Asset Price Dynamics in Partially Segmented Markets *

Robin Greenwood
Harvard University and NBER

Samuel G. Hanson
Harvard University and NBER

Gordon Y. Liao
Cambridge Square Capital

September 2017

Abstract

We develop a model in which capital moves quickly within an asset class, but slowly between asset classes. While most investors specialize in a single asset class, a handful of generalists can gradually reallocate capital across markets. Upon the arrival of a large supply shock, prices of risk in the directly impacted asset class become disconnected from those in others. Over the long run, capital flows between markets and prices of risk become more closely aligned. While prices in the directly impacted market initially overreact to the supply shock, we show that prices in related asset classes underreact under plausible conditions. We use the model to assess event-study evidence on the impact of recent large-scale asset purchases by central banks.

*We thank Daniel Bergstresser, Darrell Duffie, Yueran Ma, Andrei Shleifer, Jeremy Stein, Adi Sunderam, and seminar participants at Berkeley Haas, Brandeis, MIT Sloan, NYU Stern, University of Minnesota Carlson, University of North Carolina Kenan-Flager, and University of Texas McCombs for useful feedback. We thank David Biery for research assistance. Greenwood and Hanson gratefully acknowledge funding from the Division of Research at Harvard Business School.

1 Introduction

How do large supply shocks in one financial market affect asset prices in other markets? For example, in recent years the Federal Reserve and other major central banks have purchased large quantities of long-term government bonds in an attempt to drive down long-term borrowing rates throughout the economy. In response, there has been an active debate amongst academics and policymakers about the extent too which these "quantitative easing" policies have impacted the prices of financial assets—such as corporate bonds and equities—outside of the market for government bonds.

If all investors can frictionlessly trade each different asset class, then financial markets are fully integrated and all investors will price exposure to common risk factors in the same way. Furthermore, a shock to the supply of one asset class will lead all investors to identically adjust the way they price risk. However, if markets are more segmented—i.e. if some assets are held by specialists who are unable to trade across different markets, then in the wake of a large supply shock the way that specialists in one market price risk may become disconnected from the way it is priced by those in other markets. For example, interest rate risk may be priced differently by specialists in the government and corporate bond markets following a large shock to the supply of government bonds.

Over the longer run, however, the forces of arbitrage should ensure that capital will flow from markets where prices of risk are low to markets where prices of risk are high, narrowing these differences. But the process of market integration can be slow, because generalist investors with the flexibility to trade across asset classes can do so only gradually. For example, investment committees at pension funds and endowments—who have this flexibility—typically only reallocate capital annually or biannually. And, cross-market arbitrage is risky because risks cannot be easily unbundled from assets, so generalist investors may not aggressively trade across markets, even when they can. As a result, it may take time for a large supply shock in one market to fully propagate into other markets.

In this paper, we explore how the pricing of two risky asset classes, denoted assets A and B, responds to a large shock to the supply of one asset class, say asset A. In our model, risk-averse investors absorb shocks to the supply of these two assets, which are exposed to a common risk factor. Since the two assets are close substitutes, in the absence of asset-pricing frictions, their prices would be tightly linked by cross-market arbitrage. Our key contribution is to show how a shock to the supply of asset A is reflected over time in the price of asset B in a setting where there are two key asset pricing frictions. First, the markets for the two asset classes are partially segmented as in Gromb and Vayanos (2002). Specifically, a first group of investors ("A-specialists") can only trade asset A, a second group ("B-specialists") can only trade asset B, and only a final group ("generalists") can trade both asset A and asset B. Second, the generalist investors can only rebalance their portfolios gradually over time as in Duffie (2010), so capital moves slowly between the two markets. In combination, these two frictions imply a rich and realistic set of asset price dynamics following a supply shock that only directly impacts one market.

For concreteness, we focus our main analysis on a setting where asset A represents long-term default-free government bonds and asset B represents long-term defaultable corporate bonds. We have chosen this setting since we believe that our model provides a useful framework for understanding how quantitative easing policies might impact asset classes not directly purchased by central banks. However, we show that the insights from our model are quite general and emerge so long as the two assets are exposed to a common risk factor. For example, we show that our model can be easily adapted to understand how a shock to the demand for U.S. stocks would impact the prices of European stocks.

In the setting we have just described, what happens when there is an unanticipated supply shock in one market? Suppose, for example, that the Federal Reserve announces that it will sell a large portfolio of long-term U.S. Treasury bonds, permanently expanding the amount of interest rate risk that investors must bear in equilibrium. Treasury market specialists react immediately to this shock, absorbing the increased supply into their inventories. The risk premium on long-term Treasury bonds will rise, lifting their yields. However, Treasury yields will overreact—the short-run price impact will exceed the long-run impact—because the amount of capital that can initially accommodate the shock is limited to specialists and a handful generalists. Over the long-run, generalist investors will allocate additional capital to the Treasury market, reducing the price impact of the supply shock at longer horizons. Price dynamics of this sort are similar to those described in Duffie (2010), who explores the pricing implications of slow-moving capital for a single asset.

Our key contribution is to show how prices evolve in the second related market that is not directly impacted by the supply shock. Specifically, consider how corporate bond prices will react to a large shock to the supply of long-term Treasuries. Corporate bonds are indirectly affected by this shock because generalist investors respond by increasing their holdings of long-term Treasuries and reducing their holdings of long-term corporate bonds. These cross-market capital flows drive down the prices of corporate bonds and push up corporate bond yields. In this way, the trading of generalist investors transmits the supply shock across markets, serving to increase market integration.

Under plausible parameter values, we show that corporate bond yields will underreact to this shock to the supply of government bonds: the short-run price impact is less than the long-run impact. This stands in contrast to the initial overreaction of Treasury yields. The simultaneous overreaction and underreaction in different markets is driven by the fact that generalists only reallocate capital slowly. As a result, it takes time for financial markets to fully digest large supply shocks.

The price dynamics in our model depend critically on the fractions of specialists in each market, the number of time periods it takes generalists to rebalance their portfolios, and the degree of substitutability between the two assets. The fraction of specialists and generalist investors play an especially important role. When there are a small number of slow-moving generalists, the Treasury market overreacts while the corporate market underreacts to the shock to Treasury supply. However, if there are many slow-moving generalists, both markets overreact to the Treasury supply shock.

Importantly, we show that *both* partial segmentation and slow-moving capital between markets are needed to generate short-run underreaction in the indirectly affected market. Specifically, if there is only slow-moving capital—i.e., if all investors can trade both assets, but some investors can only rebalance their portfolios gradually—then we always obtain short-run overreaction in both markets.[1] This is because, in this case, prices in both markets reflect a common set of factor risk prices charged by fast-moving generalist investors. To be sure, these prices of risk will initially overreact to a large supply shock because some investors are slow-moving, but the overreaction is the same in both markets.

A natural application of our model—and one that motivates our analysis—is the assessment of large-scale asset purchases by central banks. A key question about these "quantitative easing" policies is whether they impacted prices outside the market in which the central bank is directly transacting. The favored methodology for answering this question—and one used in dozens of recent papers—has been a short-run event study approach that examines price changes on days when purchases are announced. Some studies have concluded that the impact is most pronounced in the directly impacted market, with only modest spillovers to other related markets (Krishnamurthy and Vissing-Jorgensen [2011, 2013] and Woodford [2012]). Others have suggested that, at longer horizons, the spillovers to other markets are more significant (Mamaysky [2014]).

The short-run event study methodology was originally developed in the 1960s to assess whether prices rapidly impound news about fundamentals (Fama, Fisher, Jensen, and Roll [1969]). However, in recent years, event studies have increasingly been used to measure changes in risk premia. Our model crystallizes the limitations of this approach, particularly when one is interested in assessing price impact across partially segmented markets. Specifically, even if prices immediately impound fundamental news—as they do in our model—it may take time for risk premia to normalize across markets following a large supply shock.

Our model is closely related to three strands of research in financial economics. The idea that front-line arbitrageurs in financial markets are highly specialized traces back to Merton (1987) and Grossman and Miller (1988), and is a central tenet of the theory of limited arbitrage (De Long et al [1990], Shleifer and Vishny [1997], and Gromb and Vayanos [2002]). A small literature in finance describes asset prices and returns in partially segmented markets (Stapleton and Subrahmanyam [1977], Errunza and Losq

[1] This corresponds to a multi-asset version of Duffie's (2010) model which is developed in Bogousslavsky (2016).

[1985], Merton [1987]). More recently, a number of researchers have demonstrated downward-sloping demand curves for individual financial asset classes, which would be puzzling if markets were fully integrated (Gabaix, Krishnamurthy, and Vigneron [2007], Gârleanu, Pedersen, and Poteshman [2009], Greenwood and Vayanos [2014], and Hanson [2014]). These researchers often motivate their analysis by positing an extreme form of market segmentation, in which a different pricing model is used to price the securities in each distinct asset class. In our model, trading by slow-moving generalists links these market-specific pricing models together over time, offering a middle ground between models positing extreme segmentation and traditional models featuring perfect integration.

Second, our paper is related to research on "slow-moving capital," which is the idea that capital does not flow as quickly towards attractive investment opportunities as textbook theories might suggest (Mitchell, Pedersen, Pulvino [2007], Duffie [2010], Acharya, Shin, and Yorulmazer [2013]). Here, our model draws most heavily from Duffie (2010), who studies the implications of slow-moving capital for price dynamics in a single asset market. Compared to his paper, our key contribution is to analyze the price impact of supply shocks across multiple partially-segmented asset markets as well as the resulting patterns of cross-market arbitrage.[2]

Third, our model is a overlapping-generations, rational-expectations model in which mean-variance investors with finite investment horizons trade an infinitely-lived asset that is subject to supply shocks.[3] From a technical standpoint, we show how to incorporate both partial market segmentation and slow-moving capital into this workhorse class of asset pricing models.

2 Model

We develop the model in three steps. First, we develop a stylized, linear model for pricing long-term default-free bonds and long-term defaultable bonds. Second, we introduce our two key asset pricing frictions, namely, partial market segmentation and slow-moving capital across markets. Finally, we solve for the rational expectations equilibrium in this setting. While we focus on two fixed-income asset classes in our main analysis, the insights from our model are completely general and emerge so long as the two assets are exposed to a common risk factor. Indeed, in the Internet Appendix, we develop a model for pricing two risky equity claims that shares the same formal structure and delivers the same insights as the model we develop below.

[2] Duffie and Strulovici (2012) present a model of the movement of capital across two partially segmented markets, but their focus is on the endogenous speed of capital mobility, which we take as exogenous.

[3] Thus, many of the technical issues that we must address are foreshadowed in this literature, including De Long, Shleifer, Summers, and Waldmann (1990), Spiegel (1998), Bacchetta and van Wincoop (2010), Watanabe (2008), Banerjee (2011), Greenwood and Vayanos (2014), and Albagli (2015).

2.1 Long-term bonds

We consider a model with two perpetual risky bonds, A and B. Both A and B bonds are exposed to interest rate risk. The B bonds are also exposed to default risk, making them a partial substitute for the A bonds.

A-bonds The A perpetuities are default-free and pay a coupon of C_A each period, so the gross return on A bonds from time t to $t+1$ is

$$1 + R_{A,t+1} = \frac{P_{A,t+1} + C_A}{P_{A,t}}, \tag{1}$$

where $P_{A,t}$ in the price of A bonds at time t.

To maintain tractability, we must assume that asset prices (or yields) and expected returns are linear functions of a vector of state variables. Although stylized, this linearity assumption is common in rational expectations models like ours where mean-variance investors absorb stochastic supply shocks. In order to apply our model to fixed income assets, we (i) substitute log returns for simple returns throughout and (ii) use a Campbell-Shiller (1988) linearization of log returns. This modelling approach is also used in Hanson (2014) and Hanson and Stein (2015). We view (i) and (ii) as linearity-generating modelling devices that do not impact the qualitative conclusions that we draw from the model. Together these modelling devices allow us to develop a discrete-time analog to the continuous-time term structure models in Vayanos and Vila (2009) and Greenwood and Vayanos (2014).

Consider the Campbell-Shiller (1988) log-linear approximation to the return on A perpetuities. Specifically, defining $\theta_A \equiv 1/(1+C_A) < 1$, the one-period log return on A bonds is approximately

$$r_{A,t+1} \equiv \ln(1 + R_{A,t+1}) \approx \overbrace{\frac{1}{1-\theta_A}}^{D_A} y_{A,t} - \overbrace{\frac{\theta_A}{1-\theta_A}}^{D_A - 1} y_{A,t+1}, \tag{2}$$

where $y_{A,t}$ is the log yield-to-maturity on A bonds at time t and

$$D_A = \frac{1}{1-\theta_A} = \frac{C_A + 1}{C_A} \tag{3}$$

is the Macaulay duration when the bonds are trading at par.[4] Thus, the log excess return on A bonds

[4] This approximation for default-free coupon-bearing bonds appears in Chapter 10 of Campbell, Lo, and MacKinlay (1997) and is an approximate generalization of the fact that the log-return on an n-period zero-coupon bond from t to $t+1$ is exactly $r_{n,t+1} = ny_{n,t} - (n-1)\, y_{n-1,t+1}$ where, for instance, $y_{n,t}$ is the log yield on n-period zero-coupon bonds at time t. We review the derivation in the Internet Appendix. Our approximation for defaultable bonds in equation (6) then follows trivially since we assume that default losses are a random fraction of market value.

over the short-term interest rate from time t to $t+1$, denoted r_t, is

$$rx_{A,t+1} \approx \frac{1}{1-\theta_A} y_{A,t} - \frac{\theta_A}{1-\theta_A} y_{A,t+1} - r_t. \tag{4}$$

B-bonds The B bonds are also exposed to default risk. Specifically, consider a homogenous portfolio of perpetual, defaultable bonds each of which promises to pay a coupon of C_B each period. Suppose that a random fraction h_{t+1} of the B bonds default at $t+1$ and are worth $(1-L_{t+1})(P_{L,t+1}+C)$ where $0 \leq L_{t+1} < 1$ is the loss-given-default as a fraction of market value. The remaining fraction $(1-h_{t+1})$ of the B bonds do not default and are worth $(P_{B,t+1}+C)$. Thus, the return on the portfolio of B bonds is

$$1 + R_{B,t+1} = \frac{(1-Z_{t+1})(P_{B,t+1}+C_B)}{P_{B,t}}, \tag{5}$$

where $Z_{t+1} = h_{t+1}L_{t+1}$, satisfying $0 \leq Z_{t+1} < 1$, is the random portfolio default realization at time $t+1$. This stylized formulation of default risk follows Duffie and Singleton's (1999) "recovery of market value" assumption which has become standard in the credit risk literature.

As a above, we use a Campbell-Shiller (1988) log-linear approximation to the return on these perpetual bonds. Specifically, the log excess return on B bonds from time t to $t+1$ is

$$rx_{B,t+1} \approx \frac{1}{1-\theta_B} y_{B,t} - \frac{\theta_B}{1-\theta_B} y_{B,t+1} - z_{t+1} - r_t, \tag{6}$$

where $\theta_B \equiv 1/(1+C_B)$ and $z_t \equiv -\ln(1-Z_t)$ is the log default loss at time t. Relative to equation (4), the additional z_{t+1} term in equation (6) reflects the default realization that is specific to B bonds.

