A Long and a Short Leg Make For a Wobbly Equilibrium

Nicolae Gârleanu*[1], Stavros Panageas†[2], and Geoffery Zheng‡[3]

[1]WashU, Olin Business School and NBER
[2]UCLA, Anderson School of Management and NBER
[3]NYU Shanghai

March 2023

Abstract

The interaction between the spot and lending markets for stocks can lead to abrupt changes in short selling. Furthermore, rational short sellers may choose to abandon the market even as mispricing widens. The model can help explain the fat-tailed dynamics of short selling in the data, and further provides conditions identifying stocks that are more likely to experience large and abrupt changes in short selling. We verify these predictions empirically. We also apply the theory to understand curious patterns in the behavior of short sellers during one of the historically worst periods for short selling, November 2020–January 2021.

Keywords: Asset Pricing with Frictions, Short Selling, Runs, Limits to Arbitrage

JEL Classification: G11, G12, G14

*garleanu@wustl.edu

†stavros.panageas@anderson.ucla.edu

‡geoff.zheng@nyu.edu. We would like to thank Itamar Dreschler, Sergei Glebkin, Dan Greenwald, Paymon Khorrami, Alberto Teguia, Adrien d'Avernas, and seminar participants at the BI-ShoF Conference, the NBER Summer Institute Asset Pricing Program, the 2022 SFS Cavalcade, the 2022 WFA, the 16th Annual Cowles Conference on General Equilibrium, Boston University Questrom School of Business, BYU Marriott School of Business, CKGSB, Duke Fuqua School of Business, Florida International University, Oxford SAID Business school, Warwick business school, UW Foster School of Business, Shanghai Advanced Institute of Finance, and UNC Kenan-Flagler Business School for their useful comments on the paper.

Short interest in individual stocks is unstable, exhibiting sudden and large changes. We propose a theoretical explanation for this instability, relying on a feedback loop between the spot and the lending markets. Our theory not only addresses the occasional sudden retreat of short sellers, but also accounts for a simultaneous increase in prices; in other words, our theory can help explain why sellers abandon their short positions despite an increased profitability of shorting.

The novel aspect of our theory is that it does not rely on portfolio constraints, limitations to arbitrage capital, agency issues, etc. Instead, our mechanism is built on a detailed modeling of stock-lending income and its implications for spot-market clearing. We show that lending income gives rise to a feedback mechanism between a stock's expected return and short interest that can generate a "backward-bending demand," and accordingly sudden equilibrium shifts in short interest and expected returns.

We document that large and sudden changes in shorting activity are a broad phenomenon. Viewed as a time series, the short interest process of many stocks exhibits jump-like features. These features appear linked to the backward-bending demand channel that we highlight: the stocks that satisfy an empirically verifiable condition for a backward-bending demand curve are also the ones that show the highest incidence of large and sudden changes in short interest.

We next outline the ingredients of our model and summarize the basic intuitions and the supporting empirical evidence.

The model features investors with heterogeneous beliefs about the expected return of a positive-supply risky stock: one group is optimistic, while the other holds rational beliefs.[1] This difference of opinion between investors prompts them to trade with each other, with the rational investors having an incentive to short the stock whenever the expected excess return becomes negative. Shorting stock requires borrowing it, for a fee determined endogenously in the lending market as a result of bargaining.

As is recognized, the presence of lending fees modifies the returns experienced by both long and short investors. The equilibrium risk compensation (the ratio of excess return to

1. Motivated by the empirical fact that stocks with high short interest tend to have low subsequent returns, we assume that the comparatively pessimistic investors are actually rational, but this is not an essential assumption for our results.

volatility, or "Sharpe ratio") is impacted both by the magnitude of the lending fee and the fraction of a representative lender's shares that are shorted. Following common terminology, we refer to the ratio of shorted-to-lendable shares as the "utilization" ratio.

All else equal, a higher utilization ratio acts as an increased subsidy for long positions, since a larger fraction of a representative long position is lent out to short sellers. This basic property of the model is responsible for equilibrium multiplicity, since the increased incentive to purchase the stock ends up reducing its Sharpe ratio and consequently inducing more shorting, thus supporting the high utilization ratio as an equilibrium outcome. Symmetrically, another equilibrium may exist with a lower utilization ratio, thus lower subsidy for long positions, and consequently a higher Sharpe ratio and smaller short positions.

The discussion in the above paragraph takes the wealth share of short sellers at a given point in time as fixed. One advantage of our dynamic setup is that we can study the evolution of the wealth shares depending on whether investors coordinate on a high or a low shorting equilibrium. We show that the wealth growth of short sellers is higher in the high shorting equilibrium than in the low-shorting equilibrium. An implication is that the (stochastic) steady state fraction of wealth controlled by short sellers is lower in the equilibrium with low shorting. Since the Sharpe ratio is increasing in the wealth share of these investors, the *steady state* Sharpe ratio may well be lower if investors coordinate on the low-shorting equilibrium than if they coordinate on the high shorting equilibrium.

To fully explore the implications of this dynamic effect in a more realistic setup, we extend the model to allow for multiple stocks with endogenous participation. After showing that our main conclusions from the single-stock economy extend to the multi-stock economy, we focus on the case of a large and a small stock, with disagreement affecting only the small stock.[2] We assume that only a small fraction of investors pay attention to the small stock and incur a small participation cost in doing so. In this extended version of the model, we show that the shift to a low shorting equilibrium causes rational investors to exit the market for the small stock, since remaining in a market without a trading opportunity is not worth

2. With this assumption, the interest rate becomes essentially fixed and therefore fluctuations in the Sharpe ratio are mirrored in the price-dividend ratio of the small stock. By contrast, in the baseline model the assumption of log utility and i.i.d. dividend growth imply that fluctuations in the Sharpe ratio are exactly offset by fluctuations in the interest rate, leaving the price-dividend ratio unaffected.

paying that participation cost. The exit of rational short-sellers causes a simultaneous rise in the price of the stock — consistent with the empirical observation that bad returns of shorting strategies coincide with drops in short interest.[3]

While the main focus of the paper is theoretical, our theory may help explain some salient time-series properties of the utilization ratio. According to the model, short-run changes in the utilization ratio should be small and (locally) normally distributed most of the time; but when there are changes in equilibrium, this ratio should exhibit jumps. Therefore, the distribution of the changes in the utilization ratio should be fat-tailed. Empirically, the utilization-ratio changes are remarkably fat tailed. The kurtosis of the residuals of an AR1 model fitted to weekly utilization data is 78.

The model also helps predict which stocks are most likely to exhibit jumps in shorting activity. As part of Proposition 4, we identify a necessary and sufficient condition for the incidence of jumps. This condition involves observable quantities, namely lending fees and utilization data. Empirically, we confirm that the stocks that satisfy this condition are the ones that are the most likely to exhibit jumps in utilization.

The paper concludes with a "case study" that illustrates the fickle behavior of short sellers. We document that the period between November 2020 and January 2021 saw an abrupt decline in short interest across hundreds of highly shorted stocks, and was also the worst period for a "betting against the short sellers" strategy, i.e., a strategy that goes long the top decile of most shorted stocks and shorts the market portfolio.

It is tempting to attribute this episode to the highly mediatized events involving the company GameStop, which saw online-forum-coordinated retail purchases resulting in a short squeeze of its stock. However, the broad based short-seller retreat that we focus on started eight weeks before the GameStop episode and impacted stocks that were not particularly discussed online by retail traders and did not experience an appreciable change in retail purchase volume.

There are three aspects of this episode that are pertinent for our model. First, the episode helps illustrate how abruptly and dramatically short selling can decline. Second, the retreat

3. This was the case, for example, during the period November 2020–January 2021, which we discuss below.

of the short sellers coincides with a rise in prices. As we highlighted above, the change in market composition induced by an equilibrium shift is central to our explanation for why short sellers are repelled, rather than attracted, by rising prices.[4] Finally, the retreat of the short sellers predated the spike in online discussion. One possible explanation for this retreat was the fear of an impending change in retail-investor behavior. Absent the backward-bending demand feature of our model, however, an impending rise in irrationality would raise the profitability of short selling and increase short interest, which is the opposite of what happened in the data.

The paper is organized as follows. After a brief literature review, Section 1 lays out the baseline version of the model and Section 2 presents the main analytical results. Section 3 discusses the dynamics of the investor wealth shares. Section 4 generalizes the results of Section 2 and provides necessary and sufficient conditions for equilibrium multiplicity. Section 5 presents the extension to multiple stocks and Section 6 discusses the model's empirical implications. Section 7 concludes. Proofs, detailed descriptions of the data, and additional results are contained in the appendix.

Related Literature

Our work relates to several strands of the asset-pricing literature. The most closely related one considers the joint determination of lending fees, short interest, and returns. In particular, D'Avolio (2002), Duffie, Gârleanu, and Pedersen (2002), Vayanos and Weill (2008), Banerjee and Graveline (2013), Evgeniou, Hugonnier, and Prieto (2022), and Atmaz, Basak, and Ruan (2020) consider explicit frictions to lending and borrowing shares, which translate into non-zero lending fees that in turn impact expected returns.[5] Similar to D'Avolio (2002),[6] Banerjee and Graveline (2013), and Atmaz, Basak, and Ruan (2020), the lending and spot markets clear simultaneously in our paper, but we use a different micro-foundation

4. In a static model, lower short seller demand would only be consistent with a lower price / higher expected return.

5. Such frictions also motivated the empirical studies of Geczy, Musto, and Reed (2002), Lamont (2012), Jones and Lamont (2002), Kaplan, Moskowitz, and Sensoy (2013), Porras Prado, Saffi, and Sturgess (2016), and Asquith, Pathak, and Ritter (2005) among others.

6. More precisely, to a working-paper version of this study, which contains a theoretical model that did not appear in the published article.

to obtain a positive lending fee. Specifically, we don't impose any hard constraint on the shares that a long investor can lend.[7] Instead, we obtain a positive lending fee by assuming that the process of matching share lenders and borrowers is a time-consuming activity, which requires compensation, similar in spirit to Duffie, Gârleanu, and Pedersen (2002). By taking that route, our model allows for a more general specification of the supply curve of lendable shares, which is not confined to being vertical.[8] This specification of the supply curve for lendable shares leads to a feedback loop between the Sharpe ratio and short interest that is not present in the aforementioned papers (which feature unique equilibria). In addition, our model allows us to explore the dynamic effects of an equilibrium shift, driven by the endogenous fluctuations in the wealth shares of the different types of agents.[9]

An even larger number of papers assume that shorting is prohibited and analyze implications for returns. Prominent examples here include Harrison and Kreps (1978), Miller (1977), Diamond and Verrecchia (1987), Detemple and Murthy (1997), Hong and Stein (2003), and Scheinkman and Xiong (2003). As in Harrison and Kreps (1978) and Miller (1977), we model the motive for trade in our paper in the convenient form of (dogmatic) differences of opinions among agents.

A large body of work studies the empirical relation between short interest and stock returns. Seneca (1967), Senchack and Starks (1993), Desai et al. (2002), Diether, Lee, and Werner (2009), Asquith, Pathak, and Ritter (2005), Blocher, Reed, and Van Wesep (2013), Beneish, Lee, and Nichols (2015), and Dechow et al. (2001) study the cross-sectional relation and find that stocks with higher short interest under-perform those with lower short interest. Cohen, Diether, and Malloy (2007) and Boehmer, Jones, and Zhang (2008) use proprietary data on quantities lent as well as shorting fees and find consistent results. Duong et al. (2017)

7. Evgeniou, Hugonnier, and Prieto (2022) also does not impose a hard constraint on the quantity of lendable shares. Instead, it assumes that the supply of lendable shares is adjusted by a monopolistic entity to maximize lending revenue. In our paper, investors face search frictions in the lending market that make it costly to locate lendable shares.

8. An exception is Atmaz, Basak, and Ruan (2020). In their model, individual agents' supply curves are vertical, but the aggregate supply curve has finite elasticity due to composition effects when aggregating across agents.

9. The fact that shorting requires borrowing shares and is subject to natural collateral requirements has several interesting general equilibrium implications, as explored by Fostel and Geanakoplos (2008), Simsek (2013), and Biais, Hombert, and Weill (2021). In contrast, our model focuses on the general equilibrium implications of the associated lending fees.

studies the empirical relation between lending fees and stock returns and finds that high lending fees predict lower future returns. Drechsler and Drechsler (2014b) documents that asset pricing anomalies concentrate in stocks with high shorting fees. Lamont and Stein (2004) studies the information content in aggregate short interest and finds that short interest declined as stock market valuations rose in the late 90's. Rapach, Ringgenberg, and Zhou (2016) shows that the predictive power of aggregate short interest stems predominantly from a cash-flow channel.

Our paper also relates to a sizable theoretical literature analyzing multiple equilibria in asset pricing and macroeconomics. Multiple equilibria can arise through a number of mechanisms, chief among them a) bubbles (or money) in OLG economies, b) increasing returns to scale and production externalities, and c) portfolio constraints.[10] The mechanism that gives rise to multiple equilibria in our paper is different, since it relies on the interaction between the lending and the spot markets. We also note in this context that, while Vayanos and Weill (2008) features multiple equilibria in the presence of shorting frictions and fees, the multiplicity of equilibria pertains to agents' choice of market to join, which renders one asset more liquid (that is, easier to find) and thus increases its attractiveness to future entrants. In addition, in our setup the spot market is not a search market, but is Walrasian.[11]

Finally, several recent papers target specifically the set of events involving GameStop. See, for instance, Pedersen (2022) and Allen et al. (2021).

1 Model

1.1 Agents: life-cycle and preferences

Time is continuous and infinite for tractability. To obtain a stationary wealth distribution, we follow Gârleanu and Panageas (2015) and assume that investors continuously arrive

10. We refer the reader to the survey by Benhabib and Farmer (1999), which lists and discusses the different mechanisms that lead to multiple equilibria and indeterminacies. Recent examples of papers using multiple-equilibrium models in asset pricing include Gârleanu and Panageas (2021), Khorrami and Zentefis (2020), Khorrami and Mendo (2021), Zentefis (2018), and Farmer and Bouchaud (2020).

11. Coordination issues are central in economies admitting multiple equilibria, but can also be of first-order importance in unique-equilbrium settings, as highlighted by Abreu and Brunnermeier (2002) in a model featuring binding portfolio constraints and a non-Walrasian price protocol.

("births") and depart ("deaths") from the economy. Per unit of time a mass π of investors arrives, and a mass π departs. Therefore, the population of agents born at time $s \leq t$ and still remaining at time t is $\pi e^{-\pi(t-s)}$, while the total population is constant and equal to $\int_{-\infty}^{t} \pi e^{-\pi(t-s)} ds = 1$. "Births" and "deaths" should be understood as arrivals and departures of market participants, a point that will become clearer in Section 5, where we introduce multiple stocks.

To introduce trade in equities, we assume that investors have heterogeneous beliefs. For simplicity, a fraction $\nu \in (0,1)$ of investors perceive the correct data-generating process. We refer to them as rational investors ("R" investors). The remaining fraction are overly optimistic (we model this optimism shortly), and we refer to these investors as "I" investors.

For tractability, both investors have logarithmic utilities and their expected discounted utility from consumption is

$$V_t^i \equiv E_t^i \int_t^\infty e^{-(\rho+\pi)(u-t)} \log\left(c_{u,t}^i\right) du \tag{1}$$

for $i \in \{I, R\}$, with ρ a discount factor and $c_{u,t}^i$ the time-u consumption of an agent of type i born at time $t \leq u$. The notation E_t^i reflects the different investor beliefs. Because of death, the effective discount rate is $\rho + \pi$.

Before proceeding, we note that, while we require heterogeneous beliefs to introduce a motivation for trading, the assumption that one group has correct beliefs helps mostly to save notation and can be easily relaxed. The same applies to the assumption that there are only two groups of investors, which can be relaxed to allow for multiple investor types, including a continuum (Section 4). Similarly, the overlapping-generations structure is just a technical device to ensure that no investor type disappears in the long run.[12] Finally, in setting up the model we make the (conventional) assumption that agents maximize over both their consumption and portfolio choices, which we introduce shortly. Our model is, however, equivalent to one in which agents delegate their portfolio decisions to professional managers, and managers maximize their clients' expected portfolio (logarithmic) growth according to the managers' beliefs (R or I). The investors in our model can therefore be equivalently

12. In particular, the lack of inter-generational risk sharing, which is a feature of some of these models, is not driving any of the results in this paper.

thought of as institutional investors.

1.2 Endowments

In order to support their consumption over their lives, we assume that the arriving investors at time t are equally endowed with shares of new "trees," which arrive at time t.[13] Letting $s \leq t$ denote the time of arrival of a tree, we specify its time-t dividends as

$$D_{t,s} = \delta e^{-\delta(t-s)} D_t, \tag{2}$$

where $\delta > 0$ captures depreciation and D_t follows a geometric Brownian motion with mean μ_D and volatility $\sigma_D > 0$,

$$\frac{dD_t}{D_t} = \mu_D dt + \sigma_D dB_t, \tag{3}$$

with B_t a standard Brownian motion. Accordingly, the time-t total endowment of this economy is the sum of the endowment produced by all trees born up to to time t,

$$\int_{-\infty}^{t} D_{t,s} ds = \left(\int_{-\infty}^{t} \delta e^{-\delta(t-s)} ds \right) \times D_t = D_t.$$

The arriving investors sell their shares, which become part of the market portfolio. An implication of assumption (2) is that the dividend growth, $\frac{dD_{t,s}}{D_{t,s}} = (\mu_D - \delta)dt + \sigma_D dB_t$, is the same for any vintage s, and equals the dividend growth of the market portfolio. In turn, the return of the market portfolio, dR_t, can be written as

$$dR_t = \mu_t \, dt + \sigma_t \, dB_t, \tag{4}$$

where μ_t and σ_t are stochastic processes to be determined in equilibrium.

In the real world, shorting frictions are more relevant for a small fraction of stocks rather

13. The assumption that investors are endowed with shares of newly arriving trees follows Gârleanu, Kogan, and Panageas (2012) and Panageas (2020). This assumption is just a convenient way to endow new cohorts as compared to introducing labor income (as in Gârleanu and Panageas (2015) or Gârleanu and Panageas (2020)). Since the goal of the overlapping generations structure in this paper is merely to ensure stationarity, we adopt this more convenient shortcut.

than the broad stock market. In Section 5 we extend the model to allow for multiple stocks and study the special case in which the shorting frictions are relevant for small stocks only.

1.3 Beliefs

The irrational investors are optimistic and believe that the aggregate endowment grows at the rate $\mu^I > \mu_D$. Irrational investors hold this optimistic view over their life-time and do not learn ("dogmatic beliefs"). Introducing learning would be a distraction for the purposes of this paper and therefore we omit it.

For future reference, we define

$$\eta \equiv \frac{\mu^I - \mu_D}{\sigma_D}. \tag{5}$$

1.4 Dynamic budget constraint and short-selling frictions

The main departure from a frictionless market is that selling the stock short requires paying a lending fee, f_t. Specifically, letting $W^i_{t,s}$ denote the time-t wealth of an investor of type i who was born at time $s \leq t$ and $w^i_{t,s}$ denote the fraction of wealth invested in the stock, the dynamic budget constraint is

$$dW^i_{t,s} = W^i_{t,s}\left(r_t + \pi + n_t + w^i_{t,s}\left(\mu_t - r_t + \lambda^i_{t,s}\right) - \frac{c^i_{t,s}}{W^i_{t,s}}\right)dt + w^i_{t,s}W^i_{t,s}\sigma_t dB_t, \tag{6}$$

where r_t is the equilibrium interest rate and $\pi W^i_{t,s}$ is the income per unit of time earned from annuitizing her entire wealth (since she has no bequest motives).[14] The non-standard terms in equation (6) are the $\lambda^i_{t,s}$ and n_t, which we describe next.

The term $\lambda^i_{t,s}$ captures the presence of lending fees. It is defined as

$$\lambda^i_{t,s} \equiv \lambda_t(w^i_{t,s}) \equiv f_t \times \left(1_{\left\{w^i_{t,s}<0\right\}} + \tau y_t 1_{\left\{w^i_{t,s}\geq 0\right\}}\right), \tag{7}$$

14. We follow Blanchard (1985) in assuming the existence of a competitive insurance company. Investors pledge their wealth upon death in exchange for receiving an income stream while alive. This income stream is equal to the hazard rate of death, π, times the wealth of the investor, so that the insurance company breaks even.

where y_t is the fraction of a long portfolio that is lent out by the representative "brokerage house" and τ is the fraction of the lending fees that accrues to the investor. (We discuss the determination of y_t, τ, and f_t shortly.) Equation (7) reflects that an investor with a short position $w^i_{t,s} < 0$ has to pay a proportion f_t of the value of her entire short position, $|w^i_{t,s}|W^i_{t,s}$, so that the net-of-fee excess rate of return per dollar shorted is $-(\mu_t - r_t + f_t)\,dt - \sigma_t dB_t$. Similarly, an investor holding a positive position, $w^i_{t,s} > 0$, obtains an excess rate of return equal to $(\mu_t - r_t + \tau y_t f_t)\,dt + \sigma_t dB_t$ on her stock investments.

Market clearing for share lending requires that the fraction of the representative long position that is lent out, y_t, times the aggregate long position, W^+_t, equal the value of the aggregate short position, W^-_t :

$$y_t W^+_t = W^-_t, \tag{8}$$

where

$$W^-_t \equiv \sum_{i\in\{I,R\}} \int_{-\infty}^{t} |w^i_{t,s}| W^i_{t,s} 1_{\left\{w^i_{t,s}<0\right\}} ds \tag{9}$$

$$W^+_t \equiv \sum_{i\in\{I,R\}} \int_{-\infty}^{t} w^i_{t,s} W^i_{t,s} 1_{\left\{w^i_{t,s}>0\right\}} ds. \tag{10}$$

Following industry terminology, we henceforth refer to the quantity y_t, as the utilization ratio (or utilization for short), since it captures the fraction of lendable shares that are utilized by shorters.

To close the model, we must specify the lending frictions and derive the fee. In the text, we specify f_t through a supply curve $f_t = f(y_t)$ given by a non-decreasing function f. In Appendix A, however, we model explicitly a search-and-bargaining friction yielding such a supply curve. Specifically, we introduce competitive firms specializing in servicing either borrowers ("brokers") or lenders ("security lenders"). Brokers are faced with a demand from would-be short sellers, while security lenders obtain investors' long portfolios. Brokers and security lenders are matched pairwise subject to a "labor cost" and engage in bilateral negotiations that result in a lending fee f_t. In equilibrium, the fee is the same for all shares

that are lent, and therefore the total revenue from lending shares equals the fee multiplied by the value of all shares lent. This revenue is shared between the stock owners (a fraction τ of the lending revenue) and the households as compensation for their labor (the remaining $1-\tau$ fraction). These shares are driven by the relative bargaining powers of stock borrowers and lenders.

The term n_t in equation (6) captures the compensation for the labor cost in operating the matching technology. Denoting aggregate wealth at time t by W_t, we have $n_t = \frac{(1-\tau)f_t W_t^-}{W_t}$. We note that aggregate share lending fees, $f_t W_t^-$, accrue back to the households as the sum of lending income to long portfolios, $\tau f_t y_t W_t^+ = \tau f_t W_t^-$, and compensation for operating the matching technology, $n_t W_t = (1-\tau) f_t W_t^-$.

1.5 Equilibrium definition

Equilibrium in the lending market requires that the supply of lendable shares $y_t W_t^+$ is equal to the demanded short interest, W_t^- (Equation (8)).

The rest of the equilibrium definition is standard. We require that investors I and R maximize (1) over $c_{t,s}^i$ and $w_{t,s}^i$ subject to the budget constraint (6), and μ_t, r_t, and σ_t are such that the bond market clears, $\sum_{i\in\{I,R\}} \int_{-\infty}^t \nu^i \left(1 - w_{t,s}^i\right) W_{t,s}^i ds = 0$, the stock market clears, $\sum_{i\in\{I,R\}} \int_{-\infty}^t \nu^i w_{t,s}^i W_{t,s}^i ds = P_t$, and the goods market clears, $\sum_{i\in\{I,R\}} \int_{-\infty}^t \nu^i c_{t,s}^i ds = D_t$. By Walras' Law, market clearing of the bond market implies stock market clearing and vice versa, and accordingly the asset-market clearing requirements can be written equivalently as $W_t = \sum_{i\in\{I,R\}} \int_{-\infty}^t \nu^i W_{t,s}^i ds = P_t$.

For future reference, we note that stock market clearing implies $y_t = \frac{W_t^-}{W_t^+} = \frac{W_t^-}{P_t + W_t^-} < 1$. It also implies that there is a simple, monotone relation between the utilization ratio, y_t, and short interest, $\frac{W_t^-}{P_t}$, given by $y_t = \frac{\frac{W_t^-}{P_t}}{1+\frac{W_t^-}{P_t}}$.

2 Analysis

We analyze the model in two steps. First, we consider a special parametric case that allows us to characterize all equilibrium quantities in closed form. The special case we analyze is the "elastic supply" case, that is, the limiting case where the supply of lendable shares

is horizontal at some level $f(y_t) = \varphi$. (As we explain in Appendix A, this special case corresponds to a particular specification for the cost of lending out shares.) Section 4 extends the analysis to allow for an increasing function $f(y_t)$.

2.1 Optimal portfolio and consumption

For a log investor the wealth-to-consumption ratio is constant and equal to $\frac{c^i_{t,s}}{W^i_{t,s}} = \rho+\pi$. Given homothetic preferences, all agents of a given type choose the same portfolio independent of their cohort, s; therefore we may write w^i_t (rather than $w^i_{t,s}$). Additionally, a convenient property of logarithmic utility is that the portfolio is myopic and maximizes the logarithmic growth rate of an investor's wealth, under the investor's beliefs,

$$w^i_t = \arg\max_w \left\{ w\left(\mu_t + \eta\sigma_t 1_{\{i=I\}} - r_t + \lambda_t(w)\right) - \frac{1}{2}\left(w\sigma_t\right)^2 \right\}, \tag{11}$$

where $1_{\{i=I\}}$ is an indicator function taking the value one when $i = I$ and zero otherwise.

