Short selling in practice: Intermediating uncertain share availability

Gene M. D'Avolio
Harvard Business School

André F. Perold
Harvard Business School

October 12, 2003

Abstract

Uncertain availability of shares to borrow increases the risk of short selling. In practice, intermediaries such as custody banks and prime brokers ameliorate short selling risk through pooling loan supply and employing loan contracts that embed options to extend and fail. These and other institutional features (e.g., average cost pricing) have a significant impact on observed marginal loan rates and the quantity and timing of short sales. We explore the efficiency of these practices with a model of the intertemporal trading strategy of a short seller, where the options to extend and fail are priced into loan fees. Our model provides a framework to isolate short seller beliefs and to understand the function of intermediation in this market. The model generates several empirical predictions for security loan fees and short interest dynamics.
1. Introduction

The ability to obtain negative exposure to price movements is an important aspect of any well functioning capital market. For roughly four centuries, financial markets have utilized two general vehicles for negative exposure: short selling and derivative contracts.\(^1\) Derivatives such as options, swaps, and futures trade in markets specific to those contracts. Derivative contracts are "side bets" that can be created spontaneously - even when neither party has possession of the underlying security. In contrast, short selling operates through a combination of two markets that require the transfer of physical securities: the loan market for borrowing shares and the cash market where they are sold. The availability of actual shares is thus crucial for short selling.\(^2\)

Uncertainty regarding future availability of shares can generate ex-ante reluctance among short sellers. Having sunk transaction costs, the short seller is vulnerable to having his loan recalled and his short position closed out before he can capture an anticipated (or even certain) price decline. Alternatively, but with a similar economic consequence, the short seller might be "held up" by the lender who could threaten to terminate the loan unless the loan fee is increased to an expropriating level. Hesitation leads to squandered profit opportunities for lenders as well as short sellers and represents a dead-weight loss. Intermediaries respond to this problem by providing services and loan contracts that reduce uncertainty about future availability and loan fees. The intermediaries have comparative advantages in performing these functions and are therefore core to the securities lending industry. This paper provides a framework to quantify the impact of intermediation on the costs, risks, quantity, and timing of short selling. We show how an understanding of current institutional practices is essential to extracting stock price beliefs from observed securities lending fees and short interest dynamics.

\(^1\) De La Vega (1688) and De Marchi and Harrison (1994) describe short selling in 17th Century Amsterdam. The former also discusses options ("opsies").
\(^2\) Short selling also inherits regulation specific to the two markets that facilitate it. For example the SEC and stock exchanges have broad authority over short selling because it uses the cash market. The uptick rule is an example a resulting regulation that distinguishes short selling from derivative trading.
In the U.S. market, short sellers rarely borrow shares directly from owners. Instead, most short
sellers borrow from prime brokers (e.g., Bear Stearns, Goldman Sachs, Morgan Stanley) who, having
exhausted internally sourced supply, borrow from custody banks (e.g., Bank of New York, Northern
Trust, State Street). The custody banks, in turn, are lending agents for large institutional owners such
as mutual funds, public retirement funds, corporate pension funds, and insurance companies. Lending
among these intermediaries (the "wholesale" market) is rarely done on an overnight ("spot") basis or
for a fixed maturity ("term"). Instead, it is common practice for the borrowing broker to have the
option to extend stock loans indefinitely at the originally agreed upon fee rate until the custody bank
no longer has shares available to loan in its centralized pool of customer holdings. When lender
holdings fall below loan balances, loans are "recalled" (in part or in full). Further, custody banks
usually give prime brokers the option to temporarily fail to return recalled shares (at a pre-specified
fee, typically expressed in the form of a non-positive rebate). The option to fail to return borrowed
shares mirrors the well-known option in the cash market to fail to deliver sold shares. Lending
intermediaries have the right to use posted collateral to "buy in" recalled stocks but they rarely do so
and rely instead on the credit worthiness of the prime brokers for eventual return of the shares. One
contribution of this paper is to value these embedded options to extend and fail to better identify short
seller beliefs when observing the wholesale fees paid by their brokers.

Another important fact concerning institutional practice is that several major prime brokers do
not charge short sellers the wholesale fee on the marginal shares borrowed for that client. Instead,
short sellers pay a blended fee that equals the broker's balance weighted cost for all outstanding loans
in that issue plus some markup. For example, if a prime broker borrowed 1 million shares of an issue
for a fee of 1% (per annum) for customer A, and subsequently borrowed another million shares at a fee

---

3 Most institutional lenders face regulatory prohibitions against lending equities and corporate bonds for term. Pension
funds and mutual funds are required, by ERISA laws and the Investment Act of 1940 respectively, to retain the right to
cancel the loan at any time. Furthermore, the IRS may treat as a sale a stock loan that is not immediately callable. See ICI
(1998) and Souder (1996) for details. This is in contrast to the US Treasury REPO market where term loans are common.
4 We confirmed this pricing convention in interviews with several prime brokers and hedge funds. According to these
interviews, the markup over average cost is roughly 20 to 25 basis points on easy to borrow ("general collateral") issues
and up to several percent for hard to borrow ("special" or "hot") issues.
of 9% to facilitate the short sale of customer B, both A and B would be charged the average fee of 5% plus any markup.\(^5\) Prime brokers thus appear not to be acting as borrowing agents on specific transactions but rather to be attempting to equalize rates across their short seller customers. Note in this example that the 9% loan fee the prime broker pays the custody bank can be substantially higher than the declines anticipated by either of the short sellers. As long as both investors expect the stock to fall by more than 5% plus the markup, the resulting average fees can be sustained in equilibrium. The option to extend earlier loans transacted at lower fees (1% in this example) combines with average cost pricing to generate non-intuitive behavior in the marginal wholesale loan fees. For example, over some ranges marginal wholesale loan fees can be increasing in the total supply of loans. This is because a larger existing base of cheap borrows can more effectively subsidize (via averaging) the higher marginal fees on incremental loans. These institutional practices mean that marginal fees in the wholesale market might far exceed the (annualized) decline in share price expected by the marginal short seller (who likely is paying a fee related to the average cost).\(^6\)

By pooling the holdings of a large universe of investors, lending intermediaries rarely lose availability. The custody bank studied by D'Avolio (2002) recalled only 2% of loans per month during his sample (April 2000 to September 2001). Of course, conditional on being “special” or hard to borrow, loans are much more likely to be recalled. When recalled - and unable to locate shares from another source - a prime broker will generally allow customers to stay short for as long as the custody bank allows failure. As described in section 2, failure in the loan market appears to be tolerated for weeks and even months.

The main hypothesis of this paper is that the practices of granting options to extend and fail, along with the benefits of pooled loan supply, permit intermediaries to reduce uncertainty about availability and loan fees and to mitigate short seller reluctance. In turn, these practices induce more

---

\(^5\) Unless otherwise noted, we will quote % fees in as percent \textit{per annum}.

\(^6\) This feature of the data is important not only to practitioners, but also to the growing number of academic researchers who study wholesale stock loan fees - e.g., D'Avolio (2002), Gezcy, Musto and Reed (2002), Reed (2002), and Jones and Lamont (2002).
"efficient" quantities of short selling. To explore this hypothesis, we develop a simple three-period model of an informed agent's decision to sell short immediately or wait another period. The trade-off is between the costs of shorting immediately and waiting. The cost of selling short immediately is that the loan can be recalled before the anticipated price decline occurs, leaving the short seller with sunk transaction costs. Alternatively, there is only limited availability of shares in the next period and the lender is able to expropriate the value of the short seller's private information. The cost of waiting to sell short is the opportunity cost (net of transaction costs) of the stock price falling before negative exposure is assumed. Key results from the model include:

- Short sellers delay sales when transaction costs (price impact) are high relative to the perceived mispricing or when the probability of losing availability is high
- Embedded options to extend and fail increase the quantity of short selling relative to a simple spot loan market
- Embedded options to extend and fail generate larger fees than seen in a pure spot market
- Options to extend and fail in the wholesale market combined with average cost pricing for short sellers generate wholesale fees that can exceed the marginal benefit of the share to even the most pessimistic short seller
- Pooling of availability by lending intermediaries increases the value of the aforementioned options and makes short sellers more aggressive
- Using minimum reservation fees, lenders may approach first best outcomes - despite being uninformed about the exact size of the mispricing

1.1. Related Literature

The literature on securities lending and its relation to asset pricing is relatively new. A key theoretical paper is Duffie (1996) which contains a one-period model of the Treasury Repo market. The significant contribution of this paper is to show that, in equilibrium, a security’s price is
determined jointly with its lending fee. Another important contribution of this work is to highlight that positive security loan fees (or “specialness”) are not sustainable in the absence of frictions. In particular, if everyone could lend their securities, no one would be willing to hold them and not lend when fees are positive. To sustain the equilibrium where expected lending income is reflected in prices, one also needs non-lending investors to have some costs or inability to exchange a position in the special security for a general collateral substitute.

