Settlement Risk and Treasury Market Restructuring: Implications for Off-the-Run Trading[a]

Fabienne Schneider[b c]

Bank of Canada

September 16, 2026

Abstract

This paper develops a model to study how settlement risk affects asset prices, repo rates, and liquidity in the context of recently issued (on-the-run) and seasoned (off-the-run) Treasury securities. Higher settlement risk arises endogenously for off-the-run Treasuries because of greater inventory uncertainty. Consistent with the model, the so-called on-the-run premium is correlated in the data with an observable indicator of settlement risk, namely settlement fails. Finally, the analysis evaluates policies designed to improve off-the-run liquidity. It shows that a repo facility can restore market activity during crises only in some market segments and may even impair trading in others.

Keywords: Treasuries, on-the-run premia, settlement fails, inventory, repo facility
JEL Codes: G10, G12, G14, G18

[a]This paper should not be reported as representing the views of the Bank of Canada. The views expressed are those of the author. No responsibility for them should be attributed to the institution. I wish to thank Bruno Biais, Martin Brown, Jens Christensen, Jonathan Chiu, David Cimon, Stefania D'Amico, Michael Fleming, Sermin Gungor, Refet Gürkaynak, Kinda Hachem, Jay Kahn, Arvind Krishnamurthy, Ricardo Lagos, Antoine Martin, Plamen Nenov, Josef Schroth, Andreas Uthemann, Adrian Walton and especially Cyril Monnet, as well as the participants of the seminars of the University of Bern (2021, 2022), the Central Banker Course on Money Markets (2022), the Unibe Macro Workshop (2022), the BDP Alumni Conference (2022), the Australasian Finance & Banking Conference (2022), the Swiss Society for Financial Market Research Conference (2023), the seminars at the Bank of Canada (2024), the Bank of England (2024), and the Sveriges Riksbank (2024), the SSES Annual Congress (2024), the Essex-UCL Workshop on Decentralised Financial Markets (2024), the Bank of Canada Payments Seminar (2026) and other conferences for feedback.

[b]Bank of Canada, 234 Wellington Street, Ottawa, ON, K1A 0G9, Canada

[c]Corresponding author's e-mail address: fschneider@bank-banque-canada.ca

1 Introduction

With an average daily trading volume of more than a trillion U.S. dollars, the U.S. Treasury market is one of the most important and liquid markets in the global financial system, crucial to the conduct of monetary policy and a key pillar of the global economy.[1] Even though the Treasury market had been perceived as a stable, long-standing market across many episodes, it increasingly exhibited signs of fragility, culminating in the March 2020 crisis when it temporarily broke down.[2] The turmoil placed the U.S. Treasury market at the center of policy discussions and prompted renewed efforts to improve its resilience. These discussions brought particular attention to differences between the market for so-called on-the-run and off-the-run Treasury securities, with off-the-run securities experiencing substantially greater market disruptions than their on-the-run counterparts (Chaboud et al., 2025b, Wells, 2023).

On-the-run Treasuries are the most recently issued Treasuries of a given maturity. All other Treasuries are off-the-run. The volume of trades in on-the-run Treasuries is much larger than in the seasoned off-the-run Treasuries. For example, in the interdealer market, on-the-run Treasuries comprise over 80% of all trades. This is striking as they account for only 2% of all outstanding Treasuries (Chaboud et al., 2025b).[3] They also trade at significantly higher prices and lower yields than off-the-run Treasuries with similar cash flows and maturity dates, giving rise to a well-known phenomenon called "on-the-run" premium. In addition they exhibit lower repo rates and bid-ask spreads (Vayanos and Weill, 2008, D'Amico and Pancost, 2022, Chaboud et al., 2025). From the data it can also be observed that on average, USD 33 billion of Treasury contracts fail to settle each day, and, digging deeper, I find that off-the-run Treasuries have a median settlement fail rate that is twice of that of on-the-run Treasuries (see Figures 2a, 2b, 3, 5b in the next section for the cited data).

In the first part of the paper, I develop a model of the U.S. Treasury market to shed light on several questions: Why do we observe settlement risk in a benchmark market such as the U.S. Treasury market? Why do off-the-run Treasuries fail to settle more frequently? Do these fails reveal anything about the on-the-run premium and why does it consistently accrue

[1] The data are provided by the Securities Industry and Financial Markets Association and can be downloaded here: https://www.sifma.org/resources/research/statistics/us-treasury-securities-statistics/. In July 2026 the average daily trading volume was 1.2 trillion U.S. dollars.

[2] Asset sales, especially by non-banks that expanded their presence in the rapidly growing market, led to a rapid drying up of liquidity and a sharp decline in market depth. This was exacerbated by primary dealers' reluctance to absorb Treasuries onto their balance sheets, partly due to regulatory constraints (Eren and Wooldridge, 2021).

[3] See Figure A.6 in the Appendix for a graphical representation of the on-the-run cycle.

to the most recently issued Treasuries, a feature that standard models leave unresolved? How do an asset's distribution and its relative effective supply–two conceptually distinct channels–affect this outcome?

In the second part of the paper, the model is used to shed light on the ongoing policy discussion on restructuring the Treasury market (see Duffie, 2023), which emerged following the last crisis. I model the crisis with the breakdown in the off-the-run segment and analyze both proposed and recently implemented policy measures aimed at improving (off-the-run) liquidity, such as access to repo facilities and larger issuance sizes. I also briefly discuss all-to-all trading and tokenisation.

In the third part of the paper, I test several hypotheses derived from the theory using weekly U.S. Treasury primary dealer data–separately for on- and off-the-run Treasuries–on settlement fails, outright and financing transactions, and net positions. I further analyze the impact of settlement risk on Treasury pricing using a policy intervention.

The theoretical model incorporates key features of the U.S. Treasury market: It is an over-the-counter (OTC) market where dealers intermediate spot market trades between buyers and sellers by using repurchase agreement (repo) financing. Repos are also used to sell assets short. I abstract from Treasury auctions and instead assume that buyers and sellers are initially endowed with newly issued (on-the-run) securities, which become off-the-run over time. The model is dynamic and has two periods with two subperiods each. In the first subperiod, a dealer acquires the asset from a seller with a temporarily low valuation[4] and finances the purchase through a repo transaction with another dealer. The latter shorts the asset by selling it to a buyer with a high valuation. In the second subperiod, the short position is closed by repurchasing the asset from a buyer whose valuation has declined.[5]

Markets are characterized by random matching and bilateral bargaining over spot prices and repo rates. Because buyers are certain to hold newly issued on-the-run securities, dealers can reliably source these securities to meet repo delivery obligations. By contrast, ownership of off-the-run securities depends on trading histories, generating inventory heterogeneity and uncertainty. As a result, dealers may fail to locate a specific off-the-run security when needed to settle a repo, leading to costly settlement fails.[6]

Settlement risk induces a preference for on-the-run Treasuries, reflecting their greater

[4]Alternatively, the sale may be driven by a temporary liquidity shock. The results are unchanged.

[5]The initial dealer resells the asset to a seller whose valuation has recovered.

[6]I assume that repos are asset specific (known as specific repos) and therefore dealers are not allowed to exchange one asset for another when returning the collateral.

ease of settlement. Dealers are willing to accept lower repo rates when these securities are posted as collateral, causing them to trade "special" in the repo market.[7] Anticipation of this collateral value raises their spot prices, generating an on-the-run premium relative to otherwise identical off-the-run securities. They also exhibit lower bid–ask spreads and, as noted, lower repo rates and fewer settlement fails. Those phenomena are all consistent with the stylized facts for the U.S. Treasury market mentioned earlier.

Another central result is that a more unequal distribution of off-the-run securities among market participants compared to on-the-run securities generates an on-the-run premium.[8] Lower relative supply of off-the-run assets (introduced by buy-and-hold buyers in an extension) is not required (nor sufficient), but it can act as an amplifier by strengthening the effects of inventory uncertainty and settlement risk.[9] This insight extends beyond the on-/off-the-run setting to asset pricing more generally. Any two otherwise identical assets (in equal effective supply) can trade at different prices if one is more unequally distributed among market participants than the other. The example of off-the-run securities being more unequally distributed is natural, however, as repeated trading in a frictional OTC market over time causes holdings to become increasingly dispersed and unevenly distributed across inventories.

Settlement fails are costly not only for the failing counterparty but also for aggregate welfare. I show that a higher probability of settlement fails, for example due to a decrease in a small imbalance between supply and demand, lowers welfare because assets remain allocated to agents who do not currently value them the most. A calibration exercise shows that relatively small imbalances between supply and demand can generate the fail probabilities observed in the data and needed to explain the on-the-run premium.

In the second part of the paper, I focus on the policy discussion on how to restructure the market triggered by the last crisis. Not only settlement fails, but also on-the-run premia across all maturities rose sharply in March 2020, as shown in Figure 1.[10] Off-the-run

[7]Specialness is the premium paid to obtain a particular security in the repo market Duffie (1996). Specialness of a specific asset can be measured by the difference between the general collateral repo rate and the repo rate of that asset, as in D'Amico and Pancost (2022).

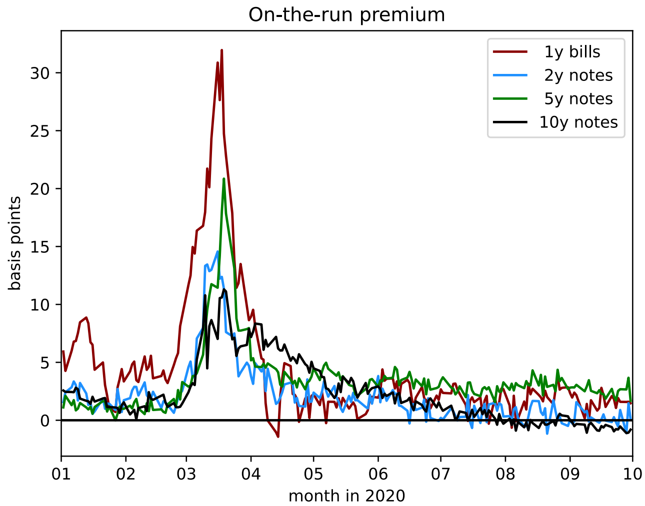
[8]"More unequally distributed" is defined relative to the on-the-run asset. In particular, the off-the-run asset is more unequally distributed if its inventory distribution across spot market participants is non-degenerate, whereas that of the on-the-run asset is degenerate by assumption.

[9]A natural conjecture from the model is that the level of scarcity per se is not what matters: as long as assets can be reliably located and sufficient supply is available at the source, scarcity alone does not generate premia. Instead, some degree of inventory uncertainty is required, with scarcity acting as an amplifier.

[10]As in Christensen et al. (2017), the on-the-run yield is subtracted from the par yield of seasoned bills and notes. The data are taken from the FED yield curve, which is an updated version of the original Gürkaynak-Sack-Wright curve (Gürkaynak et al., 2010), https://www.federalreserve.gov/data/nominal-yield-curve.htm,

Treasuries were at the epicentre of the crisis (Wells, 2023).

Figure 1: On-the-run premia during the March 2020 breakdown



Notes: This figure shows the on-the-run premium at the start of the COVID-19 pandemic. Source: FED and author's calculations.

I first explain why mainly the off-the-run segment broke down during the crisis. I model the crisis as a sudden, unexpected decline in buyer relative to seller participation that captures the "dash for cash" narrative and amplifies the pre-existing uncertainty in off-the-run trading to the point of market collapse. I then show that a central bank repo facility (as implemented by the Federal Reserve in 2021[11]) can improve liquidity in parts of the market by setting the facility rate sufficiently low. Importantly, however, it cannot fully restore intermediation in off-the-run securities and may harm trade in on-the-run assets. Even when this does not occur, the facility rate affects repo rates and spot prices in the on-the-run segment because the ceiling rate applies uniformly across assets, despite no intervention being required in that market. Also, it leaves dealers holding larger inventories of off-the-run assets on their balance sheets.

I then also analyze the following discussed policy interventions: larger issuance sizes, all-to-all trading and tokenisation.[12] I explain why larger issuance sizes improve liquidity in off-the-run but not on-the-run securities:[13] larger issuance reduces settlement fails primarily in

and the FRB-H15 Tables, https://www.federalreserve.gov/releases/h15/.

[11]See the Federal Reserve's website https://www.federalreserve.gov/monetarypolicy/standing-overnight-repurchase-agreements.htm for more info.

[12]See Duffie et al. (2021) and Chaboud et al. (2025b).

[13]Fleming and Garbade (2002) document this pattern in the data. Chaboud et al. (2025) cite it as suggestive evidence that larger issuance sizes may improve the liquidity of off-the-run securities, while noting that it is puzzling that the effect appears to be concentrated in the off-the-run segment of the market.

the off-the-run segment, where liquidity improvements are most valuable, thereby enhancing both liquidity and welfare. Finally, all-to-all trading can reduce non-strategic settlement fails and improve welfare, although this conclusion abstracts from broader market-structure trade-offs. Tokenisation can eliminate both strategic and non-strategic settlement fails, but at the cost of restricting risk taking.

Lastly, I turn to the empirical components of the paper. I first provide motivating evidence on settlement fails. I show that on-the-run securities account for only about 10% of total fails. Most observed fails are non-strategic, as a fail charge applies to all settlement fails and largely eliminates incentives for strategic failing. The associated penalty payments are economically meaningful. Based on the applicable fee schedule, I estimate that primary dealers pay roughly USD 21 million per month in fails charges. This figure captures only the direct fees and excludes other costs associated with settlement failures.

The main empirical analysis then focuses primarily on testing two predictions of the model. First, the theory delivers a joint sign restriction: higher fail rates of off-the-run Treasuries are associated with an increase in the on-the-run premium, whereas higher fail rates of on-the-run Treasuries are associated with a reduction. Using fail data for U.S. primary dealers, I regress the on-the-run premium on fail rates of both securities and find evidence consistent with the prediction. Second, the model predicts that greater inventory uncertainty is associated with higher settlement fail rates. Because individual inventory positions are not observed, I use aggregate inventory holdings as a proxy, which the model rationalizes by showing that larger inventories reduce inventory uncertainty. Consistent with this prediction, higher inventories are associated with significantly lower fail rates. Because the on-the-run fail data are available at the maturity level, I confirm this finding at the maturity level as well.

To complement the baseline evidence, I exploit the introduction of the above mentioned fail charge in 2009 as a policy intervention to provide further evidence that settlement risk is a determinant of Treasury pricing. The analysis shows that the introduction of the fail charge led to a substantial reduction in off-the-run settlement fails[14] and a simultaneous decline in on-the-run premia within a narrow window around its implementation.

To summarize, I provide a theory that links on-the-run premia to settlement risk arising from short selling. Settlement fails constitute an observable manifestation of this settlement

[14]Data on on-the-run settlement fails are only available from 2013 onward and are therefore not part of the analysis.

risk in the data and thereby provide information about how difficult it is to locate and obtain a security. The observed fail rates may even understate the true settlement risk of off-the-run Treasuries because securities with particularly high ex ante fail risk may not be sold short in the first place.[15] These avoided fails may in fact be the most costly ones, as sourcing a replacement security is likely substantially more difficult and expensive than for a highly liquid on-the-run Treasury.[16]

Related Literature First, my paper relates to the literature on on- versus off-the-run assets with Vayanos and Weill (2008) as the main reference. They provide a theoretical framework for the on-the-run premium where liquidity is self-fulfilling. The main limitation the authors point out, is that their framework features two equilibria, each with the on-the-run premium on either of two identical assets (representing the on- and off-the-run asset). It does not explain why agents systematically concentrate on on-the-run securities. I address this issue by incorporating the key dimension along which the two assets differ: time since issuance. While time since issuance is the defining feature that distinguishes on-the-run from off-the-run securities, it is typically absent from standard frameworks. Also, Vayanos and Weill (2008) point to a decline in the relative effective supply of off-the-run securities over time as a potential equilibrium-selection device for future research.[17] I show, however, that although a lower effective supply can amplify the on-the-run premium, it is not required for the premium to emerge. Lastly, in Vayanos and Weill (2008), repo contracts terminate when counterparties switch states.[18] I model repos as term contracts that create settlement pressure at a specific point in time. This generates settlement risk driven by inventory uncertainty, introducing a novel dimension to the theory of on- and off-the-run securities.

A related paper building directly on Vayanos and Weill (2008) is Corradin and Maddaloni (2020). They examine how central bank purchases affect liquidity and specialness in the repo market in the presence of short selling. They also provide evidence that older assets with lower turnover are more likely to fail, consistent with my results. With respect to

[15] Primary dealers hold, on average, a negative inventory position in on-the-run Treasuries and a positive inventory position in off-the-run Treasuries (see Table 6 in the Appendix A.8.1). This finding is consistent with the notion that dealers are more willing to short-sell on-the-run securities than off-the-run securities.

[16] In the model, assets are sold short only up to a maximum fail probability; beyond that threshold, short selling ceases.

[17] While they show that a lower supply of off-the-run assets could serve as a selection device, the static nature of the model prevents it from capturing how effective supply declines over time and, consequently, from showing how this mechanism could potentially give rise to a unique equilibrium.

[18] If the borrower of the asset does not have the asset at hand when he switches state, he simply waits until he can acquire it; if the lender switches state, the two counterparties part ways under the assumption that cash collateral is seized.

theories linking specialness and bond market pricing there is for example Duffie (1996). In Duffie (1996), consistent with this theory, higher specialness translates into higher prices.[19] Duffie (1996) is not about the on-the-run phenomenon,[20] nor does his model include fail risk, the main driver of specialness. Another example is Huh and Infante (2021) who explain the relationship between repo specialness and bid-ask spreads in bond markets, focusing on the role of leverage. Lastly, several papers study inventories in OTC markets, including Cohen et al. (2024), Lagos et al. (2011), and Weill (2007). In contrast to my focus, these models primarily examine dealers' optimal inventory management.

Empirical work on the on-the-run premium includes, for example, Strebulaev (2002), Goldreich et al. (2005), Graveline and McBrady (2011), and D'Amico and Pancost (2022), each offering different explanations for why on-the-run premia arise.[21] Closest to my work is D'Amico and Pancost (2022), who link the on-the-run premium in the spot market to the risk of unexpected fluctuations in the collateral value of Treasuries in the repo market. I also link the spot market premium to specialness in the repo market, and its magnitude moves with specialness. However, none of these studies connect the premium to inventory uncertainty and settlement risk.

Settlement risk has recently attracted increased attention in the financial literature. For example, Lee et al. (2026) provide evidence that settlement risk is a priced friction in currency markets and estimate that it increases currency excess returns by approximately 10 basis points. Related, the empirical literature on settlement fails in the Treasury market includes Fleming and Garbade (2002, 2004, 2005), Fleming et al. (2014), and Fleming and Keane (2016a, 2016b).[22] I discuss settlement fails including the literature in the following section.

The second part of the theory addresses the crisis and the current policy debate. Recent work on this topic, especially on off-the-run liquidity and how to reform the Treasury market includes Chaboud et al. (2025a, 2025b), Duffie (2020, 2023), Duffie et al. (2021), Eren and

[19]Specialness itself stems in Duffie (1996) from the inability, opportunity cost, or transaction cost of lending an asset as collateral. It increases with this inability as well as with short selling demand and liquidity in the form of lower transaction costs.

[20]Duffie (1996) mentions, however, that on-the-run assets trade frequently special and also provides numbers.

[21]Strebulaev (2002) suggests that the premium may reflect differences in tax treatment rather than liquidity premia. Goldreich et al. (2005) distinguish between current and future liquidity and argue that expected future liquidity, not just current liquidity, determines prices and is a significant driver of the on-the-run premium. Graveline and McBrady (2011) link the premium to the demand for hedging interest rate risk.

[22]New literature on settlement fails in other markets includes for example Fleming et al. (2023) and Iyer and Macchiavelli (2017).

Wooldridge (2021), Fleming and Keane (2021), He et al. (2022), Kashyap et al. (2025), Schrimpf et al. (2020), Sokolinskiy (2025), and Vissing-Jorgensen (2021).

The structure of the paper is as follows: Section 2 describes the U.S. Treasury market, with Subsection 2.1 focusing on trading structure and trading volumes and Subsection 2.2 on settlement fails. Section 3 contains the theoretical part of the paper. Subsection 3.1 describes the environment and Subsection 3.2 the maximization problems. Subsection 3.3 provides intuition for the main equilibrium. Subsection 3.4 presents the equilibrium, including all results on the relationship between on-the-run premia, specialness and settlement fail risk. Subsection 3.5 contains the discussion on scarcity and distribution, including an extension on buy-and-hold buyers. Subsection 3.6 discusses welfare. Subsection 3.7 focuses on the crisis and reform discussions. Section 4 contains the empirical part of the paper. Subsection 4.1 derives two hypotheses from the theoretical results, which are tested and analyzed in Subsection 4.2. Subsection 4.3 presents evidence from a policy intervention. Section 5 concludes.

2 Description of the Treasury market

In this section, I describe the U.S. Treasury market, its structure, and trading dynamics, and provide evidence for the stylized facts highlighted in the introduction. A particular emphasis is placed on settlement fails.

2.1 Trading structure and volumes

The Treasury spot market, particularly the dealer-to-client segment, is mostly OTC (Fleming et al., 2018, Chaboud et al., 2025a). This means that there is no all-to-all trading at a central venue and no central pricing.[23] Depending on the security traded and the trading counterparties involved, the degree of friction in the OTC market varies. For example, dealer-to-dealer trading of benchmark on-the-run Treasuries on electronic platforms such as BrokerTec is less frictional than interdealer and dealer-to-customer trading of the less liquid off-the-run Treasuries intermediated on voice and more manually assisted electronic platforms (Bessembinder et al., 2020, U.S. Department of the Treasury et al., 2015).[24]

[23]Chaboud et al. (2025a) discuss the advantages and disadvantages of introducing all-to-all trading in the U.S. Treasury market.

[24]The overall share of trading on all types of electronic platforms in the U.S. Treasury market is 70% (Bech et al., 2016).

On-the-run Treasuries are the most recently issued Treasuries of a given maturity, and all previously issued Treasuries of the same maturity are referred to as off-the-run. As Figure 2a shows, for 10-year Treasuries, on-the-run Treasuries trade at significantly lower yields than off-the-run Treasuries with very similar[25] cash flows and maturity dates.[26] On-the-run Treasuries also have higher prices and lower repo rates. D'Amico and Pancost (2022) document that on-the-run Treasuries have on average a repo rate 19bp below the general collateral repo rate. Specialness decreases to 9bp for the first off-the-run generation and 5bp for the second and older bonds.[27]

Also, despite being more expensive and far lower in outstanding amount, on-the-run Treasuries trade in much larger volumes than off-the-run Treasuries, as shown in Figure 2b.[28] This is true whether I look at dealer-to-customer or interdealer and automated trading system (ATS) trades.[29]

2.2 Settlement fails

In this section, I discuss settlement fails in the Treasury market, which play a central role in the subsequent analysis.

2.2.1 Overall fails

A surprising fact is that, on average, USD 33 billion of Treasuries are not delivered on time to settle a contract each day.[30] These events are commonly referred to as "settlement fails".

[25]The term "very similar" refers to the fact that, in general, no two Treasuries in the market have exactly the same cash flow and maturity date where one is on-the-run and the other is off-the-run. In fact, if you want to compare Treasuries with the same overall maturity, it is impossible to do so. In practice, therefore, one compares Treasuries with the same overall maturity using an estimated off-the-run yield curve (see, e.g., the first figure in Christensen et al., 2017), abstracts from small differences in cash flows and maturity dates, or compares on- and off-the-run Treasuries with the same maturity date that have a different overall maturity (see, e.g., Christensen et al., 2020). In Figure 2a, as in Christensen et al. (2017), the on-the-run yield is subtracted from the par yield of seasoned bills and notes.

[26]The data are taken from the FED yield curve, which is an updated version of the original Gürkaynak-Sack-Wright curve (Gürkaynak et al., 2010), https://www.federalreserve.gov/data/nominal-yield-curve.htm, and the FRB-H15 Tables, https://www.federalreserve.gov/releases/h15/.

[27]Keane (1996) documents the specialness pattern over the on-the-run period of a Treasury. Specialness first increases steadily before it decreases near the end of the on-the-run period.

[28]Trading volumes in on-the-run and off-the-run Treasuries are the volumes reported to TRACE between January 2024 to December 2025. The TRACE data is available here:
https://www.finra.org/finra-data/browse-catalog/about-treasury/monthly-file.

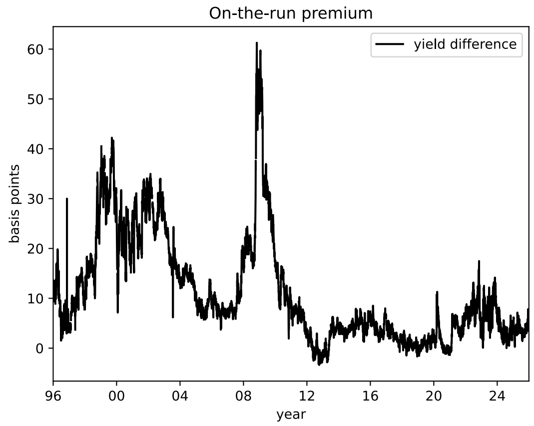
[29]See Figure 1 in Barclay et al. (2006) for a graphical representation of how the trading volume collapses on the day the Treasury goes off-the-run.

