-decreasing, and strictly convex and does not depend on δ. Then, the optimal search probability λ^* which is characterized by $c'(\lambda)=DEU$ is increasing in δ, i.e., $\lambda_\delta \equiv \partial\lambda^*/\partial\delta>0$. The price distortion in the underlying increases in the noise of the derivative payoff (i.e., δ), if λ_δ is small enough (as characterized in the proof), or if $\delta=0$.

Proof. Note that then the underlying is scarce,

$$DEU=-e^{-(1/\tau)(W_0+m\rho+(1/8\tau)(\nu/(\nu+\delta))Q\nu(Q-4\rho))}(1-e^{(\rho^2/2\tau^2)(\nu\delta/(\nu+\delta))}), \tag{131}$$

while when the underlying is not scarce,

$$DEU=-e^{-(1/\tau)(W_0+m\rho+(1/8\tau)(\nu/(\nu+\delta))(Q\nu(Q-4\rho)-4\delta\rho^2))} \\ (e^{-(1/2\tau^2)(\nu/(\nu+\delta))\delta(Q-2\rho)^2/(1+\lambda)^2}-1). \tag{132}$$

In both cases, one can show that $(\partial/\partial\delta)DEU>0$. Since $c(\cdot)$ does not depend on δ, this implies that for all else equal, an increase in δ (i.e., more noise in the derivative) shifts the DEU curve up, which implies λ^* will be higher, i.e., $(\partial/\partial\delta)\lambda^*(\delta)>0$. As a result, δ can have an ambiguous effect on the price distortion. In particular, note that when the underlying is scarce, the price distortion is given by

$$\Delta P=\frac{\delta}{\delta+\nu}\frac{\nu}{\tau}\left(\rho-\frac{1}{1-\lambda}Q\right)+\frac{\nu}{2\tau}\left(\frac{\delta}{\delta+\nu}Q\right) \tag{133}$$

$$\Rightarrow \frac{\partial}{\partial\delta}\Delta P=\frac{1}{\tau}\frac{\nu}{\delta+\nu}\left[\underbrace{\frac{\nu}{\nu+\delta}\left(\rho+Q\left(\frac{1}{2}-\frac{1}{1-\lambda}\right)\right)}_{>0}-\underbrace{\frac{\delta}{1-\lambda}Q\lambda_\delta}_{>0}\right]. \tag{134}$$

This implies that the effect of δ on ΔP depends on the relative magnitude of λ_δ, which in turn depends on the cost function c. Moreover, note that the derivative is positive when $\delta=0$ (as long as λ_δ is bounded) or if

$$\lambda_\delta<\frac{\frac{\nu}{\nu+\delta}\left(\rho+Q\left(\frac{1}{2}-\frac{1}{1-\lambda}\right)\right)}{\frac{\delta}{1-\lambda}Q}.$$

Similarly, when the underlying is not scarce, the price distortion is

$$\Delta P=\frac{\nu}{2\tau}(2\rho-Q)\frac{\delta}{\nu+\delta}\frac{1-\lambda}{1+\lambda} \tag{135}$$

$$\Rightarrow \frac{\partial}{\partial\delta}\Delta P=\frac{\nu}{2\tau}(2\rho-Q)\frac{(\nu(1-\lambda)-2\delta(\delta+\nu)\lambda_\delta)}{(\nu+\delta)^2(1+\lambda)^2}. \tag{136}$$

Again, the derivative is positive when $\delta=0$, or if $\lambda_\delta<\nu(1-\lambda)/(2\delta(\delta+\nu))$. □

Appendix D. Position limits

Consider a version of the model in Section 3, in which the payoff of the derivative is F (i.e., there is no noise), but there are position limits that restrict the size of derivative positions to $\overline{y}$, i.e., $y_L=-y_S\le\overline{y}$. In this case, the first-order conditions for investor i are given by

$$\nu(x_i+y_i+\rho_i)=\tau(m-(P-\gamma_i R)) \quad \text{and} \tag{137}$$

$$\nu(x_i+y_i+\rho_i)=\tau(m-D), \tag{138}$$

where $\gamma_L\le\gamma<1$ and $\gamma_S=1$. As in the model, the market clearing conditions are given by

$$\sum_i x_i=Q, \quad \sum_i y_i=0 \quad \text{and} \quad \gamma x_L+x_S=0. \tag{139}$$

As in the benchmark model, the optimal equilibrium allocations in the economy without any frictions (and in the absence of derivatives) is $x_L=\frac{1}{2}Q+\rho$ and $x_S=\frac{1}{2}Q-\rho$. This implies there are two cases to consider.