Duffie, Garleanu, and Pedersen (2002) show that in a multi-period setting, special equilibria can be sustained when lenders and borrowers face a search problem. The prospect of an eventual match with a paying borrower explains why rational investors who are not currently lending shares will hold the float today when its price reflects expected future loan fees. The dynamics of the model are driven primarily by two arrival processes: the random matching of borrowers and lenders, and the “day of reckoning” when differences of opinion that drive borrowing demand are resolved. The model predicts many of the stylized facts surrounding the costs of shorting IPOs.

A number of empirical papers have studied security loan data sourced from participants in the wholesale lending market. Using data on availability, fees, and recalls provided by a large U.S. custody bank, D’Avolio (2002) identifies supply and demand factors associated with the likelihood that a stock is hard to borrow. He finds that small stocks, with low institutional ownership and high turnover are more likely to be special.

Using 12 months of data on wholesale stock loan fees, Reed (2001) finds empirical support for the Diamond and Verrecchia (1987) prediction that returns of stocks with short sale constraints are more negatively skewed.

Gezey, Musto, and Reed (2002) use a data set on rebate rates from a large institutional lender to explore whether specialness in stocks can explain various patterns in returns. They find that borrowing IPOs and the stock of acquirers in risk arbitrage deals is particularly difficult and costly. This friction might be large enough to partially explain why strategies that short these stocks appear
nominally to be so profitable. They find that short sale constraints cannot explain returns generated by HML, size, or momentum strategies.

Krishnamurthy (2002) shows that from 1995-1999 much of the apparent profitability of shorting on-the-run Treasuries against buying off-the-run Treasuries would be offset by the specialness of the former. That is, on-the-runs trade at lower Repo financing rates, or equivalently higher shorting costs.

Evans, Gezcy, Musto, and Reed (2003) describe and value the exemption given to option market makers from having to locate availability before shorting. This option to sell “naked”, that is, without first locating borrowed shares, has intrinsic value whenever fees are greater than the risk free rate (i.e., when rebates are negative). By failing to deliver, the market maker’s maximum cost of short selling is foregone interest on the short proceeds.

Lamont and Jones (2002) analyze impact of short sale constraints during the period 1926-1933. Using data from the Wall Street Journal on the loan fees of selected NYSE stocks, the authors track the performance of stocks that appear among those hard to borrow names traded by the “loan crowd.” The study supports the hypothesis that lack of availability for shorting leads to overpriced stocks—these stocks exhibit size-adjusted returns that are lower by between 1% and 2% per month. The paper shows that the fees alone were not enough to explain the apparent profitability of shorting these stocks.

Lamont (2003) describes counter-offensive techniques used by companies that have been targeted by short sellers. The strategies include press campaigns, lawsuits, and technical actions that disrupt the stock loan market (e.g. having investors move their shares from lendable margin accounts to non-lendable fully paid-for accounts). He finds that stocks of companies that fight their short sellers generate significantly lower returns.

The remainder of this paper is organized as follows. Section 2 provides empirical motivation for our model. In section 3, we derive our intertemporal model of short selling. Section 4 provides
comparative statics and empirical predictions of the model. Section 5 analyzes the relative efficiency of current loan practices. Section 6 concludes.

2. Motivating the model

2.1. Shorting Allied Capital

On the evening of May 15, 2002, David Einhorn of Greenlight Capital, a New York based hedge fund, delivered his "short thesis" on Allied Capital (ALD) to an audience of investment professionals. Einhorn criticized Allied Capital's accounting practices and believed the firm to be substantially overvalued. On May 16, millions of shares of ALD were borrowed and shorted in the marketplace. Securities lending data on ALD, provided to us by a large U.S. custody bank, illustrate many of the key stylized facts that the model will try to explain.

Figure 1 contains time series data on loan balances, loan supply, and fees for ALD. Over the 10 trading days before Einhorn delivered his May 15 speech, loan balances crept up from roughly 2% to 33% of the custody bank's supply. These were general collateral or “top [rebate] rate” loans with fees of roughly 15 basis points (bps) per annum. On May 16th, balances jumped up to 95% of loan supply, leaving only about 5% available. As scarcity was imminent, many of these new loans garnered moderately higher fees, bringing the weighted average fee across all ALD balances up to 50 bps per annum. Note that shares borrowed over the previous days were not re-priced, but rather extended at the initially agreed fee (15 bps per annum). Within the next 10 trading days loan balances reached 100% of supply and even higher when the custody bank's supply began to contract as lenders sold their positions. Faced with negative availability - more shares on loan than held by lenders – the custody bank issued recall notices to various prime brokers. As illustrated in the figure, virtually none of the lenders returned the requested shares. Instead, they exercised their option to fail.

---

7 Perold and Chase (2003) provide a detailed description of the short thesis as well the responses made by Allied managers.
Between June 14th and 18th, the custody bank, following common practice, increased fees by dropping rebate rates to zero on all recalled loans that were not returned. The value-weighted fee on this cohort of loans was now equal to the opportunity cost on overnight money (e.g. fed funds) or about 1.75% per annum. For roughly the next 100 trading days, loan balances exceeded supply, sometimes by a factor of 2 to 1. While the custody bank's availability declined by 50% in the 30 days following the positive demand shock, loans were returned much more slowly. Figure 2 shows that it took roughly 60 trading days before 50% of the recalled loans were actually returned.

Five months after the first recall, the custody bank once again showed positive availability for a small amount of shares. Beginning on 11/12/02, new loans were originated at a rebate of -25% for a fee of 26.25% given the then prevailing risk free rate of 1.25%. These loans amounted to less than 10% of all balances - the prime brokers enjoyed a relatively attractive 0 rebate (1.25% per annum fee) on the remaining 90%. It is possible (depending on the broker's markup) that short sellers who received these newly available shares were paying less than 26.25%. With an average cost agreement, the broker would charge a hedge fund a fee equal to 3.75% (0.1*26.25%+0.9*1.25%) plus some markup - regardless of whether the customer shorted in May or November. As we will explore further in this paper, the marginal wholesale fee alone may not convey enough information to infer the beliefs of the marginal short seller. It is essential to understand how prime brokers pass these costs on to their customers in practice.

2.2 The large sample evidence

In order to document that this example is not atypical of hard to borrow stocks, we examine three years of security loan data from a large U.S. custody bank. Over the period April 2000 to April 2003 we identify 703 different U.S. stocks that were lent at high fees (or, more precisely, those with

---

8 The rebate is the amount of interest a short seller receives on the cash collateral (102% of market value in the wholesale U.S. stock loan market) posted when borrowing shares. The effective fee to on the stock loan is the difference between the risk free rate and rebate. For example, if the Fed Funds rate is 5% but only 2% is rebated to the borrower, then the fee is 3%. In deals where collateral other than cash is posted, an explicit fee is passed from borrower to lender.
negative rebate rates). Table 1 provides data on the relation between the high marginal fees and the balance weighted average fee for these stocks at the time those more costly loans were originated. Panel A of Table 1 shows that average (median) loan is extended at a fee that is 41% (33%) of the top marginal fee for that stock. Panel B provides data on this ratio broken out by loan "vintage" or age. Loans originated a week before the highest marginal fee is transacted are extended at fees that are 48% lower and loans originated between 90 and 180 days prior are extended at fees 78% lower than the highest marginal fee. This evidence provides some measure of the degree to which loans of hard to borrow stocks are extended at previously contracted lower rates.

There were 648 different U.S. stocks (about 10% of the sample universe) that had loans that were recalled at some point during the three-year sample. Among these stocks, the propensity to fail was 42%. Figure 3 provides information on the duration of loans that failed. More than 60% of shares that were dropped to a zero rebate (for not returning the requested amount) remain outstanding even a month after the rate drop.

3. An intertemporal model of short selling

3.1 Basic setup and assumptions

There are three dates (0, 1, 2) and shares of stock in a single company. The risk free rate of interest is zero, eliminating the need to discount future payoffs. The true value of a share $V$. $V$ is revealed at time 1 with probability $q$. If true value is not revealed at time 1 (probability 1-$q$), then it will be revealed at time 2 with certainty. Prior to revelation, uninformed investors overestimate true value of a share by an amount $\delta > 0$. They believe that true value is equal to $V + \delta$.