[30]This value represents the daily average over the period 2022-2025. The data are provided by the Depository Trust and Clearing Corporation (DTCC) and can be downloaded here: https://www.dtcc.com/charts/daily-total-us-treasury-trade-fails. In times of stress, the daily value can spike. For more information on settlement fails, see Fleming and Garbade (2005).

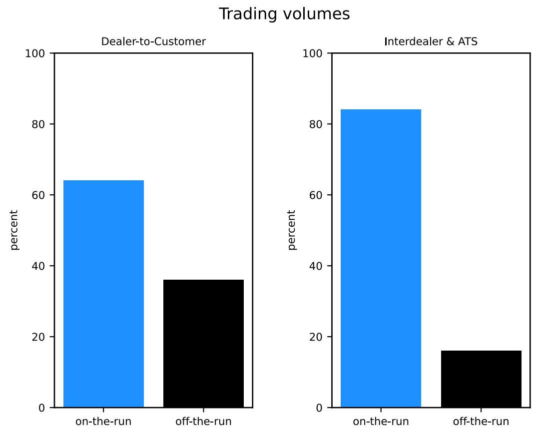
Figure 3 plots the total value of daily settlement fails in Treasuries.

Figure 2: Yields and volumes

(a) On-the-run premium on 10-year Treasury notes

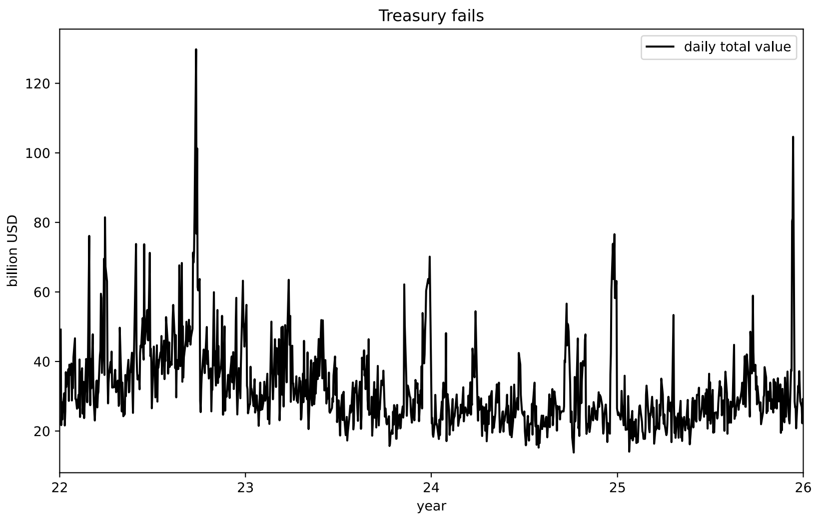


(b) Trading volumes by segment



Notes: The left figure shows the on-the-run premium for 10-year Treasuries. The right figure shows the trading volumes for on- and off-the-run Treasuries separately in the Dealer-to-Customer and Interdealer & Alternative Trading System segments. Source: a) FED and author's calculations, b) TRACE and author's calculations.

Figure 3: Daily settlement fails



Notes: The figure shows the value of Treasuries that were not delivered on time to fulfill a contract. Source: DTCC.

The OTC structure implies that there are search costs that can explain settlement fails. In particular, search costs become relevant when financial contracts include delivery constraints. For example, to sell a particular Treasury short, it is borrowed using a special repo and sold in the market today. It is called "special" because the collateral of the repo is fixed (even before the rate is negotiated) and determined by its number, called ISIN or CUSIP.[31] The next day, a Treasury with the same ISIN or CUSIP is bought in the spot market, preferably at a lower price than it was sold for on the previous day, and returned to the lender in the repo transaction. If such a Treasury cannot be found, a settlement fail occurs.

Note that arbitrage to exploit the price difference between on- and off-the-run Treasuries involves short selling, but is mostly absent as relatively expensive Treasuries also have lower repo rates (Krishnamurthy, 2002).

2.2.2 Fails in on- and off-the-run assets

Figure 4 plots the share of on-the-run fails relative to total fails. The data stemming from the FEDs Primary Dealer Statistic include outright transaction and financing fails and both, fails to deliver and receive. The series is available at weekly frequency and is aggregated to monthly averages.[32] The figure shows that the share of on-the-run fails is slightly above 10% on average[33], but exhibits pronounced spikes at certain points in time. One can show that these spikes typically coincide with spikes in off-the-run Treasuries. The spikes in on-the-run Treasuries are consistent with the occurrence of so-called daisy chain fails, which are more prevalent for on-the-run securities. In such cases, one fail can trigger subsequent fails as trades are linked within a settlement chain (Fleming and Keane, 2021). Fleming et al. (2014) and Fleming and Keane (2016a) show that gross fails are much higher in seasoned Treasuries (issued more than 180 days ago) than in others (including on-the-run Treasuries). Fleming et al. (2014) also write in 2014 that fails in well off-the-run securities have risen steadily, with notable quarter-end spikes.[34] Figure 4 additionally shows a sharp increase in off-the-run fails from 2014 to 2018, while on-the-run fails remained stable over this period. More recently, between 2021 and 2022, off-the-run fails again experienced prolonged episodes of significant growth. In both episodes, daily fails increased from about USD 10–20 billion

[31] As pointed out earlier, these repos may have a rate that differs from the general collateral rate.

[32] As the Federal Reserve's Primary Dealer dataset does not report fails for on-the-run Treasury bills, the estimated share may be slightly understated. Also, the share is not an exact measure. This is mainly because one part of the time series used in the calculation is an average over the reporting week and the other part of the time series reports a value as of the reporting weekday.

[33] The median is 10% and the mean is 11%.

[34] While this may reflect increased balance sheet management around reporting dates, the upward trend appears to predate the binding regulatory changes often cited as its cause.

to as much as USD 70 billion.

While fails in off-the-run Treasuries are only available in aggregate, Figure 5a zooms in on on-the-run fails and decomposes their value by maturity.[35] Fails were elevated in 2022 and 2023. The series also exhibits sharp, short-lived spikes, particularly in the 10-year Treasury. This is not surprising, as the 10-year bond is, even within Treasuries, an important benchmark maturity, serving as a widely used reference rate for pricing other assets and assessing economic conditions, and is also among the most heavily traded securities (FRED, 2026, Fleming and Ruela, 2020, Carlson, 2026). In addition, in 2020 the 20-year maturity was reintroduced, which displays similar spike patterns although for other reasons (Fleming and Ruela, 2020).[36]

Figure 5b shows the fail rate, which is calculated by dividing the value of Treasuries that failed to be delivered on time by the value of all Treasuries traded. Interestingly, the fail rates differ depending on whether a Treasury is on- or off-the-run, and Figure 5b shows that fails involving on-the-run Treasuries are less frequent than those involving off-the-run Treasuries.[37]

2.2.3 Strategic and non-strategic settlement fails and their costs

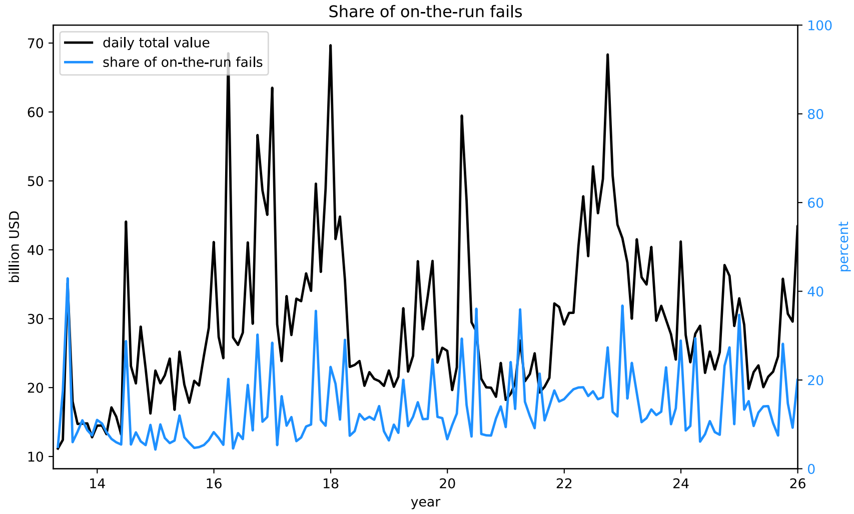
When a fail to deliver a security occurs, the intended recipient does not pay (or deliver its leg) until the failing counterparty delivers the securities. As a result, the failing counterparty loses and the recipient gains the time value of the transaction proceeds over the fail interval. This time value can be quantified as the interest earned in the general collateral repo market. Absent any explicit fail fee, this opportunity cost provides an incentive to not fail and to borrow the security at a fee in order to deliver it on time. However, when the general collateral rate is very low, failing can become privately optimal, as the borrowing cost may exceed the opportunity cost of failing (Fleming and Garbade, 2005). Such so-called strategic

[35]The data are from the FED's Primary Dealer Statistics. The data include outright transaction and financing fails to deliver and receive by primary dealers. The series is available at weekly frequency and is aggregated to monthly averages. For outlier correction, I exclude the top 5% of the observations.

[36]The U.S. Treasury brought back the 20-year bond to expand financing capacity and create an additional liquidity point along the yield curve. However, early evidence suggests that its trading activity is considerably lower than that of other on-the-run securities and that its liquidity is more comparable to that of the 30-year bond. This may help explain why this maturity also exhibits relatively high fail volumes (Fleming and Ruela, 2020).

[37]The data are from the FED's Primary Dealer Statistics. They include outright and financing fails. The median fail rate for on-the-run Treasuries is 0.36% and the median fail rate for off-the-run Treasuries is 0.59%. The rates are not an exact measure. This is mainly because one part of the time series used in the calculation is an average over the reporting week and the other part of the time series reports a value as of the reporting weekday. Each series is outlier adjusted. The highest 5% of values are excluded and values need to be non-negative.

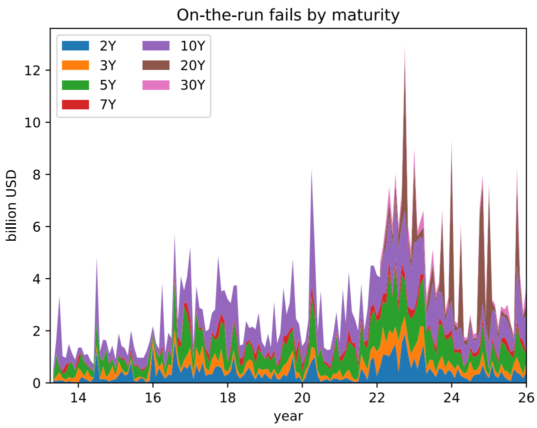
Figure 4: Share of on-the-run fails



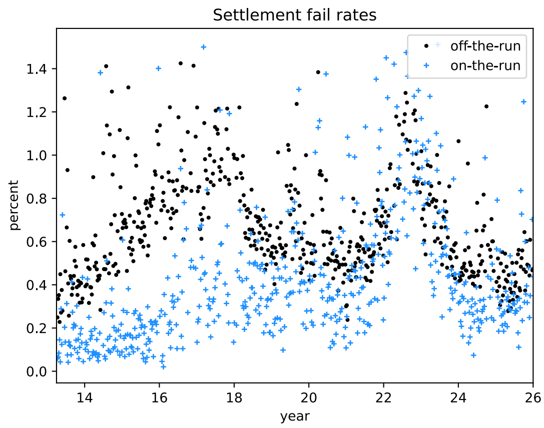
Notes: The figure shows the share of on-the-run Treasury fails relative to total Treasury fails.

Figure 5: Fails in on- and off-the-run Treasuries

(a) Settlement fails in on-the-run Treasuries



(b) Fail rates in on- and off-the-run Treasuries



Notes: The left figure shows the value of on-the-run Treasury fails by maturity. The right figure shows the percentage of Treasuries that failed to settle separately for on- and off-the-run Treasuries. Source: FED and author's calculations.

fails were common prior to the introduction of the Treasury Market Practice Group (TMPG) fails charge in 2009. The TMPG fails charge is an interest rate applied to the cash value of the transaction that accrues over the duration of the fail. To discourage strategic fails, the fails charge is tied to the federal funds rate: it is set equal to 3% per annum minus the federal funds target rate, or, since the introduction of the target range, the lower bound of that range. Initially, the fails charge was bounded below by a floor of 0%, which was later raised to 1% (TMPG, 2018, 2020).[38]

Since the introduction of the fail charge fails declined, particularly the extreme spikes. We can see this in Figure 6a.[39] Figure 6b plots the applied fail charge.[40] Nevertheless, we have seen that there still remains a substantial amount of fails. These must be predominantly non-strategic fails. Traders may be temporarily unable to locate or borrow the required securities in decentralized markets due to search frictions (Duffie et al., 2005, Vayanos and Weill, 2008). This can be one source of frictions leading to fails. Fleming and Keane (2021) further list that non-strategic fails can occur due to miscommunication, operational problems and daisy chain fails. A direct cost of these non-strategic fails are the fail fees paid. As an estimate of their size,[41] one can multiply the fails to deliver value of primary dealers by the applied fail charge. Figure 7 plots the estimated charged fees to primary dealers over time. On average they were 21 million USD per month.

In addition to those direct costs, there are other costs. Fleming and Garbade (2005) point out that an ancillary cost of increased fails is a rise in counterparty credit risk. In the event of default, the non-defaulting counterparty faces the risk of having to replace the security at a higher price if its value has increased. Settlement fails give rise to capital requirements under the Basel framework, which requires banks to hold capital against exposures from failed trades (Basel Committee on Banking Supervision, 2019).[42] In addition, there are operational costs, as market participants seek to avoid situations in which their counterparties involuntarily finance their short positions and bear the associated risk of fail-

[38]The exact formula is

$$C = P \cdot \frac{1}{360} \cdot \max(1\%,\ 3\% - R)$$

where C denotes the daily fails charge in dollars, P the principal (cash value of the transaction in dollars), and R the federal funds rate (or the lower bound of the target range), expressed in percent.

[39]This figure also shows that the value of fails to receive and fails to deliver by primary dealers are almost of equal size. The data are from the FED's Primary Dealer Statistics.

[40]The fail charge is calculated using the lower bound of the FED's target range for the federal funds rate. Data on the lower bound are obtained from FRED.

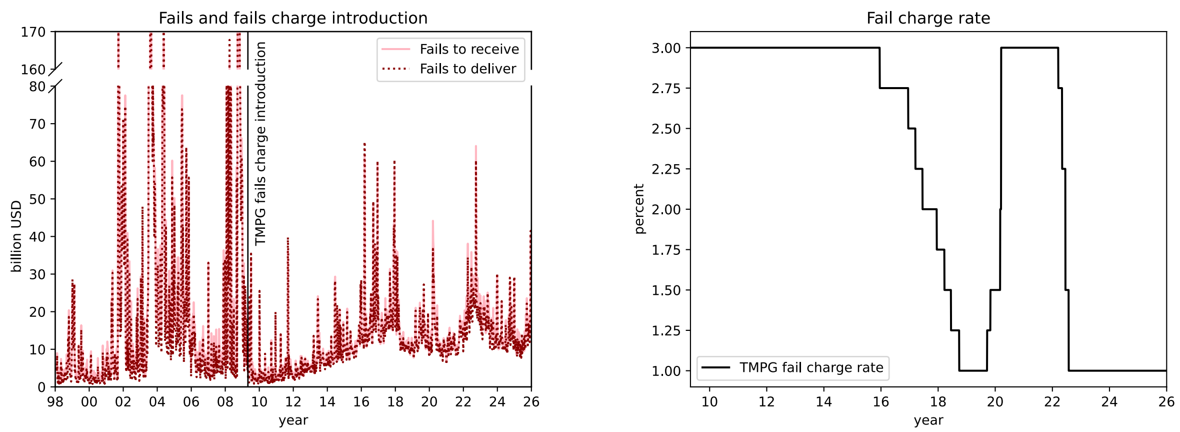
[41]It is an estimate because fails to deliver on outright transactions are measured at the par value of the transaction, whereas fails to deliver on financing transactions are measured at the amount payable on the scheduled settlement date.

[42]Under Basel III, capital requirements depend on the settlement structure and the duration of the fail. For delivery-versus-payment (DvP) transactions, only the net exposure (i.e., replacement cost) is at risk. While this exposure arises immediately (e.g., \$2M on a \$100M trade), capital charges are typically imposed only after a grace period of five days, after which they increase with the length of the delay, reaching up to 100% for prolonged fails (beyond 46 days). In contrast, for non-DvP transactions, the full principal is at risk from the outset (e.g., \$100M on a \$100M trade). If settlement fails persist beyond five business days, the entire notional exposure becomes subject to capital charges, and additional replacement risk may apply, with risk weights rising substantially (e.g., up to 1,250%) for long delays. Note that securities financing transactions are regulated separately (Basel Committee on Banking Supervision, 2019).

ing to deliver themselves. Finally, market liquidity may deteriorate as participants adjust their behavior to avoid costly fails (Fleming and Garbade, 2005).[43] The overall costs are difficult to assess. However, given the persistently elevated level of off-the-run fails observed in Figure 4, they may play an important role in the Treasury market.

Figure 6: Fails and fail charge

(a) Decrease in fails since the fail charge introduction (b) Fail charge rate since the introduction



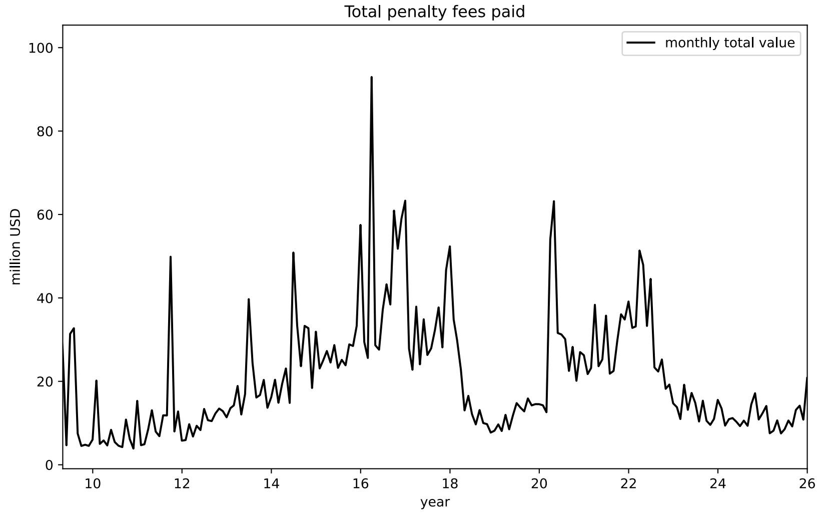
Notes: The left figure shows the daily value in fails to receive and fails to deliver before and after the introduction of the TMPG fails charge in 2009. The right figure shows the fails charge, the size of which depends on the federal funds rate (or the lower bound of the target range). Source: FED and author's calculation.

3 Theory

In the theoretical part of the paper, I develop a dynamic model of the U.S. Treasury market to explain how settlement risk impacts spot prices, repo rates and liquidity and why the impact is different for on- and off-the-run Treasuries. I distinguish between two potential drivers of settlement risk: the relative supply of an asset and its cross-sectional distribution in the inventories of market participants. Using the model, I also analyze a crisis scenario and examine the implications of several policy proposals that have either recently been implemented or are currently under discussion. Finally, the model provides the foundation for two hypotheses that are tested in the subsequent empirical analysis. First, it delivers a joint sign restriction: higher fail rates of off-the-run Treasuries are associated with an

[43]For example, in 2001, the U.S. Treasury intervened by reopening a Treasury note to provide additional liquidity.

Figure 7: Charged penalty fees to primary dealers



Notes: The figure plots the estimated penalty fees paid by primary dealers. It multiplies the daily fails to deliver with the fails charge rate (divided by 360). The figure plots the monthly sum of the daily values.

increase in the on-the-run premium, whereas higher fail rates of on-the-run Treasuries are associated with a decrease. Second, the model predicts a positive relationship between inventory uncertainty and settlement fail rates.

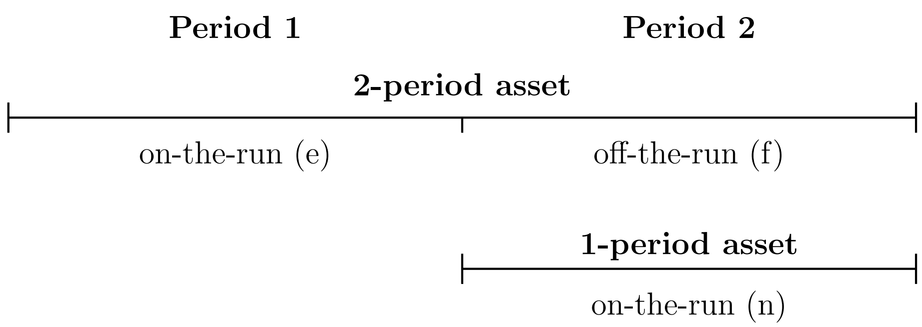
Appendix A.1 provides an overview of the notation used throughout the paper, including all parameters and equilibrium objects.

3.1 The environment

I use a deterministic two-period model. Each period consists of two subperiods, denoted by $t \in \{1, 2\}$. For simplicity, there is no discounting across periods.

There are two kinds of financial assets: *two-period assets* maturing after two periods, and *one-period assets* maturing after one period. Assets pay a coupon δ at the end of each subperiod and are storable and divisible. Two-period assets are issued at the beginning of the first period. We refer to them as on-the-run in the first period and off-the-run in the second period, after which they mature. Assets in a given period are indexed by i. We denote the two-period assets in the first period by $i = e$ and in the second period by $i = f$. One-period assets are issued at the beginning of the second period and mature after this period. We refer to them as on-the-run, and we denote them by $i = n$.[44] The letters n and f refer to the assets' o**n**-the-run and o**f**f-the-run states, respectively. I will refer to them henceforth as the on-the-run asset and the off-the-run asset. The two-period asset maturing in two periods is also on-the-run in period 1, but its state is not relevant to the analysis. Figure 8 provides an overview of the timeline.

Figure 8: Assets



Notes: The figure provides an overview of the assets available over the entire time horizon. Two-period assets are issued at the beginning of the first period; they are on-the-run in the first period (denoted by $i = e$) and off-the-run in the second period (denoted by $i = f$), after which they mature. One-period assets are issued at the beginning of the second period, are on-the-run, and are denoted by $i = n$.

In the second period the two assets are identical in terms of maturity date and coupons. The only difference is the issuance date. Therefore, I compare the prices of these two assets in period 2 to measure the on-the-run premium.[45]

In addition to assets, agents can transfer utility by producing and consuming a general good (more details below). Since this good is used to settle trade, it is natural to call it *money*.[46]

There are two types of agents: *spot traders* and *dealers*. Each spot trader is endowed

[44] The terms on-the-run and off-the-run are used here as terminology, even though no new generations of assets of the same kind are issued in the model. One can think of new issuance taking place, but this does not affect the analysis. The key distinction is that on-the-run assets are newly issued, whereas off-the-run assets have already been on the market for some time.

[45] The two assets have the same cash flow to maturity and the same maturity date as in Vayanos and Weill (2008) and Pasquariello and Vega (2009). See also, for example, Christensen et al. (2020) for how the premium can be measured.

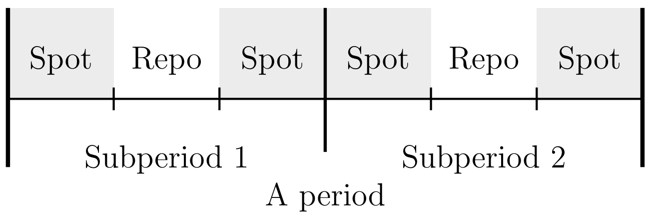
[46] This is similar to Lagos and Wright (2005), where a good can be consumed and produced with linear utility and disutility. It is used to adjust money balances, and money serves as a medium of exchange. To minimize the use of notation I directly use the good as a medium of exchange in my set-up.

with one two-period asset at the beginning of the first period and one one-period asset at the beginning of the second period. Spot traders can be divided into two subtypes: buyers (B) with mass 1, and sellers (S) with mass $\sigma \in (0,1)$.[47] They all derive linear utility from coupons but differ in the (in)convenience value they obtain from holding assets. Their per-asset utility (coupon plus (in)convenience) is as follows: At the end of the first subperiod, sellers receive utility $\delta^l < \delta$ per asset they hold, while buyers receive $\delta^h > \delta$. At the end of the second subperiod, sellers receive δ, while buyers receive $\delta^l < \delta$.[48] All spot traders have linear disutility from spending and linear utility from acquiring money, both with a marginal rate of one, and can spend and acquire money at any time.

Dealers have mass 2σ. They derive utility δ from the coupons accrued from holding assets (with no additional (in)convenience) and have linear utility from spending money with a marginal utility of one. In each first subperiod dealers can borrow money, but cannot otherwise acquire it. In the second subperiod they pay interest at rate $\bar{r}$ on the amount borrowed.

There are also two types of markets: *spot markets* and *repo markets*. In each subperiod, a repo market takes place, with one spot market preceding it and one spot market following it. Figure 9 gives an overview.

Figure 9: Markets per period



Notes: The figure provides an overview of the sequence of markets within a period.