Risk factors There are three different sources of risk in our model: interest rate risk, default risk, and supply risk. First, both the A and B bonds are exposed to interest rate risk. There is an exogenous short-term interest rate that evolves randomly, so both long-term bonds will suffer a capital loss if short-term rates rise unexpectedly. Second, the B bonds are exposed to default risk: the future period-by-period default realization is unknown and evolves randomly. Finally, both bonds are exposed to supply risk: there are random supply shocks which impact the prices and yields on long-term bonds, holding fixed the expected future path of short-term interest rates and expected future defaults. Thus, using Campbell's (1991) terminology, interest rate risk and default risk are forms of fundamental "cash flow" risk, whereas supply risk is a form of "discount rate" risk.

Concretely, we make the following assumptions:

- **Short-term interest rates:** The log short-term riskless rate available to investors between time t and $t+1$, denoted r_t, is known at time t. The short-term rate r_t follows an exogenous

AR(1) process

$$r_{t+1} = \overline{r} + \rho_r \left(r_t - \overline{r} \right) + \varepsilon_{r,t+1}, \tag{7}$$

where $Var_t \left[\varepsilon_{r,t+1} \right] = \sigma_r^2$. One can think of the short-term rate as being determined outside the model either by monetary policy or by a stochastic short-term storage technology that is available in perfectly elastic supply.

- **Default losses:** The default process z_t also follows an AR(1) process

$$z_{t+1} = \overline{z} + \rho_z \left(z_t - \overline{z} \right) + \varepsilon_{z,t+1}, \tag{8}$$

where $Var_t \left[\varepsilon_{z,t+1} \right] = \sigma_z^2$. The variance of z_{t+1} influences the degree of substitutability between the A and B bonds.

- **Supply:** Both bonds are available in an exogenous, time-varying supply. The two bonds are subject to different supply shocks which also limits their substitutability for generalists investors. Specifically, the supply that investors must hold of A bonds evolves according to

$$s_{A,t+1} = \overline{s}_A + \rho_{s_A} \left(s_{A,t} - \overline{s}_A \right) + \varepsilon_{s_A,t+1}, \tag{9}$$

where $Var_t \left[\varepsilon_{s_A,t+1} \right] = \sigma_{s_A}^2$. Similarly, the supply that investors must hold of B bonds evolves as

$$s_{B,t+1} = \overline{s}_B + \rho_{s_B} \left(s_{B,t} - \overline{s}_B \right) + \varepsilon_{s_B,t+1}, \tag{10}$$

where $Var_t \left[\varepsilon_{s_B,t+1} \right] = \sigma_{s_B}^2$.

Note that we have made no assumptions about the correlation structure among the four underlying shocks $\varepsilon_{r,t+1}$, $\varepsilon_{z,t+1}$, $\varepsilon_{s_A,t+1}$, and $\varepsilon_{s_B,t+1}$. Indeed, our model can be solved for any arbitrary correlation structure among these shocks. However, in our numerical illustrations in Section 3, for simplicity we will assume that the four underlying shocks are mutually orthogonal. Because of the clarity that it affords, this assumption is common in asset pricing models with noisy asset supply.

2.2 Market structure

There are three types of investors—A-specialists, B-specialists, and generalists—all with identical risk tolerance τ. Investors are distinguished by their ability to transact in different assets and the frequency with which they can rebalance their portfolios.[5] Because only a subset of generalist investors can trade

[5]Allowing for heterogeneity in risk aversion does not change the asset-pricing implications of our model. Indeed, an economy with arbitrary heterogeneity in risk aversion is isomorphic to the economy we describe below. With hetero-

both the A and B bonds, the markets are partially segmented. Furthermore, capital moves slowly between the two markets because the generalists can only rebalance their portfolios gradually.

A-specialists Fast-moving A-specialists are free to adjust their holdings of A bonds and the riskless short-term asset each period; however, A-specialists cannot hold B bonds. A-specialists are present in mass q_A and we denote their demand for A bonds by $b_{A,t}$. Fast-moving A-specialists have mean-variance preferences over 1-period portfolio log returns. Thus, their demand for A bonds is

$$b_{A,t} = \tau \frac{E_t \left[rx_{A,t+1} \right]}{Var_t \left[rx_{A,t+1} \right]}. \tag{11}$$

B-specialists Fast-moving B-specialists can adjust their holdings of B bonds and the riskless short-term asset each period, but cannot hold A bonds. B-specialists are present in mass q_B and their demand for B bonds is denoted $b_{B,t}$. B-specialists also have mean-variance preferences over 1-period portfolio log returns, so their demand for B bonds is

$$b_{B,t} = \tau \frac{E_t \left[rx_{B,t+1} \right]}{Var_t \left[rx_{B,t+1} \right]}. \tag{12}$$

Generalists Slow-moving generalists can adjust their holdings of both A and B bonds, as well as the riskless short-term asset, but can do so only every k periods. Generalists are present in mass $1 - q_A - q_B$. Fraction $1/k$ of these generalists investors are active each period and can reallocate their portfolios between the A and B bonds. However, they must then maintain this same portfolio allocation for the next k periods. First introduced in Duffie (2010), this is a simple way to model the frictions that limit the speed of capital flows across markets and is reminiscent of New Keynesian models of nominal price rigidity in which firms only adjust their nominal prices every k periods (e.g., Fischer [1977] and Taylor [1979]).

Slow-moving generalist investors have mean-variance preferences over their k-period cumulative portfolio excess return. Defining $rx_{A,t\to t+k} \equiv \sum_{i=1}^{k} rx_{A,t+i}$ and $rx_{B,t\to t+k} \equiv \sum_{i=1}^{k} rx_{B,t+i}$ as the cumulative k-period returns from t to $t+k$ on A and B bonds, the k-period portfolio excess return of generalists who are active at t is

$$rx_{d_t,t\to t+k} = d_{A,t} \times rx_{A,t\to t+k} + d_{B,t} \times rx_{B,t\to t+k}. \tag{13}$$

Thus, generalist investors who are active at time t choose their holdings of A and B bonds, denoted

geneous risk aversion, one should interpret τ as the aggregate risk tolerance of all investors in the economy, q_A as the fraction of aggregate risk tolerance attributable to A-specialists, q_B as the fraction attributable to B-specialists, and $1 - q_A - q_B$ as the fraction attributable to generalists.

$d_{A,t}$ and $d_{B,t}$, to solve

$$\max_{d_{A,t},d_{B,t}} \left\{ E_t\left[rx_{d_t,t\to t+k}\right] - (2\tau)^{-1}\left(Var_t\left[rx_{d_t,t\to t+k}\right]\right)\right\}. \quad (14)$$

This implies that

$$\begin{aligned}\begin{bmatrix} d_{A,t} \\ d_{B,t}\end{bmatrix} &= \tau \begin{bmatrix} Var_t\left[rx_{A,t\to t+k}\right] & Cov_t\left[rx_{A,t\to t+k}, rx_{B,t\to t+k}\right] \\ Cov_t\left[rx_{A,t\to t+k}, rx_{B,t\to t+k}\right] & Var_t\left[rx_{B,t\to t+k}\right]\end{bmatrix}^{-1} \begin{bmatrix} E_t\left[rx_{A,t\to t+k}\right] \\ E_t\left[rx_{B,t\to t+k}\right]\end{bmatrix} \quad (15) \\ &= \frac{\tau}{1-R^{2(k)}_{AB}} \begin{bmatrix} \frac{E_t\left[rx_{A,t\to t+k}\right]}{Var_t\left[rx_{A,t\to t+k}\right]} - \beta^{(k)}_{B|A}\frac{E_t\left[rx_{B,t\to t+k}\right]}{Var_t\left[rx_{B,t\to t+k}\right]} \\ \frac{E_t\left[rx_{B,t\to t+k}\right]}{Var_t\left[rx_{B,t\to t+k}\right]} - \beta^{(k)}_{A|B}\frac{E_t\left[rx_{A,t\to t+k}\right]}{Var_t\left[rx_{A,t\to t+k}\right]}\end{bmatrix},\end{aligned}$$

where, for example, $\beta^{(k)}_{A|B}$ is the coefficient from a linear regression of $rx_{A,t\to t+k}$ on $rx_{B,t\to t+k}$ and $R^{2(k)}_{AB}$ is the goodness of fit from this regression.[6,7]

Equation (15) says that, all else equal, generalist investors allocate more capital to market B when the expected return on B bonds is higher. Furthermore, assuming that A and B bonds comove positively (i.e., $\beta^{(k)}_{A|B} > 0$), generalists allocate less capital to market B when the expected return on A bonds is higher. In this way, the response of generalist investors transmits shocks to the supply of A bonds to the B market, promoting cross-market integration over time. Cross-market capital flows become more responsive to differences in expected returns between the two markets when the two bonds are closer substitutes (i.e., when $R^{2(k)}_{AB}$ is higher). In the limit as the two bonds become perfect substitutes (i.e., as $R^{2(k)}_{AB}$ approaches 1), generalist investors become extremely aggressive in exploiting any cross-market pricing differences.

Remarks on market structure The market structure we have described is a natural way to capture the industrial organization of real world asset management. Due to agency and informational problems, savers are only willing to give asset managers the discretion to adjust their portfolios quickly if the manager accepts a narrow, specialized mandate. These same agency and informational frictions

[6] All investors in our model are myopic—i.e., they make a one-time decision to maximize expected utility over terminal wealth at some future date—which means that we are shutting off intertemporal hedging motives that stem from time variation in risk premia. However, introducing these intertemporal hedging motives would not qualitatively alter our main conclusions. This is because, in settings like ours where risk premia are mean-reverting (here due to mean-reverting supply shocks), these intertemporal hedging motives tend to push in the same direction as myopic, mean-variance motives for holding risky assets (see e.g., Campbell and Viceira (1999) and Albagli (2015)).

[7] We obtain similar results if we alter equation (14) to reflect the fact that the cumulative return from rolling over an investment at the short rate for k periods, $\sum_{i=0}^{k-1} r_{t+i}$, is unknown at time t. As in Campbell and Viceira (2001), this adds a hedging motive for holding long-duration assets that have high excess returns when short rates turn out to be lower than expected. Formally, this means that generalists solve $\max_{d_{A,t},d_{B,t}}\left\{E_t\left[r_{P,t\to t+k}\right] - \frac{1}{2\tau}\left(Var_t\left[r_{P,t\to t+k}\right]\right)\right\}$ where $r_{P,t\to t+k} = (\sum_{i=0}^{k-1} r_{t+i}) + d_{A,t}\times(\sum_{i=1}^{k} rx_{A,t+i}) + d_{B,t}\times(\sum_{i=1}^{k} rx_{B,t+i})$. The solution takes the same form as (15), replacing $E_t[\sum_{i=1}^{k} rx_{A,t+i}]$ with $E_t[\sum_{i=1}^{k} rx_{A,t+i}] - \tau^{-1}\ Cov_t[\sum_{i=1}^{k} rx^{A}_{t+i}, \sum_{i=0}^{k-1} r_{t+i}]$ and similarly for the B bonds.

also mean that savers are only willing to give managers the discretion to adjust quickly if the manager gives them an open-ended claim (Stein [2005]). As a result, fast-moving investors often have endogenously short horizons. By contrast, most institutions, such as endowments and pensions, that have longer horizons and possess greater flexibility to reallocate capital across asset classes are subject to governance mechanisms—themselves a response to informational and agency frictions—that limit the speed of any such capital movements.

In this paper, we regard the parameters governing market structure—i.e., q_A, q_B, and k—as fixed and exogenously given. To be sure, the relevant empirical values for these parameters depend crucially on the two asset classes being considered. Taking these parameters as fixed and given allows us to draw out the asset-pricing implications arising from partial segmentation and slow-moving capital. In this regard, we are following Bacchetta and van Wincoop (2010, 2017), Duffie (2010), Chien, Cole, and Lustig (2012), and Bogousslavsky (2016) who all assume a fixed, exogenous rebalancing frequency. Endogenizing these parameters is challenging and is beyond the scope of the present paper. In principle, one could consider an equilibrium in which q_A, q_B, and k take on a set of fixed values such that A-specialists, B-specialists, and generalists all have the same expected utility in the long-run (see Oehmke [2009] for such an analysis). Relatedly, one could also allow generalists' rebalancing frequency and, hence, the speed of capital flow across markets could vary over time as in Duffie and Strulovici (2012).

2.3 Equilibrium

To solve for a rational expectations equilibrium of our model, we need to clear the market for both the A and B bonds in a way that is consistent with (i) optimization on the part of A-specialists, B-specialists, and generalists and (ii) where all agents correctly perceive the laws of motions for all exogenous and endogenous variables.

Market clearing In market A at time t, there is a mass q_A of fast-moving specialists, each with demand $b_{A,t}$, and a mass $(1 - q_A - q_B) \times (1/k)$ of active slow-moving generalists, each with demand $d_{A,t}$. These investors must accommodate the *active supply*, which is the total supply of $s_{A,t}$ less any supply held off the market by inactive generalist investors, $(1 - q_A - q_B)\, k^{-1} \sum_{j=1}^{k-1} d_{A,t-j}$. Thus, the market-clearing condition for A bonds is

$$\overbrace{q_A b_{A,t}}^{\substack{\text{Specialist} \\ \text{demand}}} + \overbrace{(1 - q_A - q_B)k^{-1} d_{A,t}}^{\substack{\text{Active generalist} \\ \text{demand}}} = \overbrace{s_{A,t}}^{\substack{\text{Total bond} \\ \text{supply}}} - \overbrace{(1 - q_A - q_B)(k^{-1} \sum\nolimits_{i=1}^{k-1} d_{A,t-i})}^{\substack{\text{Inactive generalist} \\ \text{holdings}}}. \tag{16}$$

The market-clearing condition for B bonds is analogous.

Equilibrium conjecture We solve for a rational expectations equilibrium in which equilibrium yields and generalist demands are linear functions of a state vector, $\mathbf{x}_t$, that includes the steady-state deviations of the short-term interest rate, the default realization, the supply of A bonds, the supply of B bonds, inactive generalist holdings of A bonds, and inactive generalist holdings of B bonds. Formally, we conjecture that the yields on A and B bonds are

$$y_{A,t} = \alpha_{A0} + \boldsymbol{\alpha}'_{A1}\mathbf{x}_t, \tag{17}$$

$$y_{B,t} = \alpha_{B0} + \boldsymbol{\alpha}'_{B1}\mathbf{x}_t, \tag{18}$$

and that the demands of slow-moving generalists are

$$d_{A,t} = \delta_{A0} + \boldsymbol{\delta}'_{A1}\mathbf{x}_t, \tag{19}$$

$$d_{B,t} = \delta_{B0} + \boldsymbol{\delta}'_{B1}\mathbf{x}_t, \tag{20}$$

where the $2(1+k) \times 1$ dimensional state vector, $\mathbf{x}_t$, is given by

$$\mathbf{x}_t = [r_t - \overline{r}, z_t - \overline{z}, s_{A,t} - \overline{s}_A, s_{B,t} - \overline{s}_B, d_{A,t-1} - \delta_{A0}, \cdots, d_{A,t-(k-1)} - \delta_{A0}, d_{B,t-1} - \delta_{B0}, \cdots, d_{B,t-(k-1)} - \delta_{B0}]'. \tag{21}$$

These assumptions imply that the state vector follows an AR(1) process

$$\mathbf{x}_{t+1} = \boldsymbol{\Gamma}\mathbf{x}_t + \boldsymbol{\epsilon}_{t+1}, \tag{22}$$

where $\boldsymbol{\epsilon}_{t+1} \equiv [\varepsilon_{r,t+1}, \varepsilon_{z,t+1}, \varepsilon_{s_A,t+1}, \varepsilon_{s_B,t+1}, 0, \cdots, 0]'$, $Var[\boldsymbol{\epsilon}_{t+1}] \equiv \boldsymbol{\Sigma}$, and the transition matrix $\boldsymbol{\Gamma}$ depends on generalists' demand functions (i.e., $\boldsymbol{\Gamma}$ depends on $\boldsymbol{\delta}_{A1}$ and $\boldsymbol{\delta}_{B1}$).