The shares trade in the market at prices $P_t$ ($t=0,1,2$) plus or minus a transaction cost equal to $\frac{1}{2} \tau$ (buyers pay $P_t + \frac{1}{2} \tau$ and sellers receive $P_t - \frac{1}{2} \tau$). The transaction cost is a deadweight loss intended
to capture commissions, ticket costs, and temporary price impact of trading.\(^9\) Uninformed risk neutral investors set prices according to their beliefs. Thus, \(P_0 = V + \delta; P_2 = V;\) and \(P_1 = V + \delta\) or \(V\) depending on whether true value is revealed at time 1 or time 2, respectively. At time 0, there is a single informed investor who knows that the true value of a share is \(V\). The investor can sell short a single share as long as he can first locate one to borrow. The requirement to borrow a share before selling short reflects current U.S. stock market regulations - from which only certain market makers are exempt.\(^10\) Restricting the short position to a single share can be interpreted as a reduced form of exogenous risk limits or capital constraints. Subject to this restriction, the informed investor acts in a risk neutral manner – i.e., he maximizes expected value. The actions of the informed investor do not affect the beliefs of uninformed investors – nor the share price, which the latter determine.

There are two investors who are willing to lend shares when they hold positions in them.\(^11\) Each holds either zero or one share and so the aggregate loan supply varies between 0, 1, and 2 shares. Lender positions follow an exogenous random process (probabilities to be described below) that is independent of short seller demand. These lenders can be interpreted as passive, index-tracking institutions that are subject to random inflows and redemptions.\(^12\) Descriptive evidence by D'Avolio (2002) suggests that such lenders typify wholesale loan supply.

There exist eight states of the world at time 1: two valuation states x four availability states. Valuation outcomes are \{true value not revealed; \(P_1 = V + d\}\) and \{true value revealed; \(P_1 = V\}\). Availability outcomes are \{0,0\}, \{0,1\}, \{1,0\}, and \{1,1\} where \(\{h_1, h_2\}\) is notation for the number of

\(^9\) For informed traders such as hedge funds price impact can be significant – and an order of magnitude greater than commissions. The assumption that this impact is temporary is particularly convenient for modeling the arrival of the second short seller whose decision to short is not complicated by the price pressures exerted by earlier rounds of shorting.

\(^{10}\) Evan, Gezcy, Musto, and Reed (2003) describe and value the "locate" exemption for option market makers. They are also an excellent source on settlement rules and mechanisms relevant to sell fails. Exemption from the locate requirement is essentially an option to bypass the security loan market. This institutional detail might limit the ability of researchers to identify negative stock lending fees from options prices.

\(^{11}\) One should assume these loans are collateralized with cash and marked to market each day. The fee would likely take the form of a deduction from the interest on this cash collateral - the difference between earned interest and interest "rebated". See D'Avolio (2002), Gezcy, et al (2002), Evans, et al (2003), Perold (1995a), and Perold (1995b) for more details on the mechanics of security loans and the risks of securities lending programs.

\(^{12}\) Another reason why institutional investors make temporarily remove their shares from a lending program is to vote them as they cannot vote a share that is on loan. Apfel, et al (2002) and Christoffersen, et al (2002) explores issues related to securities lending and voting.
shares held by investors 1 and 2, respectively. For simplicity, we collapse these 8 states down to the 5 that matter for our analysis. We do not need to describe loan supply in states where true value is revealed at time 1 because there will be zero borrowing demand when \( P_1 = V \). The probabilities of these five states appear in Table 2. The only structure imposed on these probabilities is that they are independent of the short seller's action.

3.2 Assumptions about intermediation and institutional practices

We assume that short sellers do not borrow shares directly from investors who are willing to lend. Instead, there exists a single custody bank (CB) that acts as a lending agent for these investors. Later in this paper we will introduce another important intermediary, the prime broker (PB). In practice, most short sellers borrow shares from PBs who, in turn, source a large portion of their loan supply from CBs. For now, however, we will assume that short sellers borrow directly from the CB. In practice, CBs perform three primary functions: They reduce search costs by pooling the supply of shares to lend; they minimize counterparty default risk by lending only to creditworthy borrowers; and they perform the operational aspects of securities lending such as the management and mark to market of collateral. In this model, we assume that the CB has eliminated the costs of matching borrowers and lenders, knows that the informed investor will not default of loans of shares, and has reduced operational costs to zero.

The CB charges borrowers per period fees \( f_0 \) and \( f_1 \), the latter fee being state contingent, depending on share availability and the timing of the revelation of true value. The CB offers borrowers different types of loan contracts. Of principal interest is a contract granting the borrower of a share the option to extend the loan at the original first period fee for another period, subject to the CB having shares available to lend. In addition, if the CB loses availability and recalls the share, the borrower has the option to fail to return it, and extend the loan for another period at the higher of the initial fee \( f_0 \) or the CB's cost to fail, \( k \). In practice, failure to return recalled stock results in the rebate rate being
dropped to zero - unless it was already zero or negative. This is to insure that the lender is made whole for lost interest on sale proceeds that are withheld while it fails to deliver the recalled shares it sold. In practice, our k parameter corresponds to the lender’s opportunity cost on delayed proceeds.

The contract between the borrower and the CB is satisfied from the pooled holdings of the two investors who participate in the lending program. Thus, if the share lent belongs to investor 1 and investor 1 sells after the first period, the CB would not recall the loan if investor 2 held an available share. Only when neither investor is long (i.e. there is no availability in the pool) does the CB recall the loan (at which point the borrower has the option to temporarily fail to return the share).

Finally, all parameters except δ (known only to the informed investor) are common knowledge.

3.3 Determination of equilibrium fees

Equilibrium fees are determined as if via declining price auctions. More formally, in states where loan supply is a binding constraint, we define the equilibrium fee as the non-negative cost that makes the marginal short seller indifferent between borrowing and not borrowing (shorting and not shorting) at that time. In this situation, the bargaining power rests with the CB. When the number of shares exceeds the number of borrowers, the bargaining power rests with the borrower, and the fee is driven down to zero (or some reservation level specified by the CB).13 If there are no shares available to borrow, then the loan market is temporarily closed and the fee is undefined in that period. When indifferent between selling short or not, the informed investor chooses to sell short.

Along with the determination of equilibrium fees, our model derives a time 0 trading rule for the short seller that specifies when it is optimal for him to short now ("GO") or delay his decision to short another period ("WAIT"). This rule is expressed as a threshold level for δ, the size of the mispricing, below which the short seller will not go in the first period. The higher the threshold, the greater is the short sale reluctance.

---

We assume that the CB is relatively passive. In particular, the CB does not trade based on any inferences about true value based on short seller demand. Further, the CB does not strategically act to constrain the supply of available shares in order to obtain higher fees (although in Section 5 we relax this assumption and allow the CB to set a positive reservation fee below which it will not supply any shares at time 0).

3.4. Solving the model

In order to separately understand the three key design aspects of the loan market (the option to extend, the option to fail, and pooled availability), we introduce them to the model incrementally. We solve for equilibrium fees and optimal trading strategies at each partial step leading up to the final setup that reflects the richness of current institutional practices. The decision to wait by the short seller results in the same payoffs and time 1 (state contingent) fees regardless of the lending contract offered by the CB at time 0. As such, we first examine the economics of this decision.

3.4.1. The decision to wait

Table 3 below lists the state contingent payoffs and lending fee resulting from the decision to wait. If the informed investor waits, he can potentially profit from his information only in the states where true value has not yet been revealed and there is positive share availability. If there are two shares available, \( f_1 = 0 \). The investor can borrow a share without cost and short it for a certain profit equal to the mispricing less round-trip transaction costs, \( \delta - \tau \). When there is just one share available, the lender extracts the full rent and \( f_1 = \delta - \tau \). The informed investor sells short but his profit is zero. The informed investor thus receives positive net payoffs only in the state where true value is revealed at time 2 and there is excess share availability. The expected payoff as of time 0 for the WAIT strategy is therefore:

\[
E_0 [\pi_{\text{wait}}] = p_{11} (\delta - \tau)
\]  

(1)
3.4.2. Case of a pure spot loan market (no options to extend or fail)

As a base case, we derive fees and a short selling trading rule (wait or go at t=0) in the absence of extendable loans, failure, or pooling. More precisely, we assume a true "spot" loan market where stock loans are for one period, after which all shares held at the CB are auctioned off again. We assume that an investor who is short at time 0 and who is able to continue to remain short at time 1 is not required to repurchase the share (or "cover") between the two periods. In practice, this is possible because positions can be transferred electronically "same-day" via the current clearing infrastructure. The borrower can have his newly borrowed share wired to his previous lender to close out that loan.