In every spot market dealers are randomly matched with either a seller or a buyer. The matching is persistent across periods and subperiods, so that each dealer is always matched with the same type (but not agent) over the entire time horizon.[49] We refer to dealers

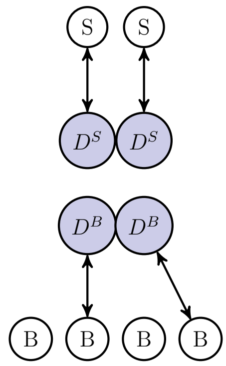
[47] In equilibrium, this imbalance between buyers and sellers induces heterogeneous trading histories and asset inventories, as not all trading demands can be satisfied at all times. The resulting dispersion in trading outcomes could, however, also arise from other market frictions.

[48] For full symmetry, one could assign a utility of δ^h to the seller in the second period. This is not necessary, however, and valuing the asset based on the coupon δ simplifies the analysis without loss of generality.

[49] This assumption is made for simplicity, as otherwise the probabilities of meeting another type in the

matched with sellers as D^S and those matched with buyers as D^B. The two dealer types differ only in the type of counterparty with whom they are matched. Otherwise, dealers are identical in all characteristics and preferences. Figure 12 gives an overview.

Figure 10: Matching in the spot markets



Notes: In the spot markets dealers are either matched with sellers S or with buyers B. We call the dealers matched with sellers D^S and the ones with buyers D^B. Each dealer is always matched with the same type over the entire time horizon.

In each subperiod, D^S–S and D^B–B pairs trade with each other once. In the first and last spot market (outer markets) of a period, D^S–S pairs trade. In the second and third spot market (middle markets) of a period, D^B–B pairs trade. Matched pairs trade assets against money, while other agents cannot trade. When trading, pairs maximize joint surplus and bargain over the price. Trade occurs if the joint surplus is (weakly) positive. For tractability I assume that maximally one asset of each type i can be traded per matched pair in a given subperiod. Dealers matched with sellers (D^S) have bargaining power $\theta^d < 1$, whereas dealers matched with buyers (D^B) have full bargaining power.[50] The bid price denotes the price at which a dealer purchases an asset, while the ask price denotes the price at which the dealer sells an asset. The bid price of asset i in subperiod t is denoted by $P_t^{i,b}$, and the corresponding ask price is denoted by $P_t^{i,a}$. The traded quantities are denoted by $A_t^{i,b}$ and $A_t^{i,a}$. Dealers can buy at most one unit of each asset i. Payments settle on a deferred net basis, so cash obligations are not immediate but instead aligned at settlement

future would need to be taken into account. While this would considerably complicate the analysis, it would not add any economic insights.

[50] Assuming full bargaining power simplifies the analysis considerably while preserving all relevant channels. First, a buyer is indifferent as to whether he sells the asset in the future because his surplus from selling is zero. Second, under partial bargaining power, the short-selling dealer would demand a higher price for the off-the-run asset than for the on-the-run asset in equilibrium due to a higher probability of settlement fail (see below), thereby blurring the main channel driving the on-the-run premium.

(see for example McLaughlin, 2023). As a result, initial financing at rate $\bar{r}$ can effectively be replaced by cheaper funding, for instance by repoing out the asset. I will denote the surplus of the dealer D^S (D^B) when trading asset $i \in \{n, f\}$ in subperiod t in the spot market by $S_t^{DS,i}$ ($S_t^{DB,i}$). The analogous surpluses for the buyer B and seller S are $S_t^{B,i}$ and $S_t^{S,i}$.

In the repo market, dealers trade repo contracts with each other as follows: In the first repo market of a period, dealers of type D^S and D^B are randomly matched with each other. They maximize their joint surplus by choosing the amount of repo contracts they trade. A repo contract is a pair (i, r^i), where one unit of asset i is exchanged for cash in the first subperiod of a period. The transaction is reversed in the repo market in the second subperiod, and the interest rate r^i is paid on the borrowed cash. The repo rate is determined through bargaining, where dealers of type D^S have bargaining power $\theta > 0$. The amount of cash exchanged is given by $P_1^{a,i}$, that is, the asset is valued at its current ask price in the spot market. (If no ask price is available because trade has ceased, the asset is valued at an exogenously specified price $\bar{P}_1^i$.) Coupons accrue to the cash borrower, who remains the legal owner of the collateral during the repo transaction. If the cash lender fails to deliver the collateral when the transaction reverses, the lender must fully compensate the cash borrower with a fail compensation C^i, which equals the prevailing spot market ask price of the asset at the time of the fail.[51] We denote by $\mathbb{P}^i$ the endogenous probability that a fail to deliver occurs in asset i, conditional on it being repoed out. I assume that cash borrowers are always committed to deliver the cash leg at settlement. Trade in the repo market occurs if the joint surplus of matched dealers is weakly positive, and the amount of repo contracts traded in asset i by matched dealers is A_R^i. I assume that matched dealer couples can maximally trade one repo contract per asset i per period to facilitate tractability. The surplus of the dealer D^S (D^B) when trading asset $i \in \{n, f\}$ in subperiod t in the repo market is denoted by $S_R^{DS,i}$ ($S_R^{DB,i}$).

[51] In equilibrium (characterized later), this payment fully compensates the non-failing dealer, D^S, for the utility loss resulting from its inability to sell the asset after receiving it back. Absent a settlement fail, the dealer would have sold the asset at the prevailing spot market ask price. We assume full compensation for simplicity. The compensation implies a linear disutility for the cash lender in equilibrium. I leave the exact interpretation of the compensation open. Next to be interpreted as a monetary compensation, it could be other economic costs associated with a settlement fail, including search costs and premiums paid for substitute securities.

3.2 Maximization problems

This subsection discusses the maximization problems and their solutions. I focus in this section on the second period, as it is the relevant period for the comparison between on-the-run and off-the-run securities. The maximization problems for the first period are very similar and presented in Appendix A.3.2. To solve the model, it will be useful to define $\mathbb{P}^{max,i}$ which will be the maximum ex ante fail probability a dealer tolerates under which he still sells an asset short.

Definition 1. *The maximum ex ante fail probability in the second period is defined as* $\mathbb{P}^{max,n} \equiv \mathbb{P}^{max,f} \equiv \frac{(\delta^h-\delta)+\bar{r}(\delta^h+\delta^l)}{\delta-\delta^l}$.

Second, I assume that $\bar{r}$ is in equilibrium sufficiently low for dealers such that it does not prohibit trade and that settlement failure with certainty is never an optimal strategy. More specifically, I assume the following:

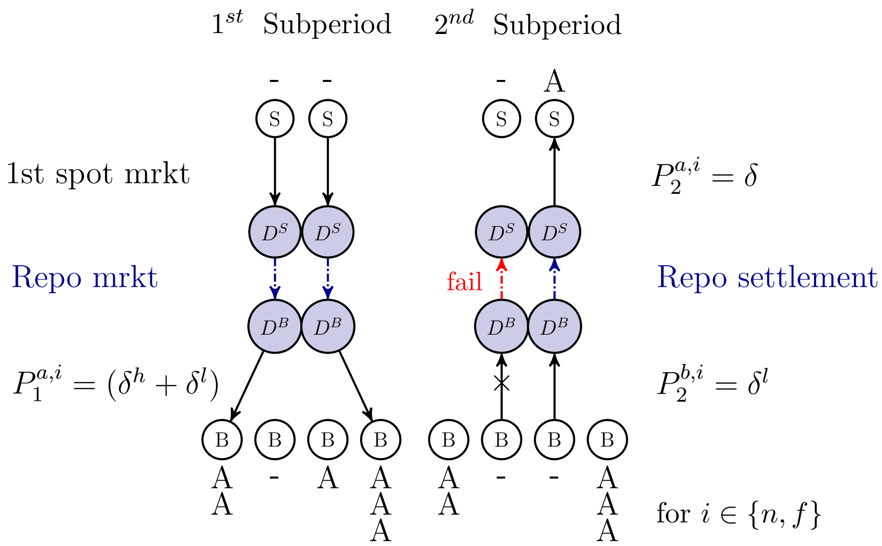
Assumption 1. $0 < \bar{r} < (\delta - \delta^h - \delta^l)/[2(\delta^h + \delta^l)]$.

In addition, traded quantities are restricted to the interval $[0, 1]$. It is straightforward to show that reversing the direction of trade is never optimal.

Lastly and importantly, I assume that in the second period, on- and off-the-run assets are bargained over separately, even when both are traded within the same trading counterparties. It does not make any difference as there is no benefit to bundling trades, as all costs and benefits in the model are linear. This goes hand in hand with the aim of determining the prices of individual assets rather than the price of a bundled portfolio.

In the second period, I concentrate on the two key optimization problems: the repo market problem and the first spot market problem at the beginning of the period. The latter is the source of the on-the-run premium. Figure 11 gives an overview. All other maximization problems are straightforward and serve primarily to close the model. It suffices to note that, whenever the fail probability of an asset is below the corresponding threshold, $\mathbb{P}^i \leq \mathbb{P}^{\text{max},i}$, one unit of the asset is traded in all markets. Figure 11 reports the corresponding equilibrium prices (except for the markets discussed below). They are identical for both assets. $P_1^{a,i} = (\delta^h + \delta^l)$ and $P_2^{b,i} = \delta^l$ can be explained by the fact that the buyer has no bargaining power, so prices reflect his valuation of the assets. $P_2^{a,i} = \delta$ because both D^S and S value the asset only for its final coupon, and the price therefore equals that coupon. I relegate the discussion of the corresponding maximization problems to Appendix A.3.1.

Figure 11: Overview markets



Notes: The figure provides an overview of the timing of markets within a period, illustrated here for the second period. If the fail probability of an asset is below the maximum fail probability, $\mathbb{P}^i \leq \mathbb{P}^{\max,i}$, one unit of the asset is traded in all markets, and the figure shows the corresponding prices for $i \in \{n, f\}$ (except for the markets described below).

3.2.1 Repo market in the second period

Although the first spot market precedes the repo market in the sequence of transactions, I analyze the repo market first and the first spot market afterwards because the model is solved by backward induction. I begin by focusing on the equilibrium in which dealers optimally establish short positions in all assets and all assets remain actively traded. This is the case whenever $\mathbb{P}^i \leq \mathbb{P}^{\max,i}$ for all assets $i \in \{n, f\}$. The threshold $\mathbb{P}^{\max,i}$ is derived from the dealer's short-selling decision in the second spot market. Since this threshold is the only result from the remaining optimization problems that is needed for the characterization of the repo rates and spot prices driving the on-the-run premium, the remaining optimization problems are as mentioned relegated to Appendix A.3.1. Let us define $M_R^i \equiv P_1^{a,i} A_R^i$, given our assumption that collateral assets are valued at the current spot market ask price during the repo transaction. For an arbitrary asset $i \in \{n, f\}$ satisfying $\mathbb{P}^i \leq \mathbb{P}^{\max,i}$, the bargaining problem between dealer D^S and dealer D^B in the repo market is

$$\max_{0 \leq A_R^i \leq 1, r^i} \left(S_R^{DS,i} + S_R^{DB,i}\right) \;\; s.t. \;\; S_R^{DS,i} = \theta\left(S_R^{DS,i} + S_R^{DB,i}\right)$$

with

$$
\begin{aligned}
S_R^{DS,i} &= (\bar{r} - r^i)M_R^i, \\
S_R^{DB,i} &= r^i M_R^i + \underbrace{[P_1^{a,i} - (1-\mathbb{P}^i)P_2^{b,i} - \mathbb{P}^i C^i]A_R^i}_{\text{expected short selling surplus in spot market}} - \delta A_R^i.
\end{aligned}
$$

The surplus of a dealer meeting a seller ($S_R^{DS,i}$) is straightforward: by lending out the asset in the repo market, he replaces his funding cost $\bar{r}$ with funding at the repo rate r^i. The surplus of a dealer meeting a buyer ($S_R^{DB,i}$) consists of the repo interest earned on the cash lent and the expected gains from establishing a short position. These gains equal the proceeds from selling the asset in the spot market at the current ask price $P_1^{a,i}$ minus the cost of covering the short position in the next subperiod. With probability $1-\mathbb{P}^i$, the dealer successfully buys back the asset at the prevailing spot market bid price $P_2^{b,i}$, and covers the short position. With probability $\mathbb{P}^i$, however, the dealer fails to obtain the asset and must compensate the cash borrower by paying the fail compensation[52] $C^i = P_2^{a,i} = \delta$. In addition, because D^S remains the legal owner of the asset throughout the repo term, D^B must compensate him for the accrued coupon payment δ, consistent with standard repo market practice.

The solution to the bargaining problem implies that dealer D^B borrows all assets i from D^S against the corresponding cash value M_R^i. The exchange of the entire asset position maximizes the gains from trade. The repo rate r^i redistributes these gains according to the bargaining power and outside options of the parties to the trade. The resulting equilibrium allocation is given by $A_R^i = 1$ (the surplus is positive given that $\mathbb{P}^i \leq \mathbb{P}^{\max,i}$), while the equilibrium repo rate is

$$
r^i = (1-\theta)\bar{r} - \theta \frac{P_1^{a,i} - (1-\mathbb{P}^i)P_2^{b,i} - \mathbb{P}^i C^i - \delta}{P_1^{a,i}}.
$$

We first analyzed the case in which short selling is profitable. We now consider the complementary case. For any $i \in \{n, f\}$ such that $\mathbb{P}^i > \mathbb{P}^{\max,i}$, short selling is not profitable (see Appendix A.3.1, second spot market maximization problem). In this case repo trade still occurs, but the repo rate equals the prefinancing rate, $\bar{r}$. The corresponding maximization

[52] By assumption, the compensation payment is equal to the prevailing spot market ask price of the asset at the time of the fail. In equilibrium, this fully compensates the non-failing dealer, D^S, for the loss resulting from its inability to sell the asset. Absent a settlement fail, the dealer would have sold the asset at the prevailing spot market ask price, $P_2^{a,i}$, which equals the value of holding the asset to maturity. Therefore, $C^i = P_2^{a,i} = \delta$. See Section 3.1 and Appendix A.3.1 for details. Full compensation is assumed for simplicity.

problem and its solution are relegated to Appendix A.3.1.

3.2.2 First spot market in the second period

Having characterized the equilibrium repo rate, we can now return to the first spot market of the second period. The repo rate affects the gains from trade and therefore the spot market equilibrium.

In the first spot market of the second period, the bargaining problem between dealer D^S and seller S for each asset $i \in \{n, f\}$ is

$$\max_{0 \leq A_1^{b,i} \leq 1, P_1^{b,i}} S_1^{DS,i} + S_1^{S,i} \text{ s.t. } S_1^{DS,i} = \theta^d (S_1^{DS,i} + S_1^{S,i})$$

with

$$\begin{aligned} S_1^{DS,i} &= -r^i M_R^i + (2\delta - P_1^{b,i}) A_1^{b,i}, \\ S_1^{S,i} &= (P_1^{b,i} - \delta^l - \delta) A_1^{b,i}. \end{aligned}$$

Concentrating first on dealer D^S, its surplus ($S_1^{DS,i}$) can be explained as follows: The dealer D^S buys the asset at the spot market bid price $P_1^{b,i}$ and receives the value of two coupons, equal to 2δ, from the asset.[53] Although the trade is prefinanced at rate $\bar{r}$, it will ultimately be financed at the asset specific repo rate r^i which is therefore the relevant rate to calculate the surplus. The surplus of the seller ($S_1^{S,i}$) is the price she receives for the asset, $P_1^{b,i}$, minus her current and future valuation of the asset, δ^l and δ.

The solution implies that the counterparties exchange the entire asset position, which maximizes the gains from trade, and bargain over the spot market price. The solution is $A_1^{b,i} = 1$ for the quantity[54] and $P_1^{b,i} = \delta + \theta^d \delta^l + (1 - \theta^d)(-r^i P_1^{a,i} + \delta)$ for the price.[55] We can see from the solution that different repo rates translate into differences in spot market prices, a mechanism that will become important for the emergence of the on-the-run premium.

[53] This holds even if he is going to resell it the next period as the reselling price equals the coupon. See Appendix A.3.1.

[54] Given $r^i \leq \bar{r}$ and Assumption 1, r^i is never too high to prohibit trade and the joint surplus is always weakly positive.

[55] We do not have to distinguish if $\mathbb{P}^i \leq \mathbb{P}^{max,i}$ holds for asset i or not. The problem and the solution are the same.

3.3 Model Intuition

Before presenting the theoretical results, I provide intuition for the model and the main equilibrium by focusing on the two-period asset. The corresponding intuition for the one-period asset follows immediately. I first describe the dynamics in the first period and then those in the second period.

3.3.1 First period

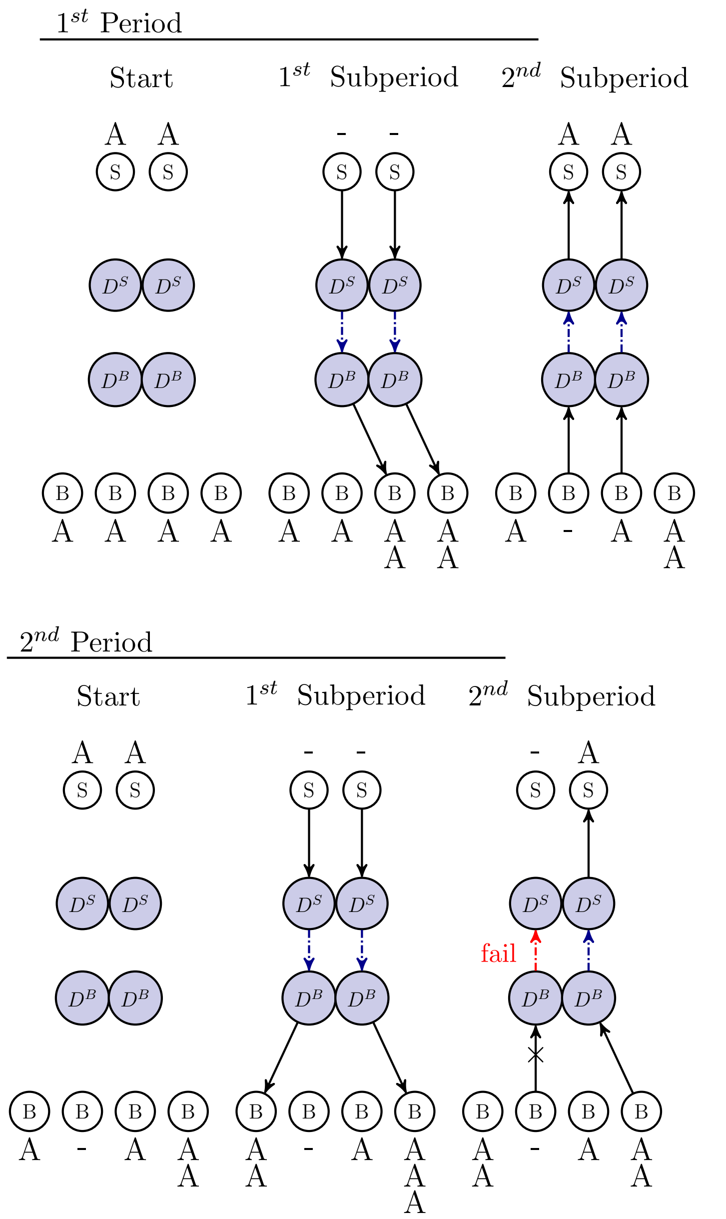
At the beginning of the first period, all buyers and sellers are endowed with one unit of the two-period asset. Figure 12 illustrates the sequence of spot and repo transactions in the model. In the first subperiod, a seller S trades with a matched dealer D^S in the spot market. Gains from trade arise because sellers temporarily have a lower valuation of the asset than dealers ($\delta^l < \delta$). In addition, dealers can use the asset as collateral in the repo market. Dealer D^S initially finances the purchase at rate $\bar{r}$, but subsequently repos out the asset to another dealer D^B, allowing him to replace this funding with repo financing. The repo transaction itself generates surplus because dealer D^B can subsequently sell the asset short to a buyer B in the spot market. Buyers are willing to purchase the asset because they currently have a high valuation, and because they can obtain cash at a zero interest rate. The repo rate is determined through bargaining between the two dealers and reflects the surplus generated by the short sale.

In the second subperiod, valuations reverse. Dealer D^B reacquires the asset from a matched buyer in the spot market and realizes a profit from the intertemporal price difference. The repo transaction is unwound as dealer D^B returns the collateral and receives back the cash plus interest. Dealer D^S then sells the asset to a seller S whose valuation has again increased.

3.3.2 Second period

The sequence of transactions in the second period is identical. The crucial difference is that the two-period asset has now become seasoned. Ownership of the asset is no longer evenly distributed across buyers. Although aggregate holdings remain unchanged, the distribution of asset holdings across market participants has become unequal. This inventory dispersion creates uncertainty about the location of the asset in the economy. As a result, a dealer that has sold the asset short may not always be able to reacquire it before settlement. These

Figure 12: Intuition



black = spot market, blue = repo financing

Notes: This figure provides intuition by focusing on the two-period asset. In the first period, buyers' inventories are identical. In the second period, buyers still hold the same aggregate quantity of the asset, but their individual inventories differ. These differences give rise to inventory uncertainty and can lead to settlement fails when the asset is sold short. In equilibrium, this uncertainty increases the repo rate and reduces the spot price of the asset.

settlement fails are non-strategic: they arise not because agents intentionally choose not to deliver, but because the asset cannot always be located through the matching process.

As long as the probability of a fail remains sufficiently small, short selling continues to be profitable in expectation. Nevertheless, fail risk reduces the value of the asset as repo collateral. Dealers accepting the collateral therefore require compensation in the form of a higher repo rate. Equivalently, cash borrowers must pay a higher borrowing cost when pledging the asset as collateral. The increase in repo financing costs is transmitted to the spot market. Anticipating higher financing costs, dealers are willing to pay less for the asset in spot transactions. Consequently, settlement risk lowers the spot price of the asset.

The example above focuses exclusively on the two-period asset. In the full model, a newly issued one-period asset also trades in the second period and serves as the on-the-run security. While both assets have identical remaining cash flows and maturity, only the seasoned off-the-run asset is subject to inventory uncertainty generated by past trading activity. Consequently, the off-the-run asset trades at a lower price than the newly issued asset, giving rise to an on-the-run premium. The next section characterizes the equilibrium and shows how this difference gives rise to distinct repo rates and prices across the two securities.

3.4 Equilibrium equations

This section presents the main equilibrium, characterizes the relationships among the key equilibrium objects, and establishes the foundational results of the model.

3.4.1 Definition

I define an equilibrium as follows.

Definition 2. *An equilibrium consists of spot market quantities and prices* $\{A_t^{a,i}, A_t^{b,i}, P_t^{a,i}, P_t^{b,i}\}_{i,t}$*, repo market quantities and repo rates* $\{A_R^i, r^i\}_i$*, and settlement fail probabilities* $\{\mathbb{P}^i\}_i$ *for assets* $i \in \{e, n, f\}$ *and subperiods* $t \in \{1, 2\}$ *such that*

1. *all agents solve their respective maximization problems; and*

2. *settlement fail probabilities are consistent with the equilibrium inventory distribution.*

I mainly focus on equilibria where there is short selling in all assets. I call those equilibria short selling equilibria.[56]

Definition 3. *A short selling equilibrium is an equilibrium where* $\mathbb{P}^i \leq \mathbb{P}^{max,i}$ *for all* i.

[56]In Appendix A.4, I additionally discuss an equilibrium in which only on-the-run assets are sold short.

3.4.2 Spot and repo market price relations

The bid and ask prices and the repo rates in the short selling equilibrium for the on- and off-the-run assets are summarized in the following lemma.

Lemma 1. *In the short selling equilibrium, the spot market prices for* $i = \{n, f\}$ *are*

$$\begin{aligned} P_2^{a,i} &= \delta, \\ P_2^{b,i} &= \delta^l, \\ P_1^{a,i} &= \delta^h + \delta^l, \\ P_1^{b,i} &= \delta + \theta^d \delta^l + (1 - \theta^d)(\delta - r^i P_1^{a,i}), \end{aligned}$$

and the repo rate for $i = \{n, f\}$ *is*

$$r^i = (1 - \theta)\bar{r} + \theta \left[\mathbb{P}^i(\delta - \delta^l) - (\delta^h - \delta)\right] \frac{1}{\delta^h + \delta^l}.$$

Proof. See Appendix A.2.1. □

All prices depend on the coupon δ the assets pay and their (in)convenience to buyers and sellers. In addition, the bid price in the first subperiod also depends on the bargaining power of the dealer θ^d and on the repo rate r^i. The lower the repo rate for asset i, the higher the bid price the dealer is willing to pay. This reflects that assets with higher collateral value command higher prices, as agents are willing to pay more for assets that imply lower repo rates when pledged.

The repo rate itself is lower when the asset can be easily sourced. This means that the higher the probability that a dealer who shorts the asset can locate it and avoid a settlement fail (i.e., the lower $\mathbb{P}^i$), the lower the repo rate. Also important to point out is that the repo rate depends positively on the ceiling rate $\bar{r}$ which is the rate of the outside option for cash funding.[57]

In summary, repo rates directly influence spot market prices, while the probability of locating the asset and avoiding settlement fail determines repo rates.

[57] The repo rate r^i is never higher than the ceiling rate $\bar{r}$ in the short selling equilibrium. This follows from $\mathbb{P}^i \leq \mathbb{P}^{max,i}$ for all i.