Letting $\hat{\mu}^i_t \equiv \mu_t + \eta\sigma_t 1_{\{i=I\}}$ denote the expected return on the stock as perceived by investor $i \in \{I, R\}$, the optimal portfolio is

$$w^i_t = \begin{cases} \frac{\hat{\mu}^i_t - r_t + f_t}{\sigma^2_t} & \text{if } \hat{\mu}^i_t - r_t + f_t < 0 \\ \frac{\hat{\mu}^i_t - r_t + \tau f_t y_t}{\sigma^2_t} & \text{if } \hat{\mu}^i_t - r_t + \tau f_t y_t > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

Figure 1 depicts equation (12), the optimal portfolio of investor i as a function of $\frac{\hat{\mu}^i_t - r_t}{\sigma^2_t}$. The figure shows the presence of an "inaction" region: for values of $\frac{\hat{\mu}^i_t - r_t}{\sigma^2_t}$ between $-\frac{f_t}{\sigma^2_t}$ and $-\frac{f_t}{\sigma^2_t}\tau y_t$, the investor optimally chooses a portfolio weight of zero.

One straightforward implication of equation (12) is that if investor R is actively shorting ($w^R_t < 0$) then the expected excess rate of return per dollar shorted is positive even after netting out the fee f_t.[15]

15. This statement uses the assumption that agent R has the correct beliefs, and is a direct consequence of the agent's risk aversion. For a precise calculation, evaluate (12) with $i = R$, impose $w^R_t < 0$, and re-arrange to obtain $-(\mu_t - r - f_t) = -(\hat{\mu}^R_t - r - f_t) = -w^R_t\sigma^2_t > 0$. The term $-w^R_t\sigma^2_t$, which equals the absolute value of the covariance of the stock's return with the short seller's portfolio, is the risk compensation to the agent for taking a short position.

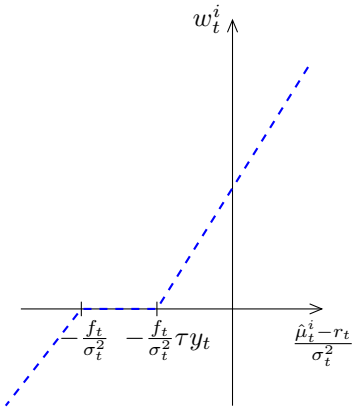


Figure 1: The optimal portfolio weight of investor i as a function of that investor's perceived value of $\frac{\hat{\mu}_t^i - r_t}{\sigma_t^2}$.

2.2 Equilibrium

It is useful to start by defining the wealth-weight ω_t^i of investors of type $i \in \{I, R\}$,

$$\omega_t^i \equiv \frac{\nu^i \int_{-\infty}^t \pi e^{-\pi(t-s)} W_{t,s}^i ds}{W_t}. \tag{13}$$

To save notation, henceforth we refer to ω_t^R simply as ω_t, and therefore $\omega_t^I = 1 - \omega_t$. Using $\frac{c_{t,s}^i}{W_{t,s}^i} = \rho + \pi$, the goods-market and stock-market clearing requirements imply

$$\begin{aligned} D_t &= \sum_{i \in \{I,R\}} \int_{-\infty}^t \nu^i \pi e^{-\pi(t-s)} c_{t,s}^i ds = (\rho + \pi) \sum_{i \in \{I,R\}} \int_{-\infty}^t \nu^i \pi e^{-\pi(t-s)} W_{t,s}^i ds \\ &= (\rho + \pi) W_t = (\rho + \pi) P_t. \end{aligned} \tag{14}$$

Taking logarithms gives $d \log D_t = d \log P_t$ and therefore the stock market volatility equals $\sigma_t = \sigma_D$. The implication of a constant stock volatility is convenient for obtaining closed-form solutions. In Section 5 we discuss extensions of the model that allow for a time-varying price-dividend ratio and volatility by introducing multiple stocks.

As mentioned earlier, in an effort to obtain a closed-form solution we assume that the supply of lendable shares is perfectly elastic at the rate φ:

Assumption 1 $f(y) = \varphi > 0$.

We maintain this assumption until Section 4. In preparation for the description of the

equilibrium, we start with the following definition and assumptions on the parameters.

Definition 1 *Define* ω_1^* *and* $F(\omega)$ *as*

$$\omega_1^* \equiv 1 - \frac{\sigma_D}{\eta - \frac{\varphi}{\sigma_D}}, \tag{15}$$

$$F(\omega) \equiv \left(\sigma_D - \omega\left((1+\tau)\frac{\varphi}{\sigma_D} - \eta\right)\right)^2 - 4\tau\frac{\omega^2}{1-\omega}\frac{\varphi}{\sigma_D}\left(\sigma_D + (1-\omega)\left(\frac{\varphi}{\sigma_D} - \eta\right)\right). \tag{16}$$

Assumption 2 *Assume that* η, φ, σ_D, *and* τ *are such that*

$$(1+\tau)\frac{\varphi}{\sigma_D} > \eta > \frac{\varphi}{\sigma_D}, \tag{17}$$

$$\omega_1^* > \frac{\sigma_D}{(1+\tau)\frac{\varphi}{\sigma_D} - \eta} > 0, \tag{18}$$

and $F(\omega)$ *has a unique root in the interval* $(0,1)$*, denoted by* ω_2^**.*

The following proposition guarantees that Assumption 2 can be satisfied.

Proposition 1 *There exists an open set of positive values* η, φ, σ_D, *and* τ *that satisfy Assumption 2.*

The next proposition describes the equilibria in our economy.

Proposition 2 *Suppose that Assumption 2 holds. Then* $\omega_2^* > \omega_1^*$ *and the equilibria in this economy can be described as follows.*

i) If $\omega_t \in (\omega_2^*, 1]$ *there is no short-selling in equilibrium. The equilibrium is unique and the Sharpe ratio* $\kappa_t \equiv \frac{\mu_t - r_t}{\sigma_D}$ *is given by*

$$\kappa_t = \begin{cases} \sigma_D - (1-\omega_t)\eta & \text{if } \omega_t > 1 - \frac{\sigma_D}{\eta} \\ \frac{\sigma_D}{1-\omega_t} - \eta & \text{if } \omega_t \in (\omega_2^*, 1 - \frac{\sigma_D}{\eta}] \end{cases}. \tag{19}$$

ii) If $\omega_t \in [\omega_1^*, \omega_2^*]$*, then there are three equilibria. The first equilibrium continues to be given by (19) and involves no short-selling. The second and third equilibria involve shorting and utilization,* y_t*, corresponds to the two roots* y^+ *and* y^- *of the quadratic equation*

$$y\left(\eta + \frac{\sigma_D}{\omega_t} - \frac{\varphi}{\sigma_D}(1-\tau y)\right) - \left(\eta - \frac{\sigma_D}{1-\omega_t} - \frac{\varphi}{\sigma_D}(1-\tau y)\right) = 0, \tag{20}$$

which has two real roots y^+ *and* y^- *in the interval* $(0,1)$*. The Sharpe ratio in the equilibria associated with* y^+ *and* y^- *are*

$$\kappa_t^{\pm} = \sigma_D - (1-\omega_t)\,\eta - \frac{\varphi}{\sigma_D}\left(\omega_t + \tau y^{\pm}(1-\omega_t)\right). \tag{21}$$

iii) If $\omega_t \in [0, \omega_1^*)$*, then the equilibrium is unique and involves shorting. In this case only the larger of the two roots* (y^+) *of equation (20) lies in the interval* $(0,1)$*, and the unique equilibrium Sharpe ratio is given by* κ^+*.*

In all three cases the interest rate is given by

$$r_t = \rho + \pi + \mu_D - \delta - \kappa_t \sigma_D. \tag{22}$$

Additionally, because κ_t*,* r_t*, and* y_t *are functions of* ω_t*, so is* w_t^R*, and the stochastic process for* ω_t*,* $d\omega_t = \mu_{\omega,t}dt + \sigma_{\omega,t}dB_t$*, is Markovian with volatility* $\sigma_{\omega,t} = \sigma_\omega(\omega_t)$ *and drift* $\mu_{\omega,t} = \mu_\omega(\omega_t)$ *given by*

$$\sigma_\omega(\omega_t) = \omega_t\left(w_t^R - 1\right)\sigma_D, \tag{23}$$

$$\mu_\omega(\omega_t) = \omega_t\left(-\mu_D + \sigma_D^2 - \pi + r_t - \rho + w_t^R\left(\mu_t - r_t + \lambda_t(w_t^R)\right) - w_t^R\sigma_D^2\right) + \nu^R\delta. \tag{24}$$

Figure 2 illustrates Proposition 2. The left graph plots $\kappa(\omega_t)$, the Sharpe ratio, as a function of the wealth share of rational agents, ω_t. As a benchmark, the line labeled "Costless shorting eqlm" depicts $\sigma_D - (1-\omega_t)\,\eta$, i.e., the Sharpe ratio that would obtain in this economy in the absence of any shorting frictions ($\varphi = 0$). The curve "No shorting eqlm" depicts the Sharpe ratio in the equilibrium that involves no shorting for the values of ω_t that such an equilibrium exists. Similarly for the curves "Med. shorting eqlm" and "High shorting eqlm," which depict equilibria with shorting for the values of ω_t that permit such equilibria. To expedite the exposition of the results, we postpone a discussion of the quantitative implications of the model until Section 5.2. The graphs in the current section are meant to illustrate qualitative properties of the model.

The figure shows that when ω_t is larger than $1 - \frac{\sigma_D}{\eta}$ the lines "Costless shorting eqlm" and "No shorting eqlm" coincide, reflecting that all investors invest strictly positive amounts

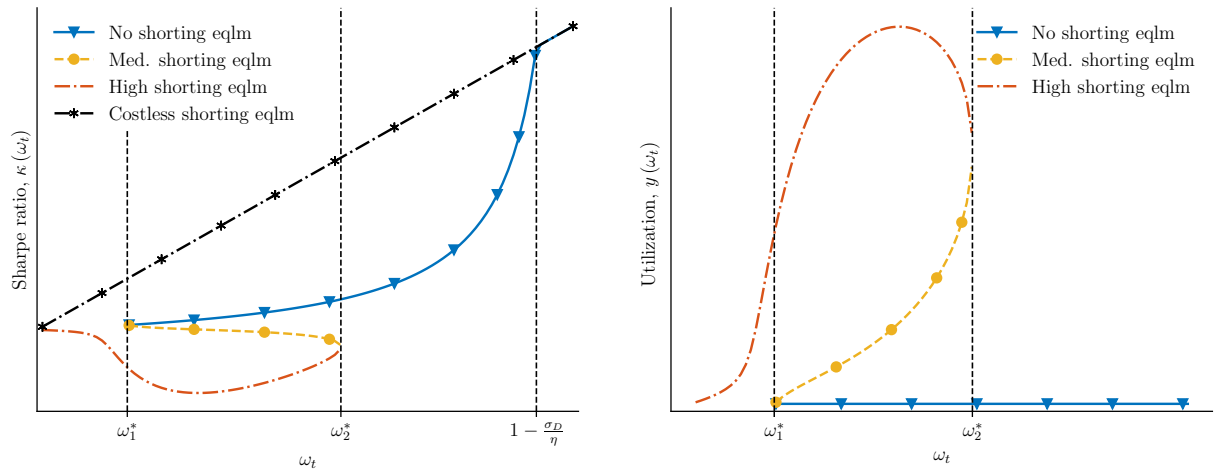


Figure 2: Left: All possible equilibrium values of the Sharpe ratio, as a function of ω_t. Right: The utilization ratio, $y(\omega_t)$, in all of the equilibria as a function of ω_t.

in the stock market in this region of ω_t.

When ω_t becomes smaller than $1-\frac{\sigma_D}{\eta}$ (but larger than ω_2^*), the rational investor puts zero weight on stocks, but the shorting fee φ deters her from actively short-selling. Since only the irrational investor is marginal in financial markets, the lines "Costless shorting eqlm" and "No shorting eqlm." deviate from each other when $\omega_t < 1-\frac{\sigma_D}{\eta}$. In this region the magnitude of the lending fee, φ, does not impact the Sharpe ratio directly (only by deterring the R investors from shorting).

If ω_t becomes smaller than ω_2^* (but larger than ω_1^*) the economy exhibits three equilibria. In the first equilibrium, there is still no shorting. In the second and third, there is active shorting by the rational investor. Across these three equilibria, the higher the extent of shorting, the lower the Sharpe ratio. This is illustrated in the right graph of Figure 2.

Finally, if ω_t becomes smaller than ω_1^*, then the equilibrium becomes unique and involves shorting.[16]

16. To see why there can be no equilibrium without shorting when $\omega_t < \omega_1^*$, assume otherwise. Indeed assume that the R investor holds zero stocks and is not marginal in the stock market $\left(w_t^R = 0\right)$. The market clearing requirement, $\omega_t w_t^R + (1-\omega_t) w_t^I = 1$, along with $w_t^I = \frac{\kappa_t+\eta}{\sigma_D}$ implies that the Sharpe ratio would be $\kappa_t = \frac{\sigma_D}{1-\omega_t} - \eta$. Under this supposition, it would therefore be the case that $\mu_t - r + \varphi = \sigma_D\left(\kappa_t + \frac{\varphi}{\sigma_D}\right) = \sigma_D\left(\frac{\sigma_D}{1-\omega_t} - \eta + \frac{\varphi}{\sigma_D}\right) < 0$, where the inequality follows from $\omega_t < \omega_1^*$. Because $\mu_t - r + \varphi < 0$, equation (12) implies that the R investor would want to short the market, contradicting the assumption that she is optimally holding zero stocks.

The presence of a region where multiple equilibria co-exist is not a very common feature of asset pricing models, especially when there is only one good and one positive-supply asset. To better understand the source of this multiplicity, it is useful to provide a concise derivation of the key statements in Proposition 1.

Specifically, suppose that we consider equilibria that involve active shorting ($w_t^R < 0$). In such equilibria, the optimal portfolio holdings can be expressed as

$$w_t^R = \frac{\kappa_t + \frac{\varphi}{\sigma_D}}{\sigma_D} \tag{25}$$

$$w_t^I = \frac{\kappa_t + \eta + \frac{\varphi}{\sigma_D}\tau y_t}{\sigma_D}, \tag{26}$$

while asset-market clearing requires

$$\omega_t w_t^R + (1 - \omega_t)\, w_t^I = 1. \tag{27}$$

Combining equations (25)–(27) leads to

$$\kappa_t = \sigma_D - (1 - \omega_t)\,\eta - \frac{\varphi}{\sigma_D}\left(\omega_t + \tau y_t(1 - \omega_t)\right), \tag{28}$$

which is equation (21) of Proposition 1. Note that the partial derivative of κ_t with respect to y_t is negative. This is intuitive: For a given ω_t, a higher value of y_t increases the effective rate of return to (long-portfolio) stock holders (I investors).

The dependence of the Sharpe ratio, κ_t, on utilization, y_t, gives rise to a feedback loop between these two quantities. A higher value of y_t increases the rate of return on a long position and strengthens investor I's demand for the asset. This increased demand lowers the Sharpe ratio to clear the market. The lower Sharpe ratio strengthens the short-sellers' appetite to borrow the stock and short it. In turn, the increased shorting demand raises the utilization ratio, y_t, increasing the effective return to I investors, which further reduces the Sharpe ratio, etc.

These self-reinforcing effects are the root cause of the multiple equilibria. The easiest way to see this is by completing the computation of the Sharpe ratio, which requires us to

determine the value of y_t that clears the lending market. Indeed, in any equilibrium involving $w_t^R < 0$ and $w_t^I > 0$ we must have

$$y_t = \frac{W_t^-}{W_t^+} = \frac{-w_t^R W_t^R}{w_t^I W_t^I} = -\frac{w_t^R}{w_t^I} \times \frac{\omega_t}{1-\omega_t}. \tag{29}$$

Using (25) to compute the ratio $\frac{w_t^R}{w_t^I}$ gives

$$\begin{aligned} y_t &= -\frac{\kappa_t + \frac{\varphi}{\sigma_D}}{\kappa_t + \eta^I + \frac{\varphi}{\sigma_D}\tau y_t} \times \frac{\omega_t}{1-\omega_t} \\ &= \frac{\eta - \frac{\sigma_D}{1-\omega_t} - \frac{\varphi}{\sigma_D}(1-\tau y_t)}{\eta + \frac{\sigma_D}{\omega_t} - \frac{\varphi}{\sigma_D}(1-\tau y_t)}, \end{aligned} \tag{30}$$

where the last line follows from (28) after collecting terms and simplifying. Rearranging (30) gives the quadratic equation (20), which is the key equation of Proposition 1. The rest of Proposition 1 is devoted to studying this quadratic equation and confirming that its roots correspond to valid equilibria (with non-zero shorting). The proposition shows that additionally there is an equilibrium with zero shorting.

Remark 1 *The fact that there are three equilibria, one of which features no shorting, is an implication of there being only two types of agents in the model. With more than two types of agents, more than three equilibria can obtain. Also, in the case of multiple equilibria, all of the equilibria can involve strictly positive short interest, as we show in Appendix B.*

Remark 2 *The presence of multiple equilibria implies that the aggregate demand curve for the stock, $D(\kappa) \equiv W_t^+(\kappa, y(\kappa)) - W_t^-(\kappa, y(\kappa))$ is a backward-bending function of κ (where $y(\kappa)$ is implicitly defined by the first line of equation* (30)*). The market-clearing requirement, $D(\kappa) = 1$, along with the fact that there are multiple values of κ such that $D(\kappa) = 1$, implies that $D(\kappa)$ is not monotonically declining, but instead is backward bending. As observed by Gennotte and Leland (1990), a backward bending demand curve gives rise to discontinuous changes in equilibrium. This necessary instability is illustrated in Figure 2: when the value of the continuous-path wealth-share process ω_t increases from below ω_1^* to above ω_2^*, the processes κ_t and y_t experience discontinuous changes on the interval $[\omega_1^*, \omega_2^*]$ irrespective of how market*

participants select between high, medium, and no shorting equilibria.[17]

2.2.1 Multiplicity and amplification

In our model multiplicity is a convenient way to illustrate a mutually reinforcing feedback loop between the Sharpe ratio, κ_t, and utilization, y_t. Before presenting general conditions that can lead to equilibrium multiplicity, in this section we confine attention to situations where the shorting market is active, but the equilibrium is unique. We show that even when the equilibrium is unique, the feedback loop between κ_t and y_t is still present and becomes the source of an "amplification" mechanism.

Specifically, assume that $\omega_t < \omega_1^*$, so that the shorting market is active and the equilibrium is unique. In this region, consider the impact of a change in the parameter η, which governs the optimism of irrational investors, on the Sharpe ratio, κ. Next, define $G(y, \kappa; \eta) \equiv y\left(\kappa + \eta + \frac{\varphi}{\sigma_D}\tau y\right) + \frac{\omega_t}{1-\omega_t}\left(\kappa + \frac{\varphi}{\sigma_D}\right)$ and note that $G(y^+, \kappa; \eta) = 0$, by Equation (30). By the implicit function theorem, $dy_t = -\frac{G_\kappa}{G_y}d\kappa - \frac{G_\eta}{G_y}d\eta$. In turn, totally differentiating Equation (28) yields $d\kappa_t = -(1-\omega_t)d\eta - \frac{\varphi}{\sigma_D}\tau(1-\omega_t)\frac{dy_t}{d\eta}d\eta$. Combining these two equations yields

$$d\kappa_t = \Lambda d\eta + \Phi d\kappa_t, \tag{31}$$

where $\Phi = \frac{\varphi}{\sigma_D}\tau(1-\omega_t)\frac{G_\kappa}{G_y}$ and $\Lambda = -(1-\omega_t) + \frac{\varphi}{\sigma_D}\tau(1-\omega_t)\frac{G_\eta}{G_y}$.

The quantity Λ captures the "direct" effect of a change in η on κ_t. The presence of the term Φ on the right-hand side of Equation (31) illustrates the presence of an "amplification" effect. Indeed, iterated substitution yields

$$\begin{aligned} d\kappa_t &= \Lambda d\eta + \Phi d\kappa_t = \Lambda d\eta + \Phi\left(\Lambda d\eta + \Phi d\kappa_t\right) \\ &= \Lambda(1+\Phi)d\eta + \Phi^2 d\kappa_t \\ &= \Lambda\left(1 + \Phi + \Phi^2 + ...\right)d\eta = \frac{\Lambda}{1-\Phi}d\eta. \end{aligned} \tag{32}$$

17. For instance, if the market participants always coordinate on the high shorting equilibrium, the jump will occur when $\omega_t = \omega_2^*$, and if market participants coordinate on the no shorting equilibrium, the jump will occur at ω_1^*.

Lemma 2 in the Appendix shows that in the region where the equilibrium is unique, Φ is between 0 and 1. Equation (32) is reminiscent of economic models that contain a "multiplier" effect. An increase in η has the direct effect of lowering the Sharpe ratio, since the optimists become more optimistic. However, this direct effect starts a "spiral" by increasing utilization, y_t, leading to a further reduction in the Sharpe ratio by a fraction $\Phi < 1$ of the original increase, further increasing y, lowering κ by a further Φ^2 of the original effect, etc. The expression for the fraction Φ is given by the product of a) how much a change in utilization, y_t, lowers the Sharpe ratio, $-\frac{\varphi}{\sigma_D}\tau\left(1-\omega_t\right)$, and b) the impact of a change in the Sharpe ratio on utilization, $-\frac{G_\kappa}{G_y}$.

The main difference between the regions of multiplicity, $\omega_t \in (\omega_1^*, \omega_2^*)$, and uniqueness, $\omega_t < \omega_1^*$, is that in the multiplicity region the feedback loop between κ_t and y_t becomes so strong that $\Phi > 1$ for some values of y.[18]

3 Dynamics of the wealth shares

When multiple equilibria are possible, both the drift rate $\mu_t^R(\omega_t)$ of the wealth share of type R investors and the expected logarithmic growth rate of their wealth are higher in equilibria that feature higher y_t, as the next proposition shows.

Proposition 3 *For a fixed wealth share of the R-agents, ω_t, consider two equilibria A and B with the following properties: (1) $w_t^R \leq 0$ in both equilibria A and B, and (2) $y_t^B > y_t^A$ (and accordingly $\kappa_t^B < \kappa_t^A$).*

Then the drift of investor R's wealth share in equilibrium $i \in \{A, B\}$, $\mu_\omega^i(\omega_t)$, satisfies $\mu_\omega^B(\omega_t) > \mu_\omega^A(\omega_t)$. In addition, the drift of the logarithmic growth rate of investor R, given by

$$g_t \equiv r_t + \max_{w \leq 0}\left\{w\left(\kappa_t\sigma_D + \varphi\right) - \frac{1}{2}\left(w\sigma_D\right)^2\right\} - \left(\rho + \pi\right), \tag{33}$$

is higher in equilibrium B than in equilibrium A, i.e., $g^B(\omega_t) > g^A(\omega_t)$.

18. For instance, one can show that $\Phi > 1$ for values of y in a neighborhood of y^-, while $\Phi < 1$ in a neighborhood of y^+. An implication is that — using the common definition of stability — the equilibria corresponding to y^+ and $y = 0$ are stable, while the equilibrium associated with y^- is unstable.

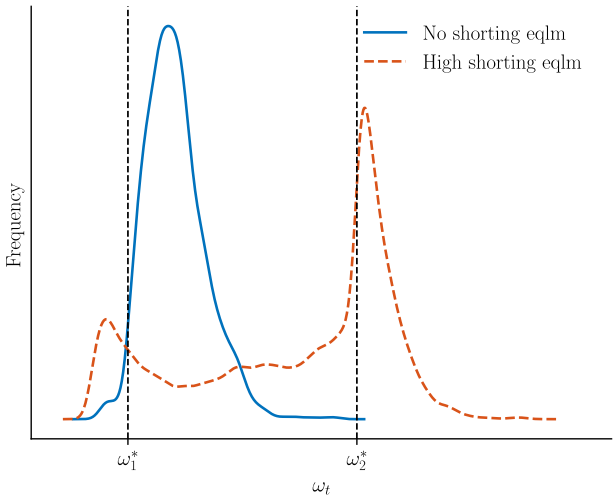


Figure 3: An illustration of Proposition 3. Simulating the model for the case in which market participants coordinate on the "high shorting" (respectively, "no shorting") equilibrium, the figure depicts the stationary distribution of the wealth share of the rational investor, ω_t, for the economy of Figure 2.

Figure 3 provides an illustration of Proposition 3. The figure shows the stationary distribution of ω_t in the equilibrium associated with no shorting for values $\omega_t \in (\omega_1^*, \omega_2^*)$ and in the equilibrium associated with the highest shorting, $y^+(\omega_t)$, for $\omega_t \in (\omega_1^*, \omega_2^*)$. The figure shows that the stationary distribution of ω_t has a higher mean in the high-shorting equilibrium rather than in the no-shorting equilibrium. This is consistent with Proposition 3, which asserts a higher (logarithmic) growth rate for the wealth of R investors in the second equilibrium.

When comparing a high-shorting to a low-shorting equilibrium, therefore, one must account for two competing effects on the stationary mean of the Sharpe ratio κ_t. On the one hand, for a fixed ω_t the Sharpe ratio is lower in the high-shorting equilibrium. On the other hand, low values of ω_t become infrequent in the high-shorting equilibrium. The first channel makes the stationary mean of the Sharpe ratio lower in the high-shorting equilibrium, but the second channel has the opposite effect. The overall effect on the stationary value of the Sharpe ratio is ambiguous. This observation will become important in Section 5, when we discuss the impact of an equilibrium shift on the price-dividend ratio of a small stock.

4 Arbitrary Supply Curve for Lendable Shares

In Section 2 we assumed a perfectly elastic supply curve for lending shares ($f(y) = \varphi$), which allowed us to solve the model in terms of a simple, quadratic equation. Here we revisit our main result, namely the existence of multiple equilibria, for an arbitrary (differentiable) non-decreasing supply curve $f_t = f(y_t)$. In addition, to allay possible fears that our results are special to the the discrete nature of the two-type distribution we considered so far,[19] the following proposition allows for a continuous distribution of beliefs (with connected support).