Table 4 gives the payoffs and fees associated with the decision to sell short at time 0 in the case of the spot market. In selling short at time 0, the investor incurs a sunk borrowing fee and round-trip transaction cost, \(-(f_0 + \tau)\). If true value is revealed at time 1, he closes out the short position for a net profit of \(\delta\) less these sunk costs. If true value is not revealed at time 1, but there is excess availability (2 shares), then the short seller will keep his short on by rolling his loan over at the then prevailing zero fee and close the short at time 2 for a gain of \(\delta\) (less aforementioned sunk costs). In any other state, the short seller suffers a net loss equal to the sunk costs. If the truth is not revealed, but there are no shares held by lenders, then the short seller will be forced to cover his short at a loss equal to the sunk costs, without any way to re-establish the position before it the price drops. If the mispricing persists at time 1, but there is only a single share available to borrow, then the informed agent can only maintain his short position by borrowing at a high fee. With scarce availability, the short seller faces a classic hold-up problem: having already sunk the transaction cost, the short seller is now willing to pay up to \(\delta\) to stay short through time 2 (when the share price falls by \(\delta\) with certainty).

For any given \(f_0\), the expected payoff as of time 0 for the action GO is:

\[
E_0 [\pi_{go}] = (p_{11} + q)\delta - (f_0 + \tau)
\]  

(2)
With expressions for the expected payoff of each possible initial action (wait or go), we can calculate the time 0 loan fee and the optimal trading rule for deciding between these actions. We can solve for \( f_0 \) by equating \( E_0 [\pi_{\text{wait}}] \) and \( E_0 [\pi_{\text{go}}] \):

\[
f_0^{\text{spot}} = q \delta - (1 - p_{11}) \tau
\]

(3)

In practice we do not see negative fees. That is, lenders do not pay short sellers to borrow shares. When zero (or some other reservation fee) is a binding minimum for the time 0 fee, the short seller will strictly prefer the action “WAIT” over “GO” at that point in time. The value of \( \delta \) that results in the time 0 fee being zero is a threshold level (\( \delta^* \)) below which waiting has a higher expected value than shorting now. The short seller's optimal trading strategy \( (S_0) \), whether there is one or two shares available to borrow, is:

\[
S_0 = \{ \text{WAIT another period if } \delta < \delta^*, \text{ GO (short now) otherwise}\}
\]

where

\[
\delta_{\text{spot}}^* = \frac{(1 - p_{11}) \tau}{q}
\]

(4)

This trading rule says that short sellers are most reluctant to trade when the perceived mispricing is small relative to transaction costs, when the expected price correction is unlikely to occur in the near future, and when the probability of future scarcity in the loan market is high.

3.4.5. Option to extend (Availability not pooled)

We now introduce the option to extend into the stock loan arrangement. This institutional feature forces us to also specify whether or not the CB pools lender positions. This is because the option to extend exists as long as the CB has availability. Availability, in turn, might be defined as
being investor specific or, as is commonly practiced, pooled across customer holdings. We denote by Investor 1 the investor whose share is borrowed at time 0 if the short seller chooses to go short in that first round of trading. Without pooling of the CB's loan supply, a share borrowed at time 0 is recalled in those states where Investor 1 no longer owns a share. With pooling, the CB recalls the loan only when neither Investor 1 nor 2 hold shares. If Investor 1 sells its position, the CB will try to settle Investor 1's trade by delivering from Investor 2's position before resorting to recalling the loan. When a loan is not recalled at time 1, the CB gives the short seller the option to extend the loan for another period at \( f_0 \), the time 0 fee. If the time 1 fee is lower than the time 0 fee (\( f_1 < f_0 \)), then this option to extend is "out of the money", and the short seller returns the expensive loan to refinance at the lower prevailing fee.

Table 5 summarizes the payoffs for the GO decision in the setting where the CB offers the option to extend, but does not pool supply. The expected profit at time 0 from waiting, \( E_0[\pi_{\text{wait}}] \), is the same as the spot market case (see equation 1). The expected profit at time 0 from going can be calculated by probability weighting the profits enumerated in Table 5:

\[
E_0[\pi_{\text{go}}] = (p_{11} + q)\delta - (f_0 + \tau) + p_{10}(\delta - f_0)
\]  

This is equal to the expected value of GO in the spot contract (equation 2) plus the expected value of the option to extend. As before, if supply at time 0 is not a binding constraint, then \( f_0 = 0 \); if there is no supply at time 0, the loan market is closed and the fee is undefined. When there is one share available (i.e. the stock is special), the equilibrium time fee at time 0 equates \( E_0[\pi_{\text{wait}}] \) and \( E_0[\pi_{\text{go}}] \):

\[
f_{0,\text{extend, no pooling}} = \frac{(q + p_{10})\delta - (1 - p_{11})\tau}{1 + p_{10}}
\]  

\( f_{0,\text{extend, no pooling}} \)
Alternatively, highlighting the fact that this fee is at least as large as the spot fee, the fee can be expressed as:

\[ f_{0}^{\text{extend}, \text{no pooling}} = f_{0}^{\text{spot}} + \frac{p_{10}}{1 + p_{10}}(\delta - f_{0}^{\text{spot}}). \]

As derived for the spot regime, the GO vs. WAIT threshold, \( \delta^* \), is the \( \delta \) that equates the initial fee to zero. When \( \delta \) is less than this threshold, the short seller should optimally WAIT.

\[ \delta^*_{\text{extend, no pooling}} = \frac{(1 - p_{11})}{q + p_{10}} \tau \]  

(7)

3.4.6. Option to extend (Pooled availability)

In practice, custody banks pool availability. As can be seen by comparing Tables 5 and 6, the only important difference between pooling and not pooling for the purpose of this analysis is that the short seller has a more valuable option to extend - one that exists in more states of the world:

\[
E_0 [\pi_{go}] = (p_{11} + q)\delta - (f_0 + \tau) + (p_{10} + p_{01})(\delta - f_0)
\]

(8)

When there are two shares available at \( t=0 \), \( f_0 = 0 \). When no shares are available, \( f_0 \) is undefined. With one share available, the fee must make the short seller indifferent between WAIT and GO. Equating equations 1 and 8 yields the equilibrium fee:

\[
f_{0}^{\text{extend, pooling}} = \frac{(q + p_{10} + p_{01})\delta - (1 - p_{11})\tau}{1 + p_{10} + p_{01}}
\]

(9)

Rearranging, the fee can be expressed as:

\[ f_{0}^{\text{extend, pooling}} = f_{0}^{\text{spot}} + \frac{p_{10} + p_{01}}{1 + p_{10} + p_{01}}(\delta - f_{0}^{\text{spot}}). \]

The latter expression, while equivalent to (9), isolates the incremental effect of the option to extend has on the equilibrium fee.
The optimal trading rule for time 0 is WAIT when $\delta < \delta^*$, GO otherwise; where,

$$\delta_{\text{extend, pooling}}^* = \frac{(1 - p_{11})}{q + p_{10} + p_{01}} \tau$$

(10)

3.4.7. Options to extend and fail (Pooled availability)

In a more institutionally accurate description of the security loan market, the borrower has both the option to extend the loan at the initial fee whenever the CB has availability in its pool as well as the option to fail to return recalled shares by paying the lender a fee no smaller than $k$. The payoffs for this case are presented in Table 7. Depending on the value of $k$, there are three possible equilibrium regimes that may obtain:

(i) When $k > \delta$, the option to fail is never exercised and so can be ignored in deriving the fee and trading rule threshold, which are identical to the extendable with pooling case derived above (see equations 9 and 10).

(ii) When $k < f_0$ (i.e., the equilibrium $t=0$ fee exceeds the CB's cost of failing), we can ignore $k$ and proceed assuming that the borrower has the option to extend the loan at $f_0$ in any state of availability. The short seller's expected payoff by shorting now is:

$$E_0 [\pi_{\text{go}}] = (p_{11} + q)\delta - (f_0 + \tau) + (p_{10} + p_{01} + p_{00})(\delta - f_0)$$

(11)

When there are two shares available at $t=0$, $f_0 = 0$. When no shares are available, $f_0$ is undefined. When there is one share available, $f_0$ must make the short seller indifferent between actions WAIT and GO. Setting equation 1 equal to equation 11 renders:
\[ f_{0}^{extend, fail, pooling} = \frac{(1 - p_{11})(\delta - \tau)}{1 + p_{10} + p_{01} + p_{00}} \]  

(12)

The optimal trading rule for time 0 is WAIT when \( \delta < \delta^* \), GO otherwise; where, \( \delta^*_{extend, fail, pooling} = \tau \)

(iii) When \( k \) is less than \( \delta \) but greater than the extendable fee derived in case ii (see equation 12), then the short seller will need to pay \( k \), not \( f_0 \), when opting to fail and extend the loan in the state with no availability. The expected profit from shorting now (GO) is:

\[ E_0 [\pi_{go}] = (p_{11} + q)\delta - (f_0 + \tau) + (p_{10} + p_{01})(\delta - f_0) + p_{00}(\delta - k) \]  

(13)