3.4.3 On-the-run premia, specialness and fail probabilities

In this paragraph I discuss how on-the-run premia, specialness, and liquidity are linked with each other and how they depend on settlement fail probabilities. For this I first define them.

Definition 4.

- *The average spot market price for any asset i is defined by $P^i \equiv \frac{1}{4}(P_1^{a,i} + P_1^{b,i} + P_2^{a,i} + P_2^{b,i})$.*
- *The average bid-ask spread for any asset i is defined by $ba^i \equiv \frac{1}{2}[(P_1^{a,i} - P_1^{b,i}) + (P_2^{a,i} - P_2^{b,i})]$.*
- *Specialness in the on-the-run asset is defined by $s \equiv r^f - r^n$.*
- *The on-the-run premium is defined by $\Delta \equiv P^n - P^f$.*

Given these definitions, it is straightforward to derive specialness, the on-the-run premium, and bid-ask spread differences in the short selling equilibrium. I derive the following proposition, which characterizes their equilibrium values and shows how they depend on differences in fail probabilities.

Proposition 1. *Iff $\mathbb{P}^f > \mathbb{P}^n$*

- *the repo market exhibits positive specialness in the on-the-run asset, $s = (\mathbb{P}^f - \mathbb{P}^n)\frac{\theta(\delta - \delta^l)}{\delta^h + \delta^l} > 0$,*
- *the spot market exhibits a positive on-the-run premium, $\Delta = \frac{1}{4}(1 - \theta^d)(\delta^h + \delta^l)\, s > 0$,*
- *the off-the-run asset has a higher bid-ask spread than the on-the-run asset, $(ba^f - ba^n) = \frac{1}{2}(1 - \theta^d)(\delta^h + \delta^l)s > 0$*

in the short selling equilibrium.

Proof. See Appendix A.2.2 □

According to Proposition 1, a lower fail probability implies that the on-the-run asset trades "special" in the short selling equilibrium. This is consistent with the data: on-the-run U.S. Treasury securities frequently trade "on special" in repo markets, with repo rates significantly below the general collateral rate (D'Amico and Pancost, 2022).

The resulting financing advantages are capitalized into higher spot prices. The previous subsection showed that lower repo rates translate directly into higher spot market prices. Proposition 1 further establishes that, in the short selling equilibrium, differences in repo rates–captured by the specialness of the on-the-run asset–translate into an on-the-run premium in the spot market. The close link between specialness and spot prices is also documented by D'Amico and Pancost (2022), who show that a special collateral risk premium explains around 80% of the on-the-run premium. Duffie (1996) likewise presents a model in which specialness is linked to spot market prices.

To provide further insight into the relationship among the model's endogenous variables, I report representative values from the data. According to D'Amico and Pancost (2022), average specialness across all coupon-bearing maturities, corresponding to s in the model, is 19 bp. In addition, Table 6 in Appendix A.8.1 reports summary statistics for the on-the-run premium and fail rates. The average on-the-run premium across all coupon-bearing maturities, corresponding to Δ, is 2.6 bp, although it exhibits substantial variation over time. Figure 1 shows that the premium can exceed 30 bp during periods of market stress, such as the Covid crisis. The average values of $\mathbb{P}^f$ and $\mathbb{P}^n$ are 67 bp and 41 bp, respectively. Finally, Chaboud et al. (2025b) report average bid-ask spread differences between on- and off-the-run Treasuries corresponding to $(ba^f - ba^n)$ in the model. For example, for the 10 year sector, they report an average bid-ask spread of about 2 bp for the on-the-run Treasury and about 7 bp for the fifth off-the-run generation, implying a difference of 5 bp. Given that the model predicts a bid-ask spread difference equal to twice the size of the on-the-run premium, this number is in line with the model. The difference is similar for the 2 year and 5 year sectors, while bid-ask spread differences widen substantially at the long end. Moreover, the differences can increase markedly during periods of market stress. For example, in the 10 year sector, the bid-ask spread difference rose to roughly 26 bp in March 2020.

3.4.4 Settlement fails

In this paragraph I link the probability to fail with the actual fails. I define the mass (and therefore also the share) of buyers B who have a lower inventory of asset i at the beginning of the second subperiod than the amount of asset i sold short (i.e. repoed in and sold) by each dealer D^B in the preceding subperiod by s_0^i. I refer to these inventories as fail inventories in asset i. By definition, a fail in asset i occurs if the dealer D^B who sold short asset i in the first subperiod meets, in the second subperiod, a buyer who has a fail inventory in asset i.

This implies that $\mathbb{P}^i = s_0^i$. I define the number of fails in asset i by $F^i \geq 0$ and derive the following lemma:

Lemma 2. $1 > s_0^i > 0$ *if* $F^i > 0$ *and* $F^i > 0$ *if* $\mathbb{P}^{max,i} \geq s_0^i > 0$ *for all* i.

Proof. See Appendix A.2.3 □

The first part of the lemma says that for fails in asset i to occur, there must be a positive mass of fail inventories in asset i and a non-degenerate distribution of inventories over asset i. The second part of the lemma states that if failing in asset i is not too probable ex ante, but the probability is also not zero, then fails occur.

It is important to highlight that a non-degenerate distribution of inventories over asset i is necessary for fails in asset i, yet no notion of scarcity or similar has been introduced up to this point. I discuss scarcity in Section 3.5.2. Next I derive the fail rates.

Corollary 1. *In the short selling equilibrium* $\mathbb{P}^f = [\sigma(1-\sigma)]^2 > \mathbb{P}^e = \mathbb{P}^n = 0$.

Proof. See Appendix A.2.4 □

Based on this corollary I prove the following:

Corollary 2. *In the short selling equilibrium* $\Delta > 0$ *iff* $F^f > F^n$. *This condition is satisfied in the short selling equilibrium.*

Proof. See Appendix A.2.5 □

A positive on-the-run premium arises in the short selling equilibrium only if there are more fails in off-the-run assets than in on-the-run assets, which is indeed the case.[58] In the short selling equilibrium, $F^i = \mathbb{P}^i$ for all i, since the condition $\mathbb{P}^i \leq \mathbb{P}^{\max,i}$ holds for all i, implying that dealers sell every asset i short. Consequently, there are more fails in off-the-run assets than in on-the-run assets, i.e., $F^f > F^n$. This is consistent with the empirical evidence presented in Section 2.2.

3.4.5 Existence and uniqueness

For completeness, I finally establish the existence and uniqueness of the short selling equilibrium when $\mathbb{P}^i \leq \mathbb{P}^{\max,i}$ for all $i \in \{e, f, n\}$ by building on the preceding results.

[58]For simplicity, there are no settlement fails in on-the-run assets in the model. As most results are framed in terms of their dependence on $\mathbb{P}^n$, it is straightforward to conjecture the effect of on-the-run fails. Their impact would arise through an increase in $\mathbb{P}^n$, affecting the results through this channel.

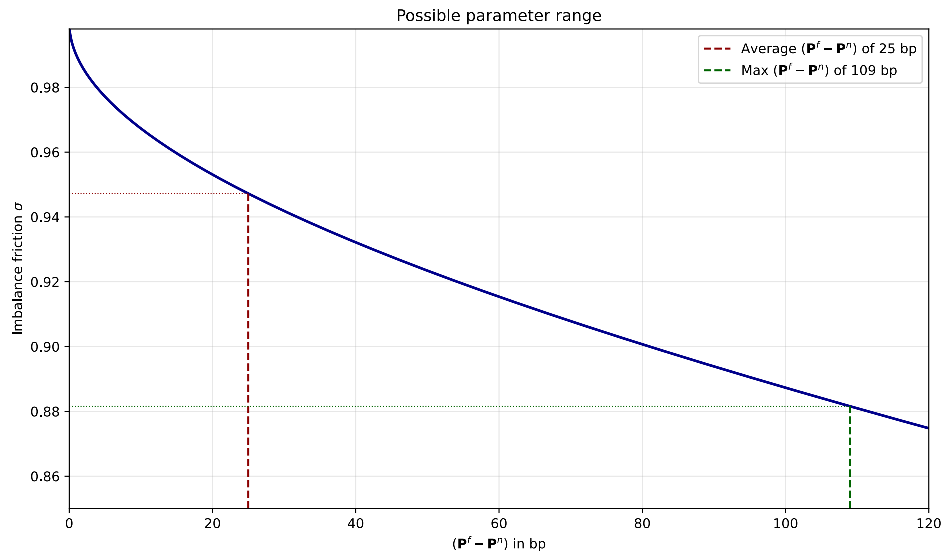
Lemma 3. *If $\mathbb{P}^i \leq \mathbb{P}^{\max,i}$ for all $i \in \{e, f, n\}$, then a short selling equilibrium exists and is the unique equilibrium.*

Proof. See Appendix A.2.8 □

3.4.6 Calibration

Corollary 1 implies that $(\mathbb{P}^f - \mathbb{P}^n) = [\sigma(1-\sigma)]^2$. This relationship allows the underlying friction σ to be inferred from observed values of $(\mathbb{P}^f - \mathbb{P}^n)$. Using the same data as in Figure 5b over the period from April 2013, when the on-the-run series begins, to the end of 2025, the average value of $(\mathbb{P}^f - \mathbb{P}^n)$ is 25 bp. This corresponds to a value of σ of approximately 0.95. This is a plausible value, as one would expect only small imbalances between buyers and sellers in the market. The exercise nevertheless demonstrates that even modest imbalances can generate the settlement fail differentials observed in the data. The maximum fail rate differential observed in the sample is 109 bp, which implies a value of σ of approximately 0.88. This corresponds to a market imbalance of slightly more than 10%.[59]

Figure 13: Calibration



Notes: The figure shows the values of $(\mathbb{P}^f - \mathbb{P}^n)$ over a range of values of σ. The average observed value of $(\mathbb{P}^f - \mathbb{P}^n)$ in the data is 25 bp, which corresponds to $\sigma \approx 0.95$. The maximum observed value is 109 bp, corresponding to $\sigma \approx 0.88$.

In addition to this calibration exercise, Appendix A.7 uses simulated endowment and

[59] There is also a range of low values of σ (below 0.15) that can generate the observed fail rate differentials. These values imply very large imbalances between buyers and sellers and are therefore considered implausible and excluded from the calibration.

trade allocations to show how different auction allocations affect settlement fails and can rationalize the delayed transition to benchmark ("on-the-run") status employed by some smaller sovereign issuers.

3.5 Distribution and relative supply

This section examines the role of relative asset supply in determining equilibrium outcomes. The motivation is the frequently proposed view that the relative lower supply, i.e. scarcity of off-the-run securities is an important driver of the on-the-run premium.

3.5.1 Distribution

First, I show that a more unequal distribution of the off-the-run asset relative to the on-the-run asset is necessary and sufficient for a positive on-the-run premium. The concept of a more unequal distribution is defined as follows.

Definition 5. *An asset l is more unequally distributed than an asset k if the buyers' inventory distribution of asset l is non-degenerate at the beginning of the first subperiod, whereas that of asset k is degenerate by assumption.*

Using this definition I derive the following proposition.

Proposition 2. *In the short selling equilibrium $\Delta > 0$ iff the off-the-run asset is more unequally distributed than the on-the-run asset. This condition is satisfied in the short selling equilibrium.*

Proof. See Appendix A.2.6. □

Proposition 2 shows that a more unequal distribution in off-the-run assets relative to on-the-run assets is necessary and sufficient for a positive on-the-run premium.

3.5.2 Scarcity of off-the-run assets

Scarcity is often proposed as a necessary condition for a unique equilibrium with an on-the-run premium. In Vayanos and Weill (2008), an established model to explain on-the-run premia, there are two equilibria, with the premium arising on either of the two otherwise identical assets through coordination. Vayanos and Weill (2008) suggest that introducing scarcity, defined as a lower supply of one asset relative to the other, may restore equilibrium uniqueness.[60] However, as they point out, their stationary model lacks the dynamics required to account for the decrease in effective supply over time. They note that their "theory can explain price differences between on- and off-the-run bonds, but it does not explain why short sellers are more likely to concentrate in on-the-run bonds." They leave this question open for future research.

I close this gap with the dynamic model and show that scarcity in the sense of Vayanos and Weill (2008) is *not* necessary for a unique equilibrium with an on-the-run premium.

Definition 6. *I define the average inventory of asset $i \in \{n, f\}$ by sellers (buyers) at the beginning of the first subperiod in the second period by $\mathcal{A}^{S,i}$ ($\mathcal{A}^{B,i}$). Asset $i \in \{n, f\}$ is defined to be relatively scarce iff $\mathcal{A}^{S,i} \leq \mathcal{A}^{S,-i}$ and $\mathcal{A}^{B,i} \leq \mathcal{A}^{B,-i}$ while the condition must hold strictly in at least one of them.*

Proposition 3. *For a given parametrization, the equilibrium is unique. In this unique equilibrium the on-the-run premium is positive and the off-the-run asset is not relatively scarce.*

Proof. See Appendix *A*.2.9 □

Proposition 3 goes further than showing that, in the unique short selling equilibrium, there is a premium and no relative scarcity; it shows that this holds in *any* equilibrium and the equilibrium is unique for a given parametrization.[61]

Drawing a more general conclusion from our analysis so far, one can say that it does not matter how scarce assets are, as long as they can be reliably located and sufficient supply is available at that source. I discuss in the policy section what happens if the latter is not the case and uncertainty becomes sufficiently high that markets break down. Moreover, while the analysis here has focused on showing that scarcity is not necessary for an on-the-run premium to emerge, Appendix A.5 discusses why scarcity is also not sufficient to generate one.

[60]Within their framework, they show that if off-the-run bonds are in lower effective supply, they are less likely to attract short sellers and are therefore less liquid. They point out that this is "a first step toward reconciling our theory based on multiple equilibria with the empirical fact that liquidity in the U.S. Treasury market concentrates systematically in just-issued bonds."

[61]In Section A.4, I show that the only other equilibrium that can exist is one in which $\mathbb{P}^i \leq \mathbb{P}^{\max,i}$ for $i \in \{e, n\}$ and $\mathbb{P}^f > \mathbb{P}^{\max,f}$. In this equilibrium only on-the-run assets are sold short.

3.5.3 Extension: buy-and-hold buyers and scarcity

I have shown that scarcity is not necessary for an equilibrium with an on-the-run premium. Still, scarcity might scale up its size. I analyze this conjecture in this paragraph. Also, I address the discussion around buy-and-hold buyers in seasoned Treasuries. As pointed out by Chaboud et al. (2025b) off-the-run Treasuries are increasingly held by buy-and-hold investors as they get more seasoned.

To address both points, I assume that a random share $0 < (1 - \sigma_H) < (1 - \sigma)$ of buyers are not going to sell any off-the-run assets. We call them buy-and-hold buyers in off-the-run assets. Given their assets are not available for sale, we do not count their assets as part of their inventory and they have therefore no inventory in off-the-run assets. Therefore buy-and-hold buyers reduce the effective supply of an asset, here the two-period asset, over time. We look at two cases: Either buy-and-hold buyers are participating in the spot market in the second period (i.e. they are matched and can sell on-the-run assets) or they are not (i.e. they are not matched and do not affect the matching probability for the remaining buyers).

Proposition 4. *With buy-and-hold buyers, the off-the-run asset is relatively scarce in the short selling equilibrium. If the buy-and-hold buyers are participating in the spot market, the on-the-run premium increases in scarcity. Otherwise, it is independent of scarcity.*

Proof. See Appendix A.2.10. □

This exercise reveals how the premium can scale with scarcity and highlights that buy-and-hold buyers may or may not affect the premium, depending on their market participation. It is important to point out that, as long as buy-and-hold buyers stay outside the market and therefore do not increase uncertainty in off-the-run assets (i.e., their fail probability), the on-the-run premium is not affected, even though they diminish the effective supply of off-the-run assets.

3.6 Welfare

This section discusses welfare and examines how settlement fails and the underlying market frictions affect it. I define welfare as the mass weighted sum of the lifetime value of all agents

$$W = V_B + \sigma(V_S + V_{DB} + V_{DS}),$$

where the lifetime value of an agent is the sum of utility and disutility an agent has over its lifetime.

The value functions of the agents in the short selling equilibrium are

$$\begin{aligned}
V_B &= 3(\delta^l + \delta^h) + \sum_i [S_1^{B,i} + \sigma(1 - \mathbb{P}^i)S_2^{B,i}], \\
V_S &= 3(\delta^l + \delta) + \sum_i [S_1^{S,i} + (1 - \mathbb{P}^i)S_2^{S,i}], \\
V_{DB} &= \sum_i [S_R^{DB,i} + S_1^{DB,i} + (1 - \mathbb{P}^i)S_2^{DB,i}], \\
V_{DS} &= \sum_i [S_1^{DS,i} + S_R^{DS,i} + (1 - \mathbb{P}^i)S_2^{DS,i}].
\end{aligned}$$

The value functions depend on the coupon payments and the (in)convenience yield from holding endowed assets, as well as on the expected surplus from trade weighted by its probability of occurrence.

Proposition 5. *In the short selling equilibrium,*

$$\frac{\partial W}{\partial \sigma} = \frac{\partial W}{\partial \mathbb{P}^f}\frac{\partial \mathbb{P}^f}{\partial \sigma} + \sum_k V^k,$$

where $k \in \{S, DS, DB\}$. Since

$$\frac{\partial W}{\partial \mathbb{P}^f} < 0$$

and

$$\frac{\partial \mathbb{P}^f}{\partial \sigma} \begin{cases} > 0, & \text{if } \sigma < \frac{1}{2}, \\ = 0, & \text{if } \sigma = \frac{1}{2}, \\ < 0, & \text{if } \sigma > \frac{1}{2}. \end{cases}$$

it follows that

$$\frac{\partial W}{\partial \sigma} > 0$$

for all $\sigma \geq \frac{1}{2}$. For $\sigma < \frac{1}{2}$, the sign of $\partial W/\partial \sigma$ is generally ambiguous.

Proof. See Appendix A.2.7. □

A fail arising from a buyer with a fail inventory requires the sequence no match (no buy) – match (sell) – no match (no buy) – match (fail) for the buyer. For high σ ($\sigma > 0.5$), an increase in σ raises the probability of buying more than it reduces the probability of selling, thereby reducing the likelihood of this sequence and thus the fail probability. For low σ ($\sigma < 0.5$), the opposite holds.

Welfare decreases in the probability of fail in off-the-run assets because each fail with a buyer implies that the asset remains in the inventory of another buyer rather than being reallocated to a seller who currently derives higher utility from holding it.

Consequently, welfare is increasing in σ for $\sigma \geq \frac{1}{2}$, whereas for $\sigma < \frac{1}{2}$ the welfare effect of a change in σ is generally ambiguous.

3.7 Policy: Restructuring the Treasury Market

The Treasury market crisis in March 2020 revealed that one of the largest and most important markets of the world is fragile (Duffie, 2020). Various ideas for how to restructure the market are discussed. The G30 Working Group on Treasury Market Liquidity discusses the creation of a standing repo facility that provides broad access to repo financing for Treasuries. They also suggest central clearing of certain Treasury trades, and adjustments to regulation and data transparency (see Duffie et al., 2021). The Federal Reserve indeed implemented around the same time as the publication of their proposals a standing repo facility but with less broad access as proposed.[62] Chaboud et al. (2025b) further discuss buyback programs, as currently undertaken by the Treasury, larger concentration of Treasury issuance in fewer securities, and all-to-all trading as policy measures. I use the model to analyze three of these policies: repo facility access, issuance concentration, and all-to-all trading. In addition, I discuss tokenisation, another proposed reform of financial market infrastructure (see e.g. Lee et al., 2024).

3.7.1 Crisis and repo facility access

During the Covid crisis the off-the-run segment broke down due to a dash for cash (Chaboud et al., 2025b). I model the crisis through the lens of my model as follows: Suppose that, in a crisis occurring in the second period, a mass of buyers is no longer willing to trade. This fits the narrative of the crisis in 2020 where clients wanted to sell their Treasuries but not buy them as everyone wanted to hold cash (Chaboud et al., 2025b). More specifically, I assume that there is a random mass of buyers, denoted by γ_c, who remain in the spot market but are unwilling to trade when matched with a dealer in the second period. I call

[62] See the Federal Reserve's website https://www.federalreserve.gov/monetarypolicy/standing-overnight-repurchase-agreements.htm for more info.

them non-trade willing buyers. The crisis, represented by the emergence of these non-trade willing buyers, is not anticipated in the first period. For simplicity, I assume that, before entering into a repo transaction in the first subperiod in the second period, dealers observe whether the buyer with whom they are matched in this subperiod is willing to trade.

In addition I assume that in the crisis, i.e. in the second period, the repo ceiling rate might be different and I call it $\bar{r}_c$. Especially I assume that an analogous assumption to Assumption 1 for $\bar{r}$ must not hold for $\bar{r}_c$. The implications of this are discussed below.

Also, I define $\mathbb{P}_c^{max} \equiv \frac{\delta^h - \delta + \bar{r}_c(\delta^h + \delta^l)}{\delta - \delta^l}$. I call the analogous failing probabilities to $\mathbb{P}^i$ in the crisis set-up $\mathbb{P}_c^i$. It is straightforward to show that the crisis failing probabilities in on- and off-the-run assets are $\mathbb{P}_c^n = \gamma_c$ and $\mathbb{P}_c^f = \sigma^2(1-\sigma)^2 + \gamma_c[1 - \sigma^2(1-\sigma)^2]$. The emergence of non-trade-willing buyers increases market uncertainty because dealers no longer know whether a matched buyer will be willing to trade. This additional uncertainty is layered on top of the existing distributional uncertainty, which remains unchanged.

Lemma 4. *If the mass of non-trade willing buyers in the crisis, γ_c, is as such that*

$$\mathbb{P}_c^{\max} \geq \gamma_c > \frac{\mathbb{P}_c^{\max} - \sigma^2(1-\sigma)^2}{1 - \sigma^2(1-\sigma)^2}$$

then $\mathbb{P}_c^f > \mathbb{P}_c^{max} \geq \mathbb{P}_c^n$. This implies that no off-the-run assets are sold by dealers to buyers. The corresponding trade in on-the-run assets still occurs if these assets can be financed in the repo market.

Proof. See Appendix A.2.11 □

Dealers do not sell off-the-run assets anymore to buyers because there are too few trade willing buyers and short selling becomes too risky. Uncertainty is too high. As observed in the crisis, the market breaks down first for off-the-run assets because $\mathbb{P}_c^f > \mathbb{P}_c^n$ for any γ_c.

If the repo ceiling rate $\bar{r}_c$ is low enough (which is the case if $\bar{r}_c = \bar{r}$), then sellers can still sell their off-the-run assets to the dealers in this situation. Spot trades in off-the-run assets are repo financed at the ceiling rate $\bar{r}_c$ in that case. It can very well be though that during a crisis repo financing becomes expensive and the ceiling rate rises above the threshold at which no spot trades in off-the-run assets are financed anymore because the surplus from trade is negative. At the same time it is possible that trade in on-the-run assets is still ongoing as their trade is financed at the lower repo rate $r^n \leq \bar{r}_c$.

Corollary 3. *If $\mathbb{P}_c^f > \mathbb{P}_c^{max} \geq \mathbb{P}_c^n$ and the repo ceiling rate in the crisis $\bar{r}_c$ is such that*

$$\frac{(\delta - \delta^l) - \theta[\gamma_c(\delta - \delta^l) + (\delta^h - \delta)]}{(\delta^h + \delta^l)(1 - \theta)} \geq \bar{r}_c > \frac{\delta - \delta^l}{\delta^h + \delta^l},$$

then dealers purchase on-the-run assets from sellers and finance these positions in the repo market, whereas no such trading occurs for off-the-run assets.

Proof. See Appendix A.2.12 □

Lemma 4 (focusing on the buyer-dealer side) and Corollary 3 (focusing on the seller-dealer side) show how trade in crises may break down in the off-the-run segment while remaining functional in the on-the-run segment as observed in March 2020. One discussed policy to ease strains in a crisis is (broad) repo facility access. Let us assume there is a central bank which can influence the ceiling rate $\bar{r}_c$ by giving access to a repo facility during the crisis.

Lemma 5. *By providing access to a repo facility, the central bank can always restore repo-financed trade in off-the-run assets such that sellers can sell off-the-run assets to dealers while preserving trade in on-the-run assets across all markets. To achieve this outcome, the facility rate must be sufficiently low to restore trade in off-the-run assets, but not so low as to disrupt trade in on-the-run assets. Specifically,*

$$\min\left\{\frac{\delta - \delta^l}{\delta^h + \delta^l}, \frac{(\delta - \delta^l) - \theta\big[\gamma_c(\delta - \delta^l) - (\delta^h - \delta)\big]}{(\delta^h + \delta^l)(1 - \theta)}\right\} \geq \bar{r}_c \geq \frac{\gamma_c(\delta - \delta^l) - (\delta^h - \delta)}{\delta^h + \delta^l}.$$

However, the provision of the repo facility cannot induce the sale of off-the-run assets from dealers to buyers.

Proof. See Appendix A.2.13 □

The central bank can keep the sellers' spot market alive by providing the necessary cash cheap enough if needed. Repo rates are pressed down and spot prices increase for all assets, not just the off-the-run assets.

However, lowering repo rates can be harmful for the rest of the market. Cheap funding provided through the facility can impair the dealer-to-buyer segment of the market by making alternative funding obtained through short selling relatively less attractive. As a result, this part of the market breaks down earlier, or remains impaired if it is already inactive, because the maximum settlement-fail probability consistent with trade becomes lower.