Equilibrium bond yields Equilibrium bond yields in our model take a natural form. To understand the equilibrium yield on asset A, rewrite equation (4) as

$$y_{A,t} = E_t[(1-\theta_A)(r_t + rx_{A,t+1}) + \theta_A y_{A,t+1}] = E_t[(1-\theta_A)(r_t + \tau^{-1}V_A^{(1)} b_{A,t}) + \theta_A y_{A,t+1}],$$

where the second equality follows from equation (11) and $V_A^{(1)} \equiv Var_t[rx_{A,t+1}] = Var_t[(\theta_A/(1-\theta_A)) \times y_{A,t+1}] = (\theta_A/(1-\theta_A))^2 \boldsymbol{\alpha}'_{A1}\boldsymbol{\Sigma}\boldsymbol{\alpha}_{A1}$ is the equilibrium variance of 1-period excess returns on asset A. Iterating this expression forward, we find that

$$y_{A,t}=(1-\theta_A)\sum_{i=0}^{\infty}\theta_A^i E_t[\overbrace{r_{t+i}}^{\text{Short rate}}+\overbrace{\tau^{-1}V_A^{(1)}b_{A,t+i}}^{\text{Risk premium}}]. \tag{23}$$

The equilibrium yield on the default-free A perpetuity is a weighted average of expected future short rates and future risk premia, where expected risk premia depend on the expected future bond holdings of A-specialists. (The equilibrium bond holdings of A-specialists are a linear function of the state vector, $\mathbf{x}_t$.) Proceeding similarly for asset B, we obtain

$$y_{B,t}=(1-\theta_B)\sum_{i=0}^{\infty}\theta_B^i E_t[\overbrace{r_{t+i}}^{\text{Short rate}}+\overbrace{z_{t+i+1}}^{\text{Default loss}}+\overbrace{\tau^{-1}V_B^{(1)}b_{B,t+i}}^{\text{Risk premium}}]. \tag{24}$$

The equilibrium yield on the defaultable B perpetuity is a weighted average of expected future short rates, future default losses, and future risk premia.

Finally, making use the assumed AR(1) dynamics for the exogenous state variables and the market-clearing condition for asset A, the equilibrium yield on asset A satisfies

$$\begin{aligned} y_{A,t} &= \overbrace{\left\{\overline{r}+\left(\frac{1-\theta_A}{1-\rho_r\theta_A}\right)(r_t-\overline{r})\right\}}^{\text{Expected future short rates}} \\ &+\overbrace{\left[(q_A\tau)^{-1}V_A^{(1)}\left(\overline{s}_A-(1-q_A-q_B)\,\delta_{A0}\right)\right]}^{\text{Unconditional term premia}} \\ &+\overbrace{\left[(q_A\tau)^{-1}V_A^{(1)}\left(\begin{array}{c}\frac{1-\theta_A}{1-\theta_A\rho_{s_A}}\left(s_{A,t}-\overline{s}_A\right)\\ -(1-\theta_A)(1-q_A-q_B)\,k^{-1}\sum_{i=0}^{\infty}\theta_A^i E_t[\sum_{j=0}^{k-1}(d_{A,t+i-j}-\delta_{A0})]\end{array}\right)\right]}^{\text{Conditional term premia}}. \end{aligned} \tag{25}$$

Thus, for instance, the unconditional, steady-state term premium on the A market perpetuity, depends on the steady-state bond holdings of A-specialists, namely $\left(\overline{s}_A-(1-q_A-q_B)\,\delta_{A0}\right)/q_A$. The yield for asset B has an extra term relating to expected future defaults, but is otherwise similar:

$$y_{B,t} = \overbrace{\left\{\overline{r} + \left(\frac{1-\theta_B}{1-\rho_r\theta_B}\right)(r_t - \overline{r})\right\}}^{\text{Expected future short rates}} + \overbrace{\left\{\overline{z} + \frac{1-\theta}{1-\rho_z\theta}\rho_z (z_t - \overline{z})\right\}}^{\text{Expected future default losses}} \tag{26}$$
$$+ \overbrace{\left[(q_B\tau)^{-1} V_B^{(1)} (\overline{s}_B - (1-q_A-q_B)\,\delta_{B0})\right]}^{\text{Unconditional term/credit premia}}$$
$$+ \overbrace{(q_B\tau)^{-1} V_B^{(1)} \left[\begin{array}{c} \frac{1-\theta_B}{1-\theta_B\rho_{s_B}} (s_{B,t} - \overline{s}_B) \\ -(1-\theta_B)(1-q_A-q_B)\,k^{-1}\sum_{i=0}^{\infty}\theta_B^i E_t[\sum_{j=0}^{k-1}(d_{B,t+i-j} - \delta_{B0})] \end{array}\right]}^{\text{Conditional term/credit premia}}.$$

Although equations (25) and (26) show that yields in markets A and B take a similar algebraic form, the risk premia in the two markets will not be the same because of the different risks that market specialists must bear in equilibrium.

Equilibrium solution As shown in the Internet Appendix, a rational expectations equilibrium of our model is a fixed point of a specific operator involving the "price-impact" coefficients, $(\boldsymbol{\alpha}'_{A1}, \boldsymbol{\alpha}'_{B1})$, which show how bond supply and inactive generalist demand impact bond yields, and the "demand-impact" coefficients, $(\boldsymbol{\delta}'_{A1}, \boldsymbol{\delta}'_{B1})$, which show how bond supply and inactive generalist demand impact active generalist demands. Specifically, let $\boldsymbol{\omega} = (\boldsymbol{\alpha}'_{A1}, \boldsymbol{\alpha}'_{B1}, \boldsymbol{\delta}'_{A1}, \boldsymbol{\delta}'_{B1})$ and consider the operator $\mathbf{f}(\boldsymbol{\omega}_0)$ which gives (i) the price-impact coefficients that will clear the two markets and (ii) the demand-impact coefficients consistent with optimization on the part of generalists when agents conjecture that $\boldsymbol{\omega} = \boldsymbol{\omega}_0$ at all future dates. Thus, a rational expectations equilibrium of our model is a fixed point $\boldsymbol{\omega}^* = \mathbf{f}(\boldsymbol{\omega}^*)$.

In any rational expectations equilibrium of our model, bond yields always reflect the expected path of future fundamentals (short rates and default losses). As a result, equilibrium bond holdings do not depend on fundamentals. This implies that an equilibrium of our model is a solution to a system of $8k$ nonlinear equations in $8k$ unknowns. Specifically, we need to determine how equilibrium yields and active generalist demand in markets A and B respond to shifts in the supply of A and B bonds: this generates 8 unknowns and 8 corresponding equations. We also need to determine how equilibrium yields and active generalist demand in markets A and B respond to the holdings of inactive generalists: this generates $8(k-1)$ unknowns and $8(k-1)$ corresponding equations. We solve the relevant system of $8k$ nonlinear equations numerically using the Powell hybrid algorithm which performs a quasi-Newton search to find solutions to a system of nonlinear equations starting from an initial guess.[8] To find all of the solutions, we apply this algorithm by sampling over $10,000$

[8] Rational expectations models with noisy asset supply can only be solved in closed-form in very simple, special cases. For instance, we have only been able to obtain closed-form solutions to our model if we turn off the two key asset pricing

different initial guesses.

As we detail in the Internet Appendix, when asset supply is stochastic, an equilibrium solution only exists if investors are sufficiently risk tolerant (i.e., for τ sufficiently large). When an equilibrium exists, there are multiple equilibrium solutions. Equilibrium non-existence and multiplicity of this sort arise in overlapping-generations, rational-expectations models such as ours where risk-averse investors with finite investment horizons trade an infinitely-lived asset that is subject to supply shocks.[9] Different equilibria correspond to different self-fulfilling beliefs that investors can hold about the price-impact of supply shocks and, hence, the risks associated with holding the two assets.

The intuition for equilibrium multiplicity can be understood most clearly in the relatively simple case when there is just a single risky asset and when there are only fast-moving investors. To clarify the relevant issues, we treat this special case in the Internet Appendix. If investors are sufficiently risk tolerant there are two equilibria in this special case: a low price impact (or low return volatility) equilibrium and a high price impact (or high return volatility) equilibrium. If investors believe that supply shocks will have a large impact on prices, they will perceive the asset as being highly risky. As a result, investors will only absorb a positive supply shock if they are compensated by a large decline in prices, making the initial belief self-fulfilling. However, if investors believe that prices will be less sensitive to supply shocks, they will perceive the asset as being less risky and will absorb a supply shock even if they are only compensated by a modest decline in prices.

Things are slightly more complicated in our more general model. Specifically, the addition of multiple risky assets, the partial segmentation of markets, and the introduction of slow-moving capital, give rise to additional unstable equilibria. However, we always find a unique equilibrium that is stable in the sense that equilibrium is robust to a small perturbation in investors' beliefs regarding the equilibrium that will prevail in the future. Formally, letting $\boldsymbol{\omega}^{(1)} = \boldsymbol{\omega}^* + \boldsymbol{\xi}$ for some small $\boldsymbol{\xi}$ and defining $\boldsymbol{\omega}^{(n)} = \mathbf{f}(\boldsymbol{\omega}^{(n-1)})$, an equilibrium $\boldsymbol{\omega}^*$ is stable if $\lim_{n\to\infty} \boldsymbol{\omega}^{(n)} = \boldsymbol{\omega}^*$ and is unstable if $\lim_{n\to\infty} \boldsymbol{\omega}^{(n)} \neq \boldsymbol{\omega}^*$. Let $\{\lambda_i\}$ denote the eigenvalues of the Jacobian $\mathbf{D}_{\boldsymbol{\omega}}\mathbf{f}(\boldsymbol{\omega}^*)$. If $\max_i |\lambda_i| < 1$, then $\boldsymbol{\omega}^*$ is stable; if $\max_i |\lambda_i| > 1$, then $\boldsymbol{\omega}^*$ is unstable.

Consistent with Samuelson's (1947) "correspondence principle," which says that the comparative statics of stable equilibria have certain properties, this unique stable equilibrium has comparative statics that accord with standard economic intuition. By contrast, the unstable equilibria have comparative statics that run contrary to standard intuition.[10] Since we will be interested in comparative

frictions—partial segmentation and slow-moving capital—that are the focus our of paper.

[9] For previous treatments of these issues, see Spiegel (1998), Bacchetta and van Wincoop (2003), Watanabe (2008), Banerjee (2011), Greenwood and Vayanos (2014), and Albagli (2015).

[10] For instance, in the simple case discussed above, the low price-impact equilibrium is stable and the high price-impact equilibrium is unstable. At the stable equilibrium, an increase in the volatility of fundamental shocks or the volatility of supply shocks is associated with an increase in the price-impact coefficient and an increase in the volatility of returns.

statics, we focus on this unique stable equilibrium in our numerical illustrations in Section 3. However, the Internet Appendix discusses the main properties of the model's unstable equilibria.

2.4 Defining market integration

What do we mean by "market integration"? In well-integrated financial markets where all investors can frictionlessly adjust their holdings of each asset, each investor prices exposure to common risk factors in the exact same way.[11] For instance, if a set of investors differ in their risk preferences and their exposures to background risks, then, so long as they all can frictionlessly buy or sell each risky asset, they will all price exposure to common risk factors in the same way.[12] However, all investors will not price risk in the same way if, as in our model, different investors are bound by different frictional constraints—e.g., different constraints on the assets they can trade. Below we clarify what it means for markets to be integrated at different horizons.

Thus, we define markets to be *fully integrated in the short run* if, at each date, all investors price exposures to common risk factors in the same way. For example, the pricing of interest rate risk is integrated in the short run if, at each date, all investors demand the same the expected return per unit of marginal exposure to short-term interest rate shocks. Similarly, we define markets to be *fully integrated in the long-run* if all investors price exposures to common risk factors in the same way on average—i.e., in the model's long-run, steady-state. Long-run integration is therefore a weaker form of market integration than short-run integration. Since investors have the same risk preferences and face no background risks in our model, investors will price exposures to common risk factors in the same way if and only if they all hold the same portfolios, which is only possible if all investors hold the market portfolio.[13]

The degree of market integration depends on which investors can bear risk at different horizons.

By contrast, these comparitive statics take the opposite sign at the unstable equilibrium.

[11] In frictionless markets the only binding constraint that investors face are their budget constraints. Investors cannot face binding short-sales constraints, borrowing constraints, or constraints on the set of assets they can trade.

[12] More generally, if markets are complete (the set of traded assets spans the set of risk factors) and all investors can frictionlessly adjust their asset holdings, then all investors must share the same stochastic discount factor (SDF)—risk sharing will be perfect—and this unique SDF will price all assets. If markets are incomplete and all investors can frictionlessly adjust their asset holdings, then each investor's SDF will price all assets. While investors' SDFs may not be equated state-by-state in incomplete markets—risk sharing may be imperfect, every investor's SDF must share the same linear projection onto the space of excess returns—i.e., there is a unique "SDF-mimicking portfolio." Thus, any investor's SDF equals this unique SDF-mimicking portfolio plus a residual term that is orthogonal to all of the excess returns. This implies that all agents price risk in the same way—i.e., all agents charge the same prices of risk for bearing additional factor risk at the margin, irrespective of their risk preferences or exposures to other background risks.