The terms of the expression correspond to the expected profit from choosing GO in the spot case plus the option to extend at \( f_0 \) when there is pooled availability plus the option to fail and extend at \( k \) when there is no availability at time 1. If two shares available at time 0, \( f_0 = 0 \). If there are no shares available at time 0, \( f_0 \) is undefined. Otherwise, \( f_0 \) equates the expected value of WAIT and GO:

\[ f_{0}^{extend, fail, pooling} = \frac{(1 - p_{11})(\delta - \tau) - p_{00}k}{1 + p_{10} + p_{01}} \]  

(14)

When \( \delta < \delta^* \), the optimal strategy is to WAIT and not decide on shoring until time 1. Otherwise, the short seller should GO and short now. In this case,

\[ \delta^*_{extend, fail, pooling} = \tau + \frac{p_{00}}{1 - p_{11}}k \]  

(15)

3.5 The Prime Broker - intermediary for the borrower
In practice, short sellers rarely borrow directly from CBs. Most borrow shares from their prime broker (PB) who, in turn, borrows from CBs when customer short demand exceeds internal “box” supply. PBs provide valuable intermediation services in securities lending. They are credit intermediaries with expertise in taking the counterparty risk that most CBs would not assume. CBs are comfortable lending shares to large PBs, even giving them the option to fail to return recalled loans for weeks and months. Most CBs would not have this level of comfort if lending directly to hedge funds. PBs also enjoy economies of scale in overcoming the search process involved with locating supply. As such, PBs not only make it possible for short sellers to exploit the benefits of a CB’s pooled supply, but also provide diversification across CBs to reduce recall risk further.

When there are multiple borrowers and scarce supply, PBs face the issue of whether to price discriminate among clients. To avoid the appearance of favoritism, some of the major PBs adopt pricing policies by which they charge the same fee to all of their customers with short positions in a given stock. As described in the introduction, this fee is generally a function of the PB's average cost in sourcing the supply on the wholesale market (i.e., from its internal supply, from CBs, and from other PBs). In order to understand the implications of average cost pricing on the behavior of observed loan fees, we will incorporate a version of this practice into our modeling framework.

To see the effect of cost averaging, we need to expand the model to include multiple borrowers. We assume that, if true value is not revealed at time 1, there is a probability that another (otherwise identical) short seller learns about the mispricing and considers shorting a share. We denote by $\alpha$ the exclusivity of the original short seller's information, and so $1-\alpha$ is the probability that a second short seller competes for shares next period. Only in the state when there are 2 shares available at time 1 will this new short seller, who has not yet sunk any costs, influence the marginal loan fee. (When there are no shares, there is no lending market at time 1; when there is only 1 share, the short seller with the existing position and sunk costs will have a higher willingness to pay).
We now assume an institutional setting in which short sellers borrow from a PB who borrows from the CB. The CB pools supply and grants the PB the option to extend at previously contracted rates. The PB passes the option to extend on to the end borrower, but reprices the loan in the second period according to the average fee it pays to the CB. In order to facilitate exposition, we will not here consider the option to fail.

At time 1, in the state where a second short seller appears and the CB has two shares available to lend, the following occurs: If the initial informed investor chose to WAIT at time 0, then the short sellers would each be willing to pay a fee of up \( \delta - \tau \) in order to sell short at time 1. In equilibrium, the PB charges each short seller \( f_1 = \delta - \tau \) and pays the combined fees, \( 2(\delta - \tau) \), to the CB.

If the initial informed investor chose to GO at time 0, he extends his loan for another period but pays a new fee determined by average cost. Given that his transaction cost is sunk, the initial short seller would be willing to pay up to \( \delta \) to extend the loan. However, the second short seller would only be willing to pay \( \delta - \tau \). In equilibrium (with a supply of 2 shares and no price discrimination), the PB charges each short seller \( \delta - \tau \). In the wholesale market, the PB needs to borrow one share, given that he already controls one share by extending the time 0 loan at \( f_0 \). To borrow this marginal share, the PB is willing to pay a fee that brings its average borrowing cost per share to \( \delta - \tau \). The PB pays the CB:

\[
f_1 = 2(\delta - \tau) - f_0
\]

The option to extend in combination with average cost pricing can result in PBs paying much higher wholesale rates on marginal loans than the marginal fees paid by its short-selling customers. Table 8 lays out the full set of state contingent payoffs and fees.

4.1. Implications and predictions of the model

4.1.1. Comparative statics

The equilibrium loan fees and trading rules derived in the previous section allow us to quantify the incremental effects of certain institutional features of the loan market (e.g. the option to extend, the
option to fail, and pooling of supply) and to articulate the sensitivity of these effects to the model's various parameters.\textsuperscript{14} In addition, there are some relations that transcend all versions of the loan contracts considered:

\textit{(i) Loan fees are increasing in }\delta\textit{, the difference in opinion between short sellers and other investors and decreasing in }\tau\textit{, the transaction costs that short sellers face (outside the loan market)}

The former prediction is common to models by D'Avolio (2001) and Duffie, Garleanu, and Pedersen (2002). It also finds significant empirical support in D'Avolio (2002) who proxies for differences of opinion with variables such as trading volume and dispersion in analyst forecasts. The relation to transaction costs does not appear in other models. While transaction and liquidity costs appear in securities lending models by Duffie (1996) and Krishnamurthy (2002), these authors are concerned with those costs faced by the owners and lenders of securities - not by short sellers. Duffie (1996) uses transaction costs to explain why not all owners lend and also to explain why owners who do not lend would hold securities whose prices reflect a lending income premium. In his model, fees are increasing in these lender transaction costs. Krishnamurthy (2002) assumes that some owners prefer more liquid bonds (those with lower transaction costs) and that this explains why they might hold bonds that reflect a specialness premium even when they do not participate in the repo market. In his model, specialness is decreasing (increasing) in transaction costs (liquidity) of on-the-run bonds.

\textit{(ii) Short seller reluctance is decreasing in }\delta\textit{ and increasing in }\tau\textit{ }

Reluctance - the tendency to defer short selling - is an intertemporal phenomenon. The only other intertemporal model of securities lending to our knowledge is DGP (2002). In their model, short sellers face no transaction costs nor any uncertainty about availability: once the search process results

\textsuperscript{14} The statics explored in this section are with regards to the time 0 loan fee. Statics for the time 1 loan fee are less interesting (although easily observed in the tables). This is because in a more realistic multi-period setting, where time steps are at a daily frequency, it is far more likely that any observed fees reflect some uncertainty about the probability of price correction. On average, observed fees behave more like "time 0" fees than "time 1" fees.
in a match between borrower and lender, the short seller is guaranteed to stay short until the price correction is realized. In fact, short interest dynamics are driven entirely by the search intensity parameter of their model.

Differences of opinion are positively correlated to trading volume and transaction costs are negatively correlated to trading volume. As such, our model makes the following empirical prediction: short interest should amass at earlier horizons (relative to the eventual revelation of bad news or overpricing) for stocks with higher share turnover. For example, in a large cross-section of risk arbitrage deals, our model predicts that (ceteris paribus) acquirers with higher average share turnover are be shorted more aggressively immediately upon a merger announcement compared to acquirers with less share turnover.

(iii) When a stock is special, the option to extend results in (initial) fees being larger than would result in a pure spot market

The incremental effect of the option to extend on the time 0 fee is

\[ \frac{p_{10}}{1 + p_{10}^2} \left( (1-q)\delta + (1 - p_{11})\tau \right). \]

By inspection, it is clear that the value of this option is increasing in \( \delta \) and \( \tau \). The intuition is that this option is like a call option or “cap” on the time 1 fee (\( f_1 \)) with a strike price equal to the time 0 fee (\( f_0 \)). When \( \delta \) is high, \( f_1 \) can be high and so this option is more valuable (because \( f_1 \) is more sensitive to movements in \( \delta \) than \( f_0 \)). When \( \tau \) is high, \( f_0 \) is low and so the option to extend at that fee is worth more. This option is also increasing in the probability that availability is scarce and decreasing in the probability that price correction is imminent. Furthermore, the option to extend combines with the practice of average pricing to allow the marginal wholesale rate to exceed beliefs about mispricing, \( \delta \), by arbitrarily large amounts (see equation 16).
(iv) Pooling makes the option to extend more valuable and further increases the fee a short seller will initially pay to borrow shares by an amount equal to 

\[
(\frac{p_{10} + p_{01}}{1 + p_{10} + p_{01}} - \frac{p_{10}}{1 + p_{10}})((1 - q)\delta + (1 - p_{11})\tau)
\]

(v) The value of pooling is increasing in \(\delta\) and \(\tau\) and decreasing in the probabilities that the price will correct or that there will be excess availability.