Since dealers can obtain funding cheaply through the facility, it becomes less worthwhile to bear the risk associated with raising cash through short sales. The facility rate therefore cannot be set arbitrarily low without harming trade in on-the-run assets. Moreover, trade in off-the-run assets between dealers and buyers cannot be restored. Dealers continue to hold these assets on their balance sheets rather than passing them on to buyers. Although balance-sheet costs are not part of the model, this limitation may be particularly relevant during crises, when elevated balance-sheet constraints could accumulate over time and potentially weaken the restorative effects of the facility.

3.7.2 Larger issuance sizes

Another discussed policy for improving off-the-run liquidity is to concentrate its issuance into fewer and larger-sized issues (Chaboud et al., 2025b). Fleming (2002) provides evidence for this which Chaboud et al. (2025b) cite as follows: "*Fleming (2002), in fact, finds little relationship between issue size and liquidity when U.S. Treasury bills are on-the-run, but significant liquidity improvements when bills are off-the-run. That is, even if larger issue sizes did not improve liquidity of already highly liquid on-the-run Treasuries, additional size could benefit liquidity when securities go off-the-run.*" Through the lens of the model, this observation can be explained: In the model, larger endowment only affects off-the-run liquidity. Larger issuance sizes decrease fails in off-the-run assets while not affecting on-the-run fails as they are already seldom (respectively inexistent). We can demonstrate this with the following example: Assume that the endowment in each type of asset is not one, but two. It is straightforward to show that in this case $\mathbb{P}^n = \mathbb{P}^f = 0$. In this stylized example, the off-the-run asset becomes as liquid as the on-the-run asset, i.e. $ba^n = ba^f$ because $\mathbb{P}^f$ decreases to the size of $\mathbb{P}^n$. Liquidity in the on-the-run asset is not affected as $\mathbb{P}^n$ does not change. I showed in Proposition 5 that a decrease in fails in off-the-run assets is welfare improving. Because $\mathbb{P}^f = F^f$ in the short selling equilibrium, we have indeed a decrease in fails in this example. Therefore this analysis supports the view that larger issuance sizes are liquidity improving in off-the-run, but not on-the-run assets, and are overall welfare improving.

3.7.3 All-to-all trading and tokenisation

Recent policy discussions have highlighted both all-to-all trading and tokenisation as potential reforms of financial market infrastructure. All-to-all trading expands trading opportunities beyond traditional intermediary-based markets and may reduce reliance on dealers, particularly for off-the-run assets (Chaboud et al., 2025a). Tokenisation, by contrast, represents traditional assets as digital tokens on a programmable platform that enables ownership transfers and settlement through a common ledger and programmable rules (Bank for International Settlements and Committee on Payments and Market Infrastructures, 2024; Aldasoro et al., 2023). While these reforms differ substantially in implementation, both have important implications for settlement risk. I focus on qualitative insights rather than a formal analysis.

All-to-all trading has the potential to significantly reduce non-strategic settlement fails by largely eliminating inventory uncertainty and, according to our analysis, thereby increase welfare. Because the availability of securities is observable, short sellers can generally locate the asset when required to close their positions.

Tokenisation can similarly reduce settlement risk, but through a different mechanism. In particular, tokenised systems enable the use of smart contracts and atomic settlement. Smart contracts specify the conditions under which transactions are executed, for example by requiring that both counterparties are able to fulfill their contractual obligations. Under atomic settlement, securities and payments are exchanged simultaneously, such that a transaction is executed only if both parties can deliver their respective legs of the trade (Bank for International Settlements and Committee on Payments and Market Infrastructures, 2024). For repo transactions, this feature can eliminate settlement fails if the smart contract requires that both the initial exchange and the reverse transaction satisfy the conditions for atomic settlement. Tokenisation can, compared to all-to-all trading, eliminate both non-strategic and strategic settlement fails by requiring settlement to be guaranteed before a transaction is executed. The cost is that market participants can no longer take positions that require obtaining the security only after the trade has been agreed upon, thereby limiting short selling and other forms of risk taking, even though such activities may be welfare-enhancing.

These considerations capture only one aspect of the broader debate, abstracting from many other benefits and costs associated with these alternative market structures.

4 Empirics

Based on the theoretical results, I derive two hypotheses that I test empirically and complement with evidence from a policy intervention to further assess the impact of settlement risk on Treasury pricing.

4.1 Hypothesis development

In my theory inventory uncertainty, leading to positive fail probabilities for short sellers attempting to repurchase the asset and close their positions, and the resulting settlement fails, are central to explaining the on-the-run premium. Higher fail risk and more frequent fails for off-the-run securities depress their value, thereby generating a premium for on-the-run securities.

As a first motivating evidence I show the scatter plot between the 10-year on-the-run premium and the net outright positions of primary dealers in Treasuries in Figure 14.[63] I observe that there is a negative correlation between the on-the-run premium and the net outright positions of the primary dealers (with tight 95% confidence bands). If we assume that lower overall inventory is positively correlated with higher inventory uncertainty (which is a result following from Section 3.7.2) then this aligns with my explanation in the theoretic part that inventory uncertainty leads to positive fail probabilities and settlement fails, which imply the premium.

4.1.1 Fail rates and on-the-run premia

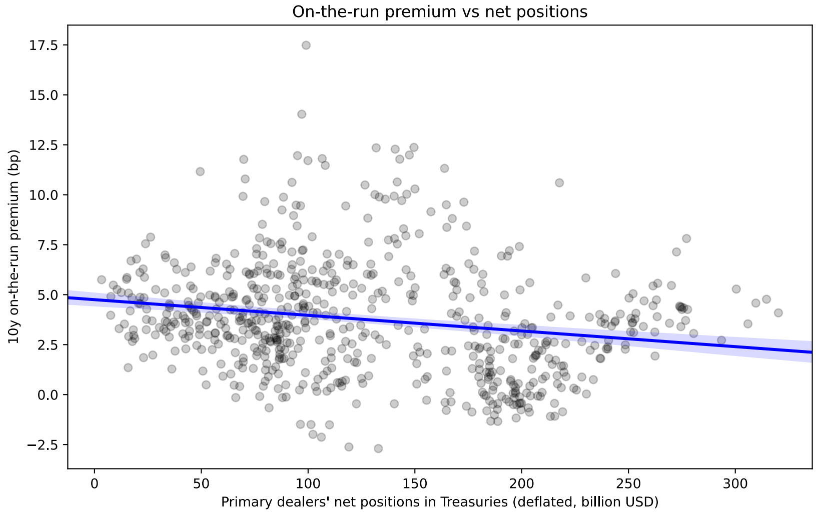
The first hypothesis is about a core part of my explanation: the relation of fail probabilities and fails with the on-the-run premium. I conjecture a joint sign restriction.

Hypothesis 1. *A higher (realized) fail rate of off-the-run Treasuries is associated with an increase in the on-the-run premium while a higher (realized) fail rate of on-the-run Treasuries is associated with a decrease in the on-the-run premium.*

The first hypothesis follows directly from Proposition 1 and Corollary 2. Proposition 1 shows that if the ex ante probability to fail in off-the-run assets is higher than in on-the-run

[63]The data sources for the on-the-run premium are the same as in Figure 2a. The net positions of primary dealers can be downloaded from the FED's Primary Dealer Statistics, https://www.newyorkfed.org/markets/counterparties/primary-dealers-statistics. The frequency is weekly. The data are deflated using the Consumer Price Index for All Urban Consumers: All Items in U.S. City Average, which can be downloaded from FRED. I set the index to 1 when the time horizon starts in April 2013.

Figure 14: The correlation between the premium and the net positions



Notes: The figure plots the on-the-run premium on 10-year Treasury notes against primary dealers' net positions in Treasuries using data from April 2013 to 2025. Source: FED and author's calculations.

assets we observe an on-the-run premium. Also, in the short selling equilibrium the ex ante probabilities equal the realized (share) of fails $\mathbb{P}^i = F^i$ and therefore there is a positive on-the-run premium if the observed fail share is higher in off-the-run assets than in on-the-run assets (see Corollary 2). A direct implication of Proposition 1 and Corollary 2 is that if the ex ante fail probabilities (and realized fail share) in off-the-run assets increases, then the premium increases. The opposite holds for on-the-run assets. An increase decreases the on-the-run premium. This directly gives rise to the hypothesis.

4.1.2 Inventories

The second part of the hypothesis development focuses on inventories and inventory uncertainty. Although the theoretical analysis focuses on inventory distribution, no data are available to observe individual inventories and therefore measure the distribution directly. Consequently, the empirical analysis must rely on aggregate inventories as a proxy. I conjecture that lower inventories lead to settlement fails. As outlined at the beginning of this section, lower inventories are expected to capture and correlate with inventory uncertainty. This conjecture is motivated by the issuance-size exercise in Section 3.7.2, which shows that larger aggregate inventories reduce settlement fail rates. Intuitively, larger inventories make it less likely that market participants completely run out of the asset and are therefore unable to satisfy delivery obligations. Conversely, lower inventories increase the likelihood that a market participant does not hold the asset when needed, thereby increasing inventory uncertainty and settlement fail risk. Since individual inventories are not observed in the data, aggregate inventories provide a natural proxy for the inventory uncertainty emphasized by the model.

Hypothesis 2. *Lower inventories in Treasuries are associated with higher fail rates. In particular, lower inventories in on-the-run Treasuries are associated with higher fail rates in on-the-run Treasuries, and lower inventories in off-the-run Treasuries are associated with higher fail rates in off-the-run Treasuries.*

In the following subsections I test and analyze the hypotheses in the data. An overview of the data sources and a summary statistic of the data used in the regressions can be found in Appendix A.8.1. In addition, Appendix A.8.2 presents further empirical evidence consistent with the theory by documenting that off-the-run Treasury inventories exhibit greater volatility than on-the-run Treasury inventories.

4.2 Empirical evidence

4.2.1 Empirical evidence for Hypothesis 1

I test the first hypothesis by examining the relationship between fail rates and the on-the-run premium. As the dependent variable in the first regression (see Table 1), I use the average on-the-run premium for U.S. Treasuries on a given day. To test the joint sign restriction implied by the model, I include the realized fail rates of both on- and off-the-run Treasuries as explanatory variables in the same regression. The primary dealer data from which the fail rates are estimated are available every Thursday, and I therefore use for my regression a dataset with one observation per week. Each date is denoted by t.[64]

I run two regressions (see Table 1). In the first, I regress the premium on the fail rates. In a second specification, I add standard control variables, including the logarithms of the VIX and MOVE indices, the effective federal funds rate, and the 10-year minus 2-year Treasury

[64] I cannot run any maturity specific regression because the fail rate for off-the-run Treasuries is not available by maturity. Also, Treasury bill data are used solely to construct the realized off-the-run fail rate, as Treasury bills are included in the aggregate measure of off-the-run fails. They are not otherwise included in the regressions. This restriction is imposed because corresponding bill-level data are unavailable for on-the-run Treasury securities.

yield spread. I further include crisis fixed effects. The crisis fixed effects capture the months October 2014, September 2019 and March 2020. In these months there was a crisis in the Treasury market (see U.S. Department of the Treasury et al., 2015, Anbil et al., 2020 and Schrimpf et al., 2020). As I use weekly data, the crisis dummies are equal to one in each week within these months. I use the time horizon from April 2013[65] to the end of 2025.

The regression equation to the second regression in Table 1 is given by

$$\begin{aligned}\text{On-the-run premium}_t = \alpha &+ \beta_1 \text{ Off-the-run fail rate}_t + \beta_2 \text{ On-the-run fail rate}_t \\ &+ \beta_3 \, ln(\text{MOVE}_t) + \beta_4 \, ln(\text{VIX}_t) + \beta_5 \text{ Effective federal funds rate}_t \\ &+ \beta_6 \text{ 10Y–2Y spread}_t + \beta_7 \, \mathbb{I}_{10/14,t} + \beta_8 \, \mathbb{I}_{09/19,t} + \beta_9 \, \mathbb{I}_{03/20,t} + u_t.\end{aligned}$$

The results are shown in Table 1. In line with Hypothesis 1, the coefficient on the off-the-run fail rate is positive, whereas the coefficient on the on-the-run fail rate is negative when control variables are included.[66] Both coefficients are highly statistically significant. Economically, the effects are sizeable. A one percentage point increase in the off-the-run fail rate is associated with a 1.7 basis point increase in the premium, while a one percentage point increase in the on-the-run fail rate is associated with a 0.6 basis point decrease. In standardized terms, a one standard deviation increase in the off-the-run (on-the-run) fail rate corresponds to a 0.30 standard deviation increase (0.13 standard deviation decrease) in the premium.

4.2.2 Empirical evidence for Hypothesis 2

Next, I test the second hypothesis by examining the relationship between fail rates and dealer inventories. I estimate two regressions. First, I regress the off-the-run fail rate on primary dealers' net outright position in off-the-run Treasuries. Second, I repeat the analysis using the on-the-run fail rate and primary dealers' net outright position in on-the-run Treasuries (see Table 2). Each specification is estimated both with and without the control variables used in the previous analysis. As before, the data is weekly and each date is denoted by t.[67]

[65] Fails in on-the-run assets were first published in April 2013.

[66] Including the MOVE index changes the sign of the effective federal funds rate coefficient. This result suggests that the positive unconditional association between policy rates and the on-the-run premium is largely captured by Treasury market volatility.

[67] As in the previous specification, Treasury bill data are used solely to construct the realized off-the-run fail rate.

Table 1: On-the-run premia and fail rates regression

	(1) Premium	(2) Premium
Off-the-run fail rate	2.3859***	1.6794***
	(0.421)	(0.328)
On-the-run fail rate	0.2035	-0.5625**
	(0.301)	(0.220)
ln(MOVE)		3.1981***
		(0.352)
ln(VIX)		-0.3210
		(0.320)
Effective federal funds rate		-0.3535***
		(0.081)
10Y–2Y spread		-0.3560**
		(0.142)
Oct 2014 fixed effect		0.8458***
		(0.166)
Sep 2019 fixed effect		-0.2574
		(0.248)
Mar 2020 fixed effect		1.4241***
		(0.307)
Constant	2.5748***	2.5661***
	(0.115)	(0.082)
No. Observations	563	563
Adj. R^2	0.189	0.526

Notes: I use Newey-West standard errors with 6 lags. *** indicates significance at the 1% level. ** indicates significance at the 5% level. * indicates significance at the 10% level. The sample period is April 2013–2025, and the frequency is weekly. The data source for the premium (in basis points) is identical to that used in Figure 2a. The premium is defined as the average on-the-run premium across all coupon-bearing Treasuries on a given day. The fail rate data (in percent) are the same as in Figure 5b; outliers are removed by excluding negative observations and the top 5% of observations. The federal funds rate data are obtained from FRED. The MOVE and VIX data are taken from FactSet. The 10-year and 2-year yields (par yields of seasoned notes in percent) are taken from the FED yield curve, which is an updated version of the original Gürkaynak-Sack-Wright curve (Gürkaynak et al., 2010), https://www.federalreserve.gov/data/nominal-yield-curve.htm. The crisis fixed effects capture the months October 2014, September 2019 and March 2020. All independent variables, except the crisis dummies, are demeaned.

The regression equation for the second regression in Table 2 is given by

$$\begin{aligned}\text{Off-the-run fail rate}_t = \alpha &+ \beta_1 \text{ Off-the-run primary dealer net position}_t \\ &+ \beta_2\, ln(\text{MOVE}_t) + \beta_3\, ln(\text{VIX}_t) + \beta_4 \text{ Effective federal funds rate}_t \\ &+ \beta_5 \text{ 10Y–2Y spread}_t + \beta_6\, \mathbb{I}_{10/14,t} + \beta_7\, \mathbb{I}_{09/19,t} + \beta_8\, \mathbb{I}_{03/20,t} + u_t.\end{aligned}$$

The regression equation for the fourth regression in Table 2 is given by

$$\begin{aligned}\text{On-the-run fail rate}_t = \alpha &+ \beta_1 \text{ On-the-run primary dealer net position}_t \\ &+ \beta_2\, ln(\text{MOVE}_t) + \beta_3\, ln(\text{VIX}_t) + \beta_4 \text{ Effective federal funds rate}_t \\ &+ \beta_5 \text{ 10Y–2Y spread}_t + \beta_6\, \mathbb{I}_{10/14,t} + \beta_7\, \mathbb{I}_{09/19,t} + \beta_8\, \mathbb{I}_{03/20,t} + u_t.\end{aligned}$$

The results are as follows. As expected, the coefficients on primary dealers' net positions are negative and statistically significant in both the off-the-run and on-the-run specifications when controls are included. Because net positions are measured in dollars, the coefficients are most meaningfully interpreted by scaling them by the standard deviation of net outright positions. In the specifications with controls, a one standard deviation increase in off-the-run (on-the-run) net positions is associated with a decline in the corresponding fail rate of about 0.25 (0.08) standard deviations of the fail rate.

For robustness, I also report the corresponding specification using first differences of net positions in Appendix A.8.4, as off-the-run net positions may be non-stationary. The results remain qualitatively the same. Moreover, in the first-difference specification, the coefficient on on-the-run net positions is negative and statistically significant even in the regression without controls.

Specifications (3) and (4) in Table 2 are the only regressions in which both the dependent variable and the key explanatory variable, namely the on-the-run fail rate and on-the-run net positions, are observed at the maturity level. I therefore estimate separate regressions for each maturity $m \in \{2\text{Y}, 3\text{Y}, 5\text{Y}, 7\text{Y}, 10\text{Y}, 20\text{Y}, 30\text{Y}\}$, matching maturity-specific fail rates with maturity-specific dealer inventories. The results are reported in Table 8 in Appendix A.8.3. Consistent with the inventory hypothesis, the estimated coefficients are negative across all maturity sectors except the 3Y sector. The coefficients are statistically significant for the 2Y, 5Y, 7Y, and 10Y sectors, indicating that increases in dealer inventories are associated with lower fail rates in these maturities.

Table 2: Fail rates and net positions regression

	(1) Off-the-run fail rate	(2) Off-the-run fail rate	(3) On-the-run fail rate	(4) On-the-run fail rate
Off-the-run net positions	-0.0012***	-0.0009***		
	(0.000)	(0.000)		
On-the-run net positions			0.0008	-0.0043**
			(0.003)	(0.002)
ln(MOVE)		0.1419**		0.4259***
		(0.055)		(0.104)
ln(VIX)		-0.1306**		-0.0207
		(0.056)		(0.097)
Effective federal funds rate		-0.0809***		-0.1110***
		(0.015)		(0.029)
10Y–2Y spread		-0.1938***		-0.3316***
		(0.021)		(0.051)
Oct 2014 fixed effect		0.1085**		-0.1520***
		(0.047)		(0.050)
Sep 2019 fixed effect		-0.1087*		0.1153
		(0.056)		(0.090)
Mar 2020 fixed effect		0.0691		1.1976***
		(0.067)		(0.263)
Constant	0.6439***	0.6432***	0.5189***	0.5136***
	(0.016)	(0.012)	(0.034)	(0.024)
No. Observations	594	594	609	609
Adj. R^2	0.108	0.297	-0.001	0.212

Notes: I use Newey-West standard errors with 6 lags. *** indicates significance at the 1% level. ** indicates significance at the 5% level. * indicates significance at the 10% level. The sample period is April 2013–2025 and the frequency is weekly. The net positions of primary dealers can be downloaded from the FED's Primary Dealer Statistics, https://www.newyorkfed.org/markets/counterparties/primary-dealers-statistics. On-the-run net outright positions (in bn USD) are measured as the sum of net outright positions in coupon-bearing on-the-run Treasuries. Off-the-run positions are constructed by subtracting aggregate on-the-run positions from total net outright positions in coupon-bearing Treasuries. The data are deflated in the same way as in Figure 14. The fail rate data (in percent) are the same as in Figure 5b; outliers are removed by excluding negative observations and the top 5% of observations. The federal funds rate data are obtained from FRED. The MOVE and VIX data are taken from FactSet. The 10-year and 2-year yields (par yields of seasoned notes in percent) are taken from the FED yield curve, which is an updated version of the original Gürkaynak-Sack-Wright curve (Gürkaynak et al., 2010), https://www.federalreserve.gov/data/nominal-yield-curve.htm. The crisis fixed effects capture the months October 2014, September 2019 and March 2020. All independent variables, except the crisis dummies, are demeaned.

4.3 Evidence from a policy intervention

I exploit the introduction of the TMPG fail charge in May 2009 (see Section 2.2 for details) as a policy intervention that changed the incentives associated with settlement fails. Specifically, I examine changes in both the on-the-run premium and off-the-run settlement fail rate within a narrow window surrounding the implementation of the fail charge. A significant response of the on-the-run premium to an institutional change designed to reduce settlement fails provides evidence that settlement risk contributes to the determination of on-the-run premia. Note that data on on-the-run settlement fails are only available from 2013 onward and can therefore not be used.

I analyze how on-the-run premia and settlement fails respond to the introduction of the TMPG fail charge within four months before and after its implementation (see Table 3). This time window ensures a sufficient number of observations while circumventing the preceding crisis period. As in the previous analysis, I use the average on-the-run premium for U.S. Treasuries as the dependent variable in the first set of regressions (Table 3). Each point in time is denoted by d. The data are daily.[68] As the main explanatory variable, I use a dummy variable that equals one following the introduction of the TMPG fail charge in May 2009. In addition, I run a second regression where I include the same control variables as in the previous regressions. The regression equation to the second regression in Table 3 is given by

$$\begin{aligned}\text{On-the-run premium}_d = {} & \alpha + \beta_1 \,\text{Fail charge introduction dummy}_d + \beta_2 \, ln(\text{MOVE}_d) \\ & + \beta_3 \, ln(\text{VIX}_d) + \beta_4 \,\text{Effective federal funds rate}_d \\ & + \beta_5 \,\text{10Y–2Y spread}_d + u_d\end{aligned}$$

In the second set of regressions, I regress the off-the-run settlement fail rate on the fail-charge dummy and, in an additional specification, on the same control variables as before. Data on off-the-run settlement fails are available by week, denoted by t, and are not available by maturity. The regression equation to the fourth regression in Table 3 is given by

$$\begin{aligned}\text{Off-the-run fail rate}_t = {} & \alpha + \beta_1 \,\text{Fail charge introduction dummy}_t + \beta_2 \, ln(\text{MOVE}_t) \\ & + \beta_3 \, ln(\text{VIX}_t) + \beta_4 \,\text{Effective federal funds rate}_t \\ & + \beta_5 \,\text{10Y–2Y spread}_t + u_t\end{aligned}$$

The introduction of the fail charge reduced the on-the-run premium by approximately 3 basis points (which equals 1.2 times the pre-event standard deviation) and off-the-run

[68] Compared to the previous regressions, all data is available at daily frequency. Given the short time window I take advantage of this high frequency.

Table 3: Fail charge regression

	(1) On-the-run premium	(2) On-the-run premium	(3) Off-the-run fail rate	(4) Off-the-run fail rate
Fail charge introduction dummy	-3.7264***	-2.9474**	-0.3954***	-0.6803***
	(0.695)	(1.162)	(0.099)	(0.151)
ln(MOVE)		1.4559		1.1445*
		(2.196)		(0.590)
ln(VIX)		4.8049**		-0.0577
		(1.928)		(0.450)
Effective federal funds rate		11.4475*		-1.5984
		(5.966)		(2.664)
10Y–2Y spread		3.0786		0.3740
		(2.423)		(0.606)
Constant	9.0096***	8.6131***	0.7281***	0.8706***
	(0.669)	(0.691)	(0.063)	(0.088)
No. Observations:	167	167	32	32
Adj. R^2	0.516	0.678	0.240	0.340

Notes: I use Newey-West standard errors with 6 lags. *** indicates significance at the 1% level. ** indicates significance at the 5% level. * indicates significance at the 10% level. The sample period is January 2009–August 2009 (four months before and after the introduction of the fail charge). The frequency is daily in Columns (1) and (2) and weekly in Columns (3) and (4).
The data source for the premium (in basis points) is identical to that used in Figure 2a. The premium is defined as the average on-the-run premium across all coupon-bearing Treasuries on a given day. The fail rate data (in percent) are the same as in Figure 5b; outliers are removed by excluding negative observations and the top 5% of observations. The federal funds rate data are obtained from FRED. The MOVE and VIX data are taken from FactSet. The 10-year and 2-year yields (par yields of seasoned notes in percent) are taken from the FED yield curve, which is an updated version of the original Gürkaynak-Sack-Wright curve (Gürkaynak et al., 2010), https://www.federalreserve.gov/data/nominal-yield-curve.htm.

settlement fail rate by roughly 0.7 percentage points (around 1.9 times the pre-event standard deviation) in the specification with controls. The estimated coefficients are highly statistically significant.

The pronounced decline in the off-the-run settlement fail rate following the introduction of the fail charge indicates that the policy substantially reduced settlement risk. The simultaneous decline in the on-the-run premium within the narrow window surrounding the introduction of the fail charge provides evidence that settlement risk is an important determinant of on-the-run premia. In Appendix A.8.4 I also present the same table for a twelve months time window before and after the introduction. The results remain qualitatively the same.