[13] To see the factor risk prices charged by A-specialists, write $rx_{A,t+1} - E_t\left[rx_{A,t+1}\right] = \boldsymbol{\phi}_A' \boldsymbol{\epsilon}_{t+1}$ where $\boldsymbol{\epsilon}_{t+1}$ are the four factor innovations and $\boldsymbol{\phi}_A = -\left[\theta_A / \left(1-\theta_A\right)\right] \boldsymbol{\alpha}_{A1}$ are the equilibrium factor loadings for A bonds. Since $E_t\left[rx_{A,t+1}\right] = \tau^{-1} Var_t\left[rx_{A,t+1}\right] b_{A,t} = \boldsymbol{\phi}_A' \left(\tau^{-1} \boldsymbol{\Sigma} \boldsymbol{\phi}_A b_{A,t}\right) = \boldsymbol{\phi}_A' \boldsymbol{\lambda}_{A,t}$, the risk prices charged by A-specialists are $\boldsymbol{\lambda}_{A,t} \equiv \tau^{-1} \boldsymbol{\Sigma} \boldsymbol{\phi}_A b_{A,t}$. Proceeding similarly for B-specialists, we have $E_t\left[rx_{B,t+1}\right] = \tau^{-1} Var_t\left[rx_{B,t+1}\right] b_{B,t} = \boldsymbol{\phi}_B' \left(\tau^{-1} \boldsymbol{\Sigma} \boldsymbol{\phi}_B b_{B,t}\right) = \boldsymbol{\phi}_B' \boldsymbol{\lambda}_{B,t}$ where $\boldsymbol{\phi}_B = -\left[\theta_B / \left(1-\theta_B\right)\right] \boldsymbol{\alpha}_{B1} - \mathbf{e}_z$. Thus, the risk prices charged by B-specialists are $\boldsymbol{\lambda}_{B,t} \equiv \tau^{-1} \boldsymbol{\Sigma} \boldsymbol{\phi}_B b_{B,t}$.

In partially segmented markets the risk prices charged by A-specialists are not consistent with the pricing of asset B and vice versa (i.e., when $(1 - q_A - q_B) < 1$, $E_t\left[rx_{B,t+1}\right] \neq \boldsymbol{\phi}_B' \boldsymbol{\lambda}_{A,t}$ and $E_t\left[rx_{A,t+1}\right] \neq \boldsymbol{\phi}_A' \boldsymbol{\lambda}_{B,t}$).

It is driven by two parameters: $(1 - q_A - q_B)$ and k.[14] The first parameter, $(1 - q_A - q_B)$, is the population share of generalists. This parameter determines the degree of long-run integration between markets. For instance, if $(1 - q_A - q_B) = 1$, markets will be well-integrated in the long-run even if k is large. When $(1 - q_A - q_B) < 1$, specialists in the two markets price risk differently on average; this is the sense in which the two markets are *partially segmented* in the long-run. By contrast, active generalists price risk consistently across the two markets on average; this is the sense in which the two markets are *partially integrated* in the long-run.

The second parameter, k, indexes the speed with which generalist capital can flow between markets. Thus, k determines the degree of short-run integration between the two markets, holding fixed the degree of long-run integration. Markets are perfectly segmented if $(1 - q_A - q_B) = 0$ or $k \to \infty$. If either of these conditions holds, the two markets operate independently of each other.

Formally, collect all of the 1-period returns in a vector $\mathbf{rx}_{t+1} \equiv [rx_{A,t+1}, rx_{B,t+1}]'$ and the asset supplies in a vector $\mathbf{s}_t \equiv [s_{A,t}, s_{B,t}]'$. Letting $rx_{M_t,t+1} = \mathbf{s}_t'\mathbf{rx}_{t+1}$ denote the excess returns on the current market portfolio, we define markets to be *fully integrated in the short-run* if

$$E_t[\mathbf{rx}_{t+1}] = \tau^{-1} Var_t[\mathbf{rx}_{t+1}]\,\mathbf{s}_t = \boldsymbol{\beta}_t[\mathbf{rx}_{t+1}, rx_{M_t,t+1}]\,E_t[rx_{M_t,t+1}], \qquad (27)$$

where $\beta_t[\mathbf{rx}_{t+1}, rx_{M_t,t+1}] = Var_t[\mathbf{rx}_{t+1}]\,\mathbf{s}_t / (\mathbf{s}_t' Var_t[\mathbf{rx}_{t\to t+j}]\,\mathbf{s}_t)$. In other words, markets are fully integrated in the short-run if, at each date, a conditional-CAPM based on the current market portfolio prices both the A and B assets. In our model, markets are fully integrated in the short-run if and only if $(1 - q_A - q_B) = 1$ and $k = 1$. By contrast, when $(1 - q_A - q_B) \neq 1$ or $k \neq 1$, this conditional-CAPM will not price the 1-period returns on the A and B assets.[15]

Similarly, we define markets to be *fully integrated in the long-run* if

$$E[\mathbf{rx}_{t\to t+k}] = \tau^{-1} Var_t[\mathbf{rx}_{t\to t+k}]\,E[\mathbf{s}_t] = \boldsymbol{\beta}[\mathbf{rx}_{t\to t+k}, rx_{\overline{M},t\to t+k}]E[rx_{\overline{M},t\to t+k}], \qquad (28)$$

where $\beta[\mathbf{rx}_{t\to t+k}, rx_{\overline{M},t\to t+k}] = Var_t[\mathbf{rx}_{t\to t+k}]\,E[\mathbf{s}_t] / (E[\mathbf{s}_t']\,Var_t[\mathbf{rx}_{t\to t+k}]\,E[\mathbf{s}_t])$ and $E[rx_{\overline{M},t\to t+k}] = E[\mathbf{s}_t']\,E[\mathbf{rx}_{t\to t+k}]$. In other words, markets are fully integrated in the long-run if the same unconditional-

[14] In our model, news about fundamentals—short-term rates and default losses—is reflected immediately in both markets. Thus, market integration has nothing to do with the speed at which fundamental news is reflected in prices.

[15] Of course, there is always a unique SDF-mimicking portfolio $\mathbf{c}_t^* = \tau\,(Var_t[\mathbf{rx}_{t+1}])^{-1}\,E_t[\mathbf{rx}_{t+1}]$ such that $E_t[\mathbf{rx}_{t+1}] = \tau^{-1} Var_t[\mathbf{rx}_{t+1}]\,\mathbf{c}_t^* = \boldsymbol{\beta}_t[\mathbf{rx}_{t+1}, rx_{C_t^*,t+1}]\,E_t[rx_{C_t^*,t+1}]$ where $rx_{C_t^*,t+1} = \mathbf{c}_t^{*\prime}\mathbf{rx}_{t+1}$ and $\beta_t[\mathbf{rx}_{t+1}, rx_{C_t^*,t+1}] = Var_t[\mathbf{rx}_{t+1}]\,\mathbf{c}_t^* / (\mathbf{c}_t^{*\prime} Var_t[\mathbf{rx}_{t+1}]\,\mathbf{c}_t^*)$. However, unless $k = 1$ and $(1 - q_A - q_B) = 1$, we will have $\mathbf{c}_t^* \neq \mathbf{s}_t$ (i.e., the SDF-mimicking portfolio will differ from the market portfolio) and different agents will price risk in different ways. When $k > 1$ and $(1 - q_A - q_B) < 1$, no individual investor will hold the portfolio $\mathbf{c}_t^*$. When $k = 1$ and $(1 - q_A - q_B) < 1$, generalists will hold $\mathbf{c}_t^*$ (i.e., $\mathbf{d}_t = \mathbf{c}_t^*$) but A- and B-specialists will clearly not hold $\mathbf{c}_t^*$.

To see the prices of risk implied by $\mathbf{c}_t^*$, let $\boldsymbol{\Phi} = [\boldsymbol{\phi}_A\ \boldsymbol{\phi}_B]'$ denote the matrix of equilibrium factor loadings and note that $E_t[\mathbf{rx}_{t+1}] = \tau^{-1} Var_t[\mathbf{rx}_{t+1}]\,\mathbf{c}_t^* = \boldsymbol{\Phi}\left(\tau^{-1}\boldsymbol{\Sigma}\left(c_{A,t}^*\boldsymbol{\phi}_A + c_{B,t}^*\boldsymbol{\phi}_B\right)\right) = \boldsymbol{\Phi}\boldsymbol{\lambda}_{C^*,t}$. Thus, the implied risk prices are $\boldsymbol{\lambda}_{C^*,t} \equiv \tau^{-1}\boldsymbol{\Sigma}\,(c_{A,t}\boldsymbol{\phi}_A + c_{B,t}\boldsymbol{\phi}_B)$.

CAPM based on the average market portfolio ($rx_{\overline{M},t\to t+k} = E[\mathbf{s}_t']\,\mathbf{rx}_{t\to t+k}$) prices k-period returns on both the A and B assets on average—i.e., in the model's steady state where asset supply equals the long-run average $E[\mathbf{s}_t]$. In our model, markets are fully integrated in the long-run if and only if $(1 - q_A - q_B) = 1$, irrespective of k. By contrast, when $(1 - q_A - q_B) < 1$, this unconditional CAPM will not price the k-period returns on the A and B assets in the model's steady state.

The reason markets are not integrated in our model is because cross-market "arbitrage" is risky for generalists. Unless $(1 - q_A - q_B) = 1$ and $k = 1$, full short-run integration fails because generalists demand compensation for the risk associated with the short-run trades they place to exploit the differences in the way that A-specialists and B-specialists price risk. Similarly, unless $(1 - q_A - q_B) = 1$, full long-run integration fails because generalists are engaged in a risky cross-market arbitrage trade even in the model's long-run steady state. Specifically, when $(1 - q_A - q_B) < 1$, generalists will not hold the market portfolio in the steady state (i.e., $E[d_{A,t}] \neq E[s_{A,t}]$ and $E[d_{B,t}] \neq E[s_{B,t}]$). Relative to the market portfolio, generalists' portfolio will incorporate a tilt that reflects average cross-market pricing differences. And, generalists will demand compensation for bearing the risks stemming from this portfolio tilt.[16] Thus, in the general case where $(1 - q_A - q_B) < 1$ and $k > 1$, we obtain neither full short-run nor long-run market integration.

3 Market integration following large supply shocks

In this section, we use our model to illustrate the asset price dynamics following a large supply shock in one asset market. We relate these dynamics to the model's two key asset pricing frictions, namely partial market segmentation and slow-moving capital.

3.1 Parameters for baseline illustration

Table 1 lists the set of parameter values that we use in our numerical illustrations. For the purposes of these illustrations, it will be helpful to think of market A as the U.S. Treasury market and market B as the corporate bond market and we have picked parameter values to approximate the behavior of these markets. Thus, for example, we choose σ_r and ρ_r to approximately match the volatility and persistence of short-term nominal interest rates. Each period in our numerical illustrations corresponds to one quarter of a year. However, we report annualized values for both risk premia and bond yields by multiplying the corresponding quarterly risk premia and yields by four.

[16] For instance, when $q_A = q_B$ and $\overline{s}_A = \overline{s}_B$, generalists will be overweight B bonds and underweight A bonds relative to the steady-state market portfolio. The reason is that B bonds are riskier than A bonds since the former are exposed to default risk. As a result, B-specialists will hold less of the B bonds than A-specialists hold of the A bonds.

Our goal is to use the model to draw structural impulse response functions that trace out the causal impact on asset prices following a shock to asset supply. Thus, for simplicity and clarity of interpretation, we assume that the four exogenous shocks in the model—$\varepsilon_{r,t+1}$, $\varepsilon_{z,t+1}$, $\varepsilon_{s_A,t+1}$, and $\varepsilon_{s_B,t+1}$—are mutually orthogonal.

We begin our analysis by choosing $k = 8$ quarters and $q_A = q_B = 45\%$, but later show comparative statics for these parameters. Based on these values, our illustrations assume that most risk-bearing capital is controlled by specialists, with 10% being controlled by flexible generalist investors, one-eighth of whom reallocate their portfolios each quarter. Our choice of $k = 8$ quarters is somewhat arbitrary, but we think of this as capturing the empirically relevant case of pension funds or endowments who typically review their asset allocations on an annual or biannual basis. As shown below in Table 2, all the main results carry through if we instead assume that $k = 4$ quarters, but the dynamics are easier to see in a graphical format when we assume $k = 8$ quarters.

The parameter τ should be interpreted as the aggregate risk tolerance of all investors in the economy and should not be equated with the risk tolerance (i.e., inverse CARA parameter) of any individual investor. We normalize the volatility of shocks to the supply of A and B to one ($\sigma_{s_A} = \sigma_{s_B} = 1$) and then choose an aggregate risk tolerance of $\tau = 1.75$ to obtain realistic values for equilibrium risk premia.[17] We assume a steady-state supply of both A and B of five units ($\overline{s}_A = \overline{s}_B = 5$).

3.2 An unanticipated supply shock

3.2.1 Baseline illustration

We first consider the impact of an unanticipated supply shock that doubles the supply of asset A in quarter 5—i.e., $s_{A,t}$ jumps from 5 to 10 units between quarters 4 and 5. To make the illustration as stark as possible, we study a near-permanent supply shock and set $\rho_{s_A} = 0.999$.

Figure 1 illustrates the impact of this shock, plotting the evolution of expected annual returns on both bonds in Panel A, the holdings of specialists and generalist investors in Panel B, the active supply of both bonds in Panel C, and bond yields in Panel D. Prior to the supply shock, Panel A shows that the steady-state risk premium in market B is 1.07% per annum versus a risk premium of 0.79% in market A. The additional risk premium of 0.28% arises because market B is subject to default risk, which exposes investors to an additional source of fundamental risk and amplifies the supply risk facing B bond holders. The initial yield in market B is 5.47% per annum versus a yield of 4.79% in market A. The steady-state yield in market A equals the average short-term riskless rate of

[17] As is common in models where mean-variance investors face stochastic supply shocks, equilibrium prices and risk premia only depend on the ratio of asset supply to investors' aggregate risk tolerance. Thus, papers studying models of this sort either (i) normalize $\tau = 1$ and then choose the volatility of supply shocks to obtain realistic results or (ii) they normalize the volatility of supply shocks to 1 and then choose τ. We opted for the latter approach.

$\overline{r} = 4.00\%$ plus a steady-state risk premia of 0.79%. The 0.68% steady-state yield spread between the B and A markets equals the difference in steady-state risk premia of 0.28% plus market B's expected default losses of $\overline{z} = 0.40\%$ per annum.

When the supply shock hits the market A at $t = 5$, expected returns and yields in both markets react immediately. Panel A shows that expected returns in market A *overreact:* they jump from 0.79% to 1.59% before ultimately falling back to a new long-run level of 1.35%. The overreaction of expected returns on A bonds reflects the relative steepness of short-run demand curves and relative flatness of long-run demand curves. Unlike textbook theories in which asset prices are determined solely by the stock of risky assets supplied, our approach suggests that supply flows—the rate at which the stock is changing—also matter in the short run. As explained in Duffie (2010), these are general properties of models featuring slow-moving capital and are not unique to our model.

In contrast, Panel A shows the key novel implication of our model: expected returns in market B actually *underreact* to the shock to the supply of A, rising slowly from 1.07% prior to the shock to a new long-run level of 1.30%. Why does market A overreact to the supply shock while market B underreacts? Panel B shows how the positions of different market participants evolve over time. Following the initial supply shock in market A, both specialist holdings of A ($b_{A,t}$) and active generalist holdings of A ($d_{A,t}$) spike upwards. As a partial hedge against their increased holdings of A, active generalists reduce their holdings in market B. This reduction in generalists' B holdings is motivated by a need to reduce the common short-term interest rate risk across their holdings in both markets. As time passes and more generalists reallocate their portfolios in response to the shock, B-specialists must hold more of the B bonds to fill the void left by the generalists. Thus, the expected returns on B bonds rise gradually over time.