(vi) The option to fail increases the fee a short seller will initially pay to borrow shares

(vii) The option to extend, pooling, and the option to fail all reduce short seller reluctance

This is easily seen by noting that the short sellers trading threshold decreased by a non-negative amount with the introduction of each of these features. In our model, these features reduce short seller reluctance by reducing the number of states in which the lender will force to short seller to cover at a loss (of sunk costs) or pay "hold up" level fees. In a more general model in which short sellers are risk averse, their demand for short positions will be decreasing in the volatility of the expected payoffs that are enumerated in various tables. The volatility of these payoffs is increasing in the volatility of loan supply. The features that intermediaries embed in the loan contracts reduce return volatility by "capping" fees and reducing the volatility of availability. While not explicitly modeled in this paper, this may be an important mechanism for reducing short sale reluctance.

4.3 Coordination or crowding out among short sellers?

In our model the first short seller has an incentive not to share information and coordinate with other short sellers. This is because the PB charges the average cost of all loan balances in a given issue. The arrival of another short seller raises the expected average cost and lowers expected profits. This is not to say that coordination does not have benefits. For example, in a richer model than we pursue in this paper, the probability that a mispricing is resolved, \(q\), may be an endogenous function of the number of short sellers. Recent work by Abreu and Brunnermeier (2002) highlights the benefit of
short seller coordination. Our model, distinguished by its emphasis on current institutional practices, complements existing research. It shows that, via the security loan market, short sellers can increase their costs and risks by sharing proprietary information and coordinating with other short sellers - "crowding out" effects in the loan market need to be weighed against any coordination benefits.

5. Efficiency of loan contracts: Too much or too little shorting?

Without availability constraints, each share of the mispriced stock generates a potential surplus of $\delta - \tau$. This is the largest possible pie that could be divided among the short seller who uncovers the negative information, the investors who lend him their shares, and the PB and CB that intermediate these transactions. An important question one might ask is whether institutional loan practices are "efficient" in the sense that they maximize the amount of surplus that is captured by these players (the division of the resulting pie can be negotiated separately). Using the framework derived in Section 3, we show that current practices move the market closer to efficient outcomes - or at least toward outcomes that the CB prefers to those arising under a spot market.

5.1 Current Practice relative to pure spot market

The test of efficiency of a loan market regime is not whether it captures the entire surplus, $\delta - \tau$, but that it provides incentives consistent with the agents getting as close to $\delta - \tau$ as possible subject to availability constraints. One explanation as to why the stock loan market is not transacted on a strict overnight or spot basis is that such a contract would be inefficient in this sense. To be more precise:

Proposition 1.

In a spot market, there is too little short selling when $\frac{P_{00} + q}{q} \tau < \delta < \frac{P_{00} + q + P_{01} + P_{10}}{q} \tau$

Proof
We derive the short seller's trading rule in a spot loan market in Section 3. We show that the short seller chooses WAIT whenever \( \delta < \frac{(1 - p_{11})}{q} \tau \), or identically, whenever \( \delta < \frac{p_{00} + q + p_{01} + p_{10}}{q} \tau \).

We can also derive the “efficient” shorting threshold that maximizes total surplus. If the short seller chooses WAIT, the expected total surplus is \( (1 - p_00 - q)(\delta - \tau) \). If the short seller chooses to GO, the expected total surplus is \( (1 - p_{00})\delta - \tau \). These actions are equally attractive to collective interest when \( \delta = \frac{p_{00} + q}{q} \tau \). This threshold on the level of the mispricing is lower than the rule followed by the short seller and so there is not enough short selling (or too much reluctance) when \( \delta \) falls somewhere between these thresholds.

The options to extend and fail embedded in current wholesale market loans lower the short seller's trading threshold and induce earlier and ultimately more shorting relative to a spot loan regime. While this reduces the deadweight costs from too little shorting, it may in fact lead to too much shorting - that is, instances where expected aggregate profit would be higher had the short seller waited. Before looking at the incentives it generates for the short seller, it is important to first note that the option to fail represents a Pareto improvement over the spot or extendable regimes. The potential surplus that can be captured by the agents is now defined by the envelope of the payoffs associated with waiting, shorting now and never failing, and shorting now and failing if recalled (respectively, \( (1 - p_{00} - q)(\delta - \tau) \), \( (1 - p_{00})\delta - \tau \), and \( \delta - \tau - p_{00}k \)). Costless failure, in fact, allows the entire surplus to be realized. However, the short seller's incentive under the resulting fee mechanism is not necessarily aligned with maximizing the total pie, which is equivalent to maximizing lender income).\(^{15}\)

We will compare the short seller’s trading rule relative to the rule that maximizes lender income in each of the three ranges of \( k \) that sustain equilibrium fees.

\(^{15}\) This is easy to see. If the pie could be increased by an alternative action that would make the short seller worse off, then it is a simple accounting identity that the lender would be made better off.
Proposition 2.

In an extendable (pooled) loan market that allows failure, there is "too much" short selling when
\[ f_{0}^{\text{extend, fail, pooling}} < k . \]

Proof

(i) When \( k > \delta \), the short seller will never fail. As derived above, the short seller will GO whenever
\[
\delta \geq \frac{(1 - p_{11})}{q + p_{10} + p_{01}} \tau ,
\]
identically, when \( \delta \geq \tau + \frac{p_{00}}{q + p_{10} + p_{01}} \tau \). The efficient trading rule (from a lender or global perspective) is to short when the total expected profit for GO is greater than the total expected profit for WAIT, \((1 - p_{00})\delta - \tau > (1 - p_{00} - q)(\delta - \tau)\). The efficient threshold for \( \delta \) is
\[
\tau + \frac{p_{00}}{q} \tau ,
\]
which is clearly higher than the short seller's threshold. When \( \delta \) falls between these two thresholds, the short seller is not reluctant enough and inflicts a dead weight loss on the lender.

(ii) When \( k < f_{0}^{\text{extend, fail, pooling}} \), the short seller will fail whenever recalled. Aggregate expected profits are equal for WAIT and GO when \((1 - p_{00} - q)(\delta - \tau) = \delta - \tau - p_{00}k\), or when \( \delta = \tau + \frac{p_{00}}{q + p_{00}} k \). This threshold is higher than the short seller's, who will WAIT only when \( \delta < \tau \) (see case ii in Section 3.4.7). However, when \( k \) is in the range described for this case, \( \delta \) is higher than both the short seller’s and the efficient threshold: the short seller always chooses GO and this is the efficient action.

(iii) When \( f_{0}^{\text{extend, fail, pooling}} < k < \delta \), the short seller will fail whenever recalled. The efficient threshold for \( \delta \) equates the aggregate expected payoffs possible given a WAIT decision to the aggregate expected profits possible given a GO decision. \((1 - p_{00} - q)(\delta - \tau) = \delta - \tau - p_{00}k\) when
\[
\delta = \tau + \frac{p_{00}}{q + p_{00}} k .
\]
This threshold is higher than the threshold followed by the short seller when \( k \) is in
this range (see equation 15), \( \delta = \tau + \frac{p_{00}}{1 - p_{11}} k \). When \( \delta \) falls somewhere between these two thresholds, the short seller chooses GO when lenders would be better off if he chooses WAIT. In sum, when \( f_{0}^{\text{extend, fail, pooling}} < k \), the short seller is too eager in choosing to GO and generates a dead weight loss. ■

5.2. GC rate as strategic reservation price

The inefficiency of the spot market occurs because there is too little shorting. This is because the CB cannot pay the borrower in the first period. The short seller needs up-front compensation in cases where he is likely to incur transaction costs without realizing the benefits of his information due to diminished availability in the loan market. The non-negativity of fees makes WAIT the dominant action for the short seller, even though all parties (including the intermediaries) could divide a greater portion of \( \delta - \tau \) if he chooses the action GO (i.e. shorting now). A solution to this problem is to allow the CB to "pay" the short seller (via his PB) in the form of options to extend and fail. Alternatively, the intermediaries provide the short seller with some insurance against recall and hold-up risks. This enhanced contract can be non-negative over a greater range of \( \delta \) than could fees for a simple spot loan.

However, the current mechanism of embedding options may work "too well" in the sense that it overshoots in reducing the short seller's trading threshold (\( \delta^{*} \)). Current practice might provide an incentive for the short seller to short now even when waiting would maximize the aggregate pie. In these situations where the fee is too low, the CB and lenders would benefit from a minimum or reservation lending fee. One potential limitation to implementing a reservation fee is the CB's information set. In our model, the CB does not know \( \delta \). The CB needs to know the efficient threshold (i.e., the cutoff below which the lender should withhold shares) and the relevant range of \( k \) (see cases i - iii). The reservation fee in each case should bind precisely when \( \delta \) equals the efficient threshold. To see this, recall that as soon as the fee stops decreasing with \( \delta \) as prescribed by the equilibrium fee function, the short seller's expected profits to waiting will exceed the expected profits to shorting now.
The reservation fee ($R_0$) for each case is simply the equilibrium fee with the efficient threshold substituted for $\delta$.

