ERASMUS UNIVERSITEIT ROTTERDAM
Erasmus School of Economics

Corporate Bond Transactions and Liquidity: The Probability of Execution

Jean-Paul Guyon van Brakel

A thesis submitted in partial fulfilment of the requirements for the degree of

MASTER IN ECONOMETRICS AND MANAGEMENT SCIENCE

November 8, 2016

Supervisor:
Dr. Michel van der Wel

Co-reader:
Dr. Bart Diris

External supervisors:
Jeroen van Zundert
Mark Whirdy

Student ID: 370549

Abstract

To capture the multiple dimensions of liquidity in the corporate bond market, we combine a model of imputed transaction costs with a model that estimates the fraction of transactions that involve dealer inventory. We connect this to a third model that estimates the arrival rates of buyers and sellers in the market. The resulting expression captures the probability of successfully executing a trade on a bond within a day, for a given target execution cost. We propose this measure, abbreviated as PEX, the Probability of Execution, as a new bond-specific liquidity proxy. By investigating various liquidity determinants that influence the individual models, we are able to identify the bond characteristics and market conditions that influence the PEX. We find that a bond's total amount outstanding, the yield spread, duration and transaction volume of the previous month are the most influential variables. This study provides first evidence that the PEX is a viable alternative to more naive liquidity estimators and can benefit applications ranging from bond selection to trade execution and post-trade transaction cost analysis.

Keywords – Corporate bond market, liquidity, proxy, probability of execution, transaction costs, warehousing rate, arrival rate, cross-sectional.

Acknowledgements

This thesis completes my four-year journey as a student of Econometrics at the Erasmus School of Economics in Rotterdam. This journey leaves me with not only a great set of new skills, knowledge and experience, but has also taught me humility, creativity and persistence. Even more, it has developed me into a better version of myself, with new aspirations and an eagerness to learn more. But I would not have completed this thesis without the help of my supervisor, Michel van der Wel, who I thank for being critical of my decisions and improving my work by providing valuable feedback and numerous good ideas. I also thank, Bart Diris, my co-reader, for questioning my assumptions and making me validate my claims. In addition, I would not have been able to make my thesis practical for investors if it were not for my two external supervisors at Robeco, Jeroen van Zundert and Mark Whirdy. Not only did they give me a warm welcome into the world of asset management in the Investment Research department at Robeco, they also showed me the ropes of fixed income trading, with incredible patience and understanding as I caught up to speed with fixed income terminology and practice. Additionally, they provided me with a short feedback loop by which I could run my ideas and results whenever I needed to. I would also like to thank the other investment professionals at Robeco, for their ideas, enthusiasm and useful insights during my various presentations. Among others, I thank Patrick Houweling and Frederik Muskens for providing feedback on my ideas, and Paul van Overbeek for being my go-to fixed income trader for any questions I had regarding the process of transacting bonds and negotiating with dealers. Furthermore, I thank the credit portfolio managers at Robeco for helping me improve the applicability of my findings through our various meetings.

Finally, I want to thank my parents for their endless support and love, for being a big inspiration to me, and for showing me that nothing is out of reach if you work hard enough to accomplish it. I thank my sisters for making me smile when I needed it, and I thank Rosanne Heeren for her unconditional support and patience throughout my studies.

Completing this part of my life as a student, I hope to never stop learning, for my current knowledge is but the tip of the iceberg. But if there is one thing that I have learned so far, it is that progress is a combination of questioning the status quo and imagining the impossible.