Proposition 4 *Let $h(y) \equiv f(y)\,(1-\tau y)$. If there exists a value $y \in [0,1)$ with (a) $h'(y) < 0$ and (b) $\sigma_D^2 < \frac{1}{4}(1-y)^2|h'(y|$, then there exist wealth distributions over beliefs for which multiple equilibrium values of y_t (and κ_t) obtain.*

A key role in Proposition 4 is played by the function $h(y)$. This function captures the difference between the (proportional) fee paid by a short seller, $f(y)$, minus the (proportional) lending income received by a long investor, $\tau f(y)\, y$. To understand why the condition that $h'(y) < 0$ for some $y \in [0,1)$ is necessary for multiple equilibria, suppose that there are (at least) two equilibria with $y_{t,1} < y_{t,2}$. From market clearing of the spot market, it must be that, in the second equilibrium, both long and short investors choose a portfolio of larger absolute value than in the first equilibrium: $y_{t,1} < y_{t,2} \implies |w_{t,1}^R| < |w_{t,2}^R|$ and $w_{t,1}^I < w_{t,2}^I$.[20]

An immediate consequence is that $w_{t,1}^I - w_{t,1}^R < w_{t,2}^I - w_{t,2}^R$. From the expressions for the portfolio weights of investors, (25) and (26), this inequality is equivalent to $f(y_{t,1})\,(1-\tau y_{t,1}) > f(y_{t,2})\,(1-\tau y_{t,2})$, that is, $h(y_{t,1}) > h(y_{t,2})$. Since f is assumed differentiable, there exists $y \in (y_{t,1}, y_{t,2})$ such that $h'(y) < 0$ (Mean Value Theorem).

Proposition 4 shows that condition (a) — when combined with condition (b) — is not just necessary, but also sufficient for the existence of multiple equilibria, in the sense that there exists (an open set of) wealth distributions over beliefs that ensure the existence of multiple equilibria. In the next section, where we present a version of the model with a large

19. Note also that Appendix B illustrates multiple equilibria obtaining with three agent types.

20. The fact that $w_t^R < 0$ along with Equations (27) and (29) imply that $y_t = \frac{\omega_t |w_t^R|}{1+\omega_t |w_t^R|}$, which is an increasing function of $|w_t^R|$ (for any given ω_t). Accordingly, $y_{t,1} < y_{t,2}$ implies that $|w_{t,1}^R| < |w_{t,2}^R|$. In turn, by market clearing, $w_t^I = \frac{1+\omega_t |w_t^R|}{1-\omega_t}$, and therefore $|w_{t,1}^R| < |w_{t,2}^R|$ implies $w_{t,1}^I < w_{t,2}^I$.

and a small stock, we show that condition (b) is satisfied as long as investor disagreement pertains to the idiosyncratic risk of a small stock. (See Remark 3 in the next section).

5 Multiple Risky Assets and the Price-Dividend Ratio

In the baseline model, the price-dividend ratio and the volatility of the stock market are both constant. This is an implication of a) logarithmic utility over intermediate consumption (which implies a constant wealth-to-consumption ratio) and b) a single asset in positive net supply. As is typical of models with similar setups, fluctuations in the interest rate offset the fluctuations of the risk premium, thus rendering the overall discount rate — and by implication the price-dividend ratio[21] — constant.

We next introduce a second risky asset to study the model implications for the price-dividend ratio. After extending Proposition 2 to this setting — a result of independent interest — we consider a limiting case of the multi-asset model that permits simple computations. Specifically, we study the limit in which there is a "small" stock subject to shorting costs and a "large" stock that can be shorted costlessly. In that limit, only the endowment of the large stock matters for the interest rate and thus the price-dividend ratio of the small stock is time varying and reflects variations in its risk premium.

5.1 Multiple risky assets

In this section we introduce an additional Lucas tree (stock 2) to our baseline model, which is not subject to any trading frictions, and accounts for a potentially large part of the total market capitalization. We continue to assume that borrowing stock 1 requires lending fees, as in the baseline model.

We allow one more feature, in the spirit of the "limited recognition hypothesis" of Merton (1987). Specifically, while all investors participate in the markets for stock 2 and the risk-free asset, only a fraction of investors pays any "attention" to stock 1. The remaining fraction of investors simply optimize their portfolio over the risk-free asset and stock 2 and assign zero weight to stock 1.

21. Note also that the expected dividend growth is constant.

To ease the comparison of the results of this section with Proposition 2, we maintain the assumption that the lending supply curve is horizontal, and the lending fee is constant and equal to φ.

We assume that the equilibrium returns on stocks 1 and 2 follow a (possibly correlated) vector diffusion process of the form

$$dR_{1,t} = \mu_{1,t}dt + \sigma_{1,t}dB_{1,t} + b_t\sigma_{2,t}dB_{2,t} \tag{34}$$

$$dR_{2,t} = \mu_{2,t}dt + \sigma_{2,t}dB_{2,t}, \tag{35}$$

where $B_{1,t}$ and $B_{2,t}$ are independent Brownian motions, and $\mu_{1,t}$, $\mu_{2,t}$, $\sigma_{1,t}$, $\sigma_{2,t}$, and b_t are determined in equilibrium. We assume that investors I believe that Brownian motion 1 follows the dynamics[22] $dB_{1,t} + \eta dt$, while no investor has any belief distortions pertaining to Brownian motion 2.

To facilitate the statement of equilibrium returns, we define $\widetilde{m}_{1,t} \equiv \frac{m_{1,t}}{\widehat{\omega}_t}$ as the ratio of the stock-1 market capitalization share, denoted by $m_{1,t}$, to the wealth share of all investors participating in the market for stock 1, denoted by $\widehat{\omega}_t$. We also define $\kappa_{1,t} \equiv \frac{(\mu_{1,t}-r)-b_t(\mu_{2,t}-r)}{\sigma_{1,t}}$ as the Sharpe ratio of a portfolio long 1 unit of asset 1 and short b_t units of asset 2.

Proposition 5 *In an equilibrium with shorting in asset 1 ($y_t > 0$), y_t is given by the root(s) of the quadratic equation*

$$0 = y\left(\eta + \frac{\widetilde{m}_{1,t}}{\omega_t}\sigma_{1,t} - \frac{\varphi}{\sigma_{1,t}}(1-\tau y)\right) - \left(\eta - \frac{\widetilde{m}_{1,t}}{1-\omega_t}\sigma_{1,t} - \frac{\varphi}{\sigma_{1,t}}(1-\tau y)\right) \tag{36}$$

that lie(s) in the interval $(0,1)$, and the Sharpe ratio is given by

$$\kappa_{1,t} = \widetilde{m}_{1,t}\sigma_{1,t} - (1-\omega_t)\eta - \frac{\varphi}{\sigma_{1,t}}(\omega_t + (1-\omega_t)\tau y_t). \tag{37}$$

Similarly, in an equilibrium without shorting in asset 1 we have $\kappa_{1,t} = \sigma_{1,t}\widetilde{m}_{1,t} - (1-\omega_t)\eta$

22. More formally, the Radon-Nikodym derivative of the true probability measure with respect to the subjective one is given by

$$Z_t^I \equiv e^{-\frac{\eta^2}{2}t + \eta B_{1,t}}.$$

if investor R holds an interior position in asset 1 *and* $\kappa_{1,t} = \frac{\sigma_{1,t}\widetilde{m}_{1,t}}{1-\omega_t} - \eta$ *otherwise.*

The excess return to asset 2 is given by the conventional CAPM relation

$$\mu_{2,t} - r_t = b_t \sigma_{2,t}^2 m_{1,t} + \sigma_{2,t}^2 m_{2,t}. \tag{38}$$

Equations (36) and (37) are the same as (20) and (21), respectively, except that the volatility, σ_D, is replaced by $\widetilde{m}_{1,t}\sigma_{1,t}$. The reason for this replacement is intuitive: In the case of a single stock, the risk of that stock, σ_D, is aggregate (by construction) and commands a risk premium. When there are multiple stocks, the risk compensation for bearing the idiosyncratic risk[23] of stock 1, $\sigma_{1,t}$, is multiplied by $\widetilde{m}_{1,t}$, i.e., the stock market capitalization of stock 1 as a fraction of the wealth share of investors actively participating in the stock. An implication is that when $\widetilde{m}_{1,t}$ approaches zero, the idiosyncratic risk becomes diversifiable, and there is no compensation for bearing that risk (the first term on the right-hand side of (37) disappears).

Remark 3 *Since the equations determining* $\kappa_{1,t}$ *and* y_t *are essentially the same as (21) and (20), Proposition 4 remains unchanged when there are multiple stocks, except that now condition (b) becomes* $(\widetilde{m}_{1,t}\sigma_{1,t})^2 < \frac{1}{4}(1-y)^2|h'(y)|$. *This condition therefore does not require that the total volatility of stock 1, or even its idiosyncratic part* $\sigma_{1,t}$, *be small, but rather that the risk of stock 1 be diversifiable by the agents trading it (small* $\widetilde{m}_{1,t}$*).*

5.2 A limiting economy with a small and a large stock

The CAPM-style formulae provide equilibrium expected returns conditional on the equilibrium covariance matrix and the investor wealth shares. To fully solve the model, we consider a limiting two-stock economy in which trees of type 1 are small compared to trees of type 2 and also the fraction of investors that pay attention to trees of type 1 is small. Since this section involves some detailed modeling assumptions, we relegate the full presentation to Appendix C. In the text we simply summarize the setup and the main findings.

23. Recall that the Sharpe ratio in Proposition 5 pertains to a portfolio that invests one dollar in asset 1 and shorts b_t units of asset 2, hedging out the exposure of the portfolio to the second Brownian shock.

Specifically, assume that there are two types of trees, namely "small" trees (type-1 trees) and "large" trees (type-2 trees). Type-2 trees have dividends similar to the baseline model, namely $D_{2,t,s} = \phi_2 \delta_2 D_{2,t} e^{-\delta_2 (t-s)}$, where $\phi_2 > 0$, $\delta_2 > 0$, and $D_{2,t}$ follows a geometric Brownian motion, $\frac{dD_{2,t}}{D_{2,t}} = \mu_{2,D} dt + \sigma_{2,D} dB_{2,t}$, with drift $\mu_{2,D} > 0$. Type-1 trees produce dividends $D_{1,t,s} = \phi_1 \delta_1 D_{2,s} e^{-\delta_1 (t-s) + \sigma_{1,D}(B_{1,t} - B_{1,s})}$, with $\phi_1 > 0$ and $\delta_1 > 0$. With the above dividend specifications, the dividend ratio $\frac{D_{1,t}}{D_{2,t}}$ is stationary and given by

$$\frac{D_{1,t}}{D_{2,t}} = \frac{\int_{-\infty}^{t} D_{1,t,s} ds}{\int_{-\infty}^{t} D_{2,t,s} ds} = \frac{\phi_1}{\phi_2} \int_{-\infty}^{t} \frac{D_{2,s}}{D_{2,t}} \delta_1 e^{-\delta_1 (t-s) + \sigma_{1,D}(B_{1,t} - B_{1,s})} ds. \tag{39}$$

When type-1 trees are small compared to type-2 trees $\left(\frac{\phi_1}{\phi_2} \approx 0 \right)$, aggregate consumption, $D_{1,t} + D_{2,t}$, is approximately equal to the aggregate dividends of the large, type-2 trees, and therefore aggregate consumption follows a geometric Brownian motion. The implication is that the interest rate and the risk premium for type-2 trees both converge to constants as the ratio $\frac{\phi_1}{\phi_2} \to 0$ goes to zero.

In the baseline model, entry and exit of investors into the single stock market was tied to the arrival and departure of agents in the economy and was essentially exogenous. The extension to two risky stocks requires that we model the entry and exit into the market for stock 1. Specifically, we assume that investors of both types (R and I) gain and lose interest in stock 1 at the rate χ per unit of time dt. Of the arriving investors a fraction ν is of type R, as in the baseline model. In addition, investors may choose to exit because they incur a small, non-pecuniary, disutility flow, ε, from paying attention to stock 1. Specifically, an investor of type $i \in (I, R)$ chooses to keep paying attention to stock 1 if and only if her expected discounted utility from remaining attentive to stock 1 is above the present value of the disutility cost of attention, ε.

Assuming that ε is sufficiently close to zero, it is irrelevant for investors of type I: these investors' perceived benefit from being able to invest in stock 1 is bounded away from zero. For investors of type R, however, there are regions of ω_t where the optimal holding of stock 1 is zero, and even a small disutility can lead them to exit the market.

Formally, the net value that an investor of type R derives from being in the the market

for stock 1 equals

$$V^R(\omega_t) \equiv \mathrm{E}_t\left[\max_{w_u,T}\int_t^T e^{-\rho(u-t)}\left(w_u\left(\mu_{1,u}-r_u+\lambda_u(w_u)\right)-\frac{1}{2}\left(w_u\sigma_{1,u}\right)^2-\varepsilon\right)du\right], \quad (40)$$

where $T \geq t$ is the stochastic time of exit from the market for stock 1 (be it endogenous or exogenous). Equation (40) uses the assumption of logarithmic preferences — along with the simplifying assumption that stocks 1 and 2 are independent — to express the net expected utility gain from continued presence in market 1 as the increase in investor R's logarithmic growth rate of wealth, $w_{1,u}^R\left(\mu_u - r_u + \lambda_u^R\right) - \frac{1}{2}\left(w_{1,u}^R\sigma_u\right)^2$, net of the flow disutility ε of presence in the market. The investor strictly prefers to remain in the market if and only if $V^R(\omega_t) > 0$. This requirement implies that, for given equilibrium functions $\kappa(\omega_t)$ and $y(\omega_t)$, there is a critical boundary $\bar{\omega}$, typically lying in the region of ω_t where $w_{1,u}^R(\omega_t) = 0$, that acts as a "reflecting barrier" for ω_t. Specifically, if the process ω_t were to ever exceed $\bar{\omega}$, there would be enough exit to restore ω_t to $\bar{\omega}$.[24]

Some further technical assumptions on investor entry and exit are detailed in Appendix C. In that appendix (Proposition 6) we also show that ω_t is a Markov diffusion (reflected at $\bar{\omega}$). The price-dividend ratio of stock 1 is a function of ω_t, and can be obtained as the solution of a non-linear ordinary differential equation (ODE), which has to be solved numerically. In the remainder of this section we discuss the (numerical) solution of this ODE.

Figure 4 presents the solution for the price-dividend ratio. We are interested in situations where the disagreement is large ($\eta = 0.9$), and the speed of investor churn in market 1 is quite large ($\chi = 2$), to capture short-termism. The idiosyncratic dividend volatility is not too large, $\sigma_{1,D} = 7\%$, and the shorting fee is at the high levels that one encounters for stocks that are "on special" ($\varphi = 5.7\%$). A proportion $\nu = 0.7$ of new investors are of type R. In equilibrium, this value of ν ensures that the endogenous exit decision is meaningful, that is, under any equilibrium there would a possibility that ω_t "spends time" in a region where a zero holding of asset 1 is optimal for investor R. Finally, we assume that the sum of interest rate and depreciation for stock 1, $r + \delta_1$, is 0.1. We choose a value of $\tau = 0.8$ based on the industry practice of rebating about 80% to the mutual funds that provide their shares for

24. This behavior is reminiscent of models of industry equilibrium with endogenous entry and exit (e.g., Leahy (1993), Baldursson and Karatzas (1996).)

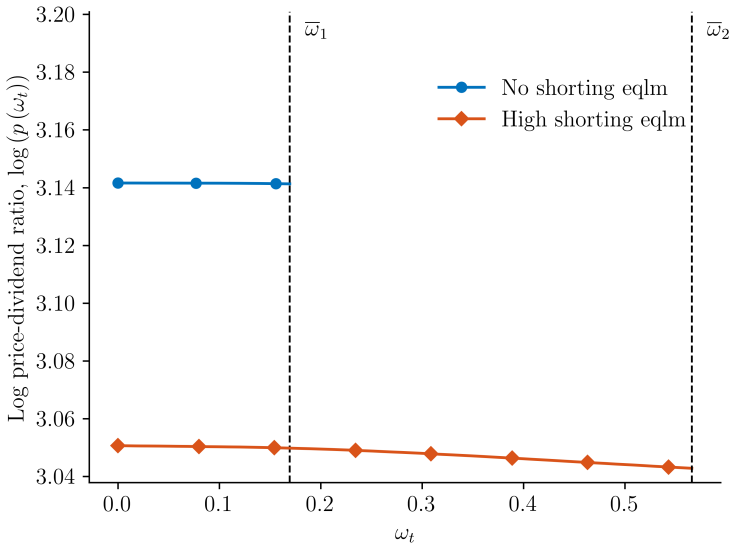


Figure 4: The price-dividend ratio in the equilibria involving the highest and the lowest (zero) extent of shorting.

lending.[25] Finally, for the disutility ε we intentionally choose a very small amount (2.5 basis points on an annual basis).

Figure 4 shows the price-dividend ratio under two different assumptions on the equilibrium that investors coordinate on (when multiple equilibria are possible). Specifically, the line "zero shorting" assumes that investors always coordinate on the equilibrium with zero shorting, if it exists. By contrast, the line "high shorting" assumes that investors always coordinate on the equilibrium with the highest possible shorting. Note that both lines extend only until the levels $\bar{\omega}_1$ and $\bar{\omega}_2$, respectively, which are the levels of ω_t at which R investors exit in the two equilibrium with zero and high shorting, respectively.

There are several noteworthy features of Figure 4. First, the price-dividend ratio for the zero shorting equilibrium is *higher* than the price-dividend ratio for the high shorting equilibrium. This may seem counter intuitive, since the high shorting equilibrium implies a lower Sharpe ratio *for a fixed* ω_t. The reason why the price-dividend ratio is higher in the zero shorting equilibrium is that the dynamics of the wealth shares of I and R investors differ depending on whether the economy coordinates on the high or zero shorting equilibrium. We already showed in Section 3 that, when the economy coordinates on the high shorting

25. Source: "Unlocking the potential of your portfolios: iShares Security Lending." Blackrock, 2021. Available at `https://www.ishares.com`.

equilibrium, the wealth dynamics favor R investors. As a result, their stationary wealth shares are higher, which in turn tends to raise the stationary Sharpe ratio.

Participation costs accelerate these wealth dynamics: To see this, suppose that the market coordinates on the high shorting equilibrium and that $\omega_t > \overline{\omega}_1$. If a coordination shock shifts the economy to the zero shorting equilibrium, short sellers exit instantaneously until $\omega_{t^+} = \overline{\omega}_1$, where ω_{t^+} is the value of ω_t after the equilibrium shift. At the new value $\omega_{t^+} = \overline{\omega}_1$, the shorting market is still inactive ($\overline{\omega}_1 > \omega_1^*$) and the remaining short sellers pay the (flow) participation cost despite holding a zero position in stock 1. However, the exit of a sufficient number of short sellers has increased the probability that future values of ω_t will be sufficiently low to activate the shorting market again. (Recall that for sufficiently low values of ω_t, the equilibrium is unique and involves shorting.) This increased probability of an active shorting market incentivizes the R investors to keep paying the participation cost.

In terms of quantities, Figure 4 implies that an unanticipated shift in equilibrium (from the "high shorting" to the "zero shorting") will make the price-dividend ratio jump upward by about 10%. As we discuss in Section 6.4, this value is similar in magnitude to the monthly return on a broad "betting against the shorts" strategy during the period covering November 2020 to January 2021, a time when short sellers abandoned their positions abruptly.

6 Empirical Evidence

6.1 Overview

The novel intuition of our model is the presence of a feedback effect between utilization (y) and the Sharpe ratio (κ). This feedback loop can lead to equilibrium shifts that empirically manifest themselves as a jump in utilization, irrespective of how agents coordinate on one equilibrium or another (Remark 2). Moreover, Proposition 4 derived a key condition for the possibility of such jumps to occur, namely that the function $h(y) = f(y)(1 - \tau y)$, which reflects the difference between the fee paid by the short investor and the income received by the long investor, be declining for some y.

The next section shows that the assumption of a declining $h(y)$ over some range of y is

empirically plausible. In addition, we present evidence that the time series for utilization does exhibit jump-like behavior. Moreover, the incidence of utilization jumps for a given stock is significantly correlated with whether the estimated $h(y)$ (for that stock) has a declining segment.

6.2 Data description

Daily returns and market capitalization data are from the Center for Research in Security Prices (CRSP). Our source for stock lending fees and short interest is IHS Markit. These data start in January 2006.[26] Markit collects self-reported data on actual rates on security loans from active participants in the securities lending market. The data set covers roughly 30,000 securities, and contains 16 million unique stock-day observations.

We match the Markit data to the CRSP database and retain only common stocks of domestic companies. Furthermore, to ensure that our results are not driven by micro-cap stocks, for our main empirical results we only retain observations that correspond to stocks that are Russell 3000 constituents (on the day of observation), which we identify using the Datastream Monthly Index Constituents file. This reduces our number of observations to 10 million.

We follow Daniel, Klos, and Rottke (2018) and use the quantity "Indicative Fee" as our measure of the marginal cost of borrowing, which is the expected borrowing cost (expressed in percentage points per year) on a given day.[27] In addition to these data on fees, we use two additional data variables from Markit: a) "Daily Cost of Borrow Score" (DCBS) and b) daily utilization. The DCBS takes integer values between one and ten and is a "bucketed score (1-10) that reflects the cost to borrow the stock charged by the lenders from the Prime

26. The Markit data of other studies (Daniel, Klos, and Rottke (2018) and Drechsler and Drechsler (2014a)) starts in 2004 and contains observations at a weekly frequency. The data set that was provided to us by Markit contains daily observations that start in 2006. Markit confirmed in an email that the pre-2006 data are no longer available.

27. Markit uses both borrowing costs between Lenders and Prime Brokers to produce this estimate of the current market rate. As discussed in Daniel, Klos, and Rottke (2018), the Indicative Fee can be interpreted as a proxy for the marginal cost of short selling. Markit also reports the "Simple Average Fee", which is the average fee over all outstanding contracts for a particular security. Following Daniel, Klos, and Rottke (2018), on each-stock day, we take the Indicative Fee as our measure of the stock's lending fee and (in the very rare instances) where the Indicative Fee is not reported, we use the Simple Average Fee. This substitution applies to only 676 observations out of the roughly 10 million observations.

Brokers in the wholesale market, where 1 reflects a cheap or a GC ("general collateral") stock and 10 reflects an expensive or a special stock." [28] The literature has used this score as a way of identifying stocks that are on special. In the data, DCBS values equal to one are by far the most prevalent ones (74% of our sample) and tend to exhibit a high degree of persistence. [29] For some of our empirical results, we focus on stocks that are hard to borrow and we drop observations with DCBS equal to 1, since our model applies predominantly to stocks where lending frictions are non-trivial. Markit's "Utilization by Quantity" is computed as the fraction of assets on loan from lenders divided by the total lendable quantity. This variable takes values between 0 and 1 and corresponds to the variable y_t in our model.

6.3 Figures and tables

Tables E.1 and E.2 in Appendix E provide some summary statistics on the lending fees. To expedite the presentation of the results that pertain to our paper, here we simply summarize our main findings from these tables as follows: When we sort stocks into five quintiles by market capitalization the median lending fee ranges between 0.35% per annum (for stocks in the large market capitalization quintile) to 0.41% per annum (for the lowest market capitalization quintile). However, lending fees exhibit substantial cross-sectional and time-series variation. The 99^{th} percentile of all lending-fee observations exceeds 7% for stocks in four out of the five market-capitalization quintiles. When we examine lending fees at the individual stock level, we find that 31% of Russel 3000 constituents exhibit a lending fee in excess of one percent for 5 out of 100 trading days, while 18% of Russel constituents exhibit lending fees in excess of 3% for 5 out out of 100 trading days. In addition, 45% of Russel constituents exhibit a fee in excess of 5% at some point in the sample. This is consistent with results reported in Engelberg, Reed, and Ringgenberg (2018), who write "loan fees can increase to levels that significantly decrease the profitability of nearly any trade."

We start the presentation of our main empirical findings with Figure 5, which examines the relationship between utilization and shorting fees. Specifically, the solid line pools all

28. IHS Markit Securities Finance Quant Summary, July 2020 edition. Available on WRDS.

29. If a stock has a DCBS value of one on any given day, the probability that it has a DCBS value of one the next day is 98.83%

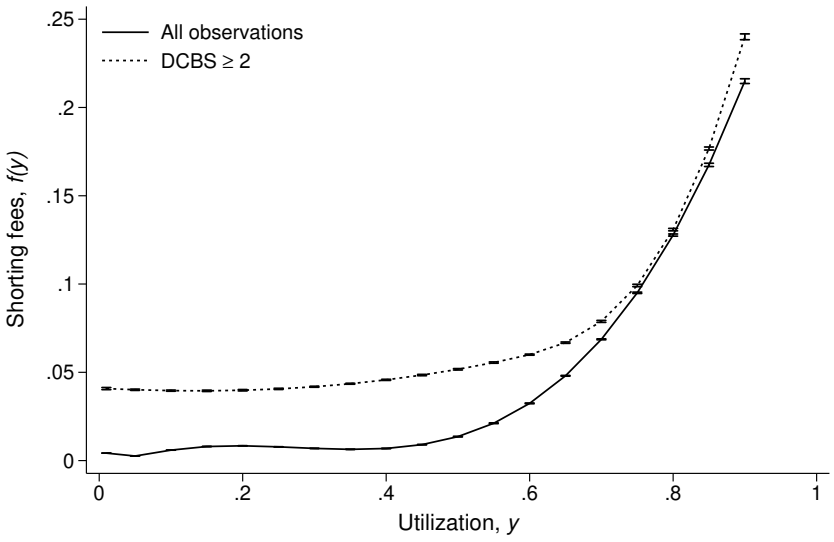
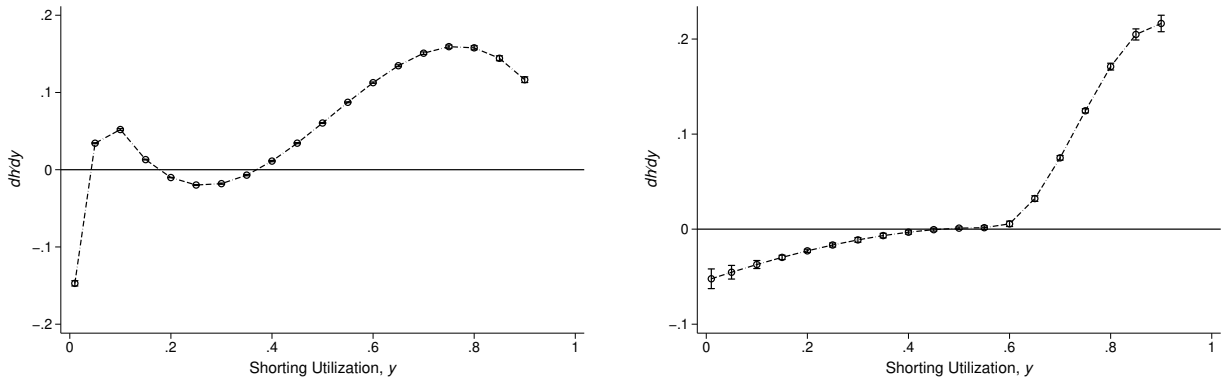


Figure 5: Relationship between shorting fees and utilization. Non-parametric series regression of daily shorting fees on utilization, pooled across Russell 3000 constituents. Daily shorting fees from 2006 to 2021 are from Markit and are reported as annualized percentage rates. (For instance, 0.05 on the y axis means 5% per annum.) Error bars denote 95% confidence intervals. Because this estimation utilizes several millions of observations, the standard errors of the estimates are negligible.

daily observations across all Russel-3000 stocks and depicts the estimates of a non-parametric regression of daily shorting fees (expressed in annual percentage terms) on utilization.[30] The dashed line depicts results from the subsample that only includes observations of stocks with a DCBS code of 2 or above, stocks that we refer to as being on special. As both plots show, the relationship between shorting fees and utilization is non-linear, with a region that is approximately constant for low and intermediate values of utilization and a steeply increasing region for high values of utilization.