The dynamics of risk premia are closely tied to the dynamics of the "active supply" of A and B that must be absorbed by active market participants each period. For instance, the active supply of A bonds is given by the right-hand-side of equation (16): $s_{A,t} - (1 - q_A - q_B)(k^{-1} \sum_{i=1}^{k-1} d_{A,t-i})$—i.e., the total asset supply less the assets that are being held off the market by inactive generalists. The evolution of the active supplies is shown in Panel C and mirrors the path of bond risk premia shown in Panel A. The initial supply shock to A at $t = 5$ immediately increases the active supply of A bonds but has no immediate effect on the active supply of B bonds. This is because slow-moving generalists have yet to reduce their holdings in market B. Over the ensuing periods, generalists gradually increase their holdings of A and reduce their holdings of B. Therefore, the active supply in A gradually declines while the active supply in B gradually rises, paralleling the movements in bond risk premia.

By $t = 12$ all generalist investors have reallocated their portfolios in response to the supply shock at $t = 5$ (recall that $k = 8$ in this illustration). However, the gradual adjustment of generalists gives rise to modest echo effects after $t = 12$, generating a series of damping oscillations that converge to the new long-run equilibrium. As in Duffie (2010), these oscillations arise because generalists who reallocate soon after the supply shock lands take large opportunistic positions. These large positions temporarily reduce the active supply of A and then need to be absorbed in later periods.[18]

In market A, risk premia charged by A-specialists are the sum of risk premia related to shocks to short-term interest rates, shocks to the supply of A, and shocks to the supply of B. Changes in total risk premia are primarily driven by the pricing of interest rate risk. In market B, the risk premia charged by B-specialists can be similarly decomposed into its components, which also includes a premium for default risk. Following the supply shock, the risk premia associated with interest rate risk differs significantly between the two sets of specialists. As generalists react to this pricing discrepancy, the difference in interest rate risk premia charged by the two sets of specialists gradually narrows. However, the difference does not vanish in the long run because of the permanent risks associated with cross-market arbitrage.

In Panel D, we shift from risk premia to bond yields. The overreaction of the A market and the underreaction of the B market are more muted in yield space than in risk premium space: market A yields overreact by 7% and market B yields underreact by 12%. This is natural since, as shown in equations (23) and (24), bond yields reflect the weighted average expected risk premia over the life of the bond with risk premia in the near future having a larger effect on the weighted average than those in the distant future.

Panel D also shows that shock to the supply of A bonds leads the yield spread between defaultable and default-free bonds, $y_B - y_A$, to decline. However, because yield A overreacts and yield B underreacts to the shock, the yield spread overreacts even more (19%) than the yield on A bonds.

3.2.2 Unpacking the role of the two key frictions

Our model combines two key asset pricing frictions, namely partial market segmentation and slow-moving capital. In this subsection, we show that *both* of these frictions are necessary to generate underreaction in market B following a shock to the supply of the A bonds.

To understand how these two frictions interact, Figure 2 shows asset price dynamics following the same near-permanent shock to the supply of the A bonds considered above: the supply of the A bonds unexpectedly doubles in quarter $t = 5$. The four panels of Figure 2 show the risk premium dynamics as we switch these two asset pricing frictions on and off.

- **Panel A: Benchmark with no frictions (Upper left):** Here all investors are fast-moving

[18] Exploiting this feature of Duffie's (2010) model, Bogousslavsky (2016) argues that infrequent rebalancing can explain the echo effects found intra-day and daily stock returns.

generalists who can invest in both assets and can rebalance their portfolios each period. This is a special case of our model where $q_A = q_B = 0$ and $k = 1$—i.e., where both markets are fully integrated in the short run. In this case, there is no overreaction and the risk premia on the A and B bonds move in near-perfect lockstep following the shock to the supply of the A bonds. This is because the price of interest rate risk jumps and both bonds have the same exposure to this risk factor.[19]

- **Panel B: Only partial segmentation (Upper right):** Here all investors can rebalance each period, so capital is not slow-moving. However, markets are partially segmented: there are specialists who can trade each bond and as well as a subset of generalists who can trade both bonds. This is a special case of our general model when $q_A > 0$, $q_B > 0$, $(1 - q_A - q_B) > 0$, and $k = 1$. Relative to the frictionless case, partial segmentation raises the steady-state risk premium in both markets because a large fraction of total risk is now born by undiversified specialists. And, the shock to the supply of A bonds has a larger impact on risk premia in market A and a smaller impact on those in market B—i.e., the cross-market spillover is smaller—than in the benchmark case with full short-run integration. However, since all agents can rebalance each period, there is no difference between short-run and long-run demand curves. As a result, the short-run price impact is the same as the long-run price impact.

- **Panel C: Only slow-moving capital (Lower left):** This case corresponds to a multi-asset version of Duffie's (2010) model and is considered Bogousslavsky (2016). Here all investors can trade both the A and B bonds; there is no market segmentation. However, fraction q of investors are fast-moving and can trade each period, while fraction $(1 - q)$ are slow-moving and can only trade every k periods. This is *not* a special case of our general model. When the two assets are exposed to a common risk factor, the two-asset Duffie model will *always* deliver short-run overreaction in *both* assets A and B. Factor risk prices overreact in the short-run because short-run demand curves are steeper than long-run demand curves. However, since fast-moving agents price risk similarly in both markets, we will always see similar short-run overreaction in both the directly impacted market (A) and the market that is not directly impacted (B).[20]

[19] The risk premia do not move in perfect lock step because (i) the prices of A-supply risk and B-supply risk also jump and (ii), when $\sigma_z^2 > 0$, the B bonds have a larger exposure to these two risk factors than the A bonds. Thus, the risk premium on the B bonds rises slightly more than that on the A bonds following the supply shock. However, this effect is tiny when σ_z^2 is small. (i.e., when A and B bonds are close substitutes) and is not visible in the plot.

[20] To see this formally, note that we can always write $\mathbf{rx}_{t+1} - E_t[\mathbf{rx}_{t+1}] = \mathbf{\Phi}\boldsymbol{\epsilon}_{t+1}$ for some set of equilibrium factor loadings $\mathbf{\Phi} = [\boldsymbol{\phi}_A \ \boldsymbol{\phi}_B]'$, so that $Var_t[\mathbf{rx}_{t+1}] = \mathbf{\Phi\Sigma\Phi}'$. In the multi-asset version of the Duffie (2010) model, the market clearing conditions can be written as $q\mathbf{b}_t + (1-q)\,k^{-1}\sum_{j=0}^{k-1}\mathbf{d}_{t-j} = \mathbf{s}_t$ where $\mathbf{b}_t = \tau\,(Var_t[\mathbf{rx}_{t+1}])^{-1}E_t[\mathbf{rx}_{t+1}]$ is the demand of fast-moving investors. This implies that $E_t[\mathbf{rx}_{t+1}] = (q\tau)^{-1}Var_t[\mathbf{rx}_{t+1}](\mathbf{s}_t - (1-q)\,k^{-1}\sum_{j=0}^{k-1}\mathbf{d}_{t-j}) = \mathbf{\Phi}\boldsymbol{\lambda}_t$ where $\boldsymbol{\lambda}_t = (q\tau)^{-1}\mathbf{\Sigma\Phi}'(\mathbf{s}_t - (1-q)\,k^{-1}\sum_{j=0}^{k-1}\mathbf{d}_{t-j})$ are factor risk prices charged by fast-moving investors.

- **Panel D: Both partial segmentation and slow-moving capital (Lower right):** Our model features both partial market segmentation ($q_A > 0$, $q_B > 0$, $(1 - q_A - q_B) > 0$) and slow-moving capital between markets ($k > 1$). Our model will *always* generate short-run overreaction in the A market that is directly impacted by the supply shock. However, our model can *either* generate short-run overreaction or underreaction in the B market that is not directly hit by the supply shock. Specifically, as we show below, when there is minimal segmentation between the two markets—i.e., when $(1 - q_A - q_B)$ is sufficiently large—our model delivers short-run overreaction in both markets as in the multi-asset version of the Duffie (2010) model. However, when the two markets are sufficiently segmented—i.e., when $(1 - q_A - q_B)$ is sufficiently small, our model delivers short-run underreaction in the indirectly impacted market.

One additional contrast between the models in Panels C and D is worth noting. In Panel C where there is only slow-moving capital, there is almost no abnormal trading in the B bonds following the shock to the supply of the A bonds.[21] In other words, the response of B's price to the shock to the supply of A is not mediated by trade. By contrast, in Panel D where there is also partial segmentation, the only reason that B's price responds to increase in the supply of A is because generalist investors are engaged in a long-A, short-B trade—i.e., the response of B's price to the shock is entirely mediated by trade.

In summary, both partial segmentation and slow-moving capital are necessary to generate short-run underreaction in market B. However, these two frictions are not jointly sufficient. Instead, to guarantee short-run underreaction in market B one needs both slow-moving capital and a meaningful degree of segmentation between the two markets.

3.2.3 Comparative statics

In Table 2, we perform a variety of comparative statics exercises to illustrate how the price dynamics following a large shock to the supply of A depend on the parameters of our model. We primarily focus on the parameters governing our two key asset pricing frictions: the population shares of A-specialists, B-specialists, and generalists as well as the frequency at which generalists can rebalance (k).

We again study the same supply shock considered in Figure 1: the supply of asset A unexpectedly doubles at time t. For each different set of model parameters, Table 2 summarizes the impact of this

[21] There is a tiny amount of abnormal trading in B bonds in Panel C following the shock to the supply of A bonds. This is because mean-reverting supply shocks generate negative serial correlation in returns, so the variance ratio $Var_t[rx_{A,t\to t+k}]/k$ is decreasing in k. As a result, longer-horizon investors worry less about transitory discount rate risk, leading them to take larger positions than short-horizon investors in both risky bonds following the supply shock. However, this is a second order effect in our illustration.

supply shock on both the A and B markets by listing the yields and expected annual returns in (i) at time $t-1$ before the shock arrives (labeled as "pre-shock level"), (ii) the short-run change from time $t-1$ to time t when the shock arrives (labeled as "short-run Δ"), and (iii) the long-run change from time $t-1$ to time $t+2k$ (labeled as "long-run Δ"). We define the degree to which bond yields overreact in the short run as the difference between the short-run change and the long-run change, expressed as a percentage of the long-run change[22]

$$\%\textit{Short-Run-Over-React}(y) \equiv \frac{(y_t - y_{t-1}) - (y_{t+2k} - y_{t-1})}{(y_{t+2k} - y_{t-1})}. \tag{29}$$

Our measure of short-run overreaction for risk premia is defined analogously.

The first row in Table 2 shows the results for the baseline set of parameters used in Figure 1. Risk premia in market A overreact by 42% in the short-run using our baseline set of parameters, while risk premia in market B underreact by 83% in the short-run. Similarly, yields in market A overreact by 7%, while yields in market B underreact by 12% in the short-run.

The second row in Table 2 shows that, if market participants are more risk tolerant, this reduces the price impact of the supply shock on both market A and market B. However, changing investor risk tolerance has a similar impact on the short- and long-run response to the supply shock. Thus, the degree of overreaction or underreaction in each market is unchanged in percentage terms.

Recall that q_A and q_B indicate the relative fraction of specialists in market A and B, respectively. In row 3, we set $q_A = q_B = 0.5$, so there are no generalists and the two markets are completely segmented. In this case, the supply shock in market A is no longer transmitted to the market B.

We next change the mix between A-specialists and B-specialists, holding fixed the overall mix between generalists and specialists. Row 4 shows that if we hold the total number of specialists fixed at $q_A + q_B = 0.9$, then as we increase the share of B-specialists and decrease the share of A-specialists, we get more overreaction in A. Market B is only modestly affected by this change because the supply shock is primarily being absorbed by generalists. Similarly, row 5 shows that if we increase the share of A-specialists, holding fixed the total number of specialists, we see less overreaction in A.

Recall that k is the number of quarters that it takes for generalists to fully reallocate their portfolios following a supply shock and that $k = 8$ quarters in in our baseline example. In row 6 we instead set $k = 12$ quarters. Naturally, this larger value of k increases the short-run overreaction in market A and increases the short-run underreaction in market B. Similarly, when we set $k = 4$ in row 7, there

[22] Since our supply shock is not quite permanent, we subtract off the constant $(1-\rho_{S_A}^{2k})/\rho_{S_A}^{2k}$ from %*Short-Run-Over-React* to ensure that our measure is zero if markets are perfectly integrated in the short run $(1 - q_A - q_B = k = 1)$. In this case, there is no "over-reaction" but only "reaction" and we have $[(y_t - y_{t-1}) - (y_{t+2k} - y_{t-1})] / [y_{t+2k} - y_{t-1}] = \left[\alpha_{s_A} s_{A,t} - \alpha_{s_A} s_{A,t} \rho_{s_A}^{2k}\right] / \alpha_{s_A} s_{A,t} \rho_{s_A}^{2k} = (1-\rho_{s_A}^{2k})/\rho_{s_A}^{2k}$. For $k = 8$ and $\rho_{s_A} = 0.999$, this constant is 1.6%.

is less short-run overreaction in market A and less short-run underreaction in market B.

In row 8, we continue to assume that $k = 4$ quarters, but we now assume that $q_A = q_B = 0.2$ as opposed to $q_A = q_B = 0.45$ as we did in row 7. In this case there are many generalists and few specialists, so the markets are well-integrated in the long-run. Both the A and B markets overreact to the shock to the supply of asset A in this case. Intuitively, the two markets behave like two well-integrated markets where the only asset pricing friction is slow-moving capital. As a result, the dynamics are similar to those in the multi-asset version of Duffie's (2010) model discussed above.

3.3 A pre-announced increase in supply

We next study asset price dynamics following the *announcement* of a large *future change* in asset supply. As an example of a large pre-announced supply change, consider the large-scale asset purchase programs initiated by the Federal Reserve between 2008 and 2013. The Fed's initial announcement of long-term bond purchases occurred on November 25, 2008, but asset purchases did not begin until January 2009 and continued in the months and years thereafter.

To mimic the announcement at time t of a future increase in the supply of asset A, we assume that $s_{A,t}$ jumps up and that the amount of asset A held by inactive generalists also jumps up such that the *active supply* of asset A does not change at time t.[23] (Duffie (2010) uses a similar approach to model price dynamics following a pre-announced supply shock.) This means that, unlike the case of an unanticipated supply shock, it would be possible to clear the market at time t without any increase in the holdings of A-specialists or active generalists. Thus, the only reason that prices change when a future supply change is announced is because the announcement leads active long-horizon generalists to opportunistically adjust their asset holdings. Formally, letting $\epsilon_t [X_t] = X_t - E_{t-1} [X_t]$ denote the time t *innovation* or *surprise* to some random process X_t, an anticipated supply shock is defined so that the innovation to the right-hand of equation (16) is zero when it is announced at time t:

$$\epsilon_t [s_{A,t}] - (1 - q_A - q_B) k^{-1} \sum_{j=1}^{k-1} \epsilon_t [d_{A,t-j}] = 0. \tag{30}$$

We examine the pre-announcement of a one-time, near-permanent jump in the supply of asset A: market participants learn in quarter $t = 5$ that the supply of asset A will double in quarter $t = 8$. We continue to assume that $k = 8$ quarters which means that, by the time the supply of A jumps at $t = 8$, half of all generalists investors have had the opportunity to adjust their portfolios following the

[23] Normally, the holdings of inactive generalist investors evolve deterministically and do not jump. Thus, a pre-announced supply increase is an event that agents assume is impossible, which is a drawback of this analysis. Greenwood, Hanson, and Vayanos (2016) develop a term structure model where agents receive *news* about the likely future path of bond supply. However, their model does not feature slow-moving capital.

announcement at $t = 5$.[24]

Figure 3 shows the dynamics of risk premia, investor positions, active asset supplies, and bond yields following the pre-announced supply increase. Panel A shows that annual risk premia in market A actually *decline* slightly after the announcement at $t = 5$ and before the supply rises at $t = 8$. And risk premia in market A still jump up when bond supply jumps to its new level at $t = 8$. The risk premium in market B, which is only indirectly affected due to cross-market arbitrage by generalists, rises gradually over time.