> *"I am enough of the artist to draw freely upon my imagination. Imagination is more important than knowledge. Knowledge is limited. Imagination encircles the world."*
>
> – Albert Einstein (October 26, 1929)

Contents

1. Introduction

Corporate bond liquidity, which constitutes the ease, speed and cost of transacting bonds, is unobserved and therefore difficult to measure. As dealers continue to cut back on market-making activities, the demand for comprehensive liquidity proxies is ever increasing (CGFS, 2016). Due to the poor liquidity conditions, finding liquidity has become a crucial part of investing in corporate debt. To get a grip on liquidity, investors have turned to cost models, such as the Barclays Liquidity Cost Score (LCS), and proxies of market impact, such as the Amihud measure and Roll's model (Ben Dor et al., 2012; Sommer & Pasquali, 2016). Even though such measures sketch a picture of one of the aspects of liquidity, they often neglect that liquidity conditions differ between market participants and they fail to combine other liquidity dimensions. For example, costs are known to decrease with trade size and liquidity also depends on dealer behaviour and market activity (Schultz, 2001). This study aims to bridge the gap between various liquidity aspects and shows that liquidity can be expressed in an intuitive way using probabilities.

To capture the multiple dimensions of liquidity in the corporate bond market, we merge different types of transaction costs and combine it with the arrival rate of buyers and sellers. This gives a proxy for the probability of successfully executing a trade for a given set of bond characteristics and market conditions. We propose this measure, abbreviated as PEX, the Probability of Execution, as a new bond-specific liquidity proxy that can benefit applications ranging from bond selection to trade execution and post-trade transaction cost analysis. Unlike most liquidity proxies, the PEX is purely based on cross-sectional bond characteristics and therefore allows the practitioner to do out-of-sample inference without the need for real-time transactional data.

More specifically, the PEX is based on three individual models: a cost model that estimates imputed transaction costs for different types of bond flows, a warehousing rate model that estimates the fraction of transactions that involve dealers using their inventory, and an arrival rate model that estimates the amount of incoming buyers and sellers in a market. We motivate our approach with a simplified representation of dealer markets in which buyers become sellers and liquidity is removed when investors hold bonds until maturity. Using this representation, we make our cost model conditional on different types of transaction flows. This is comparable to the 'Click-or-Call' framework of Hendershott and Madhaven (2015), in which observed costs are taken conditional on the chosen trading venue: electronic auctions or bilateral trading with a dealer. Hendershott and Madhaven also develop a count model to estimate dealer responses in electronic auctions. We employ the same technique with our arrival rate model in order to describe the arrival distributions of buyers and sellers in the cross-section.

We relate the framework to a set of bond characteristics and market conditions using Generalized Linear Models (GLM). Specifically, GLMs allow us to estimate the dispersion of liquidity, due to their flexible assumptions. Apart from developing the liquidity framework, this thesis also

aims to find the set of liquidity determinants that have the largest effect on both the individual models and the PEX. In order to do so, we derive approximations for the partial effects of such determinants and employ cluster robust standard errors to control for latent liquidity influences and potential model misspecification. The data we use to estimate the framework is the enhanced dataset from FINRA's Trade Reporting and Compliance Engine (TRACE), for the period between 2005 and 2013. We link the transactions from TRACE to corporate bond characteristics and market conditions using constituent data from the Barclays U.S. Corporate Investment Grade index. Although the proposed framework seems applicable to high yield bonds as well, this thesis investigates the liquidity of investment grade bonds only. Taking the intersect of both datasets, we end up with 15,489 unique CUSIPs and 57,280,531 filtered transactions. We also download market indices, such as the CBOE Volatility Index (VIX), to proxy overall market conditions as the VIX acts as a gauge for aggregate 'fear' in the financial markets.

The methodology we use to impute transaction costs is based on the work of Feldhütter (2012), who uses it to identify selling pressure in corporate bonds. Feldhütter imputes transaction costs by observing that corporate bond transactions are reported to TRACE in clusters: dealers often prearrange transaction flows before executing them. Once the transaction is set in motion, the dealer reports two transactions to TRACE: one record indicating the transfer of bonds from the seller to the dealer and a seperate record for the transfer between the dealer and the buyer. These transactions have the same reported transaction size and are reported in quick succession of each other. Feldhütter uses the imputed costs in two ways. First of all, he finds that the price difference between small trades and large trades at a given point in time is representative of selling pressure in the market, implying that there are more sellers than buyers. Secondly, he uses the imputed costs in a theoretical search model, where investors bargain with dealers about prices when initialised with random search intensities. By estimating the model with TRACE data, Feldhütter finds elaborate evidence of different liquidity conditions for buyers and sellers, and for different transaction sizes. Feldhütter's results bear an important lesson: given that liquidity is bifurcated between the buy and sell side and highly dependent on the bargaining power of the investor, it is ill-advised to generalise a liquidity proxy to all market participants.