We next use the estimates from the previous non-parametric regression analysis to compute estimates of $h'(y)$. Specifically, using the non-parametric point estimates and corresponding standard errors for $f(y)$ and $f'(y)$, and the formula $h'(y) = f'(y)\,(1 - \tau\, y) - f(y)\tau$, we calculate $h'(y)$. We present the estimated $h'(y)$ in Figure 6, along with an upper-bound

30. We estimate a third order basis spline of fees on utilization and depict the point estimates and standard errors at 0.05 increments of utilization, along with standard errors. For the computations we use the command (`npregress series`) in Stata. Standard errors are produced using the command `margins` and the Δ-method.



(a) Full sample (b) DCBS$_{i,t} \geq 2$

Figure 6: $h'(y)$, pooled non parametric estimates. Estimated marginal effects are derived from a non-parametric series regression of daily shorting fees f on utilization, y. Marginal effects are computed using the formula $h'(y) = f'(y)\,(1-\tau\,y) - \tau f(y)$. Short interest is based on utilization data from Markit. Sample consists of daily observations of shorting fees and utilization, pooled across Russell 3000 constituents from Markit for the period 2006 to 2021. Standard errors are derived from the standard-error estimates of $(1-\tau y)\hat{f}'(y)$ and $\tau\hat{f}(y)$, while assuming a worst-case correlation of -1 between these two quantities. The value τ is set to 0.8.

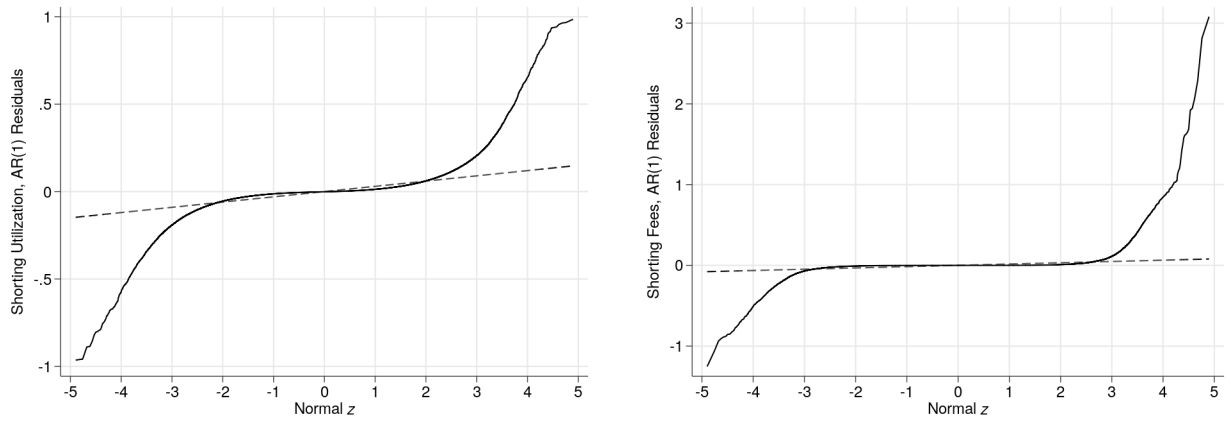
estimate of the 95% confidence interval.[31] The figure shows that $h'(y)$ is statistically significantly negative for several points between 0 and 0.4. Therefore, when we pool all observations, we can statistically reject the null hypothesis that $h'(y)$ is always positive.

The two plots of Figure 6 pool all observations together and estimate a single non-parametric regression to obtain more precise estimates. As a robustness check, in Appendix E we estimate a separate non-parametric regression between fees and utilization for each stock and compute a stock-specific $h'(y)$. Appendix E shows that the (cross-sectional) average of the estimated $h'(y)$ is negative (and statistically significant) for low values of y.[32]

Our next set of results pertains to the implications rather than the assumptions of our model. One empirical implication of the model is that utilization follows a jump-diffusion process. When there is no equilibrium change, utilization follows a diffusion process. An

31. The standard errors are computed using Stata's estimates for the variance of the estimates $f(y)$ and $f'(y)$ and a worst-case assumption that the correlation between the estimates of $\hat{f}$ and $\hat{f}'$ is -1 to provide an upper bound on the variance of the estimate.

32. As can be expected, the estimation of a separate function $h'(y)$ for each stock increases the estimation-error bounds on $h'(y)$ and therefore the range of (statistically significant) negative y-values becomes smaller than in Figure 6.



(a) Utilization (b) Lending fee

Figure 7: QQ-plots of weekly changes in utilization and lending fees. Left panel: An AR1 of utilization is estimated at the stock level at a weekly frequency. The residuals of each time series are divided by their standard deviation and then pooled across all stocks. The quantiles of the standardized-residual distribution are then plotted against the quantiles of the standard normal distribution. Right panel: Same as the left panel, but for lending fees rather than utilization. Both utilization and lending fee data from Markit over the period 2006 to 2021.

occasional equilibrium shift causes an (upward or downward) jump in utilization.

Figure 7 provides an informal way to visualize abrupt shifts in utilization in the data. For each Russell 3000 stock, we estimate a separate AR1 process for weekly utilization, so that both the long-run mean and the persistence of utilization can vary at the stock level. We then normalize the innovations (i.e., the residuals of the AR1 estimation) by their standard deviation. Assuming that utilization (at the stock level) follows an AR1 process with normal, homoskedastic increments, one would expect these normalized residuals to follow a standard normal distribution. The left panel of Figure 7 shows that this is not the case. The quantile-quantile plot of the standardized residuals clearly shows that the innovations to utilization exhibit remarkably fat tails with a non-trivial mass of the residuals in the range of 10-20 standard deviations. The kurtosis of these residuals is 78 as opposed to 3 for a standard normal.[33,34] We note that utilization is not the only fat-tailed time series. The right subplot of Figure 7 shows that weekly changes in the lending fee are also quite fat-tailed.

33. The Jarque-Bera test rejects normality with a p-value essentially equal to zero.

34. The test proposed by Aït-Sahalia and Jacod (2009), which tests whether the discretely observed utilization data emanated from a continuous-sample-path diffusion process using daily data, rejects the null hypothesis of continuous sample paths for 85% of Russel 3000 constituents.

Table 1: Regressions of Jump Rate on stock-level estimates of $h'(y)$

	Annualized Jump rate					
	$\lvert\Delta y\rvert \geq 5.5\%$		$\lvert\Delta y\rvert \geq 8.0\%$		$\lvert\Delta y\rvert \geq 10.0\%$	
Panel A: Large changes in Shorting Utilization						
$\mathbf{1}_{h'<0}$	2.339**		2.269***		2.171***	
	(4.004)		(3.823)		(3.625)	
$\mathbf{1}_{\text{Reject } h'>0}$		1.758**		1.524**		1.364*
		(3.087)		(2.636)		(2.345)
N	2249	2249	2249	2249	2249	2249
Panel B: Large changes in Shorting Utilization driven by changes in Shorting Demand						
$\mathbf{1}_{h'<0}$	1.169***		1.152***		1.113***	
	(3.560)		(3.461)		(3.313)	
$\mathbf{1}_{\text{Reject } h'>0}$		0.895*		0.787*		0.703*
		(2.823)		(2.437)		(2.156)
N	2249	2249	2249	2249	2249	2249

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Jump rate is calculated as annualized rate of detected jumps in utilization. Jumps are identified as trading weeks during which the absolute change in utilization exceeds, depending on the specification, 5.5%, 8%, or 10%. h' is based on the stock-level estimate of $h(y)$ from a non-parametric kernel regression of $h(y)$ on utilization y. h' is estimated at 5%, 10%, 20% ... 90%, and 95% and the stock-level estimate of h' is the minimum of these estimated marginal effects. $\mathbf{1}_{h'<0}$ is an indicator variable taking a value of 1 when the stock-level estimated minimum-h' is negative. $\mathbf{1}_{\text{Reject } h'>0}$, is an indicator variable taking a value of 1 when the stock-level estimated minimum h' is sufficiently low to reject the null hypothesis that $h'(y) > 0$ for all y, after applying a Bonferroni correction to account for multiple hypothesis testing.

Turning to the cross section of stocks, Table 1 shows that stocks where $h'(y) < 0$ for some y tend to exhibit a larger incidence of unusually big weekly changes in utilization. Specifically, we perform the following exercise. We fix a cutoff U above which we consider the weekly AR1 residual in a stock's utilization as economically large. We refer to a week on which the absolute value of the change in utilization exceeds U as a jump event.(Results are quite similar, if we consider the raw weekly changes in utilization rather than the AR1 residuals.) These cutoffs U correspond to unusually large (larger than 2 standard deviations) weekly changes in utilization. We define the jump rate as the ratio of the number of jump events

to the total number of weeks over which we observe the stock, which we then annualize for ease of interpretation. We then regress the obtained stock-level jump rate on two indicator functions. The first indicator function takes the value 1 for stock j if the estimated function $h'(y)$ for stock j exhibits at least one negative point estimate.[35] The second indicator function takes the value 1 for stock j if we can statistically reject that $h'(y) > 0$ for some value of y. Given our interest in stocks that are on special, we confine attention to Russell 3000 stocks that have a DCBS score larger than one for at least 50 trading days, yielding a sample of 2249 stocks.

The bottom panel of Table 1 performs a similar exercise to the top panel, except that in order to define a jump in utilization we impose an additional property of the model: we confine attention to jumps in utilization, whereby the absolute value of the percentage change in shorted shares (the numerator of utilization, y) is larger than the absolute value of the percentage change in lendable shares (the denominator of utilization, y).[36] The table shows that stocks for which $h'(y) < 0$ exhibit a higher incidence of jump events.

Inside the model, stocks with uniformly positive $h'(y)$ should not exhibit jumps, whereas stocks with some negative values of $h'(y)$ might. Of course, because of measurement error in $h'(y)$ our procedure will mis-classify some stocks as not satisfying $h'(y) < 0$, when in fact they do, and vice versa. We may also mis-classify jump events. Because of these unavoidable measurement errors, when we confront the model with the data, we should expect to find that stocks for which the estimated $h'(y)$ has a negative region have — on average — a higher incidence of jumps. This is indeed what Table 1 shows.

As a robustness check, in Appendix E (Table E.4) we also discuss a version of Table 1 using a more sophisticated econometric procedure to identify jump events. Specifically, we use a jump-robust, rolling estimator of volatility to allow us to disentangle jumps from just unusually volatile periods for each stock. This more sophisticated procedure produces the same results as the more intuitive approach of defining a jump simply as an unusually

35. For the estimation of $h'(y)$ at the stock level, we use the kernel estimator described in Appendix E.

36. Inside the model, utilization y is equal to the ratio of shorted shares to lendable shares; in turn lendable shares are equal to one plus the absolute value of shorted shares (this is the market clearing condition). Therefore, the absolute value of the percentage change in shorted shares must exceed the absolute value of the percentage change in lendable shares. By restricting attention to jumps that satisfy this requirement, we prevent that a change in utilization could be driven by —say— some institutional client giving permission to short shares.

large change in weekly utilization. Further, as a "sanity check," Table E.3 in Appendix E confirms that the incidence of utilization jumps is higher for stocks that are harder to borrow, consistent with the idea that jumps in utilization occur when shorting frictions are non-trivial.

An additional robustness check in Appendix E pertains to the estimation of $h'(y)$. In the model the only shocks are exogenous dividend shocks, which cause shifts in the wealth distribution, the Sharpe ratio, and the demand for shorting shares. Further, the relation between the fee and the utilization is deterministic: $f_t = f(y_t)$. In the data, there is a residual ε_t in that relation, $f_t = f(y_t) + \varepsilon_t$, which throughout this section we have treated as (orthogonal) measurement error in the lending fee. However, if one were to think of this ε_t as a non-orthogonal supply shock, the empirical estimates of $f'(y)$ could be biased upwards or downwards. The model offers a relatively simple approach to consistently estimate $\frac{\Delta f_t}{\Delta y_t}$ even in the presence of shocks to the lending fee. The idea is to exploit the discontinuities that occur around equilibrium shifts:[37] Assuming that dividend shocks and the shock to the lending fee, ε_t are continuous processes, jumps in y_t can only be the result of an equilibrium shift. Thus, if we assume that the residual process ε_t is continuous, jumps in y_t offer an opportunity to measure $\frac{\Delta f_t}{\Delta y_t} = \frac{f_{t^+} - f_{t^-}}{y_{t^+} - y_{t^-}} = \frac{f(y_{t^+}) - f(y_{t^-}) + \varepsilon_{t^+} - \varepsilon_{t^-}}{y_{t^+} - y_{t^-}} \approx f'(y_{t^-})$, where we used the continuity assumption $\varepsilon_{t^+} = \varepsilon_{t^-}$. In other words, around jump-events in utilization, one is able to identify $f'(y_t)$. Figure E.1 in the appendix estimates $\frac{\Delta f_t}{\Delta y_t}$ by evaluating changes in the numerator and denominator on the weeks where y_t exhibits jump events, as identified earlier. The figure shows that the derived function $h'(y)$ is small and negative for all values of y. This suggests that our conclusion $h'(y) < 0$ for some y is not the result of a bias due to endogeneity issues.

6.4 The retreat of short sellers between November 2020 and January 2021: A case study

Our theory provides a useful lens for interpreting the dramatic events that occurred in the shorting market over the three-month period covering November 2020 to January 2021. We

37. The idea is reminiscent of how Sweeting (2006) uses multiple equilibria to resolve identification issues.

start by documenting the historically dismal performance of shorting strategies over this period and the contemporaneous correlated drop in short interest across a large number of stocks. While the press focused attention on a single stock (namely GameStop), which experienced a coordinated short squeeze fueled by retail investors in January 2021, we show that the retreat of short sellers preceded the short squeeze on Gamestop by about two months, was quite broad, and occurred among stocks that neither experienced a significant change in retail trading volume, nor were the topic of intense online discussion (as was the case with GameStop). We conclude, therefore, that the very poor performance of shorting strategies was the result of an abrupt shift in the behavior of short sellers, rather than the result of coordinated short squeezes by retail investors, or the result of contagion from their losses in meme stocks.

To start, in Figure 8 we plot the cumulative returns to an equal-weighted portfolio that bets against the shorts. The portfolio is long the top decile of Russell 3000 stocks, ranked by short interest, and short the broad market. To construct this portfolio's return, we use stock return data are from CRSP and short interest data from the SEC. (Since in this section we are interested in historical comparisons, we use data on short interest from the SEC, which starts in 1973, but is available only at a monthly frequency for the full sample.[38,39]) The figure shows that the betting-against-the-shorts strategy is not particularly profitable (or unprofitable) from late June to mid-November 2020, but becomes strikingly profitable over the following three months.

To put this evidence in historical perspective, in the left panel of Figure 9 we plot a histogram of the monthly returns of this betting-against-the-shorts strategy since the beginning of the sample (1973). The left panel of Figure 9 shows that the November 2020 and January 2021 returns are the highest and second-highest (respectively) in the historical sample. (December 2020 is also in the top decile of the historically observed returns.) Figure G.1 in Appendix G further shows that November 2020 and January 2021 remain outliers if we exclude tickers that were heavily discussed online (on the WallStreetBets subreddit[40]), if

38. For additional details on the construction of this portfolio, see Appendix E.

39. SEC data on short interest are available every two weeks since January 2007, but not before.

40. The popular stocks on the WSB subreddit in January 2021 were: AMC, BBBY, GME, SPCE, TLRY, and TSLA.

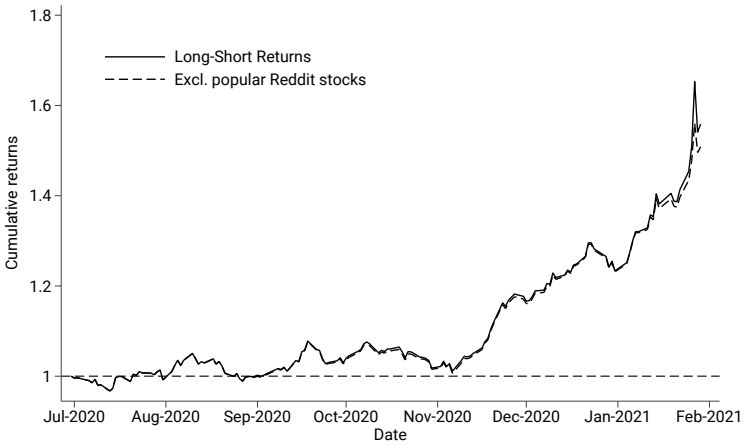


Figure 8: Cumulative Returns (July 2020 – February 2021). Cumulative returns on a strategy long the top decile of stocks by short interest and short the broad market. The cumulative returns to an equal-weighted long-short portfolio are shown by the solid black line. The returns to the same portfolio, but excluding the six most-discussed tickers on the Wallstreet-Bets subreddit (AMC, BBBY, GME, SPCE, TLRY, and TSLA), are shown in the dashed black line.

we only include S&P500 constituents (i.e., larger stocks) in the formation of the long leg of the portfolio, and if we value-weight rather than equal-weight returns.[41]

In the right panel of Figure 9, we plot the histogram of monthly innovations in short interest for highly shorted stocks. For each stock, we estimate a separate AR1 model for short interest using monthly data from January 1973 to October 2020.[42] We then extract residuals for the full sample ending in mid-February 2021 and standardize each stock's residuals by their standard deviation. Then for each month we identify the top decile of stocks in the Russell 3000 by short interest, and compute the (cross-sectional) average of the standardized residuals of these stocks for that month. As can be seen in the histogram, the three months beginning in the middle of November 2020 and extending through the middle of February 2021 saw unusually negative realizations of short interest. January 2021 was the second-most negative realization in the sample, while November was the sixth most negative in the

41. Table G.1 in Appendix G presents the results of formal statistical tests (also controlling for Fama-French factors) of whether the returns in November 2020, December 2020, and January 2021 are statistically different from the average return on the betting-against-the-shorts strategy over the full 48-year sample.

42. While data on short interest is available on a bimonthly frequency (i.e., every two weeks) from the SEC starting in January 2007, we use monthly data to include as long a sample as possible and keep results comparable across the sample. The SEC data reports short interest in stocks as of the middle of each calendar month.

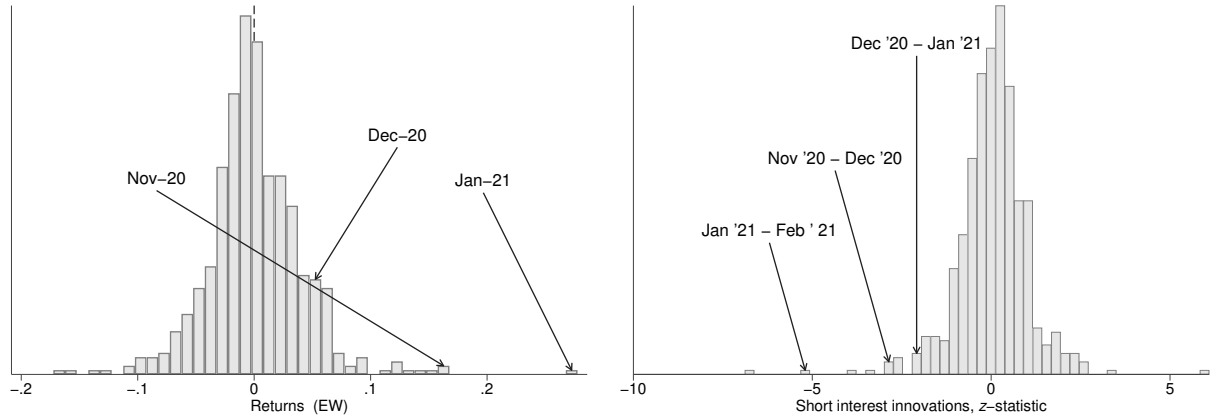


Figure 9: Left panel: Histogram of monthly returns (1973–2021). Equal-weighted, monthly returns on a portfolio long stocks in the top decile of short interest and short the market index. The arrows indicate the portfolio returns in the months of November and December 2020 and January 2021. Right panel: Histogram of monthly innovations in short interest (1973–2021). We estimate a separate AR1 process for short interest stock-by-stock, extract the residuals, and divide them by their (stock-specific) standard deviation. For each month, we compute the cross-sectional mean of standardized residuals for the top decile of Russell 3000 stocks sorted on short interest, and depict the data as a histogram. Short interest is as of the middle of each month and the arrows indicate mid-month to mid-month short-interest innovations for the period between November 15, 2020 and February 15, 2021.

sample.[43] Much like the abnormal returns to betting-against-the-shorts, the decline in short interest began prior to the meme stock events of January 2021.

The bad returns to short selling are not ripple effects of the short squeeze on GameStop, which attracted a lot of attention in the press: November 2020 was already the historically best return for the betting against the shorts strategy recorded up until that point, even though the online discussion surrounding GameStop did not start until mid-January 2020. To illustrate this point, in Figure G.2 (Appendix G), we plot the daily submissions to the WSB subreddit (which was the online forum where users posted their opinions on Gamestop and other meme stocks) on a logarithmic scale. The graph shows that the explosive growth of online submissions occurred in early January 2021; November 2020 does not stand out.

In the case of GameStop, the online discussion was highly correlated with retail trading volume. In Figure G.3 in the Appendix we plot online mentions of Gamestop on the WSB subreddit against retail trading volume, which we measure in TAQ data with the method-

43. The most negative realization occurred during the Great Financial Crisis.

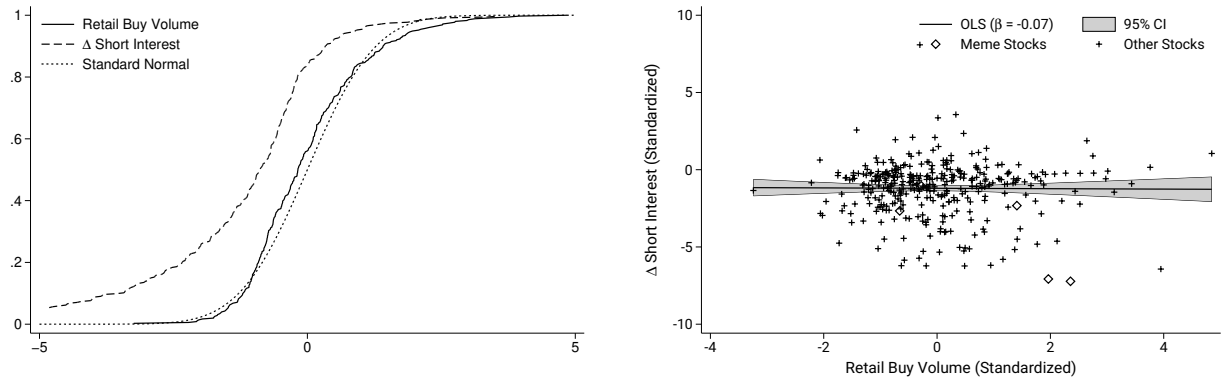
ology of Boehmer et al. (2020).[44] Both time-series are at an hourly frequency and exhibit strikingly strong comovement (0.93 rank correlation).

While for GameStop there is a clear spike in retail purchase volume, a remarkable feature of the data is that short sellers retreated across a wide range of stocks even though these stocks did not experience any unusual patterns in retail trading volume. Figure 10 plots the univariate distributions and the scatterplot of (a) changes in short interest and (b) retail purchase volume as a fraction of total volume for the most shorted stocks (top decile of stocks) ranked by short interest as of January 15, 2021. The two quantities (a) and (b) are reported as standardized z-scores using TAQ and SEC data from January 2015 through January 2021 to compute means and standard deviations. The distribution of the retail-purchase volume to total volume is centered around zero, with most values in a [-2, 2] range. By contrast, the distribution of changes in short interest is overwhelmingly negative, with most values in the [-5 0] range. This indicates that January 2021 was not an unusual month for the ratio of retail-purchase to total volume for these highly shorted stocks. This is in sharp contrast to the behavior of short interest, which saw a large decline across most of the stocks in that month. In addition the relation between the two quantities is flat, as the scatter plot in Figure 10 illustrates.