What drives these dynamics? Panel B shows how the positions of market participants evolve over time in response to this pre-announced supply increase. Upon the announcement, generalists begin to opportunistically adjust their portfolios in direction of the anticipated shock, increasing their holdings of A (d_A) and decreasing their holdings of B (d_B). A-specialists provide temporary liquidity to generalists, planning to replenish their inventories once the supply shock actually lands.[25] The gradual build up of generalist demand for asset A is responsible for the slight decline in the risk premium on asset A from $t = 5$ to 7 before its upward jump at $t = 8$. Similarly, the gradual reduction of generalists' demand for B results in a slow increase in the risk premium on asset B.

In contrast to the generalists, the specialists' demand in market A (b_A) decreases initially then increases. This is because specialists can adjust quickly, and thus they have the ability to front-run the anticipated change in supply—specialists reduce their portfolio holdings of A just before the positive supply shock and increase holding of A immediately after the shock.

Panel C summarizes the evolution of the active supplies of assets A and B. By construction, the active supplies of both assets do not change at quarter $t = 5$ when the increase in the supply of asset A is pre-announced. Instead, the active supply of asset A jumps at $t = 8$. However, the active supply of A does not jump by the full 5 units at $t = 8$ because pre-announcing the supply shock mobilizes slow-moving generalists beginning at $t = 5$.

Panel D shows the evolution of bond yields in response to the pre-announced supply shock. Following the announcement, there is now short-run underreaction in both markets: yields in both markets *gradually* rise to their new steady-state levels. The yields in market A underreact by 13% and the

[24] To implement this announcement, we assume that $s_{A,t}$ and $(1 - q_A - q_B)\, d_{A,t-5}/8$ both jump by 5 units at $t = 5$ relative to their steady-state values. Thus, if the generalists who are active at $t = 5$, 6, and 7 did not react to this pre-announced supply increase, the active supply of asset B would jump by 5 units at $t = 8$.

However, active generalists do react to the pre-announced supply increase: we have $\tilde{d}_{A,t} \equiv d_{A,t} - \delta_{A0} > 0$ at $t = 5, 6,$ and 7. As a result, when the supply shock lands at $t = 8$, the active supply that must be absorbed is $5 - (1 - q_A - q_B)\,(\tilde{d}_{A,5} + \tilde{d}_{A,6} + \tilde{d}_{A,7})/8 < 5$. As discussed below, this early mobilization of generalist investors reduces the active supply of asset A that must be absorbed when the shock lands at $t = 8$, damping the overreaction of market A.

[25] Why do generalists buy A bonds *in advance* of the supply increase? Generalists have long-horizons ($k = 8$ in this example), but can only adjust their portfolios slowly. Although generalists expect A bond yields to rise over the next several quarters and, therefore, expect to suffer a capital loss on A bonds, they expect this capital loss to be more than offset by an increase in income from holding A bonds. As a result, the expected cumulative 8-period excess return from holding A bonds rises when a future supply increase is announced.

yields in market B underreact by 16%. Why do A yields rise gradually when the supply shock is pre-announced but overreacted in Figure 1 when the same shock was unanticipated? First, risk premia in the near future have a larger effect on bond yields than those in the distant future. Second, risk-premia in the A market are only expected to rise significantly once the supply actually increases at quarter $t = 8$. In combination, these two facts imply that yields in market A must rise gradually over time following a pre-announced supply increase.

We can compare the dynamics of prices in the case of pre-announced increase in supply in Figure 3 to the unanticipated increase in supply in Figure 1. The *long-run* impact on yields and risk premia is the same in either case. However, the short-run effects can be quite different whether or not a shock is anticipated. When the shock is a unanticipated, yields and risk premia in the A market overreact more strongly at announcement. Pre-announcing the supply shock mobilizes slow-moving generalists before the supply of A actually rises. As shown in Figure 3 Panel C, this early mobilization reduces the active supply of asset A that must be absorbed when the shock lands, damping the overreaction of market A. While introducing a delay between announcement and the increase in supply limits overreaction, this also lengthens the amount of time it takes for the full impact to be reflected in prices.

To put this in more concrete terms, consider a central bank that is seeking to reduce long-term interest rates by purchasing long-term government bonds. Allowing for a long period of time between announcement and implementation of the asset purchases will result in lower profits to market specialists who must accommodate the shock in the short-run. Of course, doing so also delays the desired impact on asset prices. In this way, our model highlights an important trade-off that central bankers face when designing asset purchase programs.

4 Multiple assets per market

In describing the model so far, we have made no distinction between "asset" A and "market" A. This distinction takes on meaning when market A contains many different securities. For example, in the Treasury market, there are many different securities with different maturities and thus different exposures to interest rate risk.

In the Internet Appendix, we show how to extend our model to a more complex setting in which there are multiple risky securities trading in each market. Specifically, suppose there are multiple default-free bonds in market A which have different durations. Similarly, there are multiple bonds in market B which differ in terms of their duration or their exposure to default risk. Fast-moving A-specialists are free to adjust their holdings of all securities in the A market (and the riskless short-term asset) each period, but cannot hold securities in market B. The B-specialists face an analogous set of constraints. Generalists can invest capital in both markets, but, as in our baseline model, generalists can only gradually reallocate their portfolios over time.

Subject to some mild conditions which guarantee that cross-market arbitrage remains risky for generalist investors, the intuitions from the two asset model carry over to this richer setting with more securities.[26] Specifically, a CAPM-like pricing model based on the portfolio of A-specialists prices all assets in the A market and a different CAPM-like pricing model—with different prices of factor risk—based on the portfolio of B-specialists prices all assets in the B market. These two market-specific pricing models are linked over time by the cross-market arbitrage activities of the slow-moving generalists, who take steps to align the way that risk is priced by the two groups of specialists. And, much as before, the degree of market integration over time depends on the risks faced by cross-market arbitrageurs.

We illustrate the impact of a supply shock on two different markets that each contain multiple assets using an example. Here we use an annual calibration—i.e., one period corresponds to a year—and solve the model in the case where generalists reallocate their portfolios every $k = 2$ year and with $N = 2$ assets in both the A and B markets. The two assets in each market differ solely in their durations. The short-term bonds, denoted A_1 and B_1, have a duration of 5 years (i.e., $D_{A_1} = D_{B_1} = 5$). The long-term bonds, denoted A_2 and B_2, have a duration of 10 years (i.e., $D_{A_2} = D_{B_2} = 10$). As before, the two bonds in market A are default free whereas the two bonds in market B are exposed to default risk. We assume that B_1 and B_2 have the same exposure to the common default process, z_t, and that there is no idiosyncratic default risk.

Figure 4 shows the evolution of risk premia following an unexpected shock in year 10 which permanently increases the supply of long-term default-free bonds (A_2) by 3 units and reduces the supply of short-term default-free bonds (A_1) by the same amount. This scenario corresponds to a "reverse Operation Twist" in which the Federal Reserve sells long-term Treasuries and reinvests the proceeds in short-term Treasuries. Figure 4 shows that the risk premia for all four assets rise after impact. Naturally, this supply shock has a larger impact on the risk premium for long-maturity bonds in each market, leading both the A and B yield curves to steepen. These patterns are consistent with existing term structure models in which bond supply shocks impact bond yields because the demand for bonds is downward-sloping (e.g., Greenwood and Vayanos [2014] and Greenwood, Hanson, and Vayanos [2016]).

[26] The key condition is that generalists are unable or unwilling to use only the securities in market A and only securities in market B to construct factor-mimicking portfolios that isolate exposure to each of the underlying risk factors present in each market. If the generalists can do this in both markets—i.e., if both markets are complete, then there is a riskless arbitrage unless all agents charge the same prices of factor risk. The Internet Appendix spells out these conditions more formally.

While risk premia on all securities in each market move in lockstep, risk premia in the A and B markets do not move in lockstep. As in our simpler model, it takes time for the slow-moving generalists to integrate the A and B markets following the shock, leading the risk premia of the two bonds in market A to initially overreact and the risk premia of the two bonds in market B to initially underreact in Figure 4. Furthermore, market integration remains imperfect even in the long run. Specifically, even many period after the supply shock, Figure 4 shows that the yield curve in market A has steepened more than the yield curve in market B.

5 Application: Event studies and changes in the price of risk

In response to the rapidly evolving financial crisis and global recession in late 2008 and early 2009, central banks around the world announced their intentions to purchase substantial quantities of government bonds and other long-term debt securities. A central question in assessing the effectiveness of these large-scale asset purchase programs is whether they impacted assets prices beyond the bonds specifically targeted by central banks. Suppose, for example, that the impact of asset purchase programs was limited to asset classes in which the purchases were being made (Treasury bonds and mortgage-backed securities in the case of the Federal Reserve), perhaps because these markets are almost completely segmented from other financial markets. Such a finding should dampen central bankers' enthusiasm for these programs, and cast doubt on the idea that asset purchases could affect broader economic activity.

Our model provides a natural framework for understanding how these large-scale asset purchase programs should spill across financial markets over time. According to our model, the largest short-run effects of these quantitative easing (QE) programs should be felt in the asset classes that are directly targeted. In the long run, however, changes in risk premia should spill over to non-targeted markets. Differences between the short-run and long-run price impact should reflect the degree to which the programs were anticipated, the length of time between the announcement and implementation, the effective degree of segmentation between different markets, and the speed at which capital flows between markets.

Most empirical studies of these asset purchase programs have used an event study methodology, focusing on the one day or intraday impact on bond yields following announcements of future asset purchases. In one of the first of these event studies, Gagnon, Raskin, Remache, and Sack (2011) report the cumulative interest rate changes on a set of dates between November 2008 and January 2010 when the Federal Reserve made major announcement dates associated with its first round of quantitative easing (QE1). Looking at the two asset classes that the Fed was directly targeting, they report a

91 basis points decline in 10-year Treasury yields and a 113 basis points decline in mortgage-backed security yields. However, the find that Baa-rated corporate bond yields, which the Fed was not directly targeting, declined by only 67 basis points on these same days. Swanson (2017) also finds that the Fed's asset purchase announcements had a larger impact on Treasury yields than on Baa corporate bond yields.

Krishnamurthy and Vissing-Jorgensen (2011) extend this analysis to the Fed's second round of quantitative easing (QE2) and also discuss the impact on other assets, including high yield corporate bonds. After controlling for other factors, Krishnamurthy and Vissing-Jorgensen conclude that the effects of asset purchases were most pronounced among the assets being purchased (MBS and Treasuries in QE1 and Treasuries in QE2), suggesting a high degree of segmentation between different fixed income markets.

At the same time, some researchers have recognized that short-horizon announcement returns may not capture the full impact of these asset purchase programs. In their empirical assessment of the Bank of England's quantitative easing program, Joyce, Lasoasa, Stevens and Tong (2011) suggest that these programs may have gradually impacted corporate bonds and equities. Fratzcher, Lo Duca, and Straub (2013) suggest that the Fed's asset purchase programs triggered portfolio flows that ultimately impacted emerging market asset prices and foreign exchange rates.

Researchers have used different approaches to measure the long-run effects of large-scale asset purchases. Joyce, Lasoasa, Stevens and Tong (2011) report the cumulative change in asset prices for the longer period between March 4, 2009 and May 31, 2010 in addition to 1-day announcement returns. They show that corporate bond yields fall by a cumulative 70 basis points around the Bank of England's asset purchase announcements, but by 400 basis points over the longer period. Mamaysky (2014) takes a more tailored approach to each asset market, choosing an announcement window that maximizes the statistical power of measured returns. Using this approach, he shows that the impact of quantitative easing on both equity and high yield bond markets is much larger after 15 days than what one would measure using a 1-day window. But even this approach may significantly understate the long-run effect, because, as we have noted, it may take several quarters or even years for the full impact of a large supply shocks to register.

Our model clarifies the broader issue at stake: event studies are a useful methodology for detecting short-run price changes, but often lack the statistical power to detect changes in risk premia occurring at longer horizons. Indeed, the event study methodology was originally developed in the 1970s to tackle questions of *informational efficiency* of stock prices, not changes in risk premia, but has increasingly been used in other settings, such as in event studies assessing quantitative easing.

The limitations of event studies are particularly severe when there is noise from "cash flow" news.

For instance, consider the effects of σ_z—the volatility of fundamental cash flow shocks in market B—on our ability to detect the impact on prices in market B stemming from a supply shock that hits market A . When σ_z is large relative to σ_r, there is insufficient power to detect changes in risk premia in market B. The statistical power would increase with the number of events, but power is nonetheless decreasing in σ_z. Thus, our model suggests that short-run event studies may have a hard time detecting spillover effects on markets, such as equities and high yield bonds, where there can be significant confounding news. More generally, our framework suggests that event studies are an inappropriate methodology for measuring cross-market price impact, especially at long horizons.

In summary, our model suggests that researchers should be cautious in using event studies to assess the long-run impact of large supply shocks on market prices and risk premia. Measuring the long-run impact of supply shocks across markets is inherently difficult because the full economic impact may occur over such a long time that it is swamped by other factors.

6 Conclusion

We have studied how security prices react when there is a large shock to the supply of one asset class, such as when central banks embark on a large-scale purchase of government bonds. Our model incorporates two key asset pricing frictions: partial segmentation between markets and slow-moving capital. The presence of these two frictions means that when one market is hit with a supply shock, prices of risk in that market become disconnected from prices in other related markets. In the longer run, however, there is convergence in the pricing of risk as capital flows across markets. In summary, it takes time to digest large supply shocks and for the impact to fully spill over to related markets.

Our findings sound a note of caution for researchers who seek to measure the impact of quantitative easing policies on asset prices, a topic that has attracted considerable attention in recent years. We have shown that the standard technique for measuring price impact—through an event study of short-run price responses to central bank announcements—is poorly suited to measuring long-run changes in risk premia across markets. While prices in neighboring markets will often change at short horizons, these near-term changes are likely to reveal more about the short-run capacity of specialists to accommodate asset flows than about the ultimate long-term changes in risk premia.

References

Acharya, V., H. S. Shin, and T. Yorulmazer, 2013, "Fire-sale FDI," Korean Economic Review 27, 163-202.

Albagli, E., 2015, "Investment Horizons and Asset Prices under Asymmetric Information," Journal of Economic Theory 158,787-837.

Bachetta, P., and E. van Wincoop, 2010, "Infrequent Portfolio Decisions: A Solution to the Forward Discount Puzzle," American Economic Review 100, 870-904.