*Case*1: If $\frac{1}{2}Q+\rho<\left(1/(1-\gamma)\right)Q+\overline{y}$ (or equivalently, if $\frac{1}{2}Q-\rho\ge-\left(\gamma/(1-\gamma)\right)Q-\overline{y}$), then the underlying security is not scarce. In this case, $R=0$ and the price of the derivative and the underlying are given by

$$P=D=m-\frac{\nu}{2\tau}Q. \tag{140}$$

Since the payoffs and prices of the underlying and the derivative are identical, this implies a certain degree on indeterminacy in the equilibrium quantities held by investors. We can recover a complete characterization if we impose some type of tie-breaking rule — for instance, suppose investors trade in the underlying when indifferent between the two securities. In this case, the equilibrium positions are given by

$$x_L=Q-x_S=\min\left\{\frac{1}{2}Q+\rho,\frac{1}{1-\gamma}Q\right\} \quad \text{and} \tag{141}$$

$$y_L=-y_S=\max\left\{0,\frac{1}{2}Q+\rho-\frac{1}{1-\gamma}Q\right\}. \tag{142}$$

*Case*2: If $\frac{1}{2}Q+\rho\ge\left(1/(1-\gamma)\right)Q+\overline{y}$, then the underlying is scarce, and $R>0$. Specifically, market clearing implies:

$$R=\frac{1}{1-\gamma}\frac{\nu}{\tau}\left(2\rho-\frac{1+\gamma}{1-\gamma}Q-2y_L\right). \tag{143}$$

Given the payoff of the derivative and the underlying are identical, the price of the derivative is bounded by the net cost to L and S investors for the underlying, i.e.,

$$P-\gamma R\ge D\ge P-R. \tag{144}$$

Moreover, for any price D that is strictly in between these bounds, L and S investors take the largest derivative position they can, i.e., $y_L = -y_S = \bar{y}$. Within these bounds, the derivative price is indeterminate without additional assumptions (e.g., bargaining between L and S investors can help pin down D), but the equilibrium quantities and the price of the underlying security are pinned down as follows:

$$x_L = Q - x_s = \frac{1}{1-\gamma}Q, \quad y_L = -y_S = \bar{y}, \quad P = m - \frac{\nu}{2\tau}Q + \frac{1+\gamma}{2}R, \tag{145}$$

$$R = \max\left\{0, \frac{1}{1-\gamma}\frac{\nu}{\tau}\left(2(\rho-\bar{y}) - \frac{1+\gamma}{1-\gamma}Q\right)\right\}. \tag{146}$$

As such, this version of the model also delivers the main comparative static result of the paper. Increasing ease of trade in the derivative (in this case, by increasing $\bar{y}$), reduces the price distortion in the underlying by relaxing the scarcity that investors face.

References

Acharya, V., Pedersen, L., 2005. Asset pricing with liquidity risk. J. Financial Econ. 77, 375–410.

Allen, F., Gale, D., 1988. Optimal security design. Rev. Financial Stud. 1, 229–263.

Allen, F., Gale, D., 1994. Financial Innovation and Risk Sharing. MIT Press, Cambridge, MA.

Amihud, Y., Mendelson, H., 1986. Asset pricing and the bid–ask spread. J. Financial Econ. 17, 223–249.

Amihud, Y., Mendelson, H., Pedersen, L., 2005. Liquidity and asset prices. Found. Trends Finance 1, 1–96.

Bai, Y., Chang, E., Wang, J., 2006. Asset prices under short-sale constraints. Unpublished working paper. University of Hong Kong, Massachusetts Institute of Technology.

Banerjee, S., Graveline, J.J., 2013. The cost of short-selling liquid securities. J. Finance 68, 637–664.

Bhamra, H.S., Uppal, R., 2009. The effect of introducing a non-redundant derivative on the volatility of stock-market returns when agents differ in risk aversion. Rev. Financial Stud. 22, 2303–2330.

Bongaerts, D., De Jong, F., Driessen, J., 2011. Derivative pricing with liquidity risk: theory and evidence from the credit default swap market. J. Finance 66, 203–240.

Boyle, P., Wang, T., 2001. Pricing of new securities in an incomplete market: the Catch 22 of no-arbitrage pricing. Math. Finance 11, 267–284.

Brunnermeier, M., Simsek, A., Xiong, W., 2012. A welfare criterion for models with distorted beliefs. Unpublished working paper. Princeton University, Massachusetts Institute of Technology.

Cuny, C.J., 1993. The role of liquidity in futures market innovations. Rev. Financial Stud. 6, 57–78.

Danielsen, B.R., Sorescu, S.M., 2001. Why do option introductions depress stock prices? A study of diminishing short sale constraints. J. Financial Quant. Anal. 36, 451–484.

D'Avolio, G., 2002. The market for borrowing stock. J. Financial Econ. 66, 271–306.

Detemple, J., Selden, L., 1991. A general equilibrium analysis of option and stock market interactions. Int. Econ. Rev. 32, 279–303.

Diamond, D.W., Verrecchia, R.E., 1987. Constraints on short-selling and asset price adjustment to private information. J. Financial Econ. 18, 277–311.