Using the enhanced TRACE data, we are able improve Feldhütter's (2012) imputation methodology by identifying exactly whether a buying or selling customer is involved in a trade. This gives us the opportunity to find not just the seller-dealer-buyer roundtrips, but also identify other combinations of pre-arranged transactions. Like Feldhütter, we employ a 15 minute detection window. Our results prove to be relatively robust to the chosen window, where increasing the window leads to slightly different average costs. Using the enhanced detection method and deleting observations not covered by the Barclays data, we end up with 13,277,112 total bond flows. These include both single transactions and combinations of pre-arranged transfers. We use all of these observations in the warehousing and arrival rate models, but we can only use combinations of two or more transfers to impute transaction costs for the cost model.

The results of this thesis indicate that the PEX is able to give a full picture of bond-specific liquidity by combining the explanatory power of the three models. The bond characteristics that improve the PEX the most are the total amount outstanding and the trading volume of the previous month. As expected from Schultz (2001) and subsequent literature, we also observe that a higher transaction size yields a more liquid environment. On the contrary, an increase in the duration and age of a bond gives the largest decrease in the probability of execution, especially for lower values. Additionally, the level of the VIX and the yield spread also decrease the PEX. The effect of the yield spread is surprisingly small, caused by the fact that it coincides with both expensive transaction costs and higher market activity. In total, we find that liquidity circumstances are slightly better for sellers than for buyers. This is the result of three effects: buying is on average more expensive than selling, dealers tend to use their inventory more for buyers than for sellers, and buyers arrive more frequently than sellers.

Our cost model results confirm that of Harris and Piwowar (2006) and Edwards et al. (2007). We find similar effects for the age of a bond, its callability, the price and the amount outstanding. In the same fashion, our results confirm most of those from Harris (2015). Specifically, we find similar estimates for the size of transactions and the average transaction size of trades on a bond. The cost results of Hendershott and Madhaven (2015) also largely coincide with ours. Unfortunately, we did not have the data to include the daily absolute stock return and the treasury drift. We believe that these would also bring a valuable addition to our cost model.

This thesis differs from related literature in a couple of ways. First of all, we observe from the dealer markups that it is better to denominate transaction costs in dollar cents, not basis points. We find that denominating costs in basis points ex ante, can lead to a 'price effect' such as observed by Harris (2015). By denominating costs in dollar cents instead, the explanatory power of price is almost completely eliminated when controlling for yield spread and duration. This gives a second difference: we include bond duration and yield spread instead of maturity. This because DTS, duration times spread, is related to future volatility and thereby also bond liquidity (Ben Dor et al., 2007). Indeed, we find that DTS explains a lot of cross-sectional variation in all models. For the final framework, we split DTS into separate terms because it yields better performance. Lastly, we employ the log transform for our continuous regressors, opposed to Edwards et al. (2007) who prefer taking the square root. We find that corporate bond liquidity quickly deteriorates as bond characteristics become less favourable, but the deterioration slows down for illiquid bonds.

Apart from proposing the PEX, this thesis contributes to the corporate bond liquidity literature in two other ways. First of all, this thesis extends Feldhütter's (2012) research by shedding light on the relation of selling pressure to cross-sectional liquidity determinants. Feldhütter does not make the distinction between different types of transaction flows, although he acknowledges that the different types can lead to different costs. Because we are able to identify the difference between buy and sell costs, we are able to make transaction costs conditional on whether dealers

use inventory or not. As a result, we can discern between the liquidity conditions of buyers and sellers. The PEX is therefore consistent with Feldhütter's findings.