Bacchetta, P., van Wincoop, E., 2017, "Gradual Portfolio Adjustment: Implications for Global Equity Portfolios and Returns", NBER Working Paper 23363.

Banerjee, S., 2011, "Learning from Prices and the Dispersion in Beliefs," The Review of Financial Studies 24, 3025-3068.

Bogousslavsky, V., 2016, "Infrequent Rebalancing, Return Autocorrelation, and Seasonality," Journal of Finance 71, 2967-3006.

Campbell, J. Y., 1991, "A Variance Decomposition for Stock Returns", Economic Journal 101:157–179.

Campbell, J. Y., A. W. Lo, and A. C. MacKinlay, 1997, The Econometrics of Financial Markets, Princeton, NJ: Princeton University Press.

Campbell, J. Y. and R. J. Shiller, 1988, "Stock Prices, Earnings, and Expected Dividends,", Journal of Finance 43, 661-76.

Campbell, J. Y. and L. M. Viceira, 1999, "Consumption and Portfolio Decisions When Expected Returns are Time Varying," The Quarterly Journal of Economics 114, 433-495.

Campbell, J. Y. and L. M. Viceira, 2001, "Who Should Buy Long-Term Bonds?" The American Economic Review 91, 99-127.

Chien, Y., H. Cole, and H. Lustig, 2012, "Is the Volatility of the Market Price of Risk Due to Intermittent Portfolio Rebalancing?" American Economic Review 102, 2859-96.

DeLong, J. B., A. Shleifer, L. H. Summers, and R. J.Waldmann, 1990, "Noise Trader Risk in Financial Markets," Journal of Political Economy 98, 703-738.

Duffie, D., "Asset Price Dynamics with Slow-Moving Capital", Journal of Finance 2010, 65: 1238-1268.

Duffie, D., and B. Strulovici, 2012, "Capital Mobility and Asset Pricing", Econometrica 80: 2469-2509.

Duffie, D., and K. Singleton, 1999, "Modeling Term Structures of Defaultable Bonds," Review of Financial Studies 12: 687-720.

Errunza, V., and E. Losq, 1985, "International Asset Pricing under Mild Segmentation: Theory and Test" Journal of Finance 40, 105-124.

Fama, F., L. Fisher, M. C. Jensen, and R. Roll, 1969, "The Adjustment of Stock Prices to New Information," International Economic Review 10, 1-21.

Fischer, S., 1977, "Long-Term Contracts, Rational Expectations, and the Optimal Money Supply Rule," The Journal of Political Economy 85, 191-205.

Fratzscher, M., M. Lo Duca, and R. Straub, 2013, “On the International Spillovers of US Quantitative Easing,” European Central Bank Working Paper.

Gagnon, J., M. Raskin, J. Remache, and B. Sack, 2011, “The Financial Market Effects of the Federal Reserve’s Large-scale Asset Purchases,” International Journal of Central Banking 7 , 3–43.

Garleanu, N., L. H. Pedersen, and A.M. Poteshman, 2009, “Demand-Based Option Pricing,” The Review of Financial Studies 22, 4259-4299.

Greenwood, R., S. G. Hanson, and D. Vayanos, 2016, “Forward Guidance in the Yield Curve: Short Rates versus Bond Supply,” In Monetary Policy through Asset Markets: Lessons from Unconventional Measures and Implications for an Integrated World, edited by Albagli. E., D. Saravia, and M, Woodford, 11–62. Santiago: Banco Central de Chile.

Greenwood, R. and D. Vayanos, 2014, “Bond Supply and Excess Bond Returns,” Review of Financial Studies 27, 663-713.

Gromb, D. and D. Vayanos, 2002, “Equilibrium and Welfare in Markets with Financially Constrained Arbitrageurs,”Journal of Financial Economics 66, 361-407.

Grossman, Sanford J & Miller, Merton H, 1988. “Liquidity and Market Structure,” Journal of Finance 43, 617-37.

Hanson, Samuel G., 2014, “Mortgage Convexity,” Journal of Financial Economics 113(2), 270-299.

Hanson, S. G., and J. C. Stein, 2015, “Monetary Policy and Long-Term Real Rates,” Journal of Financial Economics 115, 429-448.

Joyce, M. A. S., A. Lasaosa, I. Stevens, and M. Tong, 2011, “The Financial Market Impact of Quantitative Easing in the United Kingdom,” International Journal of Central Banking 7, 113-161.

Gabaix, X., A. Krishnamurthy, and O. Vigneron, 2007, “Limits of Arbitrage: Theory and Evidence from the Mortgage-Backed Securities Market”, Journal of Finance, 62(2), 557-595.

Krishnamurthy, A. and A. Vissing-Jorgensen, 2011, “The Effects of Quantitative Easing on Interest Rates: Channels and Implications for Policy,” Brookings Papers on Economic Activity, Fall 2011, 215-265.

Krishnamurthy, A. and A. Vissing-Jorgensen, 2013, “The Ins and Outs of Large-Scale Asset Purchases,” Kansas City Federal Reserve Symposium on Global Dimensions of Unconventional Monetary Policy, 57-111.

Mamaysky, Harry, 2014, “The Time Horizon of Price Responses to Quantitative Easing”, Working Paper.

Merton, Robert C, 1987. “A Simple Model of Capital Market Equilibrium with Incomplete Information”, Journal of Finance 42, 483-510.

Oehmke, M., 2009, “Gradual Arbitrage,” Working paper.

Mitchell, M., L.H. Pedersen, and T. Pulvino, 2007. “Slow Moving Capital,” American Economic Review 97, 215-220.

Samuelson, P. A., 1947, Foundations of Economic Analysis, Cambridge, MA: Harvard University Press.

Shleifer, A., and R. W. Vishny, 1997. "The Limits of Arbitrage," Journal of Finance 52, 35-55.

Spiegel, M., 1998, "Stock Price Volatility in a Multiple Security Overlapping Generations Model," Review of Financial Studies 11, 419-447.

Stapleton, R.C., and M. G.. Subrahmanyam, 1977, "Market Imperfections, Capital Market Equilibrium and Corporation Finance," Journal of Finance 32, 307-319.

Stein, J. C., 2005, "Why Are Most Funds Open-End? Competition and the Limits of Arbitrage," Quarterly Journal of Economics 120, 247-272.

Taylor, J., 1979, "Staggered Wage Setting in a Macro Model," American Economic Review 69, 108-13.

Watanabe, M., 2008, "Price Volatility and Investor Behavior in an Overlapping Generations Model with Information Asymmetry," The Journal of Finance 63, 229–272.

Woodford, M., 2012, "Methods of Policy Accommodation at the Interest-Rate Lower Bound," in The Changing Policy Landscape, Federal Reserve Bank of Kansas City.

Vayanos, D. and J. Vila, 2009, "A Preferred-Habitat Model of the Term Structure of Interest Rates," NBER Working Paper No. 15487.

Figure 1: Impact of an unanticipated shock to the supply of A bonds. This figure shows the impact of an unanticipated shock that doubles the supply of A bonds in quarter 5. The model parameters for this illustration are shown in Table 1. Panel A shows the evolution of annual bond risk premia in market A, $E_t[rx_{A,t+1}]$, and market B, $E_t[rx_{B,t+1}]$, over time. Panel B shows the evolution of specialists holdings in markets A and B ($b_{A,t}$ and $b_{B,t}$) as well as the positions of active generalists ($d_{A,t}$ and $d_{B,t}$). Panel C shows the evolution of the "active supplies" of A and B bonds. The active supply of A is $s_{A,t} - (1 - q_A - q_B)k^{-1} \sum_{i=1}^{k-1} d_{A,t-i}$ and the active supply of B is defined analogously. Panel D shows the evolution of bond yields in market A, $y_{A,t}$, and market B, $y_{B,t}$, over time.

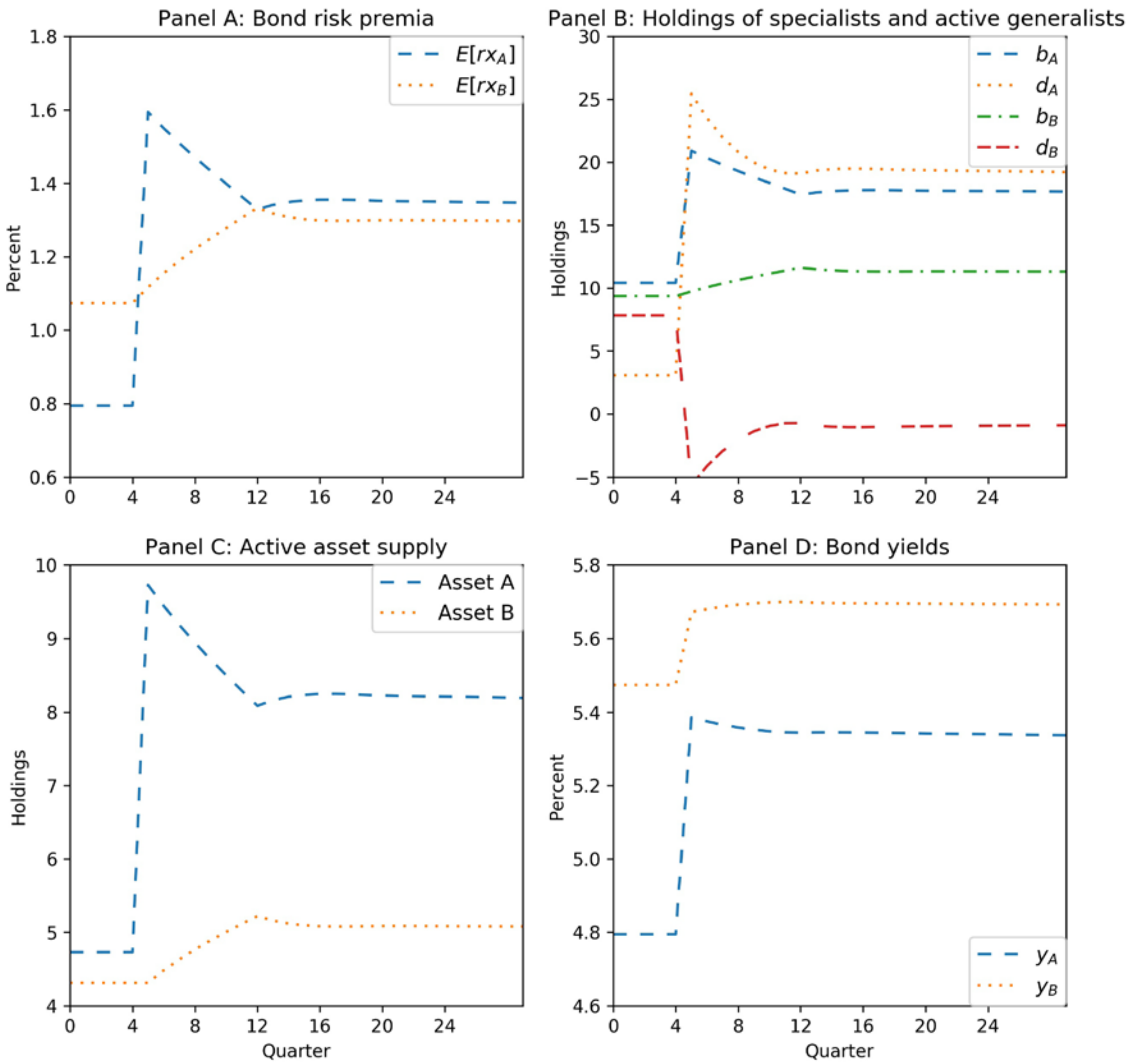

Figure 2: Unpacking the two key asset-pricing frictions. This figure shows the response of annual bond risk premia in market A, $E_t[rx_{A,t+1}]$, and market B, $E_t[rx_{B,t+1}]$, to an unanticipated shock that doubles the supply of A bonds in quarter 5. Each panel corresponds to a different configuration of the two key asset pricing frictions in our model. Panel A shows the response in benchmark case with no frictions. This is a special case of our model where $q_A = q_B = 0$ and $k = 1$. Panel B shows the results when there is only partial segmentation. This is a special case of our model where $q_A > 0$, $q_B > 0$, $1 - q_A - q_B > 0$, and $k = 1$. Panel C shows the response with only slow-moving capital. This case corresponds to a multi-asset version of Duffie's (2010) model and is not a special case of our model. We set the fraction of fast-moving investors to $q_A + q_B$. Finally, Panel D shows the response in our general model featuring both partial segmentation ($q_A > 0$, $q_B > 0$, $1 - q_A - q_B > 0$) and slow-moving capital ($k > 1$). The other model parameters for this illustration are shown in Table 1.

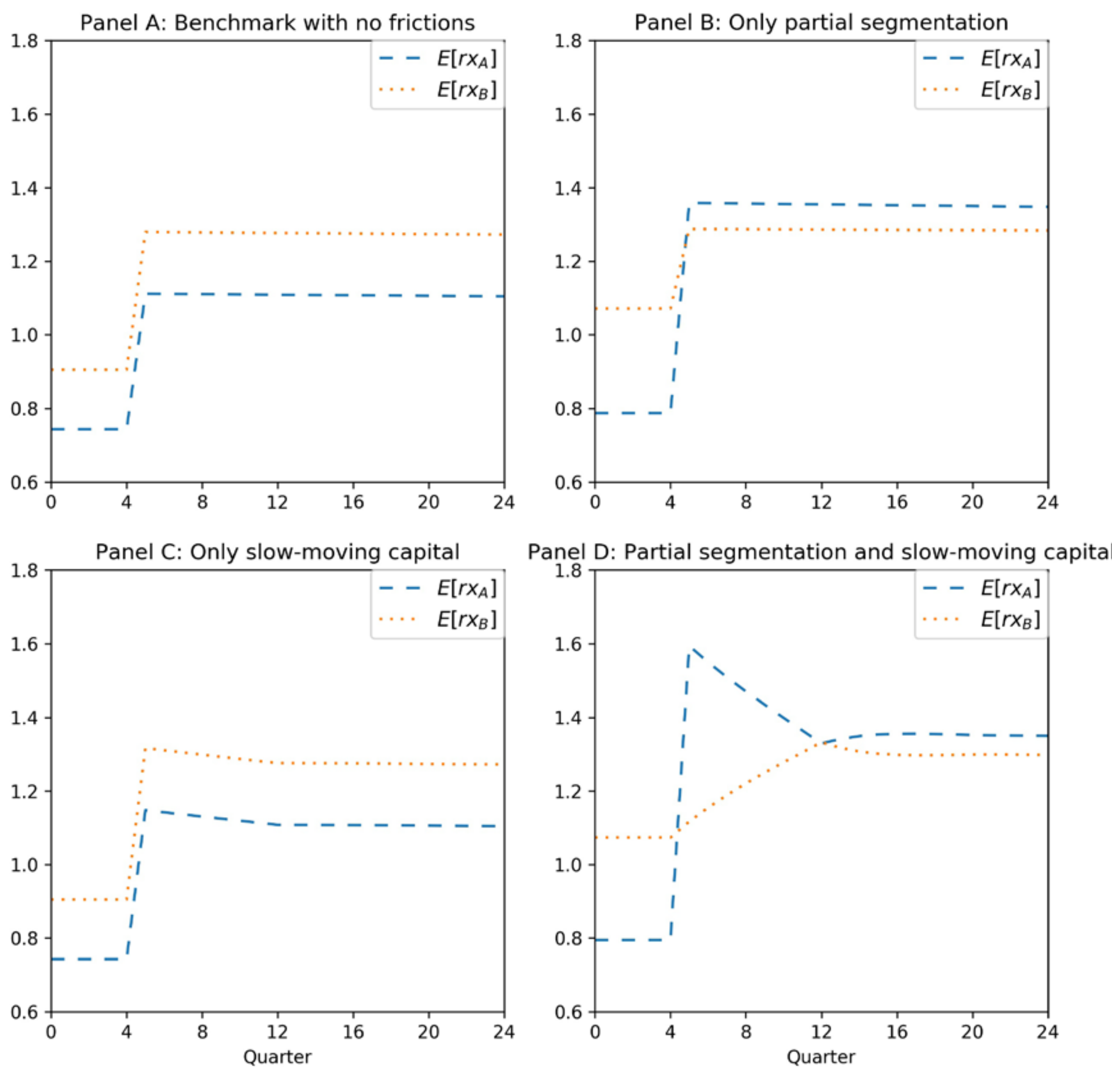

Figure 3: Impact of a pre-announced increase in the supply of A bonds. This figure shows the impact of pre-announced supply increase: there is an announcement in quarter 5 that the supply of A bonds will double in quarter 8. The model parameters for this illustration are shown in Table 1. Panel A shows the evolution of annual bond risk premia in market A, $E_t[rx_{A,t+1}]$, and market B, $E_t[rx_{B,t+1}]$, over time. Panel B shows the evolution of specialists holdings of A and B bonds ($b_{A,t}$ and $b_{B,t}$) as well as the holdings of active generalists ($d_{A,t}$ and $d_{B,t}$). Panel C shows the evolution of the "active supplies" of A and B bonds. The active supply of A is $s_{A,t} - (1 - q_A - q_B)k^{-1} \sum_{i=1}^{k-1} d_{A,t-i}$ and the active supply of B is defined analogously. Panel D shows the evolution of bond yields in market A, $y_{A,t}$, and market B, $y_{B,t}$, over time.