5 Conclusion

This paper develops a dynamic model of the U.S. Treasury market that explains the relationship between on-the-run premia, repo specialness, and settlement fails, focusing on settlement risk and inventory uncertainty as a novel mechanism. I also test the theory empirically and show as a key result that a higher realized fail rate in off-the-run assets is positively correlated with the on-the-run premium, while a higher realized fail rate in on-the-run assets is negatively correlated. Also, higher net inventory positions in Treasuries are shown to be associated with lower fail rates. I additionally exploit the introduction of the fail charge to provide further evidence on the impact of settlement risk. The simultaneous decline in settlement fails and the on-the-run premium following the introduction of the fail charge is consistent with the prediction that settlement risk contributes to the determination of on-the-run premia.

By explicitly modeling the dimension along which otherwise identical assets differ–time since issuance–the model extends standard frameworks on the on-the-run premium such as Vayanos and Weill (2008) and delivers only equilibria in which the premium always accrues to the most recently issued asset, consistent with the data.

The model also explicitly differentiates between an asset's distribution and its scarcity, i.e., its effective supply. A key result is that scarcity is not necessary to explain price differences; only a more unequal distribution of off-the-run assets among market participants relative to on-the-run assets is. Scarcity can act as an amplifier. This insight extends beyond the on-the-run setting and can explain price differences among any two otherwise identical assets. Finding empirical examples and measuring the relative importance of the two channels–scarcity and distribution–would be a fruitful avenue for future research. A related question for further research is how unequal auction allocations affect asset liquidity over time, inspired by a first simulation exercise in the appendix showing that an initially concentrated endowment first distributes more equally among participants before becoming more unequal again.

The theory also speaks to the Treasury market turmoil and the ongoing policy debate. Because off-the-run securities face greater inventory uncertainty, smaller shocks suffice to trigger a breakdown in this segment during crises, helping to explain why the March 2020 disruption was concentrated there. There is considerable scope for future work assessing the effectiveness of proposed and implemented reforms to the Treasury market. I do so for larger issuance sizes, repo facility access, as well as all-to-all trading and tokenisation. With respect to repo facility access, it would be interesting to incorporate the growing presence of non-banks in the market, which also contributed to the fragility in March 2020.[69] Whether broadening repo facility access to these participants improves market functioning is an important question that could be explored within my framework.

[69] See, for example, Eren and Wooldridge (2021).

References

Aldasoro, I., S. Doerr, L. Gambacorta, R. Garratt, and P. K. Wilkens (2023). The tokenisation continuum. *BIS Bulletin* (72).

Anbil, S., A. Anderson, and Z. Senyuz (2020). What happened in money markets in September 2019? *FEDS Notes*.

Bank for International Settlements and Committee on Payments and Market Infrastructures (2024). Tokenisation in the context of money and other assets: concepts and implications for central banks. *Report to the G20*.

Barclay, M. J., T. Hendershott, and K. Kotz (2006). Automation versus intermediation: Evidence from Treasuries going off the run. *Journal of Finance 61*(5), 2395–2414.

Basel Committee on Banking Supervision (2019). CRE 70 - Capital treatment of unsettled transactions and failed trades. https://www.bis.org/basel_framework/chapter/CRE/70.htm.

Bech, M., A. Illes, U. Lewrick, and A. Schrimpf (2016). Hanging up the phone – electronic trading in fixed income markets and its implications. *BIS Quarterly Review*, 79–94.

Bessembinder, H., C. Spatt, and K. Venkataraman (2020). A survey of the microstructure of fixed-income markets. *Journal of Financial and Quantitative Analysis 55*(1), 1–45.

Bulusu, N. and S. Gungor (2021). The life cycle of trading activity and liquidity of Government of Canada bonds: Evidence from cash, repo and securities lending markets. *Canadian Journal of Economics/Revue canadienne d'économique 54*(2), 557–581.

Carlson, S. (2026). Why is the 10-year Treasury yield so important? https://www.chase.com/personal/investments/learning-and-insights/article/why-is-the-10-yeartreasury-yield-so-important.

Chaboud, A., E. Correia-Golay, C. Cox, M. J. Fleming, Y. Huh, F. M. Keane, K. Lee, K. Schwarz, C. Vega, and C. Windover (2025). All-to-all trading in the U.S. Treasury market. *Economic Policy Review 31*(2), 1–27.

Chaboud, A., E. Correia-Golay, M. J. Fleming, Y. Huh, F. M. Keane, and O. Shachar (2025). Liquidity and trading dynamics in the off-the-run US Treasury market. *FRB of New York Staff Report* (1170).

Christensen, J. H. E., J. A. Lopez, and P. Shultz (2017). Do all new Treasuries trade at a premium? *FRB of San Francisco Economic Letter* (3).

Christensen, J. H. E., J. A. Lopez, and P. J. Shultz (2020). Is there an on-the-run premium in TIPS? *The Quarterly Journal of Finance 10*(2), 2050007.

Cohen, A., M. Kargar, B. Lester, and P.-O. Weill (2024). Inventory, market making, and liquidity in OTC markets. *Journal of Economic Theory 222*, 105917.

Corradin, S. and A. Maddaloni (2020). The importance of being special: Repo markets during the crisis. *Journal of Financial Economics 137*(2), 392–429.

D'Amico, S. and N. A. Pancost (2022). Special repo rates and the cross-section of bond prices: The role of the special collateral risk premium. *Review of Finance 26*(1), 117–162.

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

Duffie, D. (2020). Still the world's safe haven? Redesigning the U.S. Treasury market after the COVID-19 crisis. *Hutchins Center Working Paper* (62).

Duffie, D. (2023). Resilience redux in the U.S. Treasury market. *Stanford University Graduate School of Business Research Paper* (4552735).

Duffie, D., N. Gârleanu, and L. H. Pedersen (2005). Over-the-counter markets. *Econometrica 73*(6), 1815–1847.

Duffie, D., T. F. Geithner, P. Parkinson, and J. Stein (2021). U.S. Treasury markets: Steps toward increased resilience. https://group30.org/publications/detail/4950.

Eren, E. and P. Wooldridge (2021). Non-bank financial institutions and the functioning of government bond markets. *BIS Papers* (119).

Fleming, M. and F. Keane (2021). The netting efficiencies of marketwide central clearing. *FRB of New York Staff Reports* (964).

Fleming, M. J. and K. D. Garbade (2002). When the back office moved to the front burner: Settlement fails in the Treasury market after 9/11. *FRB of New York Economic Policy Review 8*(2), 35–57.

Fleming, M. J. and K. D. Garbade (2004). Repurchase agreements with negative interest rates. *Current Issues in Economics and Finance 10*(5).

Fleming, M. J. and K. D. Garbade (2005). Explaining settlement fails. *Current Issues in Economics and Finance 11*(9).

Fleming, M. J. and F. M. Keane (2016a). Characterizing the rising settlement fails in seasoned Treasury securities. *Liberty Street Economics Federal Reserve Bank of New York*.

Fleming, M. J. and F. M. Keane (2016b). What's behind the March spike in Treasury fails? *Liberty Street Economics Federal Reserve Bank of New York*.

Fleming, M. J., F. M. Keane, A. Martin, and M. McMorrow (2014). What explains the June spike in Treasury settlement fails? *Liberty Street Economics Federal Reserve Bank of New York*.

Fleming, M. J., B. Mizrach, and G. Nguyen (2018). The microstructure of a U.S. Treasury ECN: The BrokerTec platform. *Journal of Financial Markets 40*, 2–22.

Fleming, M. J. and F. Ruela (2020). How liquid is the new 20-year Treasury bond? *Liberty Street Economics Federal Reserve Bank of New York*.

Fleming, M. J., O. Shachar, and P. Van Tassel (2023). The 2022 spike in corporate security settlement fails. *Liberty Street Economics Federal Reserve Bank of New York*.

FRED (2026). How are benchmark borrowing costs measured? A close look into the 10-year Treasury yield. https://fredblog.stlouisfed.org/2026/05/how-are-benchmark-borrowing-costs-measured/.

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

Graveline, J. J. and M. R. McBrady (2011). Who makes on-the-run Treasuries special? *Journal of Financial Intermediation 20*(4), 620–632.

Gravelle, T. (1998). Buying back government bonds: Mechanics and other considerations. *Bank of Canada Working Paper* (9).

Gürkaynak, R. S., B. Sack, and J. H. Wright (2010). The TIPS yield curve and inflation compensation. *American Economic Journal: Macroeconomics 2*(1), 70–92.

He, Z., S. Nagel, and Z. Song (2022). Treasury inconvenience yields during the COVID-19 crisis. *Journal of Financial Economics 143*(1), 57–79.

Huh, Y. and S. Infante (2021). Bond market intermediation and the role of repo. *Journal of Banking and Finance 122*, 105999.

Iyer, R. and M. Macchiavelli (2017). The systemic nature of settlement fails. *FEDS Notes*.

Kashyap, A. K., J. C. Stein, J. L. Wallen, and J. Younger (2025). Treasury market dysfunction and the role of the central bank. *Brookings Papers on Economic Activity 56*, 221–296.

Keane, F. M. (1996). Repo patterns for new Treasury notes. *Current Issues in Economics and Finance 2*(10).

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

Lagos, R., G. Rocheteau, and P.-O. Weill (2011). Crises and liquidity in over-the-counter markets. *Journal of Economic Theory 146*(6), 2169–2205.

Lagos, R. and R. Wright (2005). A unified framework for monetary theory and policy analysis. *Journal of Political Economy 113*(3), 463–484.

Lee, M. J., A. Martin, and R. M. Townsend (2024). Optimal design of tokenized markets. *FRB of New York Staff Report* (1121).

Lee, S., A. Ranaldo, and T. Tsuruga (2026). Settlement risk and currency markets. *IMF Working Paper* (156).

McLaughlin, D. (2023). The trade-off between shorter settlement times and multilateral netting benefits in deferred net settlement. *Journal of Financial Market Infrastructures 11*(1), 1–17.

Pasquariello, P. and C. Vega (2009). The on-the-run liquidity phenomenon. *Journal of Financial Economics 92*(1), 1–24.

Schrimpf, A., H. S. Shin, and V. Sushko (2020). Leverage and margin spirals in fixed income markets during the COVID-19 crisis. *BIS Bulletin* (2).

Sokolinskiy, O. (2025). Trading costs v. indicative liquidity in the off-the-run Treasury market. *Finance and Economics Discussion Series* (49).

Strebulaev, I. (2002). Liquidity and asset pricing: Evidence from the U.S. Treasury securities market. Unpublished working paper.

TMPG (2018). Treasury market practices group updates its fails charge practice recommendation. https://www.newyorkfed.org/tmpg/settlement_fails.html.

TMPG (2020). Frequently asked questions: TMPG fails charges. https://www.newyorkfed.org/tmpg/settlement_fails.html.

U.S. Department of the Treasury, Board of Governors of the Federal Reserve System, Federal Reserve Bank of New York, U.S. Securities and Exchange Commission, and U.S. Commodity Futures Trading Commission (2015). Joint staff report: The U.S. Treasury market on October 15, 2014.

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

Vissing-Jorgensen, A. (2021). The Treasury market in spring 2020 and the response of the Federal Reserve. *NBER Working Paper Series* (29128).

Weill, P.-O. (2007). Leaning against the wind. *The Review of Economic Studies 74*(4), 1329–1354.

Wells, M. (2023). Averting a Treasury market crisis. *Econ Focus*, 14–17.

A Appendix

A.1 Notation

The following tables provide an overview over the used notation.

Table 4: Agents, assets, parameters and indexes

Variable	Description
Agents	
B	Buyer
S	Seller
D^B	Dealer matched with a buyer
D^S	Dealer matched with a seller
Assets	
e	Two-period Treasury in period 1
f	Two-period Treasury in period 2 ("off-the-run asset")
n	One-period Treasury in period 2 ("on-the-run asset")
Parameters	
δ	Coupon in each subperiod
δ_h	Coupon plus convenience yield (high valuation)
δ_l	Coupon plus inconvenience yield (low valuation)
σ	Mass of sellers relative to buyers
$\bar{r}$	Funding rate available to dealers outside the repo market
θ	Bargaining power of D^S in the repo market
θ_d	Bargaining power of D^S in the spot market
Indexes	
i	Generic asset index, $i \in \{e, n, f\}$
j	Generic agent index, $j \in \{B, S, DB, DS\}$
t	Subperiod index

Table 5: Equilibrium objects

Variable	Description
Prices, quantities, surpluses spot market	
$P_t^{a,i}$	Ask price
$A_t^{a,i}$	Spot market quantity sold by dealers
$P_t^{b,i}$	Bid price
$A_t^{b,i}$	Spot market quantity bought by dealers
$S_t^{j,i}$	Spot market surplus
Rates, quantities, surpluses repo market	
r^i	Repo rate
A_R^i	Repo collateral quantity
M_R^i	Repo cash quantity
$S_R^{j,i}$	Repo market surplus
Fail probabilities	
$\mathbb{P}^i$	Settlement fail probability
$\mathbb{P}^{\max,i}$	Maximum fail probability consistent with short selling
Specialness, premium and bid-ask spread	
s	Repo specialness
Δ	On-the-run premium
ba^i	Bid-ask spread
Other	
$\bar{P}_1^i$	Asset valuation used when no ask price exists
P^i	Average spot market price
s_0^i	Share of buyers with fail inventories
F^i	Settlement fails
V^j	Lifetime value
W	Aggregate welfare
$\mathcal{A}^{S,i}$	Average seller inventory at the beginning of the second period
$\mathcal{A}^{B,i}$	Average buyer inventory at the beginning of the second period
$1-\sigma_H$	Share of buy-and-hold buyers

Throughout the crisis analysis, variables with the subscript c (in particular $\bar{r}_c$, $\mathbb{P}_c^i$, and $\mathbb{P}_c^{\max}$) denote their crisis counterparts. In addition γ_c denotes the mass of so-called non-trade willing buyers.

A.2 Proofs

A.2.1 Proof of Lemma 1

The equilibrium prices and repo rates for the on- and off-the-run assets follow directly from the solutions to the spot market and repo market maximization problems in Section 3.2.

A.2.2 Proof of Proposition 1

The equilibrium expressions follow directly from Lemma 1 and Definition 4. In particular,

$$s = (\mathbb{P}^f - \mathbb{P}^n)\frac{\theta(\delta - \delta^l)}{\delta^l + \delta^h}.$$

Because $\theta > 0$, $\delta > \delta^l$, and $\delta^h + \delta^l > 0$, it follows that $s > 0$ iff $\mathbb{P}^f > \mathbb{P}^n$. Moreover,

$$\Delta = \frac{1}{4}(1 - \theta^d)(\delta^h + \delta^l)s,$$

and

$$ba^f - ba^n = \frac{1}{2}(1 - \theta^d)s(\delta^h + \delta^l).$$

Since the coefficients multiplying s are positive, $\Delta > 0$ iff $s > 0$ and $(ba^f - ba^n) > 0$ iff $s > 0$.

A.2.3 Proof of Lemma 2

Proof. I first prove the first part of the lemma. Dealers maximally sell one asset i short by the assumption that trade is limited to one unit per asset i. By the definition of a fail, this implies that fails only occur if a dealer meets a buyer with zero inventory of asset i at the beginning of the second subperiod. Therefore, $s_0^i > 0$ is a necessary condition for $F^i > 0$. Otherwise, $F^i = 0$. Moreover, if all buyers have zero inventory of asset i at the beginning of the second subperiod, then no dealer short sells asset i. Specifically, $s_0^i = 1$ implies $\mathbb{P}^i = 1$, and in this case the joint surplus satisfies $(S_1^{DB,i} + S_1^{B,i}) < 0$, implying that no trade occurs between dealers and buyers in the first subperiod. Hence, $F^i = 0$, since no short selling occurs in asset i.

For the second part of the lemma, note that a positive number of fails in asset i requires short selling in that asset. Dealers only sell asset i short if $\mathbb{P}^{max,i} \geq \mathbb{P}^i = s_0^i$. In addition, for a positive number of fails in asset i, there must be fail inventories, i.e., $s_0^i > 0$, since the probability of a fail cannot be zero: $\mathbb{P}^i = s_0^i > 0$. Combining these conditions yields $\mathbb{P}^{max,i} \geq s_0^i > 0$. □

A.2.4 Proof of Corollary 1

In Section 3.2, we have shown that in the short selling equilibrium dealers demand one unit of the e asset in the second subperiod of the first period, and one unit each of the n and

f assets in the second subperiod of the second period, as they need to deliver these assets to their repo counterparty and would otherwise fail. In all other cases where dealers match with buyers in the short selling equilibrium, dealers do not buy from buyers and cannot fail. We therefore do not need to consider these cases.

I derive first the probability $\mathbb{P}^n$. All buyers are endowed with one unit of the n asset in the second period. Therefore, whenever a buyer is matched with a dealer in the second subperiod of the second period, the buyer can always sell the asset and does so in equilibrium, allowing the dealer to meet its repo obligation. Hence, fails do not occur and $\mathbb{P}^n = 0$. The analogous reasoning applies to asset e. Therefore, $\mathbb{P}^e = 0$.

Next, I derive the probability $\mathbb{P}^f$. A fail to deliver asset f occurs only if a dealer is matched with a buyer in the second subperiod of the second period who does not hold any inventory of asset f. Such a buyer arises if and only if the buyer is not matched in the first subperiod of either period and therefore does not build up additional inventory, and is matched in the second subperiod of both periods, selling his endowment in the first instance and being unable to sell in the second. The probability of this event is therefore given by $\mathbb{P}^f = [\sigma(1-\sigma)]^2$.

A.2.5 Proof of Corollary 2

Proof. In the short selling equilibrium, dealers sell all assets i short because $\mathbb{P}^i \leq \mathbb{P}^{max,i}$ for all i. Hence, a fail occurs whenever the event associated with the fail probability $\mathbb{P}^i$ is realized. Therefore, $F^i = \mathbb{P}^i$ for all i. By Proposition 1, $\Delta > 0$ iff $F^f = \mathbb{P}^f > F^n = \mathbb{P}^n$. By Corollary 1, $\mathbb{P}^f > \mathbb{P}^n$. Therefore, $\Delta > 0$. $\square$

A.2.6 Proof of Proposition 2

Proof. By assumption, the on-the-run inventory distribution is degenerate, as all buyers are endowed with one unit of the asset n. Hence, all inventories are non-empty and $\mathbb{P}^n = 0$ (see the proof of Corollary 1).

Next, I show that a positive on-the-run premium requires the inventory distribution of off-the-run assets to be non-degenerate at the beginning of the second period. By Proposition 1, $\Delta > 0$ if and only if $\mathbb{P}^f > \mathbb{P}^n$. Since $\mathbb{P}^n = 0$, it follows that $\Delta > 0$ if and only if $\mathbb{P}^f > 0$. In the short selling equilibrium, $\mathbb{P}^f = F^f > 0$ (see the proof of Corollary 2). By Lemma 2,

$$F^f > 0 \quad \text{if and only if} \quad \mathbb{P}^{\max,f} \geq s_0^f > 0.^{70}$$

Since buyers only purchase (and not sell) assets in the first subperiod, the condition $s_0^f > 0$ implies that some inventories are already empty at the beginning of the second period. At the same time, because not all buyers are matched in the second subperiod of the first period, some inventories remain non-empty. Therefore, the off-the-run inventory distribution is non-degenerate at the beginning of the second period. □

A.2.7 Proof of Proposition 5

Proof. In the short selling equilibrium,

$$\frac{\partial S_t^{j,i}}{\partial \sigma} = \frac{\partial S_2^{j,f}}{\partial \sigma} = \frac{\partial S_R^{DB,i}}{\partial \sigma} = \frac{\partial S_R^{DS,i}}{\partial \sigma} = 0$$

for all $i \in \{e, n\}$, $j \in \{S, B, DS, DB\}$, and all t. Therefore,

$$\begin{aligned}
\frac{\partial W}{\partial \sigma} &= \frac{\partial V_B}{\partial \sigma} + \sigma \frac{\partial \sum_k V_k}{\partial \sigma} + \sum_k V_k \\
&= \sigma \left[\frac{\partial (S_R^{DS,f} + S_R^{DB,f})}{\partial \mathbb{P}^f} + \sum_k \frac{\partial S_1^{k,f}}{\partial \mathbb{P}^f} - \sum_k S_2^{k,f} \right] \frac{\partial \mathbb{P}^f}{\partial \sigma} + \sum_k V_k \\
&= \frac{\partial W}{\partial \mathbb{P}^f} \frac{\partial \mathbb{P}^f}{\partial \sigma} + \sum_k V_k,
\end{aligned}$$

where $k \in \{S, DS, DB\}$ and the second equality follows from $S_1^{B,f} = S_2^{B,f} = 0$. Moreover,

$$\begin{aligned}
\frac{\partial (S_R^{DS,f} + S_R^{DB,f})}{\partial \mathbb{P}^f} &= -(\delta - \delta^l) < 0, \\
\sum_k \frac{\partial S_1^{k,f}}{\partial \mathbb{P}^f} &= -(1+\theta)(\delta - \delta^l) < 0, \\
\sum_k S_2^{k,f} &\geq 0,
\end{aligned}$$

where $k \in \{S, DS, DB\}$. Hence,

$$\frac{\partial W}{\partial \mathbb{P}^f} < 0.$$

[70] The case $1 > s_0^f > \mathbb{P}^{\max,f}$ cannot arise in the short selling equilibrium because $\mathbb{P}^{\max,f} \geq \mathbb{P}^f = s_0^f$.

Furthermore,

$$\frac{\partial \mathbb{P}^f}{\partial \sigma} = 2\sigma(1-\sigma)(1-2\sigma),$$

which implies

$$\frac{\partial \mathbb{P}^f}{\partial \sigma} = \begin{cases} > 0, & \sigma < \frac{1}{2}, \\ = 0, & \sigma = \frac{1}{2}, \\ < 0, & \sigma > \frac{1}{2}. \end{cases}$$

Since $\sum_k V^k \geq 0$, it follows that

$$\frac{\partial W}{\partial \sigma} > 0$$

for all $\sigma \geq \frac{1}{2}$. For $\sigma < \frac{1}{2}$, the sign of $\partial W / \partial \sigma$ is generally ambiguous. □

A.2.8 Proof of Lemma 3

Proof. The candidate equilibrium consists of the spot market prices and repo rates derived in Section 3.2 for the case $\mathbb{P}^i \leq \mathbb{P}^{\max,i}$ for all i (and, in the second period, characterized in Lemma 1), the corresponding equilibrium quantities, which equal one, and the fail probabilities characterized in Corollary 1.

I first prove existence. The quantities, prices, and rates prescribed by the candidate equilibrium satisfy the agents' maximization problems. Moreover, the fail probabilities in Corollary 1 are derived from the equilibrium inventory distribution (see the proof of Corollary 1). Therefore, all conditions of Definitions 2 and 3 are satisfied, implying that the short selling equilibrium exists.

Next, I prove uniqueness. Given $\mathbb{P}^i \leq \mathbb{P}^{\max,i}$ for all i, each agent's maximization problem admits a unique solution. These solutions uniquely determine equilibrium quantities, spot market prices, repo rates, and the resulting inventory distribution. Since fail probabilities are uniquely determined by the equilibrium inventory distribution (see the proof of Corollary 1), all equilibrium objects are uniquely determined. Therefore, the short selling equilibrium is unique. □

A.2.9 Proof of Proposition 3

Proof. From Proposition 1 and Corollary 1, it follows that in the short selling equilibrium, $\Delta > 0$. Moreover, Section 3.2 shows that the solution to the maximization problems is unique whenever $\mathbb{P}^i \leq \mathbb{P}^{\max,i}$ for all i. Given that $\mathbb{P}^i = F^i$ for all i in the short selling equilibrium (see the proof of Corollary 2) and $\mathbb{P}^e = 0$ (by the same reasoning as for $\mathbb{P}^n = 0$, since all buyers hold at least one unit of asset e when dealers repurchase the asset), there are no fails in the first period, i.e., $F^e = 0$. This implies that the amount of assets in the hands of buyers is the same at the beginning and at the end of the first period. The same applies to sellers. Hence,

$$\mathcal{A}^{S,n} = \mathcal{A}^{S,f} \quad \text{and} \quad \mathcal{A}^{B,n} = \mathcal{A}^{B,f}.$$

The off-the-run asset is not relatively scarce.

In addition, Section A.4 shows that the only other equilibrium that can exist is the on-the-run short selling equilibrium, which arises if $\mathbb{P}^f > \mathbb{P}^{\max,f}$, $\mathbb{P}^{\max,e} \geq \mathbb{P}^e$, and $\mathbb{P}^{\max,n} \geq \mathbb{P}^n$. The maximization problems also imply a unique solution in this equilibrium, and the on-the-run premium is positive, i.e., $\Delta > 0$. Since $\mathbb{P}^e = 0$ (by the same reasoning as before), there are no fails in the first period, implying again that

$$\mathcal{A}^{S,n} = \mathcal{A}^{S,f} \quad \text{and} \quad \mathcal{A}^{B,n} = \mathcal{A}^{B,f}.$$

Hence, the off-the-run asset is not relatively scarce.