The second contribution of this thesis is that we make the first step towards combining multiple liquidity dimensions in a probabilistic setting. Harris (1990) divides liquidity into four dimensions: *width*, *depth*, *immediacy* and *resiliency*. A fifth dimension, *breadth*, was proposed by Lybek and Sarr (2002). These dimensions are well defined for exchange traded limit-order markets, in which everybody can observe the quoted prices and volumes (Appendix A). Such definitions are based on the premise that transactions always take place against the best prevailing price. Dealer markets have no such guarantee: two market participants can transact the same quantity of assets at the same time, but with wildly varying prices (Feldhütter, 2012; Harris, 2015). We must therefore adjust the definitions of the liquidity dimensions to fit dealer markets. Sommer and Pasquali (2016) propose to think of liquidity as a distribution of costs in a probability space conditioned on transaction size and market impact. Following their proposition, we redefine the liquidity dimensions by using our cost model to describe the probability distributions of realised transaction costs. This is visualised in Figure 1. *Width* is taken as the smallest difference between bid and ask executions, *depth* as the cost for which we observe the average probability of success and *breadth* as the shape of the distribution, measured by the skewness and kurtosis.

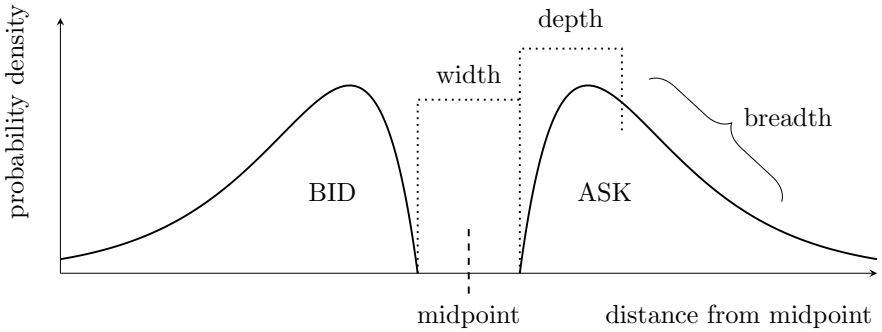


Figure 1: Liquidity dimensions for the probability density of transaction costs

There is no consensus in the liquidity literature on how to measure *immediacy*. We propose a new way of measuring immediacy with the arrival rate model. We find that the estimation of arrival rates greatly contributes to our liquidity framework, yielding additional information from the cost model. We are not able to measure *resiliency* because it depends the behaviour of individual dealers, for which we do not have information. By combining the *width*, *depth*, *breadth* and *immediacy* dimensions in the PEX measure, this thesis makes a first step towards Sommer and Pasquali's proposal (2016) of expressing multiple liquidity dimensions in a single probability.

The remainder of this thesis starts with the development of the PEX measure in Section 2. We explain estimation procedures in Section 3 and introduce our data in Section 4. We show and discuss our results in Section 5 and complete the thesis with a conclusion in Section 6.

2. Models

In this section we introduce our methodology for estimating the Probability of Execution (PEX). We base the PEX on a conceptual transaction flow framework that captures multiple dimensions of liquidity. We denote sellers as 'S', dealers as 'D' and buyers as 'B'. Immediate roundtrips are then represented as 'SDB', signifying the flow of bonds from left to right. By definition of a dealer market, all transactions involve a dealer. If buyers and sellers arrive at the same time, the dealer is able to execute instantaneous roundtrips (SDB). If not, the dealer either transacts against his own inventory (SD or DB) or transfers the bonds to another dealer (SDD or DDB). In the latter case, the bonds in the transaction either come from a dealer's inventory or end up in a dealer's inventory. Let γ_B denote the fraction of buy trades that are sold from a dealer's inventory or short-sold using the repurchase agreement (repo) market. Likewise, let γ_S be the fraction of sell trades where a dealer takes bonds into inventory. Buyers arrive in the market with rate λ_B and sellers arrive with rate λ_S. Together, this gives the framework in Figure 2a.