To interpret the events through the lens of theory, the fact that short sellers retreated before the events surrounding GameStop precludes the possibility that their retreat was a balance-sheet style contagion in reaction to losses they suffered in GameStop.[45] The fact that retail purchase volume did not change for the large number of stocks that saw declines in short interest suggests that there was no dramatic inflow of optimistic investors in these markets, as was the case for GameStop. We therefore favor the interpretation that short sellers retreated because of fears that led them to abandon their strategies in run-type fashion. One possibility is that some early signs of shifts in retail purchase volume for a few stocks may have acted as a coordination shock that led to a shift in equilibrium, resulting in a drop of short interest, a decline in the participation of short sellers, and a higher price for the previously shorted stocks — consistent with the empirical patterns of that period. We

44. Appendix F contains details on (a) how we identify online mentions, and (b) how we identify retail trading.

45. See, e.g., Kyle and Xiong (2001).



(a) Marginal Distributions

(b) Joint Distribution

Figure 10: Retail purchase volume (as a fraction of total volume) and change in short interest, January 2021. Both quantities are reported as standardized z scores using TAQ and SEC data (respectively) from January 2015 through January 2021 to compute means and standard deviations. Panel (a) plots the empirical cumulative distribution of the two quantities, alongside a standard normal for reference. Panel (b) plots their joint distribution, along with the line of best fit. Tickers that were popular discussion topics on WSB and that are also in the top decile of short interest are indicated with "◊", while all other tickers are indicated with "+".

should also note that the seemingly simpler explanation of interpreting the broad declines in short interest as a correlated increase in irrationality (η) across stocks, but assuming a unique equilibrium, would run into the problem that $\frac{dy_t}{d\eta}$ is positive rather than negative.[46] Increased irrationality would therefore lead to a higher level of short interest.

A concluding empirical observation is that the decline in short interest did not show any signs of reverting to its old levels in the six months that followed January 2021 (Figure G.4 in Appendix G). This suggests that the short-seller retreat was not just a transient reaction to let the "dust settle."

7 Conclusion

Shorting can exhibit run-type patterns. An event that prompts some short sellers to abandon their short positions can ignite a self-propagating cycle: Less shorting also implies less lending income for investors with long positions, who now need to be compensated with a higher

46. See the proof of Lemma 2.

expected return, which in turn further prompts short sellers to abandon their strategies. Thus, for the same fundamentals there can be multiple equilibria — differently phrased, the self-reinforcing nature of shorting decisions gives rise to a backward-bending demand for the asset.

Our model also provides a rationale for the simultaneous occurrence of declining short interest and rising stock prices, a phenomenon we document in our empirical analysis. At first sight, it would appear that a rise in the stock price (absent a change in fundamentals or lending fees) should attract rather than repel short sellers. We show, however, that when investors coordinate on a no-shorting equilibrium the incentive for likely shorters to participate in the market becomes low enough to prompt them to "abandon" the asset to the optimists.

We further identify a general condition on the relation between fees and utilization that is necessary for the existence of multiple equilibria, and confirm empirically that stocks that satisfy this condition are more likely to experience jump-like behavior in utilization. Finally, we argue that the retreat of the short sellers and the dismal performance of shorting strategies between November 2020 and January 2021 may have occurred as a result of self-propagating fears among short sellers that led them to abandon the market to the optimists, in line with our theoretical results.

References

Abreu, Dilip, and Markus K Brunnermeier. 2002. "Synchronization risk and delayed arbitrage." *Journal of Financial Economics* 66 (2): 341–360.

Aït-Sahalia, Yacine, and Jean Jacod. 2009. "Testing for jumps in a discretely observed process." *The Annals of Statistics* 37, no. 1 (February).

Allen, Franklin, Marlene Haas, Eric Nowak, Matteo Pirovano, and Angel Tengulov. 2021. "Squeezing Shorts Through Social Media Platforms." *History of Finance eJournal.*

Asquith, Paul, Parag A. Pathak, and Jay R. Ritter. 2005. "Short interest, institutional ownership, and stock returns." *Journal of Financial Economics* 78 (2): 243–276.

Atmaz, Adem, Suleyman Basak, and Fangcheng Ruan. 2020. "Dynamic Equilibrium with Costly Short-Selling and Lending Market." Working paper.

Baldursson, Fridrik M, and Ioannis Karatzas. 1996. "Irreversible investment and industry equilibrium." *Finance and Stochastics* 1 (1): 69–89.

Banerjee, Snehal, and Jeremy J. Graveline. 2013. "The Cost of Short-Selling Liquid Securities." *Journal of Finance* 68 (2): 637–664.

Beneish, Messod Daniel, Charles MC Lee, and D Craig Nichols. 2015. "In short supply: Short-sellers and stock returns." *Journal of Accounting and Economics* 60 (2-3): 33–57.

Benhabib, Jess, and Roger Farmer. 1999. "Indeterminacy and sunspots in macroeconomics." Chap. 06 in *Handbook of Macroeconomics,* edited by J. B. Taylor and M. Woodford, vol. 1, Part A, 387–448.

Biais, Bruno, Johan Hombert, and Pierre-Olivier Weill. 2021. "Incentive Constrained Risk Sharing, Segmentation, and Asset Pricing." *American Economic Review, forthcoming.*

Blanchard, Olivier J. 1985. "Debt, Deficits, and Finite Horizons." *Journal of Political Economy* 93 (2): 223–247.

Blocher, Jesse, Adam V Reed, and Edward D Van Wesep. 2013. "Connecting two markets: An equilibrium framework for shorts, longs, and stock loans." *Journal of Financial Economics* 108 (2): 302–322.

Boehmer, Ekkehart, Charles M Jones, and Xiaoyan Zhang. 2008. "Which shorts are informed?" *Journal of Finance* 63 (2): 491–527.

Boehmer, Ekkehart, Charles M Jones, Xiaoyan Zhang, and Xinran Zhang. 2020. "Tracking retail investor activity." *Journal of Finance, forthcoming.*

Cohen, Lauren, Karl B Diether, and Christopher J Malloy. 2007. "Supply and demand shifts in the shorting market." *Journal of Finance* 62 (5): 2061–2096.

D'Avolio, Gene. 2002. "The market for borrowing stock." *Journal of Financial Economics* 66 (2): 271–306.

Daniel, Kent D., Alexander Klos, and Simon Rottke. 2018. "Overconfidence, Information Diffusion, and Mispricing Persistence." *SSRN Electronic Journal.*

Dechow, Patricia M, Amy P Hutton, Lisa Meulbroek, and Richard G Sloan. 2001. "Short-sellers, fundamental analysis, and stock returns." *Journal of Financial Economics* 61 (1): 77–106.

Desai, Hemang, Kevin Ramesh, S Ramu Thiagarajan, and Bala V Balachandran. 2002. "An investigation of the informational role of short interest in the Nasdaq market." *Journal of Finance* 57 (5): 2263–2287.

Detemple, Jerome, and Shashidhar Murthy. 1997. "Equilibrium asset prices and no-arbitrage with portfolio constraints." *The Review of Financial Studies* 10 (4): 1133–1174.

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

Diether, Karl B., Kuan-Hui Lee, and Ingrid M Werner. 2009. "Short-sale strategies and return predictability." *The Review of Financial Studies* 22 (2): 575–607.

Drechsler, Itamar, and Qingyi (Freda) Song Drechsler. 2014a. "The Shorting Premium and Asset Pricing Anomalies." *SSRN Electronic Journal.*

Drechsler, Itamar, and Qingyi Freda Drechsler. 2014b. *The shorting premium and asset pricing anomalies.* Technical report. National Bureau of Economic Research.

Duffie, Darrell, Nicolae Gârleanu, and Lasse Heje Pedersen. 2002. "Securities lending, shorting, and pricing." *Journal of Financial Economics* 66 (2-3): 307–339.

Duong, Truong X, Zsuzsa R Huszár, Ruth S K Tan, and Weina Zhang. 2017. "The Information Value of Stock Lending Fees: Are Lenders Price Takers?" *Review of Finance* 21, no. 6 (January): 2353–2377.

Engelberg, Joseph E., Adam V. Reed, and Matthew C. Ringgenberg. 2018. "Short-Selling Risk." *The Journal of Finance* 73, no. 2 (February): 755–786.

Evgeniou, Theodoros, Julien Hugonnier, and Rodolfo Prieto. 2022. "Asset Pricing with Costly Short Sales." Working paper.

Farmer, Roger, and Jean-Philippe Bouchaud. 2020. *Self-Fulfilling Prophecies, Quasi Non-Ergodicity & Wealth Inequality.* Working Paper, Working Paper Series 28261. National Bureau of Economic Research, December.

Fostel, Ana, and John Geanakoplos. 2008. "Leverage cycles and the anxious economy." *American Economic Review* 98 (4): 1211–44.

Gârleanu, Nicolae, Leonid Kogan, and Stavros Panageas. 2012. "Displacement risk and asset returns." *Journal of Financial Economics* 105 (3): 491–510.

Gârleanu, Nicolae, and Stavros Panageas. 2015. "Young, old, conservative, and bold: The implications of heterogeneity and finite lives for asset pricing." *Journal of Political Economy* 123 (3): 670–685.

———. 2020. *Heterogeneity and asset prices: A different approach.* Technical report. National Bureau of Economic Research.

———. 2021. "What to expect when everyone is expecting: Self-fulfilling expectations and asset-pricing puzzles." *Journal of Financial Economics* 140 (1): 54–73.

Geczy, Christopher C., David K. Musto, and Adam V. Reed. 2002. "Stocks are special too: an analysis of the equity lending market." Limits on Arbitrage, *Journal of Financial Economics* 66 (2): 241–269.

Gennotte, Gerard, and Hayne Leland. 1990. "Market Liquidity, Hedging, and Crashes." *American Economic Review* 80 (5): 999–1021.

Harrison, J Michael, and David M Kreps. 1978. "Speculative investor behavior in a stock market with heterogeneous expectations." *Quarterly Journal of Economics* 92 (2): 323–336.

Hong, Harrison, and Jeremy C Stein. 2003. "Differences of opinion, short-sales constraints, and market crashes." *The Review of Financial Studies* 16 (2): 487–525.

Jones, Charles M, and Owen A Lamont. 2002. "Short-sale constraints and stock returns." *Journal of Financial Economics* 66 (2-3): 207–239.

Kaplan, Steven N, Tobias J Moskowitz, and Berk A Sensoy. 2013. "The effects of stock lending on security prices: An experiment." *Journal of Finance* 68 (5): 1891–1936.

Khorrami, Paymon, and Fernando Mendo. 2021. "Rational Sentiments and Financial Frictions." Working paper.

Khorrami, Paymon, and Alexander Zentefis. 2020. "Arbitrage and Beliefs." Working paper.

Kyle, Albert S., and Wei Xiong. 2001. "Contagion as a Wealth Effect." *Journal of Finance* 56 (4): 1401–1440.

Lamont, Owen A. 2012. "Go down fighting: Short sellers vs. firms." *The Review of Asset Pricing Studies* 2 (1): 1–30.

Lamont, Owen A, and Jeremy C Stein. 2004. "Aggregate short interest and market valuations." *American Economic Review* 94 (2): 29–32.

Leahy, John V. 1993. "Investment in competitive equilibrium: The optimality of myopic behavior." *Quarterly Journal of Economics* 108 (4): 1105–1133.

Merton, Robert. 1987. "A Simple Model of Capital Market Equilibrium with Incomplete Information." *Journal of Finance* 42 (3): 483–510.

Miller, Edward M. 1977. "Risk, uncertainty, and divergence of opinion." *Journal of Finance* 32 (4): 1151–1168.

Panageas, Stavros. 2020. *The implications of heterogeneity and inequality for asset pricing.* Technical report. National Bureau of Economic Research.

Pedersen, Lasse Heje. 2022. "Game on: Social networks and markets." *Journal of Financial Economics* 146, no. 3 (December): 1097–1119.

Porras Prado, Melissa, Pedro A. C. Saffi, and Jason Sturgess. 2016. "Ownership Structure, Limits to Arbitrage, and Stock Returns: Evidence from Equity Lending Markets." *The Review of Financial Studies* 29, no. 12 (July): 3211–3244.

Rapach, David E., Matthew C. Ringgenberg, and Guofu Zhou. 2016. "Short interest and aggregate stock returns." *Journal of Financial Economics* 121 (1): 46–65.

Scheinkman, Jose A, and Wei Xiong. 2003. "Overconfidence and speculative bubbles." *Journal of Political Economy* 111 (6): 1183–1220.

Senchack, Andrew J, and Laura T Starks. 1993. "Short-sale restrictions and market reaction to short-interest announcements." *Journal of Financial and Quantitative Analysis* 28 (2): 177–194.

Seneca, Joseph J. 1967. "Short interest: bearish or bullish?" *Journal of Finance* 22 (1): 67–70.

Simsek, Alp. 2013. "Belief disagreements and collateral constraints." *Econometrica* 81 (1): 1–53.

Sweeting, Andrew. 2006. "Coordination, Differentiation, and the Timing of Radio Commercials." *Journal of Economics & Management Strategy* 15, no. 4 (December): 909–942.

Vayanos, Dimitri, and Pierre-Olivier Weill. 2008. "A Search-Based Theory of the On-the-Run Phenomenon." *Journal of Finance* 63 (3): 1361–1398.

Zentefis, Alexander. 2018. "Self-fulfilling asset prices." Working paper.

For Online Publication – Appendix

A The Determination of the Lending Fee

In the text we assume a "flat" supply curve for lending shares. That is, we assume $f_t = f(y_t) = \varphi$. We provide here the simplest model that supports this assumption. We also discuss how to extend the model to allow for an increasing $f(\cdot)$.

All interactions considered in this section happen anew every period, where the length of the period is idealized to be "dt," that is, infinitesimal. (We could formalize this assumption by considering a discrete-time model where the length Δ of a period is taken to go to zero, and focusing on the limit of resultant equilibria.)

We start by considering the long investors, who wish to lend their shares. Each investor lends all her shares to any one of a competitive fringe of profit-maximizing "security lenders" in exchange for an income stream that is proportional to the dollar value of shares the investor lends. This income stream is determined as follows. In equilibrium, each security lender lends a proportion y_t of the shares it borrows from investors and receives a fee f_l per dollar of shares it lends out. (We omit time subscripts from now on.) Competition between the security lenders drives the income stream paid to long investors to yf_l per dollar of stock owned by the investors.[47]

At the other end of the lending transaction, desirous short sellers interact with a competitive set of "borrower's brokers." Specifically, for every borrowing fee f_b the would-be short sellers provide the dollar amount that they would like to short, and the brokers take the value f_b as a given when they attempt to fill the investors' borrowing orders.

All of the frictions in this model pertain to the interaction between security lenders and borrower's brokers. Specifically, to initiate a stock loan the representative broker must pay a cost ξ per dollar value of share "located" with a security lender, per unit of time. This cost is construed as labor cost that compensates brokers for their disutility of labor.

The interaction between the broker and the security lender takes the form of bilateral Nash bargaining in which the broker has bargaining power $1/(1+z)$ for a parameter $z \in (0, \infty)$. Given our assumption that all interactions (between investors and brokers or security lenders and between brokers and security lenders) happen anew every period, the outside option for both brokers and security lenders is the failure to transact during the period. This means that the gains from trade to the security lender equal the lending fee f_l, while to the broker the borrowing fee net of the lending one $f_b - f_l$ — the searching and matching cost ξ has been sunk at this point. The total gains from trade equal f_b, the foregone revenue from the would-be short seller. Given the bargaining protocol, it follows that

$$f_l = \frac{z}{1+z}(f_l + f_b - f_l) = \frac{z}{1+z} f_b. \tag{A.1}$$

47. In that sense, the security lenders resemble the "insurance companies" in Blanchard (1985). Similar to how insurance companies collect payments from the fraction of agents who die and rebate them to the surviving population, the security lenders collect lending fees from the proportion of a long portfolio that gets loaned out and rebate it in the form of an income stream to the representative long investor.

Brokers break even on net, meaning that

$$f_b = f_l + \xi, \tag{A.2}$$

so that

$$f_l = z\xi, \tag{A.3}$$
$$f_b = (1+z)\xi. \tag{A.4}$$

To keep the model transparent and tractable, assume that all brokers are members of the representative household, and therefore the fees that compensate them for their effort are rebated to each households as an income stream proportional to the household's wealth and independent of the composition of the household's portfolio.

Setting $\varphi = (1+z)\xi$ and $\tau = z/(1+z)$, this extended model is equivalent to the model we assumed in the text. To generalize to upward-sloping supply curves, one would simply assume an increasing cost $\xi(y)$.

B Multiple Agent Types

We illustrate here that the multiplicity of equilibria may expand with the number of agent types. In particular, adding a third group of agents can result in a third equilibrium featuring non-zero shorting; such a model may admit, in fact, up to five equilibria.

Specifically, let us assume a third group of investors characterized by beliefs that are summarized by the quantity η^P. We think of these investors as pessimists, which implies $\eta^P < 0$. The intuition we wish to capture is that, in addition to the "high-shorting" and "medium-shorting" equilibria in the base-line model, low-shorting equilibria may exist in which investor R is inactive, while investor P shorts actively.

To make the point theoretically, one may argue by continuity. Specifically, consider the zero-shorting equilibrium in the baseline model, and perturb the setting by adding a small mass of sufficiently pessimistic investors ($|\eta^P|$ large enough). These investors will want to short, but will not be sufficiently numerous to move the Sharpe ratio or lending income to a point where investors R and I are no longer in equilibrium.

It is helpful to write down the equilibrium conditions in the augmented model — both to allow for a formal argument and in the interest of a numerical illustration. We repeat the analysis in the text — letting ω^P denote the wealth share of agents P — to obtain the market clearing condition

$$\begin{aligned} 1 = \frac{1}{\sigma_D}\Bigg[&\omega^P\left(\kappa+\eta^P+\frac{\varphi}{\sigma_D}\right)1_{\left\{\kappa+\eta^P+\frac{\varphi}{\sigma_D}<0\right\}} + \omega^R\left(\kappa+\frac{\varphi}{\sigma_D}\right)1_{\left\{\kappa+\frac{\varphi}{\sigma_D}<0\right\}} + \\ &\omega^I\left(\kappa+\eta^I+\frac{\varphi}{\sigma_D}\tau y\right)\Bigg], \end{aligned} \tag{B.1}$$

where the left-hand side is the proportion of aggregate wealth represented by the supply of the stock, while the right-hand side equals the proportion of aggregate wealth invested in the stock. We restricted attention to cases in which R agents do not take a long position in

the stock.

We solve for the Sharpe ratio κ:

$$\kappa = \sigma_D - \left(\omega^P \eta^P + \omega^I \eta^I\right) - \frac{\varphi}{\sigma_D}\left(\omega^P + \omega^R + \omega^I \tau y\right) \tag{B.2}$$

if $\kappa + \varphi/\sigma_D < 0$, respectively

$$\kappa = \frac{\sigma_D}{\omega^P + \omega^I} - \frac{\omega^P \eta^P + \omega^I \eta^I}{\omega^P + \omega^I} - \frac{\varphi}{\sigma_D}\frac{\omega^P + \omega^I \tau y}{\omega^P + \omega^I} \tag{B.3}$$

if $\kappa + \varphi/\sigma_D \geq 0 > \kappa + \eta^P + \varphi/\sigma_D$.

The other equilibrium condition concerns the determination of the value of y:

$$y = -\frac{\omega^P \left(\kappa + \eta^P + \frac{\varphi}{\sigma_D}\right) 1_{\left\{\kappa + \eta^P + \frac{\varphi}{\sigma_D} < 0\right\}} + \omega^R \left(\kappa + \frac{\varphi}{\sigma_D}\right) 1_{\left\{\kappa + \frac{\varphi}{\sigma_D} < 0\right\}}}{\omega^I \left(\kappa + \eta^I + \frac{\varphi}{\sigma_D}\tau y\right)}. \tag{B.4}$$

Depending on whether κ is determined according to (B.2) or (B.3) we obtain a different quadratic equation. For appropriate parameter choices all but one combinations are possible in terms of how many solutions in the interval $(0, 1)$ each of them admits. We are particularly interested in situations in which (B.2) applies and results in two admissible solutions, in addition to which at least one solution obtains when (B.3) applies.

We illustrate such outcomes in Figure B.1. The two panels differ in terms of parameters, but depict the same objects. Specifically, the x-axis records candidate values of y that agents anticipate. Agents form demands taking such a value y and a Sharpe ratio κ as given, and clearing in the asset market determines the Sharpe ratio. With the Sharpe ratio now specified for each candidate y, we can compute the actual resulting y — the value of the right-hand side of equation (B.4). This quantity is recorded on the y-axis. An equilibrium requires that the x and y coordinates are equal.

The line "R and P short" plots y as if both R and P shorted, that is, their portfolio weights are calculated by adding the return φ to their perceived intrinsic expected return from the asset; in that case, the Sharpe ratio is given by (B.2). The line "Only P shorts" is produced similarly, except that the demand of agent R is set to zero; equation (B.3) applies. The actual resulting y is depicted by the thick continuous line, labeled "Actual response." Finally, the line "Diagonal" depicts the equilibrium condition. Equilibria are therefore represented by points of intersection between the two continuous lines. The left panel presents a situation in which four equilibria with positive amounts of shorting and one with zero shorting obtain. The right panel presents a situation with three equilibria, all of which feature positive y.

We also flesh out the theoretical argument for the existence of a third equilibrium when ω^P is close to zero and two equilibria with $y > 0$ exist with $\omega^P = 0$ — i.e., the baseline model. By assumption, with $\omega^P = 0$ and $y = 0$ equation (B.3) applies and $\kappa + \frac{\varphi}{\sigma_D} > 0$.

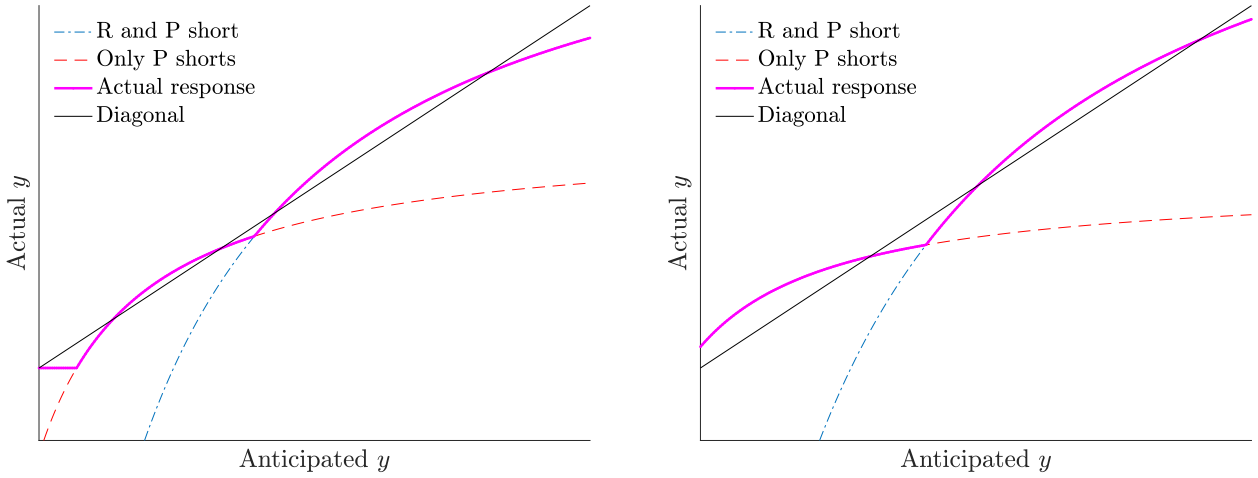


Figure B.1: The figure plots, in each panel, four lines pertaining to the model extension developed in this section. Equilibria are characterized by the satisfaction of equation (B.4), whose right-hand side is represented here by the line "Actual response" and the left-hand side by the line "Diagonal." Further details are provided in the text.

Choosing η^P so that $\kappa + \eta^P + \frac{\varphi}{\sigma_D} < 0$, we wish to conclude that equations

$$y = -\frac{\omega^P \left(\kappa + \eta^P + \frac{\varphi}{\sigma_D}\right)}{\omega^I \left(\kappa + \eta^I + \frac{\varphi}{\sigma_D}\tau y\right)} \tag{B.5}$$

and (B.3) admit a solution that satisfies $\kappa + \frac{\varphi}{\sigma_D} > 0$ even for $\omega^P > 0$, at least when it is small enough. For simplicity, we keep ω^I constant as we increase ω^P from zero. Plugging (B.3) in (B.5) we obtain a quadratic that can be written as

$$y = \frac{\omega^P}{\omega^I} \frac{\left(\eta^I - \eta^P\right)\omega^I - \frac{\varphi}{\sigma_D}\omega^I \tau\left(1-y\right) - \sigma_D}{\left(\eta^I - \eta^P\right)\omega^P - \frac{\varphi}{\sigma_D}\omega^P \tau\left(1-y\right) + \sigma_D} \equiv H(\omega^P, y). \tag{B.6}$$

Our choice of η^P is such that the numerator of the second fraction on the right-hand side is positive at $y = 0$, which implies $\frac{\partial H}{\partial \omega^P} > 0$ evaluated at $\omega^P = 0$, as well as $\frac{\partial H}{\partial y} = 0$ at $\omega^P = 0$. We therefore have

$$\frac{dy}{d\omega^P} = \left(1 - \frac{\partial H}{\partial y}\right)^{-1} \frac{\partial H}{\partial \omega^P} > 0, \tag{B.7}$$

confirming that an equilibrium with positive y exists for small ω^P. (The condition $\kappa + \frac{\varphi}{\sigma_D} > 0$ is satisfied by continuity.)

C The Price-Dividend Ratio of a Small Stock

This section provides the details of the entry-and-exit process for the model of Section 5.2.

We start with a few definitions. We let $\overrightarrow{m}_t$ denote the vector of market-capitalization weights of the two stocks, and $m_{j,t}$, $j \in \{1,2\}$, its entries. Since the analysis of interest pertains to asset 1, from now on we use W_t^i to denote the wealth of all agents of type i that participate in the market for stock 1; the relevant state variable is $\omega_t^i \equiv W_t^i/(W_t^R + W_t^I)$, and to save notation we maintain the convention $\omega_t \equiv \omega_t^R$. As stated in the text, $\widehat{\omega}_t$ denotes the wealth share of the investors who actively participate in the market for stock 1. We also let $\widehat{w}_{2,t} = \frac{\mu_{2,t}-r_t}{\sigma_{2,t}^2}$ denote the optimal portfolio holding of stock 2 by investors who don't participate in stock 1, and $\overrightarrow{w}_t^i$ is the (row) vector of portfolio holdings of an investor $i \in \{I, R\}$ that is active in the market for stock 1. Finally, $\overrightarrow{B}_t \equiv (B_{1,t}, B_{2,t})^\top$.