(i) When $k > \delta$, 
$$R_{0,\text{extend, fail, pooling}}^{\text{extend, fail, pooling}} = \frac{P_{01} + P_{10} - P_{00}}{1 + P_{01} + P_{10}} \tau$$  \hspace{1cm} (17)$$

(ii) When $k < f_0^{\text{extend, fail, pooling}}$, 
$$R_{0,\text{extend, fail, pooling}}^{\text{extend, fail, pooling}} = \frac{1 - P_{11}}{1 + P_{01} + P_{10} + P_{00}} \frac{P_{00}}{P_{00} + q} k$$  \hspace{1cm} (18)$$

(iii) When $f_0^{\text{extend, fail, pooling}} < k < \delta$, 
$$R_{0,\text{extend, fail, pooling}}^{\text{extend, fail, pooling}} = \frac{P_{01} + P_{10} - P_{00}}{1 + P_{01} + P_{10} + P_{00}} \frac{P_{00}}{P_{00} + q} k$$  \hspace{1cm} (19)$$

Case ii is irrelevant. This equilibrium is sustained only for fees greater than $k$. Before this reservation fee could become binding, we would be in the range covered by case iii. Alternatively, the equilibrium is sustained only in regions of mispricing where the short seller would short now and where shorting now is in fact the efficient action. In case ii, no reservation fee is necessary. Even if the CB does not know whether they face case i or case iii (because they do not know $\delta$), they can improve their expected lending income by using the smaller of the two reservation fees as a minimum.

In practice we do not observe stock specific minimum loan fees. Rather there is a single "GC" fee (roughly 15 bps on the wholesale market) that serves as a minimum on most stocks. This might reflect the CB's lack of information on several of the parameters that would be needed to derive stock specific minimum loan fees. The model presented here suggests that CBs should consider at least having a different minimum fee for different classes of stocks based on their $q$, $\tau$, and availability processes ($p_{00}$, $p_{01}$, $p_{10}$, and $p_{11}$). Certainly, they are in position to gather information on the latter.

5.3 Objective function of the custody bank

Another institutional feature that adds complexity to the problem is that the CB derives income based on their outstanding loan balances above and beyond the loan fees received from the PB. Institutional investors allow their CBs to keep some fraction (typically 20-30%) of the return...
generated by the reinvestment of the cash collateral posted by the PB. Thus, even at a zero fee the CB has an incentive to lend shares. The implication of this arrangement is that "too much" shorting is less of a problem than it may otherwise appear. While overly aggressive first period shorting may reduce total expected lending fee income - it provides offsetting income from profit sharing in the reinvestment of cash collateral.

6. Conclusion

A reliance on the availability of shares distinguishes short selling from derivative forms of negative exposure. Uncertainty regarding this availability and the prospect of sinking unrecoverable transaction costs generates reluctance among short sellers. Intermediaries offer a solution to these problems by pooling supply, allowing failure, and designing contracts that bound fees. This paper provides a simple framework for understanding the strategic timing of a short sale as well as the determination of lending fees in the context of this intermediation. We also show that prevalent securities lending practices can confound the interpretation of the marginal wholesale loan rate. For example, the marginal loan fee can be increasing in the supply of shares. This is because some intermediaries charge short sellers according to the average cost of keeping all its customers short in that stock. This average cost can remain attractive despite very high marginal fees because the broker can extend previous loans that were borrowed at lower fees. Early short sellers subsidize late short sellers. This generates an important incentive for short sellers not to share their information – and may represent a friction that reduces coordination among short sellers.

Our model also relates short interest dynamics to transaction costs, the nature of the arbitrage opportunity, and the volatility of loan supply. In particular, it predicts that short sellers will short stocks earlier when (non-lending) transaction costs are low and future loan supply is expected to be relatively high. Finally, we show how current intermediation works to eliminate dead-weight costs from short sale reluctance, but that there is a sense in which it can lead to too much shorting unless minimum fees are utilized.
References


Table 1. Large sample evidence on the option to lend

Using data provided by a large U.S. custody bank, we identify 703 stocks with negative rebate rates (i.e., high borrowing fees) between April 2001 and April 2003. For each stock, we record the highest (MAX) fee charged and also calculate the value-weighted (VW) fee for all open loan balances in that stock on that day. The VW/MAX fee ratio is low whenever much of the loan balance is being extended at lower fees. Panel A sorts the observations of this ratio by percentile. Panel B sorts this ratio by age of the loan.

<table>
<thead>
<tr>
<th>A. By percentile</th>
<th>VW/MAX fee ratio</th>
<th>B. By loan age (days)</th>
<th>VW/MAX fee ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>99th</td>
<td>100%</td>
<td>[0]</td>
<td>83%</td>
</tr>
<tr>
<td>95th</td>
<td>98%</td>
<td>(0,7]</td>
<td>52%</td>
</tr>
<tr>
<td>90th</td>
<td>88%</td>
<td>(7,30]</td>
<td>38%</td>
</tr>
<tr>
<td>75th</td>
<td>67%</td>
<td>(30,90]</td>
<td>28%</td>
</tr>
<tr>
<td>50th</td>
<td>33%</td>
<td>(90,180]</td>
<td>22%</td>
</tr>
<tr>
<td>25th</td>
<td>16%</td>
<td>(180,365]</td>
<td>16%</td>
</tr>
<tr>
<td>10th</td>
<td>8%</td>
<td>(365,…]</td>
<td>13%</td>
</tr>
<tr>
<td>5th</td>
<td>5%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st</td>
<td>2%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Description of the possible states at time 1

<table>
<thead>
<tr>
<th>State</th>
<th>Valuation outcome</th>
<th>Availability outcome</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Truth not revealed, $P_1=V+\delta$</td>
<td>0 shares, {0,0}</td>
<td>$p_00$</td>
</tr>
<tr>
<td>2</td>
<td>Truth not revealed, $P_1=V+\delta$</td>
<td>1 share, {0,1}</td>
<td>$p_{01}$</td>
</tr>
<tr>
<td>3</td>
<td>Truth not revealed, $P_1=V+\delta$</td>
<td>1 share, {1,0}</td>
<td>$p_{10}$</td>
</tr>
<tr>
<td>4</td>
<td>Truth not revealed, $P_1=V+\delta$</td>
<td>2 shares, {1,1}</td>
<td>$p_{11}$</td>
</tr>
<tr>
<td>5</td>
<td>Truth revealed, $P_1=V$</td>
<td>Any number</td>
<td>$q$</td>
</tr>
</tbody>
</table>
Table 3. The decision to WAIT at time 0: state contingent payoffs and loan fees

<table>
<thead>
<tr>
<th>State at time 1</th>
<th>Valuation</th>
<th>Availability</th>
<th>Probability</th>
<th>Action at time 1</th>
<th>Payoff</th>
<th>$f_1$</th>
<th>Net Profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>P_1 = V + $\delta$</td>
<td>{0,0}</td>
<td>$p_{00}$</td>
<td>None</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>P_1 = V + $\delta$</td>
<td>{0,1}</td>
<td>$p_{01}$</td>
<td>Sell short</td>
<td>$\delta - \tau - f_1$</td>
<td>$\delta - \tau$</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>P_1 = V + $\delta$</td>
<td>{1,0}</td>
<td>$p_{10}$</td>
<td>Sell short</td>
<td>$\delta - \tau - f_1$</td>
<td>$\delta - \tau$</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>P_1 = V + $\delta$</td>
<td>{1,1}</td>
<td>$p_{11}$</td>
<td>Sell short</td>
<td>$\delta - \tau - f_1$</td>
<td>0</td>
<td>$\delta - \tau$</td>
<td></td>
</tr>
<tr>
<td>P_1 = V</td>
<td>Any</td>
<td>q</td>
<td>None</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Table 4. The decision to GO at time 0: Spot loan market