Duffie, D., 1996. Special repo rates. J. Finance 51, 493–526.

Duffie, D., Gârleanu, N., Pedersen, L., 2002. Securities lending, shorting, and pricing. J. Financial Econ. 66, 307–339.

Duffie, D., Jackson, M., 1989. Optimal innovation of futures contracts. Rev. Financial Stud. 2, 275–296.

Duffie, D., Rahi, R., 1995. Financial market innovation and security design: an introduction. J. Econ. Theory 65, 1–42.

Evans, R., Ferreira, M., Porras Prado, M., 2012. Equity lending, investment restrictions and fund performance. Unpublished working paper. University of Virginia, Universidade Nova de Lisboa.

Gallmeyer, M., Hollifield, B., 2008. An examination of heterogeneous beliefs with a short-sale constraint in a dynamic economy. Rev. Finance 12, 323–364.

Gârleanu, N., 2009. Portfolio choice and pricing in illiquid markets. J. Econ. Theory 144, 532–564.

Gârleanu, N., Pedersen, L., 2004. Adverse selection and the required return. Rev. Financial Stud. 17, 643–665.

Geanakoplos, J., 2003. Liquidity, default and crashes: endogenous contracts in general equilibrium. In: Advances in Economics and Econometrics: Theory and Applications: Eighth World Congress, Cambridge University Press, Cambridge, MA, pp. 170–205.

Geczy, C., Musto, D., Reed, A., 2002. Stocks are special too: an analysis of the equity lending market. J. Financial Econ. 66, 241–269.

Gilboa, I., Samuelson, L., Schmeidler, D., 2012. No-betting pareto dominance. Unpublished working paper. HEC Paris, Yale University, Tel Aviv University.

Goldreich, D., Hanke, B., Nath, P., 2005. The price of future liquidity: time-varying liquidity in the US Treasury market. Rev. Finance 9, 1–32.

Grossman, S.J., Stiglitz, J.E., 1980. On the impossibility of informationally efficient markets. Am. Econ. Rev. 70, 393–408.

Gupta, S., Sundaram, R.K., 2011. CDS credit-event auctions. Unpublished working paper. New York University.

Hakansson, N., 1979. The fantastic world of finance: progress and the free lunch. J. Financial Quant. Anal. 14, 717–734.

Jordan, B., Jordan, S., 1997. Special repo rates: an empirical analysis. J. Finance 52, 2051–2072.

Jordan, B., Kuipers, D., 1997. Negative option values are possible: the impact of treasury bond futures on the cash us treasury market. J. Financial Econ. 46, 67–102.

Kaplan, S.N., Moskowitz, T.J., Sensoy, B.A., 2013. The effects of stock lending on security prices: an experiment. J. Finance 68, 1891–1936.

Krishnamurthy, A., 2002. The bond/old-bond spread. J. Financial Econ. 66, 463–506.

Kubler, F., Schmedders, K., 2012. Financial innovation and asset price volatility. Am. Econ. Rev. 102, 147–151.

Kyle, A.S., 1985. Continuous auctions and insider trading. Econometrica 53, 1315–1335.

Merton, R., 1989. On the application of the continuous-time theory of finance to financial intermediation and insurance. Geneva Pap. Risk Insur. Issues Pract. 14, 225–261.

Miller, E.M., 1977. Risk, uncertainty, and divergence of opinion. J. Finance 32, 1151–1168.

Ofek, E., Richardson, M., 2003. Dot-com mania: the rise and fall of Internet stock prices. J. Finance 58, 1113–1138.

Rahi, R., 1995. Optimal incomplete markets with asymmetric information. J. Econ. Theory 65, 171–197.

Rizova, S., 2011. Securities lending by mutual funds. Unpublished working paper. University of Chicago.

Shen, J., Yan, H., Zhang, J., 2012. Collateral-motivated financial innovation. Unpublished working paper. London School of Economics, Yale University.

Simsek, A., 2011. Speculation and risk sharing with new financial assets. NBER Working Paper No. 17506.

Simsek, A., 2013. Belief disagreements and collateral constraints. Econometrica 81, 1–53.

Sorescu, S.M., 2000. The effect of options on stock prices: 1973 to 1995. J. Finance 55, 487–514.

Vayanos, D., 1998. Transaction costs and asset prices: a dynamic equilibrium model. Rev. Financial Stud. 11, 1–58.

Vayanos, D., Wang, J., 2012. Theories of liquidity. Found. Trends Finance 6, 221–317.

Vayanos, D., Weill, P.O., 2008. A search-based theory of the on-the-run phenomenon. J. Finance 63, 1361–1398.

Wang, J., 1993. A model of intertemporal asset prices under asymmetric information. Rev. Econ. Stud. 60, 249–282.

Zapatero, F., 1998. Effects of financial innovations on market volatility when beliefs are heterogeneous. J. Econ. Dyn. Control 22, 597–626.