We further define

$$\overrightarrow{B}_t \equiv \begin{bmatrix} B_{1,t} \\ B_{2,t} \end{bmatrix}, \ \sigma_t = \begin{bmatrix} \sigma_{1,t} & b_t\sigma_{2,t} \\ 0 & \sigma_{2,t} \end{bmatrix}, \ \overrightarrow{\varphi} = \begin{bmatrix} \varphi \\ 0 \end{bmatrix}, \ \overrightarrow{\eta} = \begin{bmatrix} \eta \\ 0 \end{bmatrix}. \tag{C.1}$$

The entry and exit into market 1 happens either for endogenous or exogenous reasons. By "endogenous" we mean that investors conduct a cost-benefit analysis before deciding whether to keep paying attention to the market for stock 1. In addition to this optimizing choice, we assume that investors enter and exit the market for exogenous reasons. This exogenous flux of investors is modeled with the sole purpose of making the model solution more tractable and transparent.

Specifically, with W_t^i the (aggregate) wealth of type-i investors that participate in market 1, we assume

$$dW_t^i = dW_t^{i,\text{part}} + \chi\left(\nu^i(W_t^I + W_t^R) - W_t^i\right)dt - 1_{i=R} \times \frac{W_t^I + W_t^R}{1-\omega_t} dF_t + \omega_t^i\left(dL_t - dN_t\right), \tag{C.2}$$

where $dW_t^{i,\text{part}}$ is the wealth growth of all investors of type $i \in \{I, R)$ who already participate in the market for stock 1.[48] The term $\chi\left(\nu_i(W_t^I + W_t^R) - W_t^i\right)dt$ reflects entirely exogenous, non-optimizing entry, which happens at some rate χ.

As in the baseline model, we are assuming that this exogenous entry-and-exit process affects the composition, but not the sum, of $W_t^I + W_t^R$, since

$$\sum_{i\in\{I,R\}} \chi\left(\nu^i\left(W_t^I + W_t^R\right) - W_t^i\right) = 0.$$

The term $-1_{i=R} \times \frac{W_t^I+W_t^R}{1-\omega_t} dF_t$ captures the endogenous exit of R investors. As we described in the text, the (singular) process dF_t is constructed so that ω_t stays below the critical value $\bar{\omega}$ of ω_t (see (C.3) below) that ensures $V^R(\omega_t) > 0$ for $\omega_t < \bar{\omega}$.

Mostly for technical tractability reasons, we assume another source of exogenous entry and exit, which is reflected in the term $\omega_t^i\left(dL_t - dN_t\right)$ on the right-hand side of (C.2).

48. For completeness, $dW_t^{i,\text{part}} = W_t^{i,\text{part}}\mu_W^i dt + W_t^{i,\text{part}}(\overrightarrow{w}^i)^\top \sigma_t d\overrightarrow{B}_t$ where

$$\mu_W^i = r_t + \pi + n_t + (\overrightarrow{w}_{t,s}^i)^\top\left(\overrightarrow{\mu}_t - r_t 1_{2\times 1} + \lambda_{t,s}^i \begin{bmatrix} 1 \\ 0 \end{bmatrix}\right) - \frac{c_{t,s}^i}{W_{t,s}^i}.$$

This entry and exit process leaves the composition of wealth in the market (between R and I investors) unaffected, but ensures that the wealth of the investors who pay attention to the market 1 stays proportional to the "size" of market 1. Specifically, we define dL_t and dN_t as the two singular, increasing processes that control $W_t^I + W_t^R$ so that the ratio of stock market capitalization of asset 1 to the total wealth of investors participating in market 1, $\widetilde{m}_t = \frac{M_{1,t}}{W_t^I + W_t^R}$, stays constant across time ($\widetilde{m}_t = \widetilde{m}$).[49] Because $(dL_t - dN_t)$ is multiplied by ω_t^i, this exogenous entry-and-exit process does not impact the composition of wealth between R and I investors. The purpose of this exogenous entry-and-exit term is transparency and tractability: By ensuring a constant $\widetilde{m}_t$, if there were no differences of opinion ($\eta = 0$), the excess return, the price-dividend ratio, and the volatility of stock 1 would all be constant. Thus, we can eliminate a state variable from the problem, namely the ratio of market capitalization to the total wealth of investors in market 1. Economically, this means that we can abstract from the effects of limited participation (that have been studied extensively in the literature) and isolate the impact of shorting frictions. It is also worth highlighting that the term $\omega_t^i (dL_t - dN_t)$ endogenously approaches zero as δ_1 and χ approach infinity.[50] Thus, our computations would be approximately valid if we eliminated the term $\omega_t^i (dL_t - dN_t)$, as long as the analysis focuses on cases where investors are short-termist (χ is large) and the ratio of the dividends of a typical tree 1 to tree 2 mean reverts fast.

Having described the entry and exit of investors into the market for stock 1, we are ready to state a formal result describing the determination of equilibrium in this economy. For simplicity, we assume that the Brownian motions $B_{1,t}$ and $B_{2,t}$ are independent.

Proposition 6 *Using the expressions for w_t^i, $\kappa_{1,t}$ (with $b = 0$), and y_t from Proposition 5, the wealth share ω_t follows the diffusion process*

$$d\omega_t = \mu_\omega(\omega_t)dt + \sigma_\omega(\omega_t)dB_{1,t} - dF_t, \tag{C.3}$$

where F_t is an increasing (singular) process that reflects ω_t to remain below the value $\bar{\omega}$ that is the lowest value for which $V^R(\omega_t) = 0$, and $\mu_\omega(\omega_t)$ and $\sigma_\omega(\omega_t)$ are given by

$$\mu_\omega(\omega_t) = \omega_t \left(\left(w_{1,t}^R - \widetilde{m} \right) \sigma_{1,t} \left(\kappa_t - \sigma_{1,t}\widetilde{m} \right) + w_{1,t}^R \varphi + \frac{y_t \widetilde{m}}{1 - y_t} \varphi \left(1 - \tau \right) \right) + \tag{C.4}$$
$$\chi \left(\nu - \omega_t \right),$$
$$\sigma_\omega(\omega_t) = \omega_t \left(w_{1,t}^R - \widetilde{m} \right) \sigma_{1,t}, \tag{C.5}$$

where $\sigma_{1,t} = \frac{p'(\omega_t)}{p(\omega_t)} \sigma_\omega(\omega_t) + \sigma_{1,D}$ is the volatility of stock 1 and the price-dividend ratio $p_t = p(\omega_t)$ solves the ordinary differential equation

$$\frac{1}{2} \frac{\partial^2 p}{\partial \omega_t^2} \left(\sigma_\omega(\omega_t) \right)^2 + \frac{\partial p}{\partial \omega_t} \left(\mu_\omega(\omega_t) + \left(\sigma_{1,D} - \kappa_{1,t} \right) \sigma_\omega(\omega_t) \right) - p \left(r + \delta_1 + \kappa_{1,t}\sigma_{1,D} \right) + 1 = 0 \tag{C.6}$$

49. These processes can be uniquely constructed from the running maximum and minimum of the difference between $(W_t^R + W_t^I) - M_{1,t}$. For details see Karatzas and Shreve (2012, p. 210) on the Skorohod equation.

50. The reason is that the price-dividend ratio and the ratio of the dividend processes for the two trees (given in (39)) approach constants, thus implying that $\widetilde{m}_t$ approaches a constant ($\widetilde{m}$).

in the region $0 \leq \omega_t \leq \bar{\omega}$.

Remark 4 *Since there are multiple equilibrium values for* w_t^i, $\kappa_{1,t}$, *and* y_t *in Proposition 6, there exist a large set of solutions for* $p(\cdot)$ *and* $\bar{\omega}$, *depending on the equilibrium on which agents coordinate at each value of* ω_t.

The expressions for μ_ω and σ_ω in Proposition 6 coincide with (24) and (23) when $\widetilde{m} = 1$, $\chi = \pi$, and $\sigma_{1,t} = \sigma_D$.[51] Moreover, with the dividend growths of stocks 1 and 2 being independent, so are their stock-price processes (in the limit where stock 1 becomes small) and the expressions for y_t, $w_{1,t}^i$, and $\kappa_{1,t}$ in Proposition 6 (with $\widetilde{m} = 1$ and $\sigma_{1,t} = \sigma_D$) coincide with the respective expressions in the baseline model. Finally, if $\varepsilon = 0$, then $\bar{\omega} = 1$, as in the baseline model. In short, if one dropped the goods-market clearing requirement from the baseline model, the resulting expression for the price-to-dividend ratio would be given by (C.6) (with $\widetilde{m} = 1$ and $\varepsilon = 0$).

The main complications with solving (C.6) are that a) it is a non-linear ODE[52] and b) for $\varepsilon > 0$, this ODE is to be solved over a domain of values of ω_t on which $V^R(\omega_t) > 0$, with $V^R(\bar{\omega}) = 0$ as a boundary condition.

We solve (C.6) with iterated Monte Carlo. We start with the initial guess $\sigma_{1,t} = \sigma_{1,D}$ and some guess for the cutoff $\bar{\omega}$. We define the stopping time to be T the hitting time when ω_t first equals $\bar{\omega}$. With that guess we use a Monte Carlo simulation to evaluate $V_t^R(\omega_t, T)$ on a grid of ω_t values. We find the value ω_t for which $V_t^R(\omega_t, T) = 0$ and update our guess for $\bar{\omega}$ to satisfy $V_t^R(\bar{\omega}, T) = 0$. With this guess for $\bar{\omega}$ we compute the price-dividend ratio on a grid of ω values by using the Feynman-Kac theorem to express (C.6) as an expectation, which we evaluate with Monte Carlo simulation. After obtaining the price-dividend ratio on a fine grid of values, we evaluate $\frac{p'(\omega_t)}{p(\omega_t)}$, and compute $\sigma_{1,t} = \frac{p'(\omega_t)}{p(\omega_t)}\sigma_\omega(\omega_t) + \sigma_{1,D}$. Using this new guess for $\sigma_{1,t}$ we repeat the above procedure until convergence.

D Proofs

Proof of Proposition 1. Fix parameters $\eta > 0$ and $\psi > 1$ and define φ according to

$$\varphi = \sigma_D \left(\eta - \psi \sigma_D\right) \tag{D.1}$$

for any value of σ_D. Note that when σ_D is sufficiently small, φ is guaranteed to be positive.

We show next that, as σ_D gets close to zero, Assumption 2 is satisfied. Rearranging (D.1) gives

$$\frac{\eta}{\frac{\varphi}{\sigma_D}} = \frac{1}{1 - \psi\frac{\sigma_D}{\eta}}. \tag{D.2}$$

For sufficiently small σ_D we obtain

$$1 + \tau > \frac{1}{1 - \psi\frac{\sigma_D}{\eta}} > 1. \tag{D.3}$$

51. To see this, substitute the expression for the equilibrium interest rate (22) into (24).

52. Equation (C.6) is non-linear because μ_t^i and σ_t^i depend on $\sigma_{1,t}$, which in turn depends on $p(\cdot)$ and $p'(\cdot)$.

Combining (D.2) and (D.3) yields (17).

Turning to (18), we note that the definition of ω_1^* along with (D.1) implies

$$\omega_1^* = 1 - \frac{\sigma_D}{\psi\sigma_D} = \frac{\psi - 1}{\psi} > 0,$$

while also

$$\lim_{\sigma_D \to 0} \frac{\sigma_D}{(1+\tau)\frac{\varphi}{\sigma_D} - \eta} = \lim_{\sigma_D \to 0} \frac{\sigma_D}{(1+\tau)(\eta - \psi\sigma_D) - \eta} = 0.$$

Therefore, for sufficiently small σ_D, the left-hand side of (18) converges to $\frac{\psi-1}{\psi} > 0$, while the right-hand side converges to zero, and therefore the inequality holds.

We conclude the proof by showing that $F(\omega)$ has a unique root in the interval $(\omega_1^*, 1)$. To this end, it is useful to introduce the definitions

$$A(\omega) \equiv \tau\frac{\omega}{\sigma_D}\varphi, \tag{D.4}$$

$$B(\omega) \equiv \sigma_D - \omega\left((1+\tau)\frac{\varphi}{\sigma_D} - \eta\right), \tag{D.5}$$

$$C(\omega) \equiv \frac{\omega}{1-\omega}\left(\sigma_D + (1-\omega)\left(\frac{\varphi}{\sigma_D} - \eta\right)\right). \tag{D.6}$$

With these definitions, $F(\omega)$ can be written as $F(\omega) = B^2(\omega) - 4A(\omega)C(\omega)$. We start by observing that $C(\omega_1^*) = 0$ for any parametric choice (since the definition of ω_1^* in Equation (15) implies $\sigma_D + (1-\omega_1^*)\left(\frac{\varphi}{\sigma_D} - \eta\right) = 0$). Also, Inequality (18) implies that $B(\omega_1^*) \neq 0$, and thus $B^2(\omega_1^*) > 0$. Accordingly, $F(\omega_1^*) > 0$. Also $B(1) < \infty$, while $C(1) = \infty$. By continuity, there exists at least one value $\omega_2^* \in (\omega_1^*, 1)$ such that $F(\omega_2^*) = 0$.

To show that this value is unique, consider any value $\omega_2^* \in (\omega_1^*, 1)$ such that $F(\omega_2^*) = 0$. We next show that $F'(\omega_2^*) < 0$.

To this end, note that

$$\begin{aligned} F'(\omega) &= 2B(\omega)B'(\omega) - 4\left[A'(\omega)C(\omega) + A(\omega)C'(\omega)\right] \\ &= 2B^2(\omega)\frac{B'(\omega)}{B(\omega)} - 4A(\omega)C(\omega)\left(\frac{A'(\omega)}{A(\omega)} + \frac{C'(\omega)}{C(\omega)}\right). \end{aligned}$$

Since ω_2^* is a root of $F(\omega)$ it follows that $B^2(\omega_2^*) = 4A(\omega_2^*)C(\omega_2^*)$. Therefore,

$$F'(\omega_2^*) = B^2(\omega_2^*)\left(2\frac{B'(\omega_2^*)}{B(\omega_2^*)} - \frac{A'(\omega_2^*)}{A(\omega_2^*)} - \frac{C'(\omega_2^*)}{C(\omega_2^*)}\right). \tag{D.7}$$

We have

$$\frac{A'(\omega_2^*)}{A(\omega_2^*)} = \frac{1}{\omega_2^*}$$
$$\frac{B'(\omega_2^*)}{B(\omega_2^*)} = -\frac{(1+\tau)\frac{\varphi}{\sigma_D} - \eta}{\sigma_D - \omega_2^*\left((1+\tau)\frac{\varphi}{\sigma_D} - \eta\right)}$$

and

$$\frac{C'(\omega_2^*)}{C(\omega_2^*)} = \frac{1}{\omega_2^*(1-\omega_2^*)} + \frac{\eta - \frac{\varphi}{\sigma_D}}{\sigma_D + (1-\omega_2^*)\left(\frac{\varphi}{\sigma_D} - \eta\right)}.$$

Combining terms gives

$$\begin{aligned} &2\frac{B'(\omega_2^*)}{B(\omega_2^*)} - \frac{A'(\omega_2^*)}{A(\omega_2^*)} - \frac{C'(\omega_2^*)}{C(\omega_2^*)} \\ &= -\frac{2\left((1+\tau)\frac{\varphi}{\sigma_D} - \eta\right)}{\sigma_D - \omega_2^*\left((1+\tau)\frac{\varphi}{\sigma_D} - \eta\right)} - \frac{1}{\omega_2^*} - \frac{1}{\omega_2^*(1-\omega_2^*)} - \frac{\eta - \frac{\varphi}{\sigma_D}}{\sigma_D + (1-\omega_2^*)\left(\frac{\varphi}{\sigma_D} - \eta\right)}. \end{aligned} \tag{D.8}$$

For future reference, we note that using $\omega_2^* > \omega_1^*$ along with (17) and the definition of ω_1^* implies that

$$\sigma_D + (1-\omega_2^*)\left(\frac{\varphi}{\sigma_D} - \eta\right) > \sigma_D + (1-\omega_1^*)\left(\frac{\varphi}{\sigma_D} - \eta\right) = 0. \tag{D.9}$$

Using (D.1) we can write the right-hand side of (D.8) as

$$-\frac{2\left((1+\tau)(\eta - \psi\sigma_D) - \eta\right)}{\sigma_D - \omega_2^*\left((1+\tau)(\eta - \psi\sigma_D) - \eta\right)} - \frac{1}{\omega_2^*} - \frac{1}{\omega_2^*(1-\omega_2^*)} - \frac{\psi}{1 - \psi(1-\omega_2^*)}. \tag{D.10}$$

Taking the limit as σ_D approaches zero, the expression (D.10) converges to

$$-\frac{1}{1-\omega_2^*} - \frac{\psi}{1-\psi(1-\omega_2^*)} < 0,$$

where the inequality follows from (D.9) along with (D.1).[53]

The fact that the derivative $F'(\omega_2^*) < 0$ for any root of the equation $F(\omega_2^*) = 0$ in the interval $(\omega_1^*, 1)$ implies that the root ω_2^* must be unique. ■

Proof of Proposition 2. In preparation for the proof, we state and prove an auxiliary result.

53. Equation (D.1) implies $\frac{\varphi}{\sigma_D} - \eta = -\psi\sigma_D$, and therefore $0 < \sigma_D + (1-\omega_2^*)\left(\frac{\varphi}{\sigma_D} - \eta\right) = \sigma_D(1 - (1-\omega_2^*)\psi)$, where the inequality follows from (D.9).

Lemma 1 *The following statements hold for the quadratic Equation* (20).

1. $\omega_1^* < \omega_2^*$ *and the discriminant of (20) is non-negative for all* $\omega_t \leq \omega_2^*$.
2. *When* $\omega_1^* \leq \omega_t \leq \omega_2^*$, *the two roots of the equation are both in the interval* $[0,1)$.
3. *For* $\omega_t \in [0, \omega_1^*)$, *only the larger root of (20) is in the interval* $(0,1)$.
4. *If* y *is a root of (20), then* $(1-\omega_t)\eta - \sigma_D - \frac{1-\omega_t}{\sigma_D}\varphi(1-\tau y) > 0$.

Proof of Lemma 1. We start with part 1. Using the definitions (D.4)–(D.6), Equation (20) can be written in the familiar form

$$A(\omega_t) y^2 + B(\omega_t) y + C(\omega_t) = 0,$$

and the discriminant of this quadratic equation is given by $F(\omega_t)$ as defined in Equation (16).

For $\omega_t \leq \omega_1^*$, $C(\omega_t) < 0$ and the discriminant, $B^2(\omega_t) - 4A(\omega_t)C(\omega_t)$, is positive. The assumption that ω_2^* is the unique root of $F(\omega)$ along with the facts that $F(\omega_1^*) = B^2(\omega_1^*) > 0$ and $F(1) = -\infty$ imply that $\omega_1^* < \omega_2^*$.[54] The uniqueness of the root ω_2^* also implies that $F(\omega_t) = B^2(\omega_t) - 4A(\omega_t)C(\omega_t) \geq 0$ for all $\omega_t \leq \omega_2^*$.

We now turn to part 2. To economize on notation we write A rather $A(\omega_t)$ and similarly for B and C. Fix a given ω_t and let $g(y) = Ay^2 + By + C$. We have $g(1) = A + B + C = \frac{\sigma_D}{1-\omega_t} > 0$ and $g'(1) = 2A + B = \sigma_D + \omega_t\left(\eta - (1-\tau)\frac{\varphi}{\sigma_D}\right) > 0$, where the inequality follows from (17). Since $A > 0$, it follows that all roots of $g(y)$ must be smaller than one. Also, the fact that $\omega_t \geq \omega_1^*$ implies that $g(0) = C > 0$, while assumptions (17) and (18) together with the fact that $\omega_t \geq \omega_1^*$ imply that $g'(0) = B < 0$.

The facts that i) $g(y)$ is a convex, quadratic function of y, ii) $g(1) > 0$, $g(0) > 0$, $g'(1) > 0$, and $g'(0) < 0$ and iii) $B^2 - 4AC > 0$ for $\omega_t \in [\omega_1^*, \omega_2^*)$ imply that there are two roots in $(0,1)$.

For part 3, we note that, when $\omega_t < \omega_1^*$, $g(0) = C < 0$, while $g(1) = A+B+C = \frac{\sigma_D}{1-\omega_t} > 0$. Therefore there exists one and only one root in $(0,1)$.

Finally, let $y \in (0,1)$ denote a root of the quadratic equation (20). Accordingly,

$$\begin{aligned}(1-\omega_t)\eta - \sigma_D - (1-\omega_t)\frac{\varphi}{\sigma_D}(1-\tau y) &= \frac{1-\omega_t}{\omega_t} y\left(\sigma_D + \omega_t\eta - \omega_t\frac{\varphi}{\sigma_D}(1-\tau y)\right) \\ &= \frac{1-\omega_t}{\omega_t} y\left(\sigma_D + \omega_t\left(\eta - \frac{\varphi}{\sigma_D}\right) + \omega_t\frac{\varphi}{\sigma_D}\tau y\right) \\ &> 0,\end{aligned}$$

where the last inequality follows from (17). This proves property 4. ■

We now continue with the proof of the proposition. We provide expressions for r_t and κ_t that apply in any equilibrium in which $w_t^R \neq 0$. Since $\sum_i \omega_t^i = 1$, it follows that $\sum_i \sigma_t^i = 0$

54. Assumption (18) implies that $B(\omega_1^*) \neq 0$ and therefore $B^2(\omega_1^*) > 0$.

and $\sum_i \mu_t^i = 0$. Using (23) and $\sum_i \sigma_t^i = 0$ implies that $\sum_i \omega_t^i w_t^i = 1$. Combining $\sum_i \omega_t^i w_t^i = 1$ with (12) along with the definition $y_t = \frac{W_t^-}{W_t^+}$ gives

$$\kappa_t + (1-\omega_t)\,\eta + \left(\omega_t \frac{1}{\sigma_D}\varphi + (1-\omega_t)\,\tau y_t \frac{1}{\sigma_D}\varphi\right) 1_{\left\{w_t^R<0\right\}} = \sigma^D. \tag{D.11}$$

Similarly, using (24) along with $\sum_i \mu_t^i = 0$ and $\sum_i \omega_t^i \left(n_t + w_t^i s_t^i\right) = 0$ gives (22).

We next describe the equilibria for the three intervals of ω_t described in the statement of the proposition.

i) In this case, $\omega_t > \omega_2^*$. The equilibrium prescribes non-negative portfolios for both investors. If $\omega_t > 1 - \frac{\sigma_D}{\eta}$, Equation (D.11) implies that $\kappa_t > 0$ and (12) implies that both investors hold positive portfolios and the shorting market is inactive. If $\omega_t \in [\omega_1^*, 1 - \frac{\sigma_D}{\eta})$, then there exists an equilibrium that involves no shorting and a zero portfolio for investor R. We check this assertion by observing that the associated market clearing requirement becomes $(1-\omega_t)\, w_t^I = 1$, which together with $y_t = 0$ leads to (19). We then note that

$$\begin{aligned}\kappa_t + \frac{\varphi}{\sigma_D} =& \frac{\sigma_D}{1-\omega_t} - \eta + \frac{\varphi}{\sigma_D} \\ >& \frac{\sigma_D}{1-\omega_1^*} - \eta + \frac{\varphi}{\sigma_D} \\ =& 0.\end{aligned} \tag{D.12}$$

The first line follows from (19), the second line follows from $\omega_t > \omega_1^*$ and the third line follows from the definition of ω_1^*. Since $\kappa_t + \frac{\varphi}{\sigma_D} > 0$, investor R does not choose a negative portfolio. And since $\kappa_t < 0$ for $\omega_t \in [\omega_1^*, 1 - \frac{\sigma_D}{\eta})$, the investor chooses a zero portfolio.

ii) In this case, $\omega_1^* < \omega_t < \omega_2^*$. Since $\omega_t > \omega_1^*$, Equation (D.12) implies that the no-shorting equilibrium continues to be an equilibrium. There exist, however, two more equilibria. To compute them, we guess (and verify shortly) that $w_t^R < 0$. Using (12) and (D.11) gives

$$\begin{aligned} y_t &= \frac{W_t^-}{W_t^+} = \frac{-\omega_t w_{t,s}^R}{(1-\omega_t)\, w_{t,s}^I} = \frac{\omega_t}{1-\omega_t} \frac{-\left(\kappa_t + \frac{1}{\sigma_D}\varphi\right)}{\kappa_t + \eta_t + \frac{1}{\sigma_D}\varphi\tau y_t} \\ &= \frac{\omega_t}{1-\omega_t} \frac{(1-\omega_t)\,\eta - \sigma_D - \frac{1-\omega_t}{\sigma_D}\varphi\,(1-\tau y_t)}{\sigma_D + \omega_t\eta - \frac{\omega_t}{\sigma_D}\varphi\,(1-\tau y_t)}. \end{aligned}$$

Rearranging leads to (20). Statement 1 of Lemma 1 implies that, when $\omega_t \in (\omega_1^*, \omega_2^*)$, Equation (20) has two roots in $(0,1)$. Under the supposition that $w_t^R < 0$, Equation (D.11) leads to (21). In turn

$$\begin{aligned} \kappa_t^{\pm} + \frac{\varphi}{\sigma_D} &= \sigma_D - (1-\omega_t)\,\eta - \frac{\omega_t}{\sigma_D}\varphi\left(1 + \tau y^{\pm}\frac{1-\omega_t}{\omega_t}\right) + \frac{\varphi}{\sigma_D} \\ &= \sigma_D - (1-\omega_t)\left(\eta + \frac{\varphi}{\sigma_D}\left(1 - \tau y_t^{\pm}\right)\right) < 0, \end{aligned} \tag{D.13}$$

where the last inequality follows from statement 4 of Lemma 1. Combining this observation

with (12) confirms that $w_t^R < 0$. Note that in the second and third equilibria we have that

$$\kappa_t^{\pm} + \eta_t + \frac{1}{\sigma_D}\varphi\tau y_t^{\pm} = \sigma_D + \omega_t\eta - \frac{\varphi\omega_t}{\sigma_D}\left(1 - \tau y_t^{\pm}\right) > 0,$$

where the last inequality follows from (D.13) along with the fact that $y^{\pm}$ satisfy the equation (20). This implies that $w_t^I > 0$.

iii) In this case, $\omega_t < \omega_1^*$. Statement 3 of Lemma 1 implies that the quadratic equation (20) has only one solution in $(0,1)$. This shows that there can only be one equilibrium with shorting. Moreover, this is the unique equilibrium. If w_t^R were zero and the Sharpe ratio were $\frac{\sigma_D}{1-\omega_t} - \eta$, then the inequality in (D.12) reverses, i.e., $\frac{\sigma_D}{1-\omega_t} - \eta + \frac{\varphi}{\sigma_D} < 0$ and investor R would want to deviate from the equilibrium prescription and choose a negative portfolio.