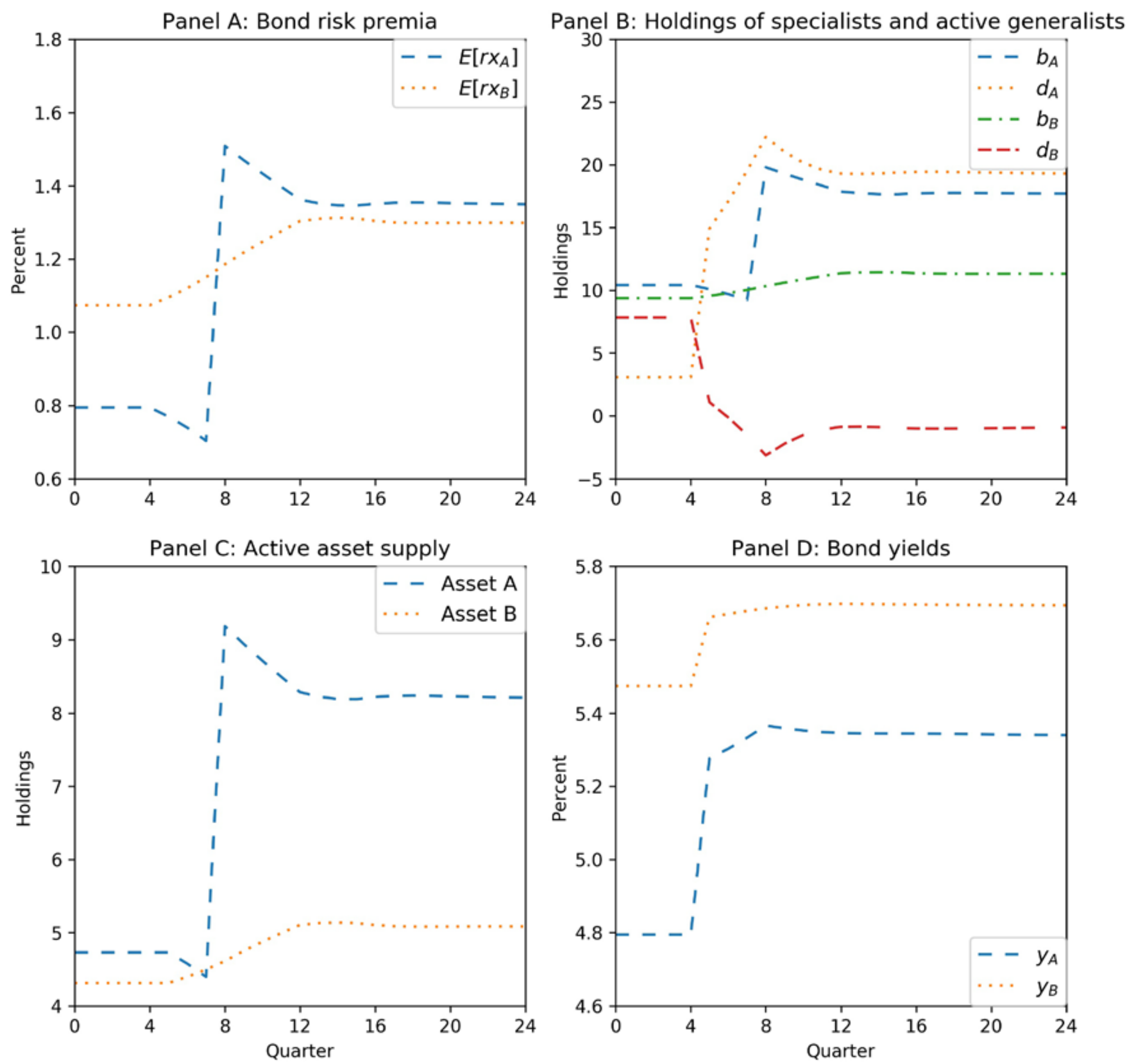

Figure 4: Price impact with multiple securities in each market. This figure shows the impact on bond risk premia of an unexpected supply shock that permanently increases the supply of long-term default-free bond (A_2) by 3 units and decreases the supply of short-term default-free bond (A_1) by the same amount in year 10. In Figure 4, one period corresponds to a year. We choose the annual persistence parameters for short rates, default losses, and bond supply to be consistent with the quarterly persistence assumed in Table 1. The volatility of annual shocks are then chosen so that the unconditional volatility of each process is the same in the annual and quarterly calibrations. We assume $\tau = 2, k = 2$, and $q_A = q_B = 40\%$ in this figure. Panel A shows the evolution of risk premia for short-term securities in each market (A_1 and B_1). Panel B shows the evolution of risk premia for long-term securities in each market (A_2 and B_2).

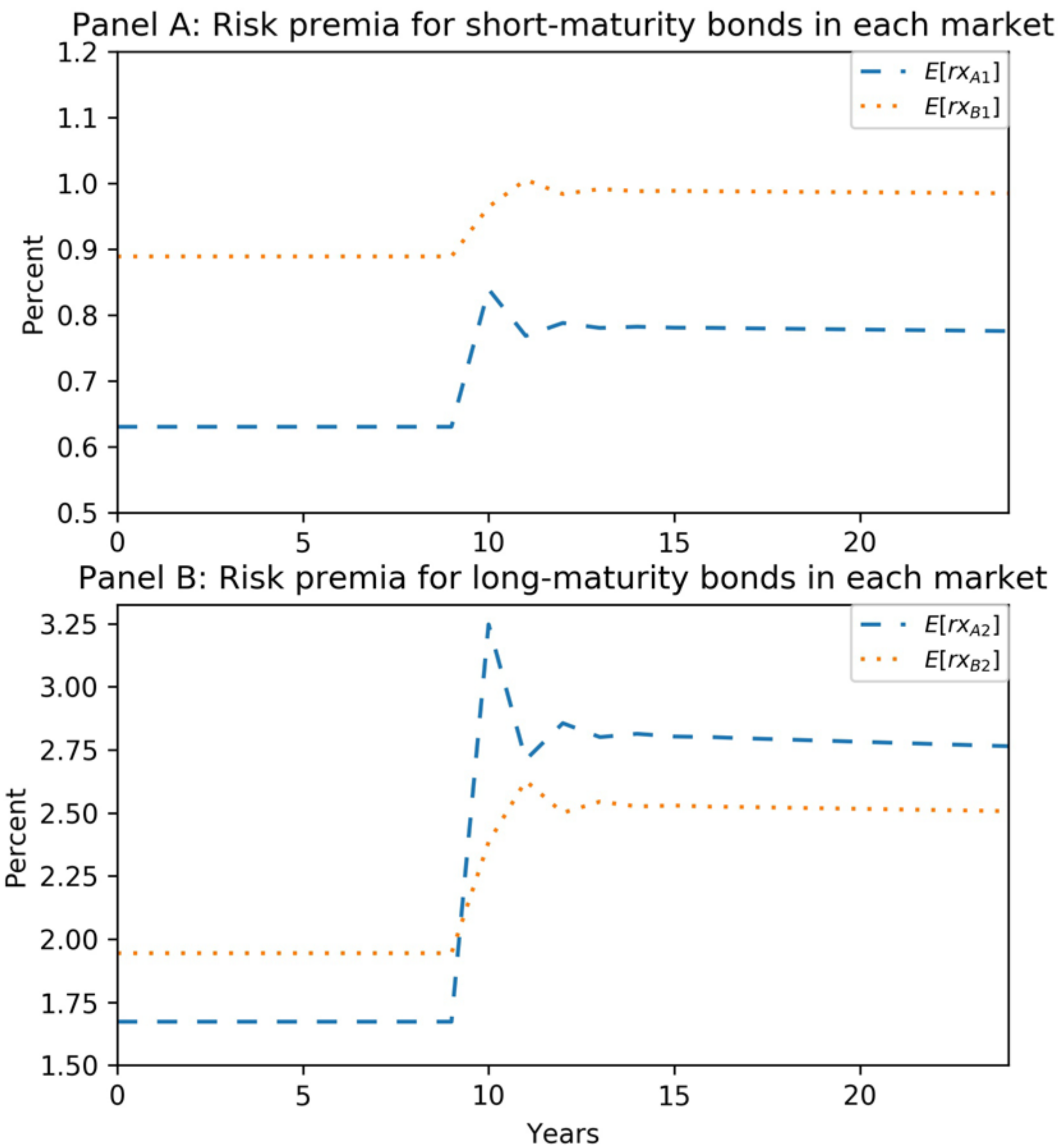

Table 1: Illustrative model parameters for numerical exercises. This table presents the illustrative model parameters that we use throughout our numerical exercises. One period corresponds to one quarter of a year. We report annualized values for the mean and standard deviation of shocks to both the short rate and default losses on asset B.

Parameter	Description	Value
q_A, q_B	Percentage of investors that are specialists in A and B	45%
k	Number of quarters between generalist portfolio rebalancing	8
$\bar{r}$	Average annualized short-term riskless rate	4%
σ_r	Volatility of quarterly shocks to annualized short-term riskless rate	0.64%
ρ_r	Quarterly persistence of short-term riskless rate	0.96
$\bar{z}$	Expected annualized default losses on asset B	0.4%
σ_z	Volatility of quarterly shocks to annualized default losses on asset B	0.4%
ρ_z	Quarterly persistence of default losses on asset B	0.96
$\bar{s}_A, \bar{s}_B$	Average asset supplies	5
$\sigma_{s_A}, \sigma_{s_B}$	Volatility of quarterly supply shocks	1
ρ_{s_A}, ρ_{s_B}	Quarterly persistence of supply shocks	0.999
D_A, D_B	Macaulay duration in years (implies $\theta_A = \theta_B$=0.95)	5 years
τ	Aggregate investor risk tolerance	1.75

Table 2: Model comparative statics. This table shows how the price impact of the same supply shock—an unanticipated shock that doubles the supply of asset A—varies as we change key model parameters one at a time. All other parameters are held constant at the values listed in Table 1. For a given set of model parameters, we summarize the impact of the supply shock on both the A and B markets by listing (i) the yields and expected annual returns in the period before the shock arrives (labeled as "pre-shock level"), (ii) the changes in yields and expected annual returns in the period when the shock arrives (labeled as "short-run Δ"), and (iii) in $2k$ periods after the shock arrives (labeled as "long-run Δ"). Finally, we report the degree to which bond yields over- or underreact as the difference between the short-run change and the long-run change, expressed as a percentage of the long-run change

$$\%\textit{Short-Run-Over-React}(y) = \frac{(y_t - y_{t-1}) - (y_{t+2k} - y_{t-1})}{(y_{t+2k} - y_{t-1})}.$$

Our measure of over-reaction for risk premia is defined analogously.

		Market A								Market B							
		Risk premia, $E_t[rx_{A,t+1}]$				Yields, $y_{A,t}$				Risk premia, $E_t[rx_{B,t+1}]$				Yields, $y_{B,t}$			
		Pre shock level	Short run Δ	Long run Δ	Over-react	Pre shock level	Short run Δ	Long run Δ	Over-react	Pre shock level	Short run Δ	Long run Δ	Over-react	Pre shock level	Short run Δ	Long run Δ	Over-react
Supply shock hits market A																	
(1) Baseline example in Figure 1		0.79	0.80	0.56	42%	4.79	0.59	0.55	7%	1.07	0.04	0.23	-83%	5.47	0.20	0.22	-12%
(2) More risk tolerant	$\tau = 200$	0.67	0.68	0.47	43%	4.67	0.50	0.46	7%	0.90	0.04	0.19	-83%	5.30	0.17	0.19	-12%
(3) No generalists	$q_A = q_B = 0.5$	0.81	0.81	0.80	0%	4.81	0.80	0.79	0%	1.33	0.00	0.00	N/A	5.73	0.00	0.00	N/A
(4) More B specialists	$q_A = 0.3$, $q_B = 0.6$	1.01	1.28	0.75	68%	5.01	0.83	0.74	10%	0.92	0.05	0.24	-82%	5.32	0.21	0.23	-11%
(5) More A specialists	$q_A = 0.6$, $q_B = 0.3$	0.71	0.59	0.46	27%	4.71	0.48	0.45	4%	1.51	0.05	0.24	-82%	5.91	0.21	0.23	-12%
(6) Slower-adjusting generalists	$k = 12$	0.80	0.82	0.55	46%	4.80	0.61	0.54	10%	1.08	0.03	0.23	-89%	5.48	0.19	0.23	-18%
(7) Faster-adjusting generalists	$k = 4$	0.79	0.76	0.56	33%	4.79	0.57	0.55	3%	1.07	0.08	0.22	-66%	5.47	0.21	0.22	-5%
(8) Many generalists	$q_A = 0.2$, $q_B = 0.2$, $k = 4$	0.77	1.12	0.41	174%	4.77	0.44	0.39	12%	0.94	0.42	0.35	18%	5.34	0.38	0.36	4%
(9) Greater default risk for B	$\sigma_Z = 0.64\%$	0.83	0.83	0.63	30%	4.83	0.66	0.62	5%	1.85	0.03	0.19	-85%	6.25	0.16	0.18	-13%