Lastly, the proof of Lemma 3 establishes the uniqueness of the short selling equilibrium. The proof for the on-the-run short selling equilibrium is analogous. □

A.2.10 Proof of Proposition 4

Proof. We have shown that in the short selling equilibrium with no buy-and-hold buyers, i.e., $(1 - \sigma_H) = 0$, the off-the-run asset is not relatively scarce, i.e.

$$\mathcal{A}^{S,n} = \mathcal{A}^{S,f} \quad \text{and} \quad \mathcal{A}^{B,n} = \mathcal{A}^{B,f}.$$

Given that the short selling equilibrium exists, quantities traded in the first period are independent of fail probabilities in the second period. Therefore, inventories at the beginning of the second period are exactly the same as in the benchmark case with $(1-\sigma_H) = 0$, except that, by assumption, the inventories of off-the-run assets held by buy-and-hold buyers are not counted. Hence,

$$\frac{\partial \mathcal{A}^{B,f}}{\partial (1 - \sigma_H)} < 0.$$

In combination with $\mathcal{A}^{B,f} = \mathcal{A}^{B,n}$ when $(1 - \sigma_H) = 0$ it follows that $\mathcal{A}^{B,f} < \mathcal{A}^{B,n}$ for all $(1 - \sigma_H) > 0$. Hence, the off-the-run asset is relatively scarce.

Consider first the case in which buy-and-hold buyers participate in the spot market. A fail in the off-the-run asset occurs if a dealer is matched either with a buy-and-hold buyer or with a regular buyer who has no inventory of the off-the-run asset. Therefore,

$$\mathbb{P}^f = \sigma\Big[(1 - \sigma_H) + \sigma_H s_0^f\Big],$$

where $s_0^f = (1 - \sigma)^2\sigma$. Hence,

$$\frac{\partial \mathbb{P}^f}{\partial(1 - \sigma_H)} = \sigma(1 - s_0^f) > 0,$$

where the inequality follows from $0 < s_0^f < 1$. The fail probability of the on-the-run asset is unaffected by the presence of buy-and-hold buyers in the off-the-run asset market, so that

$$\mathbb{P}^n = 0 \quad \text{and} \quad \frac{\partial \mathbb{P}^n}{\partial(1 - \sigma_H)} = 0.$$

From Proposition 1 it follows that,

$$\frac{\partial \Delta}{\partial(\mathbb{P}^f - \mathbb{P}^n)} > 0.$$

Applying the chain rule yields

$$\frac{\partial \Delta}{\partial(1 - \sigma_H)} = \frac{\partial \Delta}{\partial(\mathbb{P}^f - \mathbb{P}^n)}\left[\frac{\partial \mathbb{P}^f}{\partial(1 - \sigma_H)} - \frac{\partial \mathbb{P}^n}{\partial(1 - \sigma_H)}\right] > 0.$$

Therefore, the on-the-run premium increases in scarcity.

Next, suppose that buy-and-hold buyers do not participate in the market. In this case, they do not affect matching probabilities and therefore do not affect fail probabilities. Hence,

$$\mathbb{P}^f = \sigma s_0^f, \qquad \mathbb{P}^n = 0,$$

with $s_0^f = (1 - \sigma)^2\sigma$, and therefore

$$\frac{\partial \mathbb{P}^f}{\partial(1 - \sigma_H)} = \frac{\partial \mathbb{P}^n}{\partial(1 - \sigma_H)} = 0.$$

Again applying the chain rule,

$$\frac{\partial\Delta}{\partial(1-\sigma_H)} = \frac{\partial\Delta}{\partial(\mathbb{P}^f - \mathbb{P}^n)}\left[\frac{\partial\mathbb{P}^f}{\partial(1-\sigma_H)} - \frac{\partial\mathbb{P}^n}{\partial(1-\sigma_H)}\right] = 0.$$

Hence, the on-the-run premium is independent of scarcity in this case. □

A.2.11 Proof of Lemma 4

Proof. Solving $\mathbb{P}_c^f > \mathbb{P}_c^{\max}$ and $\mathbb{P}_c^{\max} \geq \mathbb{P}_c^n$ for γ_c yields $\mathbb{P}_c^{\max} \geq \gamma_c > \frac{\mathbb{P}_c^{\max} - \sigma^2(1-\sigma)^2}{1-\sigma^2(1-\sigma)^2}$. Applying the crisis fail probabilities to the second spot market maximization problem in Section 3.2 implies that dealers do not sell off-the-run assets to buyers because $\mathbb{P}_c^f > \mathbb{P}_c^{\max}$, while they continue to sell on-the-run assets because $\mathbb{P}_c^{\max} \geq \mathbb{P}_c^n$. At this stage, however, it has not yet been established that dealers acquire and repo out the on-the-run asset, since this depends on the profitability of trade in the first spot market and therefore on the crisis ceiling rate $\bar{r}_c$. Unlike in the benchmark economy, there is no assumption analogous to Assumption 1 for $\bar{r}_c$, which guaranteed trade in the first spot market for every asset. □

A.2.12 Proof of Corollary 3

Proof. Given that

$$\mathbb{P}_c^f > \mathbb{P}_c^{\max} \geq \mathbb{P}_c^n,$$

it follows from the repo market maximization problem in Section 3.2 that dealers do not sell off-the-run assets short, whereas they continue to sell on-the-run assets short. Consequently,

$$r^f = \bar{r}_c$$

and

$$r^n = (1-\theta)\bar{r}_c + \theta\frac{\mathbb{P}_c^n(\delta-\delta^l) - (\delta^h - \delta)}{\delta^h + \delta^l},$$

with $r^f \geq r^n$. This presumes that trade takes place in the first spot market, a condition that we verify next. From the first spot market maximization problem described in Section 3.2, the joint surplus[71] is negative whenever

$$r^f = \bar{r}_c > \frac{\delta - \delta^l}{\delta^h + \delta^l}.$$

[71] There is no assumption analogous to Assumption 1 for $\bar{r}_c$, which guaranteed trade in the first spot market in the benchmark economy.

Hence, there is no trade in off-the-run assets. For on-the-run assets, trade occurs if the joint surplus is positive, which requires

$$\frac{\delta - \delta^l}{\delta^h + \delta^l} \geq r^n.$$

Substituting for r^n yields

$$\frac{(\delta - \delta^l) - \theta[\gamma_c(\delta - \delta^l) - (\delta^h - \delta)]}{(\delta^h + \delta^l)(1 - \theta)} \geq \bar{r}_c.$$

Therefore, dealers do not purchase off-the-run assets from sellers and, consequently, do not repo them. In contrast, they continue to purchase and repo on-the-run assets. □

A.2.13 Proof of Lemma 5

Proof. First, I prove that a change in $\bar{r}_c$ can never restore trade in off-the-run assets between dealers and buyers if there is currently no trade. If there is no trade, then

$$\mathbb{P}_c^f = \sigma^2(1-\sigma)^2 + \gamma_c[1 - \sigma^2(1-\sigma)^2] > \mathbb{P}_c^{\max} = \frac{\delta^h - \delta + \bar{r}_c(\delta^h + \delta^l)}{\delta - \delta^l}.$$

A necessary condition for trade is

$$\mathbb{P}_c^f \leq \mathbb{P}_c^{\max}.$$

Since

$$\frac{\partial \mathbb{P}_c^f}{\partial \bar{r}_c} = 0 \qquad \text{and} \qquad \frac{\partial \mathbb{P}_c^{\max}}{\partial \bar{r}_c} > 0,$$

a decrease in $\bar{r}_c$ lowers $\mathbb{P}_c^{\max}$ while leaving $\mathbb{P}_c^f$ unchanged, making the condition for trade even less likely to hold. Hence, a reduction in $\bar{r}_c$ cannot restore trade when there is initially no trade. An increase in $\bar{r}_c$ can also not restore trade because trade only takes place if there is also trade between dealers and sellers. However, only a decrease in $\bar{r}_c$ can restore the latter trade (see below).

Next, I prove that a sufficiently large decrease in $\bar{r}_c$ restores spot trade between sellers and dealers and repo trade among dealers in off-the-run assets. The following argument is based on the maximization problems in Section 3.2, adapted to the crisis environment. Since there is no trade between dealers and buyers, off-the-run assets are not sold short. Therefore, if repo trade occurs, the repo rate of off-the-run assets is given by $r^f = \bar{r}_c$. Moreover, repo trade occurs whenever dealers purchase off-the-run assets from sellers in the spot market (which is not given as in the benchmark economy). Such trade is profitable in the first subperiod if the associated joint surplus is positive, which requires

$$\frac{\delta - \delta^l}{\delta^h + \delta^l} \geq \bar{r}_c.$$

Hence, for a sufficiently low $\bar{r}_c$, dealers purchase off-the-run assets from sellers and finance these positions in the repo market. If trade in the first subperiod takes place, then also in the second and vice versa. Therefore, both spot trade and repo trade in off-the-run assets are restored.

Third, I prove that if $\bar{r}_c$ lies within a certain range, trade in on-the-run assets continues to take place in all markets. Trade between dealers and buyers occurs if

$$\mathbb{P}_c^{\max} \geq \mathbb{P}_c^n.$$

Solving this condition for $\bar{r}_c$ yields

$$\bar{r}_c \geq \frac{\gamma_c(\delta - \delta^l) - (\delta^h - \delta)}{\delta^h + \delta^l}.$$

Moreover, from the proof of Corollary 3, we know that trade between dealers and sellers (and among dealers in the repo market) takes place if

$$\frac{(\delta - \delta^l) - \theta\left[\gamma_c(\delta - \delta^l) - (\delta^h - \delta)\right]}{(\delta^h + \delta^l)(1 - \theta)} \geq \bar{r}_c.$$

Trade in on-the-run assets in the first subperiod also implies trade in on-the-run assets in all markets in the second subperiod.

Lastly, it is always the case that

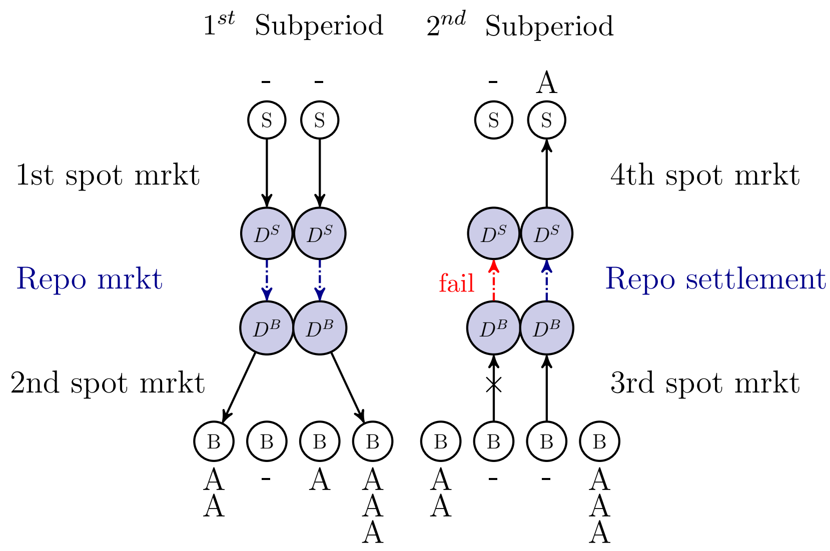
$$\min\left\{\frac{\delta - \delta^l}{\delta^h + \delta^l}, \frac{(\delta - \delta^l) - \theta\left[\gamma_c(\delta - \delta^l) - (\delta^h - \delta)\right]}{(\delta^h + \delta^l)(1 - \theta)}\right\} > \frac{\gamma_c(\delta - \delta^l) - (\delta^h - \delta)}{\delta^h + \delta^l}.$$

Therefore, there always exists a value of $\bar{r}_c$ such that trade in off-the-run assets between sellers and dealers, as well as repo trade in off-the-run assets, is restored while trade in on-the-run assets continues to take place in all markets. $\square$

A.3 Maximization problems

Figure 15 provides an overview of the sequence of the markets. I solve each period's maximization problems by backward induction, starting from the fourth spot market and moving in the direction opposite to the arrows when solving the problems sequentially.

Figure 15: Order of markets



Notes: The figure provides an overview of the timing of markets over one period (here the second). In each subperiod, exactly one repo market takes place, with one spot market preceding it and one spot market following it.

Lastly, I assume that dealers D^B never purchase assets in the first subperiod. This assumption eliminates the possibility of reverse order flow, particularly in situations where settlement risk is sufficiently high to make short selling unprofitable, i.e. when $\mathbb{P}^i > \mathbb{P}^{\max,i}$. As such cases are not the focus of the analysis, they are not considered further.

A.3.1 Second period maximization problems

4th spot market

In the final spot market of the second period, dealer D^S and seller S trade asset $i \in \{n, f\}$, provided that D^S holds the asset. The bargaining problem is

$$\max_{0 \leq A_2^{a,i} \leq 1, P_2^{a,i}} (S_2^{DS,i} + S_2^{S,i}) \quad \text{s.t.} \quad S_2^{DS,i} = \theta^d (S_2^{DS,i} + S_2^{S,i})$$

where

$$S_2^{DS,i} = (P_2^{a,i} - \delta)A_2^{a,i}, \qquad S_2^{S,i} = (\delta - P_2^{a,i})A_2^{a,i}.$$

Both agents derive the same utility from holding the asset until maturity, namely the final coupon payment δ. The solution is therefore $A_2^{a,i} = 1$ and $P_2^{a,i} = \delta$, where I assume that trade occurs even when the parties are indifferent. If dealer D^S does not hold asset i, for example because of a settlement fail in an earlier market, no trade in asset i can occur.

Repo settlement

Settlement of outstanding repo contracts in all assets $i \in \{n, f\}$ occurs automatically whenever dealer D^B holds asset i. In that case, D^B delivers the asset to dealer D^S, while D^S returns the cash, including accrued interest. If D^B does not hold asset i, a settlement fail occurs. Dealer D^S nevertheless returns the cash, including accrued interest, and dealer D^B compensates D^S by paying the fail compensation C^i.

3rd spot market

In the third spot market of the second period, dealer D^B and buyer B trade asset $i \in \{n, f\}$, provided that B holds the asset in its inventory. The bargaining problem is

$$\max_{0 \leq A_2^{b,i} \leq 1, P_2^{b,i}} S_2^{DB,i} \quad \text{s.t.} \quad S_2^{B,i} = 0$$

with

$$S_2^{DB,i} = (C^i - P_2^{b,i})A_2^{b,i}, \qquad S_2^{B,i} = (P_2^{b,i} - \delta^l)A_2^{b,i}.$$

Trade generates a surplus because dealer D^B values the asset at δ, whereas buyer B values it at only δ^l. If D^B has an outstanding repo obligation in asset i, acquiring the asset allows him to avoid paying the fail compensation $C^i = \delta$. If no such obligation exists, then there is no trade.[72]

2nd spot market

In the second spot market of the second period, the bargaining problem between dealer D^B and buyer B for each asset $i \in \{n, f\}$ with $\mathbb{P}^i \leq \mathbb{P}^{\max,i}$ is

$$\max_{0 \leq A_1^{a,i} \leq 1, P_1^{a,i}} S_1^{DB,i} \text{ s.t. } S_1^{B,i} = 0$$

[72]The dealer has no access to funding.

with

$$
\begin{aligned}
S_1^{B,i} &= (\delta^h + \delta^l - P_1^{a,i})A_1^{a,i}, \\
S_1^{DB,i} &= \bar{r}M_R^i + \underbrace{[P_1^{a,i} - (1-\mathbb{P}^i)P_2^{b,i} - \mathbb{P}^i C^i]A_R^i}_{\text{expected short selling surplus in spot market}} - \delta A_R^i.
\end{aligned}
$$

The buyer's surplus is the difference between his valuation of the asset over the next two subperiods, $\delta^h + \delta^l$, and the price he pays, $P_1^{a,i}$.[73] The dealer's surplus can be explained as follows. Dealer D^B receives the expected profit from short selling the asset (described in the repo market problem in Section 3.2) and can offset his pre-financing cost $\bar{r}$. By selling the asset, he no longer receives the coupon payment δ. Since this coupon is still owed to his repo counterparty, he must finance the payment from other sources.

Given that $\mathbb{P}^i \leq \mathbb{P}^{\max,i}$ for asset i, the joint surplus is non-negative. Therefore, the solution to the problem is $A_1^{a,i} = 1$ and $P_1^{a,i} = \delta^h + \delta^l$. Again, buyer B has no bargaining power, so the price exactly compensates him for the utility he forgoes by selling the asset.

For all $i \in \{n, f\}$ for which $\mathbb{P}^i > \mathbb{P}^{\max,i}$, there is no trade in asset i as $S_1^{DB,i} < 0$, which also implies that the joint surplus is negative.[74] The probability of failing to settle is sufficiently high that, in expectation, selling the asset short yields a loss.

Repo market

The repo market problem has two cases depending on whether $\mathbb{P}^i > \mathbb{P}^{\max,i}$ for a given asset $i \in \{n, f\}$. The case with $\mathbb{P}^i \leq \mathbb{P}^{\max,i}$ is discussed in the main text (Section 3.2); here I focus on the case with $\mathbb{P}^i > \mathbb{P}^{\max,i}$.

For all $i \in \{n, f\}$ for which $\mathbb{P}^i > \mathbb{P}^{\max,i}$, short selling is not profitable and no short position is established in asset i. Since trade in asset i has ceased, collateral is valued at $\bar{P}_1^i$ as defined in Section 3.1. The corresponding repo market surpluses are

$$
S_R^{DB,i} = (r^i - \bar{r})\bar{P}_1^i A_R^i = -S_R^{DS,i}.
$$

Otherwise, the maximization problem is identical to that discussed in Section 3.2. Since no short position is established, dealer D^B does not obtain any additional surplus from short selling. The only gains from trade arise from the difference between the repo rate r^i earned

[73] The buyer's payoff is independent of whether he is matched in the next subperiod. If unmatched, he receives utility δ^l from holding the asset. If matched, he sells the asset at $P_2^{b,i} = \delta^l$. Thus, his payoff in the next subperiod is δ^l in either case.

[74] Assumption 1 implies that $\mathbb{P}^{\max,i} < 1$, so that states with $\mathbb{P}^i > \mathbb{P}^{\max,i}$ can arise.

on the cash lent and the funding rate $\bar{r}$ at which the cash is obtained. Dealer D^S's surplus is unchanged. As dealers are assumed to trade even when indifferent, the solution is $A_R^i = 1$ and $r^i = \bar{r}$.

1st spot market

Because of its importance, the maximization problem for the first spot market and its solution are presented in the main body of the paper; see Section 3.2.

A.3.2 First period maximization problems

In this subsection, I only point out the differences between the maximization problems in the first period and those in the second period described in Section 3.2 and the previous subsection A.3.1. Analogous to Definition 1, I define

Definition 7. $\mathbb{P}^{max,e} \equiv \frac{(\delta^h-\delta)+2\bar{r}(\delta^h+\delta^l)}{(\delta-\delta^l)-(\delta^h+\delta^l)+P_1^{b,f}}$.

4th spot market

In the last spot market of the first period, the bargaining problem between dealer D^S and seller S for asset e is

$$\max_{0\leq A_2^{a,e}\leq 1, P_2^{a,e}} (S_2^{DS,e} + S_2^{S,e}) \text{ s.t. } S_2^{DS,e} = \theta^d(S_2^{DS,e} + S_2^{S,e})$$

provided that D^S holds a positive inventory of asset e, with $S_2^{DS,e} = (P_2^{a,e} - \delta - P_1^{b,f})A_2^{a,e}$ and $S_2^{S,e} = (\delta - P_2^{a,e} + P_1^{b,f})A_2^{a,e}$. Compared to the second period, the price $P_1^{b,f}$ enters the surpluses because we know that in the first subperiod of the second period (when the asset is off-the-run), the dealer will buy and the seller will sell the asset. The solution is $A_2^{a,e} = 1$ and $P_2^{a,e} = \delta + P_1^{b,f}$.

If D^S does not hold asset e, there is no trade.

Repo settlement

Settlement of outstanding repo contracts in asset e occurs in the same way as for the assets in the second period.

3rd spot market

In the third spot market, the bargaining problem between a dealer D^B who has an outstanding repo contract in asset e and a buyer B is

$$\max_{0\le A_2^{b,e}\le 1, P_2^{b,e}} S_2^{DB,e} \text{ s.t. } S_2^{B,e}=0$$

provided that B holds a positive inventory of asset e, with $S_2^{DB,e}=(C^i-P_2^{b,e})A_2^{b,e}$ and $S_2^{B,e}=(P_2^{b,e}-2\delta^l-\delta^h)A_2^{b,e}$. In the first period, the buyer's surplus also includes his valuation of the asset in the second period, which he loses if he sells the asset.[75] The solution is $A_2^{b,e}=1$ and $P_2^{b,e}=2\delta^l+\delta^h$.

In all other cases there is no trade.

2nd spot market

In the second spot market of the first period, if $\mathbb{P}^e \le \mathbb{P}^{\max,e}$, the bargaining problem between dealer D^B and buyer B is

$$\max_{0\le A_1^{a,e}\le 1, P_1^{a,e}} S_1^{DB,e} \text{ s.t. } S_1^{B,e}=0$$

with

$$\begin{aligned} S_1^{B,e} &= (2\delta^h+2\delta^l-P_1^{a,e})A_1^{a,e}, \\ S_1^{DB,i} &= \bar{r}M_R^e + \underbrace{[P_1^{a,e}-(1-\mathbb{P}^i)P_2^{b,e}-\mathbb{P}^e C^e]A_R^e}_{\text{expected short selling surplus in spot market}} -\delta A_R^e. \end{aligned}$$

As in the previous maximization problem, the buyer's surplus in the first period also includes his valuation of the asset in the second period. The solution to the problem is $A_1^{a,e}=1$ and $P_1^{a,e}=2\delta^h+2\delta^l$.

If $\mathbb{P}^e > \mathbb{P}^{\max,e}$, there is no trade in asset e because $S_1^{DB,e}<0$, implying that the joint surplus is also negative.

Repo market

The repo market maximization problem is the same for asset e as for the other assets in the second period with $C^e = P_2^{a,e}$.

1st spot market

In the first spot market of the first period, the bargaining problem between dealer D^S and

[75] It does not matter whether the buyer would resell the asset in the second period if he did not sell it in the first period. The reason is that he has no bargaining power, so the price equals his valuation.

seller S for asset e is

$$\max_{0\leq A_1^{b,e}\leq 1, P_1^{b,e}} \quad S_1^{DS,e} + S_1^{S,e} \text{ s.t. } S_1^{DS,e} = \theta^d(S_1^{DS,e} + S_1^{S,e})$$

with $S_1^{DS,e} = -r^e M_R^e + (\delta + P_2^{a,e} - P_1^{b,e})A_1^{b,e}$ and $S_1^{S,e} = (P_1^{b,e} - \delta^l - P_2^{a,e})A_1^{b,e}$. The maximization problem is the same as in the second period. The only difference is that when writing down the surpluses, we no longer substitute the price in the second subperiod (which reflects the future value of the asset), as it is no longer equal to δ (see above). The solution is $A_1^{b,e} = 1$ and $P_1^{b,e} = P_2^{a,e} + \theta^d\delta^l + (1-\theta^d)(-r^e P_1^{a,e} + \delta)$.

A.4 Other equilibria

In this section I discuss all other equilibria that can exist next to the short selling equilibrium. First, I analyze an equilibrium where there is only short selling in on-the-run assets.

Definition 8. *An on-the-run short selling equilibrium is an equilibrium where* $\mathbb{P}^i \leq \mathbb{P}^{max,i}$ *for* $i \in \{e, n\}$ *and* $\mathbb{P}^f > \mathbb{P}^{max,f}$.

From our analysis in Section 3.2 we know that in this case the prices and repo rates of the on- and off-the-run asset are the same as described in Lemma 1 except that $r^f = \bar{r}$ and there is no $P_1^{a,f}$ and $P_2^{b,f}$. The latter is the case because there is no trade between D^B and B as D^B do not want to sell the off-the-run asset short as it is too risky to fail and just keep it in their inventory.

It is straightforward to show that specialness in this equilibrium is higher than in the short selling equilibrium because the off-the-run repo rate equals the ceiling rate $\bar{r}$. Due to the difference in repo rates there is also a positive on-the-run premium in this equilibrium (when comparing the remaining prices).

Lastly note that there exists no equilibrium where there is no short selling in on-the-run assets e and n. The reason is that because buyers optimally only buy and never sell in the first subperiod and therefore always have at least an inventory of one asset in the second subperiod $\mathbb{P}^e = \mathbb{P}^n = 0$. This implies that $\mathbb{P}^{max,i} \geq \mathbb{P}^i$ for $i \in \{e, n\}$ always holds.

Therefore the short selling and the on-the-run short selling equilibrium are the only equilibria that can exist.

A.5 Extension: scarcity without inventory uncertainty

I established in Proposition 3 that scarcity is not necessary for an on-the-run premium, although (as we have seen) it can act as an amplifier. As a last step, I provide in this extension intuition that scarcity is also not sufficient for a positive on-the-run premium.