The dynamics of the wealth share follow from a straightforward application of Ito's lemma. ∎

Lemma 2 *When the equilibrium is unique,* $0 < \Phi < 1$.

Proof of Lemma 2. We start by noting that an application of the implicit function theorem to (20) gives $\frac{dy}{d\eta} = \frac{1-y}{Z'(y)}$, where $Z(y) \equiv y\left(\eta + \frac{\sigma_D}{\omega_t} - \frac{\varphi}{\sigma_D}(1-\tau y)\right) - \left(\eta - \frac{\sigma_D}{1-\omega_t} - \frac{\varphi}{\sigma_D}(1-\tau y)\right)$. $Z(y)$ is a quadratic equation in y with positive leading coefficient, and satisfies $Z(0) < 0$ when $\omega_t < \omega_1^*$. There consequently exists a unique value $y > 0$ such that $Z(y) = 0$; for this value, $Z'(y) > 0$. Hence, $\frac{dy}{d\eta} > 0$.

Next note that $G_y = \kappa + \eta + 2\frac{\varphi}{\sigma_D}\tau y > 0$, $G_\kappa = y + \frac{\omega_t}{1-\omega_t} > 0$, and $G_\eta = y > 0$. This proves $\Phi > 0$.

Finally, note that $Z(y) = G(y, \kappa(y))$. Therefore, $Z'(y) = G_y + G_\kappa\frac{d\kappa}{dy} = G_y(1-\Phi)$. Since $Z'(y) > 0$ at the equilibrium value of y, it follows that $G_y(1-\Phi) > 0$. Since $G_y = y > 0$, it follows that $\Phi < 1$. ∎

Proof of Proposition 3. We note first that, for $w \leq 0$, the function

$$\iota(w,\kappa) \equiv w(\kappa\sigma_D + \varphi) - \frac{1}{2}(w\sigma_D)^2 \tag{D.14}$$

is decreasing in κ, and therefore it attains a higher maximum for equilibrium B (since $\kappa^B < \kappa^A$).

It immediately follows that

$$g_t^B - g_t^A = -\left(\kappa_t^B - \kappa_t^A\right)\sigma_D + \max_{w\leq 0}\iota(w,\kappa_t^B) - \max_{w\leq 0}\iota(w,\kappa_t^A) \geq 0.$$

We further have, based on the expressions for g_t and μ_ω (Equation (24)),

$$\begin{aligned}\mu_\omega^B(\omega_t) - \mu_\omega^A(\omega_t) &= \omega_t\left(g_t^B - g_t^A\right) + \frac{1}{2}\omega_t\left(w_t^B(w_t^B-2) - w_t^A(w_t^A-2)\right)\sigma_D^2 \\ &= \omega_t\left(g_t^B - g_t^A\right) + \frac{1}{2}\omega_t\left(w_t^B - w_t^A\right)(w_t^B + w_t^A - 2)\sigma_D^2 \\ &> 0,\end{aligned}$$

with both of the factors in parentheses in the second term on the right-hand side of the second-to-last line being negative (since $w_t^B < w_t^A < 0$). ■

Proof of Proposition 4. We start by describing the determination of the equilibrium in this case. Fix a time t and let E denote expectations with respect to the wealth distribution over types η at time t. (For notational simplicity, we remove time-subscripts throughout the proof.) For a given Sharpe ratio κ and anticipated utilization ration y, define the following two functions, giving the aggregate long and short positions, respectively.

$$L(y,\kappa) = \mathrm{E}\left[\sigma^{-1}\left(\eta + \kappa + \sigma^{-1}\tau y f(y)\right)^{+}\right] \tag{D.15}$$

$$S(y,\kappa) = \mathrm{E}\left[\sigma^{-1}\left(\eta + \kappa + \sigma^{-1} f(y)\right)^{-}\right]. \tag{D.16}$$

An equilibrium is defined through the two market-clearing conditions

$$1 = L(y,\kappa) - S(y,\kappa) \tag{D.17}$$

$$y = \frac{S(y,\kappa)}{L(y,\kappa)}. \tag{D.18}$$

Furthermore, (D.17) defines κ uniquely as a function of y, so that we can write $S(y) = S(y,\kappa(y))$ and $L(y) = L(y,\kappa(y))$, and the equilibrium determination comes down to

$$F(y) \equiv \frac{S(y)}{L(y)} = y. \tag{D.19}$$

The remainder of the proof is organized as follows. We start by showing that, given y_1 with $h'(y_1) < 0$, a continuous distribution with connected support (thus the density does not drop to zero on an intermediate range to then become positive again) exists for which $F'(y_1) > 1$. Using this property, we show that there exist multiple equilibria for this distribution. The continuity of the problem then ensures that, for any sequence of distributions converging to the one we construct,[55] a sequence of equilibrium utilization rates $y_1^{(n)}$ obtain that converges to y_1, and consequently $F'\left(y_1^{(n)}\right) > 1$ for n large enough. In this sense, the set of type distributions admitting multiple equilibria is not "knife-edge" or even sparse, but in fact has non-empty interior.

For convenience, we define $\bar{h}(y) = \frac{h(y)}{\sigma}$ and note that $\bar{h}'(y) < 0$ is equivalent to $h'(y) < 0$. Equation (D.17) implies that

$$\kappa(y) = \frac{\sigma - \omega^S \bar{\eta}^S - \omega^L \bar{\eta}^L - \left(\omega^S \frac{f(y)}{\sigma} + \omega^L \tau y \frac{f(y)}{\sigma}\right)}{\omega^S + \omega^L}, \tag{D.20}$$

55. Convergence in the space of distribution is defined in terms of convergence of expectations of any smooth function with compact support.

where we defined the quantities

$$\omega^L = \mathrm{E}\left[1_{\{\eta+\kappa+\sigma^{-1}\tau y f(y)\geq 0\}}\right] \tag{D.21}$$
$$\omega^S = \mathrm{E}\left[1_{\{\eta+\kappa+\sigma^{-1} f(y)\leq 0\}}\right] \tag{D.22}$$
$$\bar{\eta}^L = \mathrm{E}\left[\eta \,|\, \eta+\kappa+\sigma^{-1}\tau y f(y)\geq 0\right] \tag{D.23}$$
$$\bar{\eta}^S = \mathrm{E}\left[\eta \,|\, \eta+\kappa+\sigma^{-1} f(y)\leq 0\right]. \tag{D.24}$$

(These quantities depend on y, but we suppress that dependence in our notation.)

Furthermore, one can differentiate the same equation (D.17) with respect to y to obtain

$$\kappa'(y) = -\sigma^{-1}\frac{\omega^S f'(y) + \tau\omega^L\left(f(y) + y f'(y)\right)}{\omega^S+\omega^L}, \tag{D.25}$$

where we have made use of the fact that $\frac{d}{dx}\mathrm{E}[(g(x,\eta))^+] = \mathrm{E}\left[\frac{d}{dx}g(x,\eta)1_{\{g(x,\eta)\geq 0\}}\right]$ for an arbitrary differentiable function g, given that the distribution of η is absolutely continuous.

Using equations (D.16) and (D.20) and the definitions of $h(y)$ and $\bar{h}(y)$, we compute

$$S(y) = \sigma^{-1}\frac{\omega^L\omega^S}{\omega^L+\omega^S}\left(\bar{\eta}^L - \bar{\eta}^S - \frac{\sigma}{\omega^L} - \bar{h}(y)\right) = B^{-1}(A-\bar{h}(y)) \tag{D.26}$$
$$F(y) = \frac{\bar{\eta}^L - \bar{\eta}^S - \frac{\sigma}{\omega^L} - \bar{h}(y)}{\bar{\eta}^L - \bar{\eta}^S + \frac{\sigma}{\omega^S} - \bar{h}(y)} = \frac{A-\bar{h}(y)}{A+B-\bar{h}(y)}, \tag{D.27}$$

where we also defined

$$A \equiv \overline{\eta}^L - \overline{\eta}^S - \frac{\sigma}{\omega^L} \tag{D.28}$$
$$B \equiv \frac{\sigma}{\omega^S} + \frac{\sigma}{\omega^L}. \tag{D.29}$$

Noting now, using (D.16) and (D.25), that

$$S'(y) = -\sigma^{-1}\frac{\omega^L\omega^S}{\omega^S+\omega^L}\bar{h}'(y) = -B^{-1}\bar{h}'(y), \tag{D.30}$$

we use (D.26) and (D.30), as well as $F(y) = \frac{S(y)}{L(y)} = \frac{S(y)}{1+S(y)} = 1 - \frac{1}{1+S(y)}$, to write

$$F'(y) = \frac{S'(y)}{(1+S(y))^2} = \frac{-B\bar{h}'(y)}{\left(A+B-\bar{h}'(y)\right)^2}. \tag{D.31}$$

Our intermediate goal, therefore, is to show that, given $\bar{h}'(y_1) < 0$, values A and B exist

satisfying

$$\frac{A-\bar{h}(y_1)}{A+B-\bar{h}(y_1)} = y_1 \tag{D.32}$$

$$\frac{-B\bar{h}'(y_1)}{\left(A+B-\bar{h}(y_1)\right)^2} = 1+\varepsilon > 1 \tag{D.33}$$

for some $\varepsilon > 0$. In fact, for any $\varepsilon > 0$, solutions A and B to these equations are given by

$$B = (1-y_1)^2 \frac{\left|\bar{h}'(y_1)\right|}{1+\varepsilon} > 0 \tag{D.34}$$

$$A = \bar{h}(y_1) + B\frac{y_1}{1-y_1} = \bar{h}(y_1) + y_1(1-y_1)\frac{\left|\bar{h}'(y_1)\right|}{1+\varepsilon} > \bar{h}(y_1). \tag{D.35}$$

To show the existence of a distribution yielding these desired values of A and B, we first note that the right-hand side of (D.29) can be made arbitrarily close to 4σ while keeping $\omega^L + \omega^S < 1$, and therefore condition b) of the proposition ensures that such ω^L and ω^S exist delivering B for a small enough ε. Fixing ω^L and ω^S, $\bar{\eta}^L$ and $\bar{\eta}^S$ can be chosen arbitrarily subject to (D.28) delivering the desired value of A. We therefore now have the value of $\kappa(y_1)$, which determines the sets of types that go long, respectively short, the asset. Finally, the density of the distribution on each of these two sets can be chosen freely subject to the two integrals defining ω^L and $\bar{\eta}^L$, respectively ω^S and $\bar{\eta}^S$. In the complementary, intermediate type region in which agents are inactive, the density is only subject to a total mass condition.

Finally, with $Y = \min\{1, y \,|\, \bar{h}(y) = A\}$, either $Y < 1$ and $F(Y) = 0 < Y$ or $F(Y) = F(1) < 1 = Y$. Since $F(Y) < Y$ in either case, and $F'(y_1) > 1$, a value $y_2 \in (y_1, Y)$ exists such that $y_2 = F(y_2)$. Thus, a second equilibrium exists. ∎

Proof of Proposition 5. The proof essentially repeats the steps from the one-risky asset case, so we provide only a sketch, focusing on the elements that differ.

With these definitions, the market clearing condition is

$$\widehat{\omega}_t \sum_{i\in\{I,R\}} \omega_t^i \overrightarrow{w}_t^i + (1-\widehat{\omega}_t)\begin{bmatrix} 0 \\ \widehat{w}_{2,t} \end{bmatrix} = \overrightarrow{m}_t. \tag{D.36}$$

We consider first an equilibrium with $y_t > 0$. Investor R's and I's optimal portfolios are given by

$$\overrightarrow{w}_t^R = (\sigma_t\sigma_t')^{-1}(\overrightarrow{\mu}_t - r_t 1_{2\times 1} + \overrightarrow{\varphi}), \tag{D.37}$$

$$\overrightarrow{w}_t^I = (\sigma_t\sigma_t')^{-1}(\overrightarrow{\mu}_t - r_t 1_{2\times 1} + \sigma_{1,t}\overrightarrow{\eta} + \tau y_t \overrightarrow{\varphi}). \tag{D.38}$$

Using (D.37) inside (D.36) yields

$$(\sigma_t\sigma_t')\,\overrightarrow{m}_t = \widehat{\omega}_t\left(\omega_t\left(\overrightarrow{\mu}_t - r1_N + \overrightarrow{\varphi}\right) + (1-\omega_t)\left(\overrightarrow{\mu}_t - r1_N + \sigma_1\overrightarrow{\eta} + \tau y_t\overrightarrow{\varphi}\right)\right) + (1-\widehat{\omega}_t)\,(\sigma_t\sigma_t')\begin{bmatrix} 0 \\ \frac{\mu_{2,t}-r}{\sigma_{2,t}^2} \end{bmatrix}. \tag{D.39}$$

Next we use the row selection vector $[0,1]$ to pre-multiply both sides of (D.39). Noting that $[0,1]\,\overrightarrow{\varphi} = [0,1]\,\overrightarrow{\eta} = 0$, and also

$$(\sigma_t\sigma_t')\begin{bmatrix} 0 \\ \frac{\mu_{2,t}-r}{\sigma_{2,t}^2} \end{bmatrix} = \begin{bmatrix} b_t\,(\mu_{2,t}-r) \\ \mu_{2,t}-r \end{bmatrix}, \tag{D.40}$$

leads to (38). We next note that

$$[1,-b_t]\,\sigma_t\sigma_t'\begin{bmatrix} m_{1,t} \\ m_{2,t} \end{bmatrix} = [\sigma_{1,t},0]\begin{bmatrix} \sigma_{1,t} & 0 \\ b_t\sigma_{2,t} & \sigma_{2,t} \end{bmatrix}\begin{bmatrix} m_{1,t} \\ m_{2,t} \end{bmatrix} = \sigma_{1,t}^2 m_{1,t}. \tag{D.41}$$

Pre-multiplying both sides of (D.39) with the row vector $[1,-b_t]$, using (D.40), (D.41), and the definition of $\kappa_{1,t}$, and re-arranging yields

$$\kappa_{1,t} = \widetilde{m}_{1,t}\sigma_{1,t} - (1-\omega_t)\,\eta - \frac{\varphi}{\sigma_{1,t}}\left(\omega_t + (1-\omega_t)\tau y_t\right). \tag{D.42}$$

Using the definition of $\kappa_{1,t}$ inside (D.37) gives

$$w_{1,t}^R = \frac{\kappa_{1,t}}{\sigma_{1,t}} + \frac{\varphi}{\sigma_{1,t}^2} \tag{D.43}$$

$$w_{1,t}^I = \frac{\kappa_{1,t}+\eta}{\sigma_{1,t}} + \frac{\tau y_t\varphi}{\sigma_{1,t}^2}, \tag{D.44}$$

where we used the notation $w_{1,t}^i$, $i \in \{R,I\}$, to denote the first element of w_t^i.

Using the market clearing condition $y_t = -\frac{\omega_t^R w_{1,t}^R}{\omega_t^I w_{1,t}^I} = -\frac{\omega_t w_{1,t}^R}{(1-\omega_t)w_{1,t}^I}$ leads to (36).

If agent R chooses not to short then the market clearing condition becomes

$$\widehat{\omega}_t(1-\omega_t)\,\overrightarrow{w}_t^I + (1-\widehat{\omega}_t)\begin{bmatrix} 0 \\ \widehat{w}_{2,t} \end{bmatrix} = \overrightarrow{m}_t. \tag{D.45}$$

Substituting in $\overrightarrow{w}_t^I$ from (D.38) and pre-mutiplying by $\sigma_t\sigma_t'$ gives

$$(\sigma_t\sigma_t')\,\overrightarrow{m}_t = \widehat{\omega}_t(1-\omega_t)\left(\overrightarrow{\mu}_t - r1_N + \sigma_1\overrightarrow{\eta}\right) + (1-\widehat{\omega}_t)\,(\sigma_t\sigma_t')\begin{bmatrix} 0 \\ \frac{\mu_{2,t}-r}{\sigma_{2,t}^2} \end{bmatrix}. \tag{D.46}$$

Premultiplying (D.46) by the row $[1, -b_t]$ and using (D.40) and (D.41) gives

$$\sigma_{1,t}^2 \widetilde{m}_{1,t} = (1-\omega_t)\,\sigma_{1,t}\,(\kappa_{1,t}+\eta)\,,$$

and therefore

$$\kappa_{1,t} = \sigma_{1,t}\frac{\widetilde{m}_{1,t}}{1-\omega_t} - \eta. \tag{D.47}$$

Finally, when both agents hold positive portfolios, the optimal portfolios are $\overrightarrow{w}_t^R = (\sigma_t\sigma_t')^{-1}(\overrightarrow{\mu}_t - r_t 1_{2\times 1})$, $\overrightarrow{w}_t^I = (\sigma_t\sigma_t')^{-1}(\overrightarrow{\mu}_t - r_t 1_{2\times 1} + \sigma_{1,t}\overrightarrow{\eta})$. Repeating the arguments in Equations (D.37)–(D.42), we obtain $\kappa_{1,t} = \widetilde{m}_{1,t}\sigma_{1,t} - (1-\omega_t)\,\eta$. ∎

Proof of Proposition 6. It remains to derive the differential equation in Proposition 6. Using the market clearing condition $\sum_{i\in\{I,R\}}\omega_t^i w_{1,t}^i = \widetilde{m}$, and applying Ito's Lemma to $\omega_t^i = \frac{W_t^i}{W_t^I + W_t^R}$ leads to

$$d\omega_t^i = \mu_{\omega,t}^i dt + \sigma_{\omega,t}^i dB_{1,t} \tag{D.48}$$

with

$$\begin{aligned}
\mu_{\omega,t}^i &= \omega_t^i\left[\left(w_{1,t}^i - \widetilde{m}\right)\sigma_{1,t}\left(\kappa_t - \sigma_{1,t}\widetilde{m}\right) + w_{1,t}^i f_t + \widetilde{n}_t\right] + \chi\left(\nu_t^i - \omega_t^i\right),\\
\sigma_{\omega,t}^i &= \omega_t^i\left(w_{1,t}^i - \widetilde{m}\right)\sigma_{1,t},
\end{aligned}$$

and[56]

$$\widetilde{n}_t \equiv -\sum_{i\in\{I,R\}} w_{1,t}^i {\omega_t}^i \lambda_t^i = \frac{y_t\widetilde{m}}{1-y_t} f_t\,(1-\tau)\,.$$

Since $\frac{\phi_1}{\phi_2} \approx 0$, the aggregate endowment follows a geometric Brownian motion in the limit, and the interest rate is constant $r_t = r$. Accordingly, the price of a stock of type 1 follows the dynamics

$$\frac{dP_{1,t,s} + D_{1,t,s}dt}{P_{1,t,s}} = (r + \kappa_{1,t}\sigma_{1,t})dt + \sigma_t dB_{1,t}. \tag{D.49}$$

Applying Ito's Lemma to the product $P_{1,t,s} = p\,(\omega_t)\,D_{1,t,s}$ also implies that

$$\frac{dP_{1,t,s}}{P_{1,t,s}} = \frac{dp_t}{p_t} + \frac{dD_{1,t,s}}{D_{1,t,s}} + \frac{p'\,(\omega_t)}{p\,(\omega_t)}\sigma_{\omega,t}^R\sigma_{1,D}dt. \tag{D.50}$$

56. Using $\sum_{i\in\{I,R\}} w_{1,t}^i\omega_t^i = \widetilde{m}_t$, the definition $y_t = -\frac{w_{1,t}^R\omega_t 1_{\{w_{1,t}^R<0\}}}{w_{1,t}^I\omega_t^I}$ and the definition of λ_t^i leads to

$$-\sum_{i\in\{I,R\}} w_{1,t}^i\omega_t^i\lambda_t^i = \frac{y_t\widetilde{m}}{1-y_t} f_t\,(1-\tau)\,.$$

Combining (D.49) with (D.50) and using $\sigma_{1,t} = \frac{p'(\omega_t)}{p(\omega_t)}\sigma^R_{\omega,t} + \sigma_{1,D}$ and Ito's Lemma to compute the drift of $\frac{dp_t}{p_t}$ leads to

$$\frac{1}{2}\frac{\partial^2 p}{\partial \omega_t^2}\left(\sigma^R_{\omega,t}\right)^2 + \frac{\partial p}{\partial \omega_t}\left(\mu^R_{\omega,t} + \sigma^R_{\omega,t}\sigma_{1,D}\right) - p \times (r + \delta_1 + \kappa_{1,t}\sigma_{1,t}) + 1 = 0, \tag{D.51}$$

which in turn leads to (C.6) after substituting $\sigma_{1,t} = \frac{p'(\omega_t)}{p(\omega_t)}\sigma^R_{\omega,t} + \sigma_{1,D}$. ■

E Additional Data Discussion

E.1 Summary Statistics - IHS Markit

We start by reporting some summary statistics on lending fees. In Table E.1, we group Russell 3000 constituents based on their end-of-prior-year market capitalization into five quintiles. We then fix the set of stocks in each quintile over the subsequent year and compute various statistics (median, 75th percentile, etc.) of the daily lending fees for the stocks in each quintile. We then average across the years. The table shows that the median lending fee ranges between 0.35% and 0.41%. However, the table also shows that some of the observations on lending fees can become quite large. For instance, for stocks that are in the size portfolios 1, 2, and 3, the 95-th percentile of fees exceeds 2 % and the 99-th percentile exceeds 7% for stocks in portfolios 1, 2, 3, and 4. This table suggests that sometimes even relatively large stocks (by market capitalization) can exhibit sizeable lending fees.

Table E.2 helps to illustrate this last point in greater detail. Specifically, Table E.2 reports some stock-level statistics on lending fees, and in particular the fraction of Russell 3000 constituents for which a given percentile of shorting fees across time exceeds certain cutoffs. The table shows that 96% of Russel 3000 constituents exhibit a lending fee in excess of 1% at some point between 2006 and 2021, while 45% of stocks exhibit a fee in excess of 5% at some point over that same time period. But even if we leave these extreme observations aside, and focus on — say — the 95-th percentile of the distribution of lending fees at the stock level, the numbers are large: 31% of Russel 3000 constituents exhibit a lending fee in excess of one percent for 5 out of 100 trading days, while 18% of Russel constituents exhibit lending fees in excess of 3% for 5 out out of 100 trading days.

E.2 Heterogeneous $h'(y)$

For our baseline results we pooled observations across all stocks and estimated a single function $h'(y)$. Figure E.1 shows results for the case where we allow $h'(y)$ to differ for each stock. Specifically, we focus on observations that are on special (DCBS > 1) and estimate a separate $h'(y)$ for each Russell 3000 constituent. We then evaluate $h'(y)$ for different values of y for each stock separately. Subsequently, we pool all $h'(y)$ values across all stocks and present them as a bin-scatter diagram.[57] Since stock-level estimates of $h'(y)$ are noisy, we

57. Since the observations per stock are not in the millions (as they are for the pooled regressions in the text), it is computationally feasible to use a kernel regression estimator with automatic, cross-validated, bandwidth selection. We present the results for this alternative estimation method, as a check that our

Table E.1: Summary Statistics of Shorting fees.

	Percentile				
Size quintile	50^{th}	75^{th}	90^{th}	95^{th}	99^{th}
1	0.41%	0.84%	2.96%	7.43%	28.48%
2	0.38%	0.50%	1.44%	3.97%	19.38%
3	0.36%	0.43%	0.86%	2.04%	12.30%
4	0.34%	0.39%	0.53%	1.02%	7.39%
5	0.35%	0.38%	0.43%	0.50%	1.72%
Total	0.37%	0.51%	1.24%	2.99%	13.85%

Lending fees by stock market capitalization quintile. Each year, we form 5 portfolios of Russell 3000 constituents sorted into size quintiles based on end-of-prior-year market capitalization. Within each size quintile, we compute the p^{th} percentile, $p \in \{50, 75, 90, 95, 99\}$, of daily shorting fees over the following year. We then report the time-series average of these percentiles from 2006 to 2021. Daily shorting fees from 2006 to 2021 are from Markit and are reported as annualized percentage rates.

Table E.2: Stock-level distribution of shorting fees.

	Shorting fee cutoffs				
Percentile	≥1%	≥2%	≥3%	≥5%	≥10%
90^{th}	0.23	0.16	0.12	0.09	0.05
95^{th}	0.31	0.21	0.18	0.14	0.07
99^{th}	0.50	0.32	0.26	0.21	0.13
99.5^{th}	0.62	0.38	0.30	0.23	0.14
Maximum	0.96	0.79	0.66	0.45	0.27

Fraction of Russell 3000 constituents for which the indicated percentile (first column) of daily shorting fees exceeds the cutoff noted in the header row. For example, the bottom rightmost number (0.27) means that 27% of the stocks in the Russell 3000 had a maximum daily shorting fee in excess of 10%. Similarly, the number 0.12 in the top row/ middle column indicates that 12% of the stocks have a lending fee in excess of 3 percent for one out of the ten trading days. Daily shorting fees from 2006 to 2021 are from Markit and are reported as annualized percentage rates.

trim stock-level estimates of $h'(y)$ at the 10-th and 90-th percentile levels. (Results are similar if we don't trim and instead report medians by shorting-utilization bin.) The main conclusion from Figure E.1 is similar to our conclusion in the text: for low values of y, $h'(y)$

conclusions are not driven by whether we use kernels or splines to estimate the non-parametric regression.