<table>
<thead>
<tr>
<th>State at time 1</th>
<th>Valuation</th>
<th>Availability</th>
<th>Probability</th>
<th>Action at time 1</th>
<th>Payoffs</th>
<th>f_1</th>
<th>Net Profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>P_1 = V + $\delta$</td>
<td>{0,0}</td>
<td>$p_{00}$</td>
<td>Cover</td>
<td>$-(f_0 + \tau)$</td>
<td>0</td>
<td>NA</td>
<td>$-(f_0 + \tau)$</td>
</tr>
<tr>
<td>P_1 = V + $\delta$</td>
<td>{0,1}</td>
<td>$p_{01}$</td>
<td>Stay short</td>
<td>$-(f_0 + \tau)$</td>
<td>$\delta - f_1$</td>
<td>$\delta - \tau$</td>
<td>$-(f_0 + \tau)$</td>
</tr>
<tr>
<td>P_1 = V + $\delta$</td>
<td>{1,0}</td>
<td>$p_{10}$</td>
<td>Stay short</td>
<td>$-(f_0 + \tau)$</td>
<td>$\delta - f_1$</td>
<td>$\delta - \tau$</td>
<td>$-(f_0 + \tau)$</td>
</tr>
<tr>
<td>P_1 = V + $\delta$</td>
<td>{1,1}</td>
<td>$p_{11}$</td>
<td>Stay short</td>
<td>$-(f_0 + \tau)$</td>
<td>$\delta - f_1$</td>
<td>0</td>
<td>$\delta - (f_0 + \tau)$</td>
</tr>
<tr>
<td>P_1 = V</td>
<td>Any</td>
<td>q</td>
<td>Cover</td>
<td>$-(f_0 + \tau)$</td>
<td>$\delta$</td>
<td>0</td>
<td>$\delta - (f_0 + \tau)$</td>
</tr>
</tbody>
</table>
Table 5. The decision to GO at time 0: Loan market with option to extend, availability not pooled

<table>
<thead>
<tr>
<th>State at time 1</th>
<th>Action at time 1</th>
<th>Payoffs</th>
<th>f_1</th>
<th>Net Profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valuation</td>
<td>Availability</td>
<td>Probability</td>
<td>Sunk at time 0</td>
<td>t &gt; 0</td>
</tr>
<tr>
<td>P_1 = V + δ</td>
<td>{0,0}</td>
<td>p_{00}</td>
<td>Cover</td>
<td>-(f_0 + τ)</td>
</tr>
<tr>
<td>P_1 = V + δ</td>
<td>{0,1}</td>
<td>p_{01}</td>
<td>Stay short, borrow at spot from investor 2</td>
<td>-(f_0 + τ)</td>
</tr>
<tr>
<td>P_1 = V + δ</td>
<td>{1,0}</td>
<td>p_{10}</td>
<td>Stay short, extend loan at f_0</td>
<td>-(f_0 + τ)</td>
</tr>
<tr>
<td>P_1 = V + δ</td>
<td>{1,1}</td>
<td>p_{11}</td>
<td>Stay short, refinance at spot</td>
<td>-(f_0 + τ)</td>
</tr>
<tr>
<td>P_1 = V</td>
<td>Any</td>
<td>q</td>
<td>Cover</td>
<td>-(f_0 + τ)</td>
</tr>
</tbody>
</table>

Table 6. The decision to GO at time 0: Loan market with option to extend, availability pooled

<table>
<thead>
<tr>
<th>State at time 1</th>
<th>Action at time 1</th>
<th>Payoffs</th>
<th>f_1</th>
<th>Net Profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valuation</td>
<td>Availability</td>
<td>Probability</td>
<td>Sunk at time 0</td>
<td>t &gt; 0</td>
</tr>
<tr>
<td>P_1 = V + δ</td>
<td>{0,0}</td>
<td>p_{00}</td>
<td>Cover</td>
<td>-(f_0 + τ)</td>
</tr>
<tr>
<td>P_1 = V + δ</td>
<td>{0,1}</td>
<td>p_{01}</td>
<td>Stay short, extend loan at f_0</td>
<td>-(f_0 + τ)</td>
</tr>
<tr>
<td>P_1 = V + δ</td>
<td>{1,0}</td>
<td>p_{10}</td>
<td>Stay short, extend loan at f_0</td>
<td>-(f_0 + τ)</td>
</tr>
<tr>
<td>P_1 = V + δ</td>
<td>{1,1}</td>
<td>p_{11}</td>
<td>Stay short, refinance at spot</td>
<td>-(f_0 + τ)</td>
</tr>
<tr>
<td>P_1 = V</td>
<td>Any</td>
<td>q</td>
<td>Cover</td>
<td>-(f_0 + τ)</td>
</tr>
</tbody>
</table>
Table 7. The decision to GO at time 0: Loan market with options to extend and fail, availability pooled

<table>
<thead>
<tr>
<th>Valuation</th>
<th>Availability</th>
<th>Probability</th>
<th>Action at time 1</th>
<th>Payoffs</th>
<th>f₁</th>
<th>Net Profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>P₁ = V + δ</td>
<td>{0,0}</td>
<td>p₀₀</td>
<td>If k &gt; δ, cover</td>
<td>-(f₀ + τ)</td>
<td>max(0, δ – max(k,f₀))</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Else, fail at cost of max(f₀,k)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P₁ = V + δ</td>
<td>{0,1}</td>
<td>p₀₁</td>
<td>Stay short, extend loan at f₀</td>
<td>-(f₀ + τ)</td>
<td>δ – f₁ + max(0, f₁ – f₀)</td>
<td>δ – τ</td>
</tr>
<tr>
<td>P₁ = V + δ</td>
<td>{1,0}</td>
<td>p₁₀</td>
<td>Stay short, extend loan at f₀</td>
<td>-(f₀ + τ)</td>
<td>δ – f₁ + max(0, f₁ – f₀)</td>
<td>δ – τ</td>
</tr>
<tr>
<td>P₁ = V + δ</td>
<td>{1,1}</td>
<td>p₁₁</td>
<td>Stay short, refinance at spot</td>
<td>-(f₀ + τ)</td>
<td>δ – f₁ + max(0, f₁ – f₀)</td>
<td>0</td>
</tr>
<tr>
<td>P₁ = V</td>
<td>Any</td>
<td>q</td>
<td>Cover</td>
<td>-(f₀ + τ)</td>
<td>δ</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 8. Average cost pricing by the PB: Loan market with option to extend, availability pooled

<table>
<thead>
<tr>
<th>Time 0 Action</th>
<th>State at time 1</th>
<th>Action at time 1</th>
<th>Short seller Payoffs</th>
<th>f_i paid to PB</th>
<th>Wholesale fees paid by PB to CB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Valuation</td>
<td>Availability</td>
<td>Probability</td>
<td>Sunk at Time 0</td>
<td>t &gt; 0</td>
</tr>
<tr>
<td>WAIT</td>
<td>P_1 = V + δ</td>
<td>{0,0}</td>
<td>p_{00}</td>
<td>None</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>P_1 = V + δ</td>
<td>{0,1}</td>
<td>p_{01}</td>
<td>Sell short</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>P_1 = V + δ</td>
<td>{1,0}</td>
<td>p_{10}</td>
<td>Sell short</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>P_1 = V + δ</td>
<td>{1,1}</td>
<td>p_{11}α</td>
<td>Sell short</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>P_1 = V + δ</td>
<td>{1,1}</td>
<td>p_{11} (1 – α)</td>
<td>Sell short</td>
<td>0</td>
</tr>
<tr>
<td>GO</td>
<td>P_1 = V + δ</td>
<td>{0,0}</td>
<td>p_{00}</td>
<td>Cover</td>
<td>-(f_0 + τ)</td>
</tr>
<tr>
<td></td>
<td>P_1 = V + δ</td>
<td>{0,1}</td>
<td>p_{01}</td>
<td>Stay short, PB extends</td>
<td>-(f_0 + τ)</td>
</tr>
<tr>
<td></td>
<td>P_1 = V + δ</td>
<td>{1,0}</td>
<td>p_{10}</td>
<td>Stay short, PB extends</td>
<td>-(f_0 + τ)</td>
</tr>
<tr>
<td></td>
<td>P_1 = V + δ</td>
<td>{1,1}</td>
<td>p_{11}α</td>
<td>Stay short, PB extends</td>
<td>-(f_0 + τ)</td>
</tr>
<tr>
<td></td>
<td>P_1 = V + δ</td>
<td>{1,1}</td>
<td>p_{11} (1 – α)</td>
<td>Stay short, PB refines</td>
<td>-(f_0 + τ)</td>
</tr>
<tr>
<td></td>
<td>P_1 = V</td>
<td>Any</td>
<td>q</td>
<td>Cover</td>
<td>-(f_0 + τ)</td>
</tr>
</tbody>
</table>
Figure 1. The market for borrowing Allied Capital (ALD)

- 5/16/02 - day after Einhorn reveals short thesis
- 6/14/02 - 6/18/02, Rebates on recalled loans are dropped to zero
- 11/12/02, new loans originated at -25 rebate

- Shares (Scaled)
- % Fee (Fed funds - rebate rate)

Legend:
- new loans
- loans
- supply
- value weighted fee on initial cohort of loans
- fee on new loans
Figure 2: Failing to return ALD

Recalled shares that were not returned

Recalled shares that were returned

Days relative to recall notice
Figure 3. Large sample evidence on the option to fail

Recalled shares that were not returned

Recalled shares that were returned

Trading days relative to first day of failure