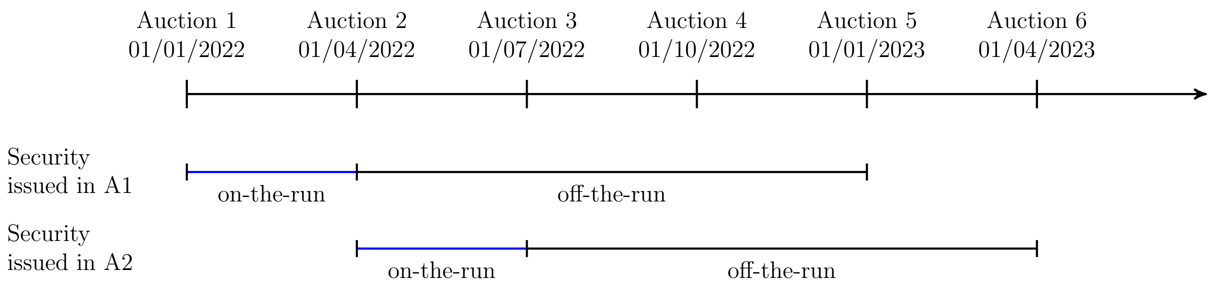
Assume the benchmark economy and suppose that the endowment of spot traders in the two-period on-the-run asset e is two, while their endowment in the one-period on-the-run asset n is three. In this case, the off-the-run asset f is scarcer than the on-the-run asset n in every equilibrium. At the same time, there is no settlement risk associated with off-the-run assets in any equilibrium because short sellers can always repurchase the asset when they need to close their position. As a result, $\mathbb{P}^n = \mathbb{P}^f = 0$, implying that there is no on-the-run premium, i.e., $\Delta = 0$. Hence, this extension shows that off-the-run asset scarcity alone is not sufficient for an on-the-run premium. There exist equilibria in which off-the-run assets are scarce, yet no on-the-run premium arises because inventory uncertainty is absent.

While this example is not a general proof that scarcity is not sufficient for a positive on-the-run premium, it provides strong intuition for why this is the case. Even when the off-the-run asset is relatively scarce, no premium arises if inventory uncertainty is absent.

A.6 On-the-run cycle

The following figure illustrates the on-the-run cycle. In this example, an auction (A) is held every quarter. The newly issued security is on-the-run until the next auction of securities with the same maturity, after which it becomes off-the-run.

Figure 16: On-the-run cycle



Notes: The figure illustrates that an asset is initially on-the-run and becomes off-the-run when the next generation of securities with the same maturity is auctioned and issued.

A.7 Simulation

I have shown in Corollary 2 that fails are increasing over time (i.e., off-the-run assets fail more often than on-the-run assets) and in Proposition 5 that those fails are welfare decreasing.

Despite this evidence, some smaller sovereign issuers (e.g., Canada and Sweden) employ a pre-benchmark period in which a security is issued through several auction rounds before it attains benchmark status. A benchmark has similar price and liquidity characteristics to U.S. on-the-run Treasuries. It stays the benchmark for its maturity group until the next bond of the same maturity becomes the new benchmark (Bulusu and Gungor, 2021).

One reason is the objective to first achieve a sufficiently large issuance size (Bulusu and Gungor, 2021). This is in line with my result that larger issuance sizes decrease settlement fails and are welfare improving (see Section 3.7.2). Another reason is that repeated issuance rounds can, under some circumstances, allow the security to become more broadly distributed across market participants before benchmark status is assigned (Gravelle, 1998). Gravelle (1998) describes how distribution and liquidity are interdependent and how a concentrated auction allocation in the hands of a few traders can lead to lower liquidity in benchmark assets.

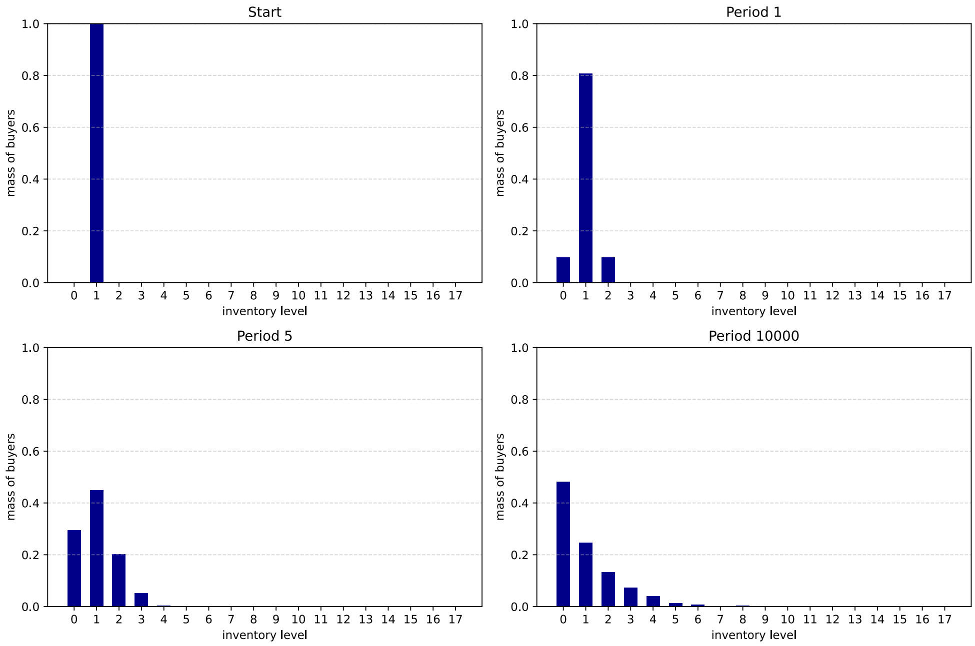
I conjecture that if the primary auction leads to an uneven allocation of assets among participants, which may be more pronounced in smaller economies, settlement fails may initially decline before increasing. To test this, I run two simulations: In one simulation, I start with an equal endowment across participants. In the other, half of the participants have no endowment, while the other half have twice the endowment of those in the first simulation, so that the total number of assets in the system is the same.

A.7.1 Equal endowment

I assume there are 100 sellers and 900 buyers. Each buyer and seller is initially endowed with one asset with infinite maturity. There are multiple periods, each with two subperiods. In each first subperiod, all sellers sell one unit of the asset, if they hold any, to a randomly matched buyer. In the second subperiod, the sellers attempt to buy back one unit of the asset, matching again with a random buyer who supplies the asset if they have a positive inventory. If they do not have a positive inventory, we say a fail occurs and no asset is bought back. In this set-up we abstract from the intermediation by dealers as this is not relevant for what we want to show in this exercise.

Figure 17 gives an overview over the inventory distribution across buyers at the very beginning and at the end of each period. We start from a degenerate distribution. Over

Figure 17: Inventories of buyers over time (equal initial endowment)



Notes: The figure shows the inventory levels over time if the initial endowment is equal for all buyers and sellers.

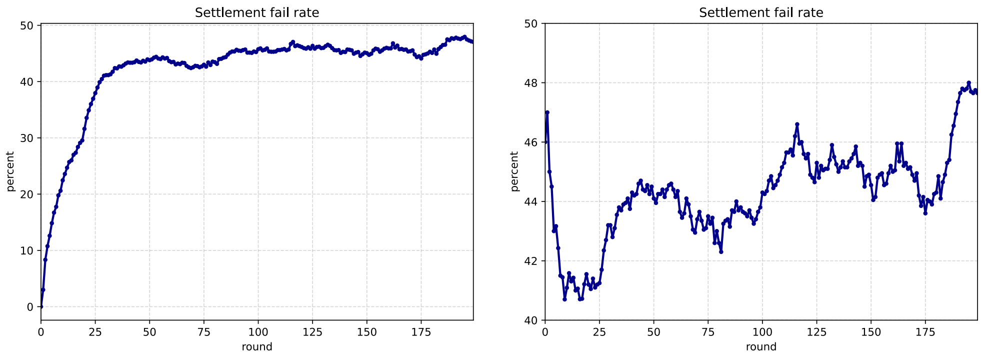
time, the distribution becomes non-degenerate, and the tail of the inventory distribution lengthens progressively. At the same time, the mass of buyers with zero inventory (i.e., agents holding no assets) increases steadily. This, in turn, implies that the fail rate rises over time, as illustrated in Figure 18a.

A.7.2 Unequal endowment

With unequal endowments, the pattern looks different. We assume that half of the agents are endowed with two units of the asset, while the other half have no endowment. As in the previous case, the tail of the right-skewed distribution becomes longer over time (see Figure 19). However, unlike before, the mass of zero-inventory agents initially decreases, leading to a corresponding initial decline in the fail rate (see Figure 18b). This may help explain why some sovereign issuers delay benchmark status until a security becomes more evenly distributed across market participants.

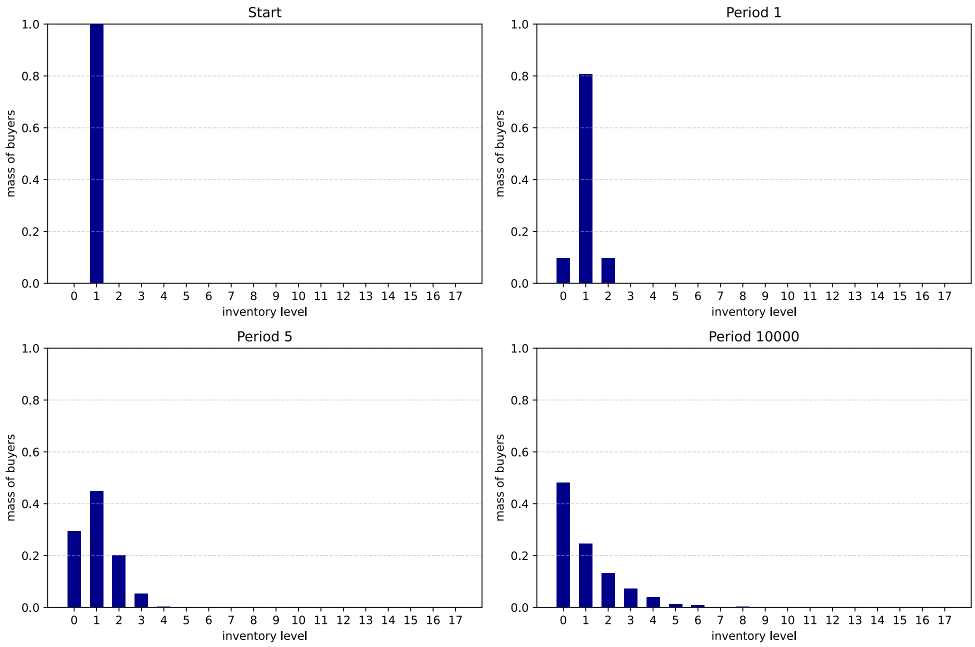
Figure 18: Fail rates over time

(a) Equal endowment (b) Unequal endowment



Notes: The figure shows simulated fail rates over time under different endowment distributions.

Figure 19: Inventories of buyers over time (unequal initial endowment)



Notes: The figure shows the inventory levels over time if the initial endowment is unequal.

A.8 Data

A.8.1 Data sources and summary statistics

This section summarizes all data sources used for the regressions and provides summary statistics for the variables.

To calculate the on-the-run premium, the on-the-run yield is subtracted from the par yield of seasoned bills and notes, as in Christensen et al. (2017). The data are taken from the FED yield curve, which is an updated version of the original Gürkaynak-Sack-Wright curve (Gürkaynak et al., 2010)[76] and the FRB-H15 Tables.[77] The fail rate data is based on data from the FED's Primary Dealer Statistics. It includes outright and financing fails.[78] The net positions of primary dealers can be downloaded from the FED's Primary Dealer Statistics as well.[79] The net positions are deflated using the Consumer Price Index for All Urban Consumers: All Items in U.S. City Average, which can be downloaded from FRED.[80] The federal funds rate data are obtained from FRED. The VIX and MOVE data can be downloaded from FactSet. The 10-year and 2-year yields (par yields of seasoned notes in percent) are as well taken from the FED yield curve mentioned above.

[76] The data can be downloaded here: https://www.federalreserve.gov/data/nominal-yield-curve.htm.

[77] The data can be downloaded here: https://www.federalreserve.gov/releases/h15/.

[78] The percentage rates are not an exact measure. I use fail data and outright and financing transaction data to calculate the series. According to the information from the NY FED, the off-the-run fails are an average over the reporting week. The on-the-run transaction data is an average over the business days of the reporting week. The rest of the used data is the value as of the reporting weekday or already a daily average. I adjust the final series for outliers and remove any negative values.

[79] The data can be downloaded here: https://www.newyorkfed.org/markets/counterparties/primary-dealers-statistics. On-the-run net outright positions are measured as the sum of net outright positions in coupon-bearing on-the-run Treasuries. Off-the-run positions are constructed by subtracting aggregate on-the-run positions from total net outright positions in coupon-bearing Treasuries.

[80] I set the index to 1 at the beginning of April 2013 as this is the start of the time horizon.

Table 6: Summary statistics

Data	Maturity	Mean	Median	St. Dev.
Average on-the-run premium in bp		2.57	2.37	1.30
Fail rate of off-the-run Treasuries in %		0.67	0.61	0.23
Fail rate of on-the-run Treasuries in %		0.41	0.33	0.30
Fail rate of on-the-run Treasuries in %	2Y	0.5621	0.2741	0.9647
Fail rate of on-the-run Treasuries in %	3Y	0.4504	0.2495	0.6222
Fail rate of on-the-run Treasuries in %	5Y	0.4953	0.3339	0.5810
Fail rate of on-the-run Treasuries in %	7Y	0.4200	0.2876	0.4537
Fail rate of on-the-run Treasuries in %	10Y	0.9644	0.3954	3.0010
Fail rate of on-the-run Treasuries in %	20Y	3.1826	0.6573	9.4184
Fail rate of on-the-run Treasuries in %	30Y	0.6342	0.3406	1.0103
PD off-the-run net positions (defl.) in USD bn		108.83	92.16	59.69
PD on-the-run net positions (defl.) in USD bn		-12.51	-12.40	9.96
PD on-the-run net positions (defl.) in USD bn	2Y	-0.6355	-0.5163	4.0860
PD on-the-run net positions (defl.) in USD bn	3Y	-2.6215	-2.6308	3.8661
PD on-the-run net positions (defl.) in USD bn	5Y	-3.7879	-3.7193	3.3768
PD on-the-run net positions (defl.) in USD bn	7Y	-0.2551	-0.2026	2.6697
PD on-the-run net positions (defl.) in USD bn	10Y	-4.2010	-3.8182	4.0179
PD on-the-run net positions (defl.) in USD bn	20Y	-2.7003	-2.6485	1.5694
PD on-the-run net positions (defl.) in USD bn	30Y	-0.5893	-0.6937	2.3312
Effective federal funds rate in %		1.76	0.91	1.93
ln(MOVE)		4.31	4.27	0.32
ln(VIX)		2.83	2.78	0.29
10Y–2Y spread in %		0.73	0.58	0.84

Notes: The time horizon used for this summary statistics table is April 2013 to December 2025. The on-the-run premium is the average premium across all coupon-bearing Treasuries on a given day. PD is the abbreviation for primary dealers.

A.8.2 Volatilities

Table 7 presents the volatility of primary dealers' aggregate net Treasury positions in on- and off-the-run notes and bonds.[81] Primary dealers' aggregate inventories in on-the-run

[81] The data can be downloaded from the FED's Primary Dealer Statistics. The data are deflated using the Consumer Price Index for All Urban Consumers: All Items in U.S. City Average, which can be downloaded from FRED. I set the index to 1 when the on-the-run net positions time series starts in April 2013. I use data from April 2013 to the end of 2025. The frequency is weekly. To calculate the volatilities, I use the average of the data over the full time horizon.

Treasuries are substantially less volatile than their inventories in off-the-run Treasuries. This holds across all maturity buckets.

Given the higher volatility of aggregate inventories in off-the-run relative to on-the-run securities, it is plausible that individual inventories of off-the-run Treasuries are also more volatile and exhibit greater cross-sectional dispersion, thereby giving rise to the distribution characterized in the theoretical framework, which in turn generates on-the-run premia.

Table 7: Primary dealer net positions volatilities

On-the-run	Maturity	2y	3y	5y	7y	10y	20y	30y
	Volatility	4.0	3.8	3.5	2.6	4.0	1.6	2.2
Off-the-run	Maturity	≤2y	(2y,3y]	(3y,6y]	(6y,7y]	(7y,11y]	>11y	
	Volatility	20.7	7.9	17.5	8.0	9.6	10.6	

Notes: The table presents the volatilities of aggregate net positions held by primary dealers in on- and off-the-run coupon-bearing Treasuries. The frequency of the used data is weekly and the time horizon is April 2013 to 2025. Source: FED.

A.8.3 Fail rates and net positions regression by maturity

Table 8 presents the same analysis as in Table 2 but using maturity-specific on-the-run fail rates and on-the-run net positions. The regression equation is given by

$$\begin{aligned}\text{On-the-run fail rate}_{m,t} = {} & \alpha + \beta_1 \text{On-the-run primary dealer net positions}_{m,t} \\ & + \beta_2 \ln(\text{MOVE}_t) + \beta_3 \ln(\text{VIX}_t) + \beta_4 \text{Effective federal funds rate}_t \\ & + \beta_5 \text{ 10Y–2Y spread}_t + \beta_6 \mathbb{I}_{10/14,t} + \beta_7 \mathbb{I}_{09/19,t} + \beta_8 \mathbb{I}_{03/20,t} + u_{m,t},\end{aligned}$$

where $m \in \mathcal{M} = \{2\text{Y}, 3\text{Y}, 5\text{Y}, 7\text{Y}, 10\text{Y}, 20\text{Y}, 30\text{Y}\}$ denotes the maturity sector.

A.8.4 Additional robustness test

Table 9 reports regressions using first differences in net positions as a robustness check to the baseline specifications in Table 2, which use levels. Table 10 presents the same analysis as in Table 3 except that the estimation window is increased to twelve months before and after the introduction of the fail charge.

Table 8: Fail rates and net positions regression by maturity

	Dependent variable: On-the-run (OTR) fail rate by maturity						
	(1) 2Y	(2) 3Y	(3) 5Y	(4) 7Y	(5) 10Y	(6) 20Y	(7) 30Y
OTR net positions by mat.	-0.0393***	0.0060	-0.0167**	-0.0382***	-0.1303**	-0.5939	-0.0474
	(0.013)	(0.009)	(0.008)	(0.013)	(0.059)	(0.473)	(0.031)
ln(MOVE)	0.7935***	0.5238***	0.4708***	0.1998**	0.1656	-0.9801	-0.1870
	(0.189)	(0.151)	(0.129)	(0.098)	(0.479)	(5.406)	(0.180)
ln(VIX)	-0.0875	-0.1338	0.0019	-0.1771**	-0.3561	1.1397	0.3039
	(0.162)	(0.090)	(0.115)	(0.087)	(0.504)	(3.246)	(0.278)
Effective federal funds rate	-0.2418***	-0.1511***	-0.1477***	-0.0773***	-0.0983	0.6585	-0.0058
	(0.057)	(0.037)	(0.032)	(0.019)	(0.122)	(0.430)	(0.045)
10Y–2Y spread	-0.4478***	-0.2986***	-0.3101***	-0.2334***	-0.4843*	-0.0599	-0.2692***
	(0.127)	(0.071)	(0.050)	(0.050)	(0.247)	(2.281)	(0.083)
Oct 2014 fixed effect	-0.4734***	0.0792	-0.2091***	-0.2059***	-0.7644**	–	-0.1922*
	(0.112)	(0.128)	(0.060)	(0.053)	(0.323)	–	(0.110)
Sep 2019 fixed effect	0.4371	-0.3407***	-0.2202***	-0.2841***	1.7421**	–	-0.0113
	(0.441)	(0.065)	(0.062)	(0.060)	(0.678)	–	(0.261)
Mar 2020 fixed effect	-0.8068***	0.1500	0.0196	0.5336***	3.1428***	–	2.4547***
	(0.129)	(0.269)	(0.180)	(0.133)	(0.609)	–	(0.800)
Constant	0.5656***	0.4508***	0.4979***	0.4207***	0.9396***	3.1826***	0.6158***
	(0.043)	(0.030)	(0.028)	(0.022)	(0.124)	(0.659)	(0.049)
No. Observations	456	469	629	385	634	174	490
Adj. R^2	0.118	0.102	0.133	0.093	0.026	-0.012	0.100

Notes: I use Newey-West standard errors with 6 lags. *** indicates significance at the 1% level. ** indicates significance at the 5% level. * indicates significance at the 10% level. The sample period is April 2013–2025 and the frequency is weekly.

The net positions of primary dealers can be downloaded from the FED's Primary Dealer Statistics, https://www.newyorkfed.org/markets/counterparties/primary-dealers-statistics. On-the-run net outright positions (in bn USD) are measured as the sum of net outright positions in coupon-bearing on-the-run Treasuries by maturity. The data are deflated in the same way as in Figure 14. The fail rate data (in percent) are the same as in Figure 5b but by maturity. The top 1% of observations within each maturity bucket are excluded. The 20-year Treasury bond was reintroduced in 2020. Maturity-specific data are only available from 2022 onward, so the 20-year sample begins in 2022. The federal funds rate data are obtained from FRED. The MOVE and VIX data are taken from FactSet. The 10-year and 2-year yields (par yields of seasoned notes in percent) are taken from the FED yield curve, which is an updated version of the original Gürkaynak-Sack-Wright curve (Gürkaynak et al., 2010), https://www.federalreserve.gov/data/nominal-yield-curve.htm. The crisis fixed effects capture the months October 2014, September 2019 and March 2020. All independent variables, except the crisis dummies, are demeaned.

Table 9: Fail rates and net positions regression

	(1) Off-the-run fail rate	(2) Off-the-run fail rate	(3) On-the-run fail rate	(4) On-the-run fail rate
Δ Off-the-run net positions	-0.0013***	-0.0012***		
	(0.000)	(0.000)		
Δ On-the-run net positions			-0.0053***	-0.0051***
			(0.002)	(0.002)
ln(MOVE)		0.2510***		0.4356***
		(0.045)		(0.101)
ln(VIX)		-0.2093***		-0.0379
		(0.049)		(0.099)
Effective federal funds rate		-0.1107***		-0.1096***
		(0.010)		(0.029)
10Y–2Y spread		-0.2043***		-0.3126***
		(0.022)		(0.049)
Oct 2014 fixed effect		0.1514***		-0.1476***
		(0.053)		(0.051)
Sep 2019 fixed effect		-0.1806***		0.1087
		(0.049)		(0.087)
Mar 2020 fixed effect		0.0003		1.2623***
		(0.062)		(0.263)
Constant	0.6449***	0.6444***	0.5198***	0.5139***
	(0.017)	(0.012)	(0.034)	(0.024)
No. Observations	593	593	608	608
Adj. R^2	0.009	0.279	0.011	0.217

Notes: I use Newey-West standard errors with 6 lags. *** indicates significance at the 1% level. ** indicates significance at the 5% level. * indicates significance at the 10% level. The sample period is April 2013–2025 and the frequency is weekly.
The net positions of primary dealers can be downloaded from the FED's Primary Dealer Statistics, https://www.newyorkfed.org/markets/counterparties/primary-dealers-statistics. On-the-run net outright positions (in bn USD) are measured as the sum of net outright positions in coupon-bearing on-the-run Treasuries. Off-the-run positions are constructed by subtracting aggregate on-the-run positions from total net outright positions in coupon-bearing Treasuries. The data are deflated in the same way as in Figure 14 and first-differenced. The fail rate data (in percent) are the same as in Figure 5b; outliers are removed by excluding negative observations and the top 5% of observations. The federal funds rate data are obtained from FRED. The MOVE and VIX data are taken from FactSet. The 10-year and 2-year yields (par yields of seasoned notes in percent) are taken from the FED yield curve, which is an updated version of the original Gürkaynak-Sack-Wright curve (Gürkaynak et al., 2010), https://www.federalreserve.gov/data/nominal-yield-curve.htm. The crisis fixed effects capture the months October 2014, September 2019 and March 2020. All independent variables, except the crisis dummies, are demeaned.

Table 10: Fail charge regression with 12 months window

	(1) On-the-run premium	(2) On-the-run premium	(3) Off-the-run fail rate	(4) Off-the-run fail rate
Fail charge introduction dummy	-2.9934***	-2.0553**	-0.9654**	-2.4006***
	(0.530)	(0.806)	(0.396)	(0.915)
ln(MOVE)		2.6158**		2.0626**
		(1.098)		(0.821)
ln(VIX)		2.1713**		-0.9437
		(1.076)		(0.850)
Effective federal funds rate		-0.7773		0.2935
		(0.641)		(0.411)
10Y–2Y spread		0.0317		2.5228**
		(0.904)		(1.048)
Constant	6.7198***	6.2517***	1.2180***	1.9725***
	(0.454)	(0.518)	(0.390)	(0.556)
No. Observations:	497	497	97	97
Adj. R^2	0.294	0.623	0.164	0.535

Notes: I use Newey-West standard errors with 6 lags. *** indicates significance at the 1% level. ** indicates significance at the 5% level. * indicates significance at the 10% level. The sample period is May 2008–April 2010 (a year before and after the introduction of the fail charge). The frequency is daily in Columns (1) and (2) and weekly in Columns (3) and (4).
The data source for the premium (in basis points) is identical to that used in Figure 2a. The premium is defined as the average on-the-run premium across all coupon-bearing Treasuries on a given day. The fail rate data (in percent) are the same as in Figure 5b; outliers are removed by excluding negative observations and the top 5% of observations. The federal funds rate data are obtained from FRED. The MOVE and VIX data are taken from FactSet. The 10-year and 2-year yields (par yields of seasoned notes in percent) are taken from the FED yield curve, which is an updated version of the original Gürkaynak-Sack-Wright curve (Gürkaynak et al., 2010), https://www.federalreserve.gov/data/nominal-yield-curve.htm.