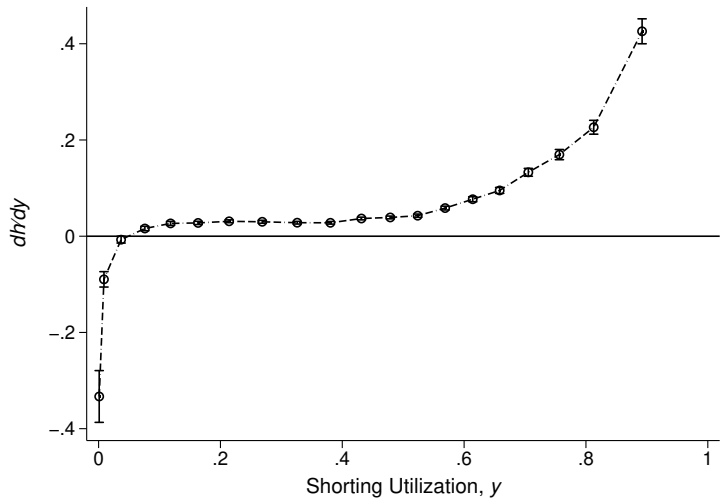


Figure E.1: $h'(y)$, binned-means from stock-level estimates using local-linear non-parametric kernel regressions. For each stock, we estimate the marginal effect $h'(y)$ at 7 points, corresponding to the stock-level 5^{th}, 10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th}, and 95^{th} percentile of utilization for observations exhibiting a Daily Cost of Borrowing Score (DCBS) over 1. Error bars represent 95% confidence intervals around bin means. Data on shorting fees and shorting utilization are from Markit over the period 2006 to 2021.

is negative.

E.3 The relation between jumps and specialness

One would expect the economic issues identified by the model to be more relevant among stocks that are harder to short. Table E.3 confirms this. The incidence of utilization jumps is larger among stocks that exhibit a higher frequency of trading days with a DCBS score above one and stocks that have a higher average DCBS score.

E.4 Alternative ways to detect jumps

In the main text we opted for a very simple and transparent definition of a jump. Specifically, we defined jumps as instances where utilization jumps by more than a given cutoff. This approach is intuitive and helps identify economically large jumps.

In this subsection we consider a more sophisticated approach to identifying jumps in utilization. To start, we note that there are various approaches in the econometric literature to test whether a given time series, which is observed at discrete time intervals, emanates from a continuous sample-path process against the alternative hypothesis that the underlying stochastic process exhibits jumps. In testing for jump-discontinuities, we opted to use the test of Aït-Sahalia and Jacod (2009), which is based on testing for some general properties of higher-order moments of diffusions at short observation intervals. To apply the test we use the highest frequency data that we have on utilization (daily) and isolate contiguous,

Table E.3: Regressions of Jump Rate on stock-level measures of specialness

	Annualized Jump rate					
	$\|\Delta y\| \geq 5.5\%$		$\|\Delta y\| \geq 8.0\%$		$\|\Delta y\| \geq 10.0\%$	
Panel A: Large changes in Shorting Utilization						
Pct. special	23.388***		20.835***		19.627***	
	(46.321)		(38.405)		(35.133)	
DCBS		4.573***		4.077***		3.841***
		(31.476)		(27.928)		(26.279)
N	6156	6156	6156	6156	6156	6156
Panel B: Large changes in Shorting Utilization driven by changes in Shorting Demand						
Pct. special	11.827***		10.352***		9.648***	
	(38.531)		(32.273)		(29.455)	
DCBS		2.171***		1.899***		1.770***
		(24.871)		(22.086)		(20.695)
N	6156	6156	6156	6156	6156	6156

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Jump rate is calculated as annualized rate of detected jumps in shorting utilization. Jumps are identified as trading weeks during which the absolute change in shorting utilization exceeds, depending on the specification, 5.5%, 8%, or 10%. Percent special is the fraction of trading days on which the stock has a DCBS greater than 1. DCBS is the simpe time-series average of daily DCBS scores for each stock.

uninterrupted samples of daily data without a change in the number of outstanding shares.[58] If the Aït-Sahalia and Jacod (2009) rejects continuity, we proceed to identify the time of the occurrence of the jump(s) using the jump-robust volatility estimator of Wang and Zheng (2022). Specifically, for a discretely-observed process, x, Wang and Zheng (2022) provides a jump-robust estimate of local diffusive volatility, $\hat{\sigma}(x)$. Using this estimate, we can compute $\left|\frac{\Delta x}{\hat{\sigma}(x)}\right|$, which can be interpreted as a local "z-score" for the daily change in x. We identify the dates when this z-score is above 4 as "jump dates," in order to isolate economically meaningful jumps.

Table E.4 repeats the analysis of Table 1, but using this alternative definition of a jump. Specifically, for each stock we count the number of jumps according to the procedure described in the above paragraph (assigning a value of zero jumps to stocks where the Aït-Sahalia and Jacod (2009) test cannot reject continuity). We then divide by the number of trading-day observarions for the respective stocks to arrive at a jump rate for each stock. Table E.4 shows that the results of Table 1 remain unchanged when we use this alternative jump-rate definition.

58. This latter restriction is done to ensure that utilization does not change due to –say– share issuance.

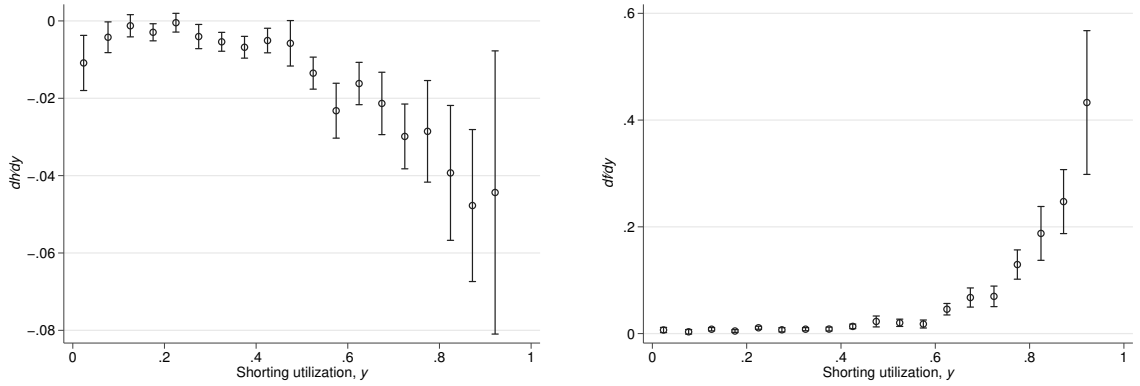
Table E.4: Alternative identification of the jump rate.

	Jump rate			
Pct. special	15.829***			
	(22.358)			
DCBS		3.682***		
		(16.738)		
$\mathbf{1}_{h'<0}$			0.177***	
			(4.793)	
$\mathbf{1}_{\text{Reject } h'>0}$				0.097**
				(2.635)

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

This table repeats the regressions of Table E.3 and Table 1, except that the jump rate is computed according to the methodology described in section E.4.



(a) $h'(y)$ (b) $f'(y)$

Figure E.1: $h'(y)$ and $f'(y)$, binned-means from jump discontinuities. Estimated marginal effects are calculated by dividing observed weekly changes in $h(y) = f(y)\,(1 - \tau\, y)$ by observed weekly changes in shorting utilization, conditional on sufficiently large weekly changes in shorting utilization, defined to be a change in shorting utilization whose magnitude exceeds 5.5%. We calibrate τ to be 0.8 based on industry literature on the pass-through of shorting fees to institutional investors. Sample consists of daily observations of shorting fees and shorting utilization for Russell 3000 constituents. Error bars represent 95% confidence intervals around bin means. Data on shorting fees and shorting utilization are from Markit over the period 2006 to 2021.

F Details and Additional Results for Section 6.4

F.1 Measuring ticker discussion on WallstreetBets

Our measure of ticker mentions on WallstreetBets is constructed as follows. We use the PushshiftAPI to collect all submissions posted on WallstreetBets subreddit from January 1, 2020 through February 7, 2021 (Baumgartner et al. 2020). For each submission, we observe the title text, the body of the submission, the author of the submission, and the time of the submission.

We then identify all cases in which these tickers are mentioned in submissions, irrespective of whether they are prefixed with a dollar sign. To address the possibility of falsely identifying tickers, we require that, if the ticker is a common word in the written English language, it must be prefaced by a dollar sign. For example, AT&T's ticker T is also a common word in written English, and thus we require that the text "$T" appear in a submission for it to be considered as mentioned AT&T. We consider a ticker as being mentioned in a submission if it appears in either the title or the body of the submission. We identify common word-stems based on the Google Trillion Word Corpus (Michel et al. 2011). In a robustness check, we account for the downward bias this restriction introduces by scaling common-word tickers by an in-sample estimated adjustment factor. This adjustment leaves the relative ranking of ticker mentions largely unchanged. We estimate the adjustment factor by comparing the frequency of tagged ticker mentions versus untagged ticker mentions for the set of tickers which do not commonly appear in written English.

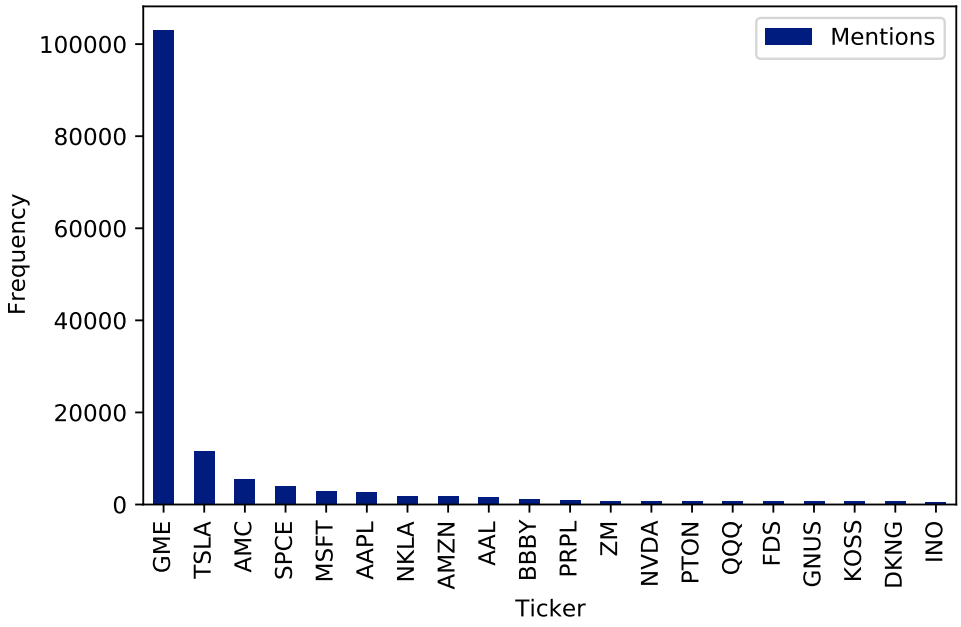
Revised submissions and comments. Authors of Reddit comments have the ability to edit their comments even after the comment has been posted. The PushshiftAPI records the comment text as of a certain day, and does not update to reflect potential revised comments. The same constraint applies to the content body of submissions. Titles of submissions cannot be revised and thus do not have this measurement problem.

Missed tickers Tickers that, for whatever reason, are never tagged with a leading dollar sign will be omitted from our dataset. Similarly, we under-count the occurrences of tickers that are common words, owing to requiring they appear with a leading "$" We attempt to correct for this by scaling the observed counts for common word tickers. For AAPL and GME, which are not common word tickers, the ticker appears with the leading "$" roughly 20% of the time. We can thus simply multiply our observed frequencies by a factor of five to adjust for the more stringent matching procedure. As can be seen in Figures F.1a and F.1b, the adjustment does not have a significant impact on the relative popularity of the top tickers.

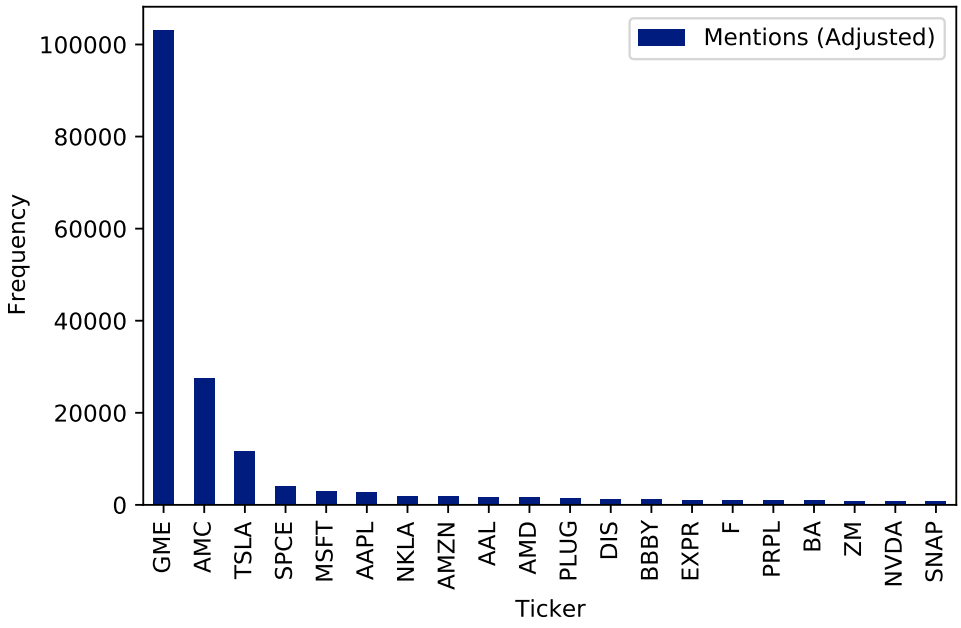
In some cases, users may choose to refer to the company by its name, rather than by its ticker. We do not attempt to identify mentions of companies by name.

F.2 Measuring retail trading

We adopt the methodology of Boehmer et al. (2020) to identify retail trades in the TAQ data. We briefly summarize the methodology here and refer readers to the paper for details.



(a) Submissions mentioning each Ticker



(b) Submissions mentioning each Ticker, adjusted for word-ticker overlap

Figure F.1: Popular Tickers on WallstreetBets (January 1, 2020 – February 7, 2021).

The intuition behind the methodology is the knowledge that retail trades are often executed by wholesalers or via broker internalization, rather than on the major trading exchanges. These trades appear in the TAQ consolidated tape data under the exchange code "D." These trades are given a small price improvement on the order of tenths of a penny as a means to induce brokers to route orders to the wholesaler. Similarly, brokers which internalize retail trades offer a subpenny price improvement in order to comply with Regulation 606T. Importantly, institutional trades are rarely, if ever, internalized or directed to wholesalers and their trades are usually in round penny prices, with the notable exception of midpoint trades.

The methodology of Boehmer et al. (2020) uses these institutional details to identify retail trades in the TAQ consolidated tape data. Trades flagged with exchange code "D" and with a subpenny amount in the set $(0, 0.40) \cup (0.60, 1.00)$ are identified as retail trades. Splitting these trades further, retail trades with subpenny amounts between zero- and forty-hundredths of a penny are labeled as "sell orders," whereas subpenny amounts between sixty- and one hundred-hundredths are considered "buy orders." The midpoint trades are excluded to avoid mis-classifying institutional trades executed at midpoints as retail trades.

F.2.1 Challenges

Derivatives The TAQ data only contains trades of equities. Options offer another way to benefit for investors to benefit from increases in the price of stock. As an added advantage for retail investors, options offer embedded leverage greater than what might otherwise be available through their broker. The Boehmer et al. (2020) methodology relies on institutional details to identify off-exchange retail trades, and thus cannot reliably identify replication trades by market makers.

F.3 Betting against the shorts portfolio

As is standard in the literature, we restrict attention to common shares of COMPUSTAT firms which trade on the NYSE, AMEX, and NASDAQ exchanges. We further exclude companies for whom no share class has a price exceeding $1. The strategy equally weights each firm in the top decile, shorts the market index, and reconstitutes 8 trading days following the disclosure date, which is the first opportunity following the public dissemination of the short interest data.

G Additional Table and Figures

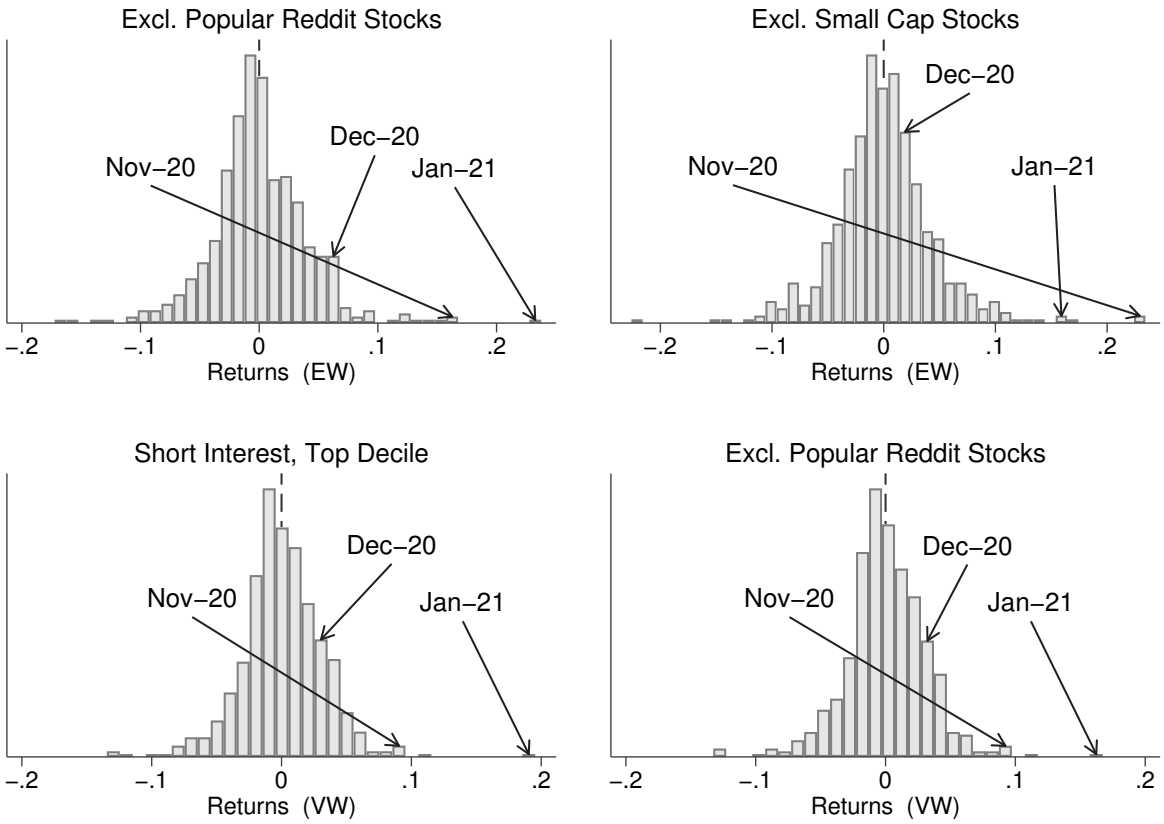


Figure G.1: Monthly returns (1973–2021). Histograms show monthly returns to a trading strategy long stocks in the top decile of short interest and short the market index. The top-left plot depicts equal-weighted returns, excluding the six most-popular stocks discussed on Reddit (AMC, BBBY, GME, SPCE, TLRY, and TSLA). The top-right plot depicts equal-weighted returns, further excluding small market capitalization stocks. The bottom-left plot depicts value-weighted returns. The bottom-right plot depicts value-weighted returns, excluding popular stocks discussed on Reddit. The arrows indicate the portfolio returns in the months of November and December 2020 and January 2021.

	Highly Shorted Stocks	Excl. Popular Reddit Stocks	Excl. Small Stocks
Panel A: November 2020			
r^{EW}	0.163	0.160	0.227
	(4.127)	(4.050)	(5.194)
r^{VW}	0.094	0.092	0.133
	(3.062)	(3.029)	(3.455)
r^{EW}_{FF3}	0.084	0.081	0.160
	(3.452)	(3.327)	(4.371)
r^{VW}_{FF3}	0.045	0.043	0.083
	(1.769)	(1.706)	(2.431)
Panel B: December 2020			
r^{EW}	0.055	0.058	0.019
	(1.385)	(1.477)	(0.437)
r^{VW}	0.033	0.036	0.021
	(1.088)	(1.191)	(0.540)
r^{EW}_{FF3}	0.012	0.016	-0.002
	(0.515)	(0.665)	(-0.056)
r^{VW}_{FF3}	0.012	0.014	0.008
	(0.466)	(0.576)	(0.244)
Panel C: January 2021			
r^{EW}	0.271	0.232	0.156
	(6.835)	(5.865)	(3.576)
r^{VW}	0.194	0.161	0.183
	(6.341)	(5.293)	(4.764)
r^{EW}_{FF3}	0.208	0.169	0.121
	(8.560)	(6.978)	(3.296)
r^{VW}_{FF3}	0.171	0.136	0.165
	(6.709)	(5.452)	(4.816)

Table G.1: Portfolio returns (November 2020–January 2021). Test of whether the monthly return to the strategy of betting against the shorts is "abnormal" in November 2020 (Panel A), December 2020 (Panel B), and January 2021 (Panel C). The table reports the coefficient and the t-statistic of the month dummy variable that takes the value of one for the month listed in the title of the panel and zero otherwise from the regression:

$$r_{\text{Betting against the shorts}} = \text{const.} + \text{month dummy} + \beta' F_t + \varepsilon_t.$$

The first two rows of each panel do not control for any factor exposures and refer to equal-weighted (EW) and value-weighted (VW) returns, respectively. The last two rows of each panel control for Fama-French 3-factor exposures.

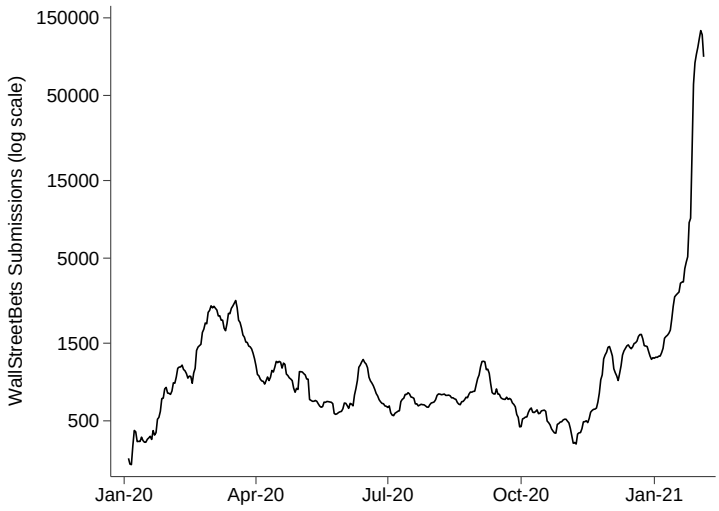


Figure G.2: Seven-day moving average of daily submissions to the WallstreetBets subreddit (January 1, 2020 – February 7, 2021). The vertical axis is on a logarithmic scale.

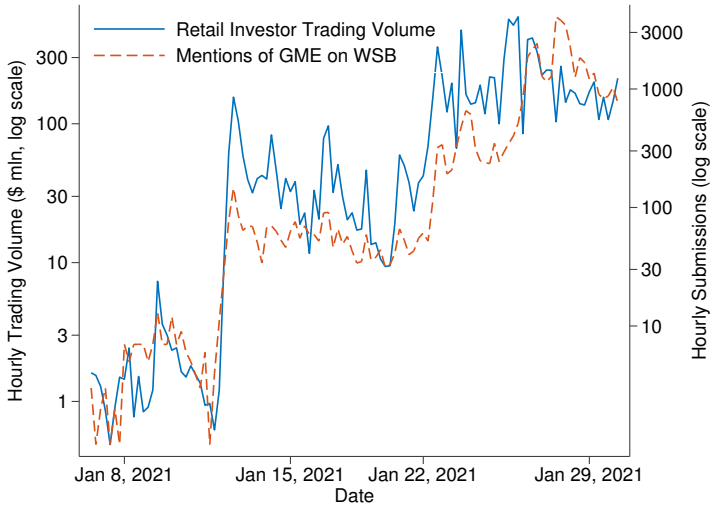


Figure G.3: Retail trading volume in GME (January 7 – January 29, 2021). Hourly trading volume in GME, measured using the methodology of Boehmer et al. (2020), plotted together with hourly mentions of the GME ticker on the WallStreetBets subreddit. Both vertical axes are on logarithmic scales.

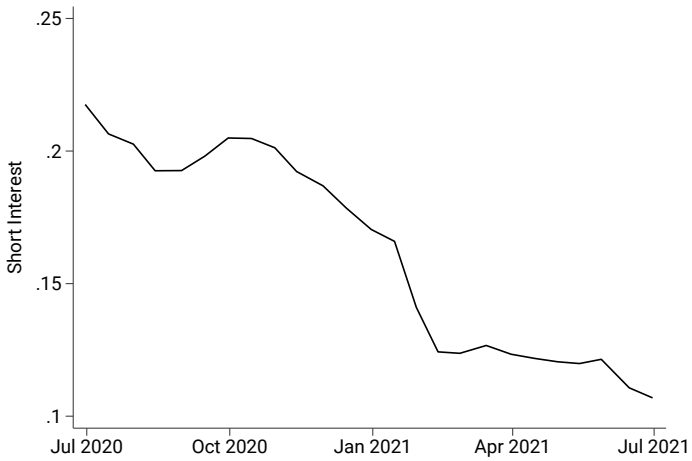


Figure G.4: Aggregate short interest (July 2020–June 2021). The figure plots value-weighted short interest for highly shorted stocks as of October 31, 2020. Highly shorted stocks are defined as the stocks in the top decile of the Russell 3000, ranked by short interest. The identities of these stocks is fixed and their short interest is plotted over the preceding four and subsequent eight months.

References - Online Appendix

Aït-Sahalia, Yacine, and Jean Jacod. 2009. "Testing for jumps in a discretely observed process." *The Annals of Statistics* 37, no. 1 (February).

Baumgartner, Jason, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. *The Pushshift Reddit Dataset.*

Blanchard, Olivier J. 1985. "Debt, Deficits, and Finite Horizons." *Journal of Political Economy* 93 (2): 223–247.

Boehmer, Ekkehart, Charles M Jones, Xiaoyan Zhang, and Xinran Zhang. 2020. "Tracking retail investor activity." *Journal of Finance, forthcoming.*

Karatzas, Ioannis, and Steven Shreve. 2012. *Brownian motion and stochastic calculus.* Vol. 113. Springer Science & Business Media.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Google Books Team, Joseph P. Pickett, et al. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331 (6014): 176–182.

Wang, Bin, and Xu Zheng. 2022. "Testing for the presence of jump components in jump diffusion models." *Journal of Econometrics* 230, no. 2 (October): 483–509.