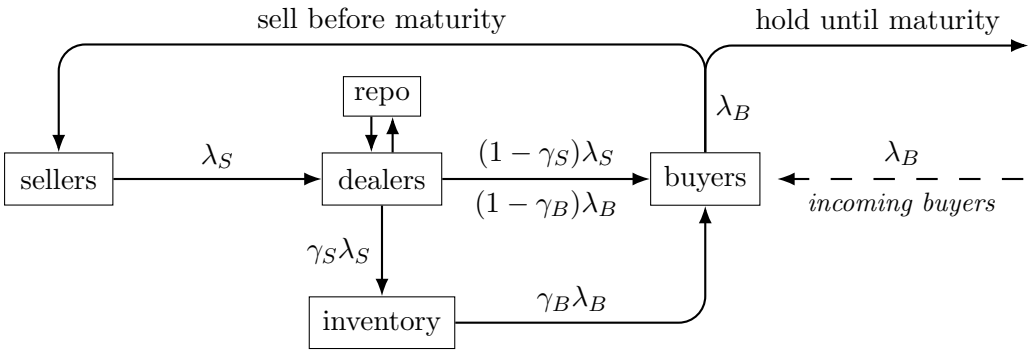


Figure 2a: Overview of bond flows in a dealer market, arrows indicate bond transfers

Given the bond flows in Figure 2a, we can now relate different flows in this framework to observed transaction combinations. This is displayed in Figure 2b. Instantaneous roundtrips appear as 'SDB' transactions. Sell transactions where a dealer absorbs bonds into inventory appear as 'SD' and 'SDD'. Buy transactions involving dealer inventory appear as 'DB' and 'DDB'.

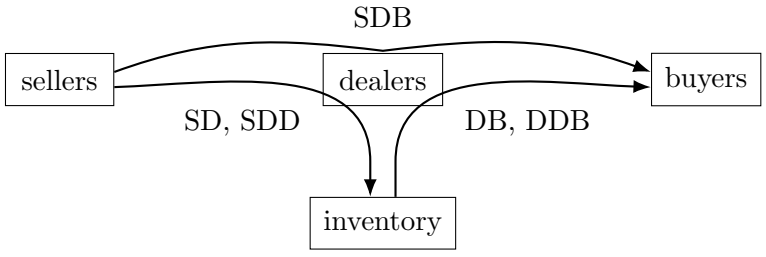


Figure 2b: Overview of transaction cost combinations, arrows indicate bond transfers

The framework is based on a couple of assumptions concerning market restrictions. We assume that a customer is able to sell his bonds in three cases: the dealer knows a buyer that is willing

to buy the bonds, the dealer is willing to take the bonds into inventory, or the dealer can transfer the bonds to another dealer. For buyers, we assume that transactions are possible if the dealer knows a seller, the dealer owns the bonds, the dealer buys the bonds from another dealer or the dealer is willing to short sell bonds through the interdealer repo market. We assume that if dealers short sell bonds to buyers, they will buy the bonds back as soon as possible[1]. Because selling customers generally do not have direct access to the interdealer repo market, we assume that short selling by customers is not possible. By this assumption, the total flow in the framework is less than or equal to the amount of incoming buyers. Incoming buyers have two options: they can keep their purchased bonds until maturity or they can sell the bonds. Over the lifetime of the bond, the amount of incoming sellers is therefore bounded from above by the amount of incoming buyers: $\lambda_S \leq \lambda_B$. If all buyers keep their bonds until maturity, liquidity dries up and λ_S converges to zero. In essence, liquidity is added to the system when buyers arrive and removed when buyers hold their bonds until maturity. This can be seen from Figure 2a, where bonds can only leave the system if buyers hold them until maturity.

The inspiration for this framework comes from the 'Click-or-Call' decision model of Hendershott and Madhaven (2015). They develop a venue selection model to estimate the difference in transaction costs when trading in an electronic auction instead of engaging in bilateral trading with a single dealer. They estimate the probability that an investor chooses a specific venue with a probit model and use the outcome in a venue-specific cost model to account for possible selection bias[2]. The different 'venues' can be compared to the different transaction types in our framework. They also estimate the amount of positive dealer responses to electronic auctions and use it as a proxy for getting a successful electronic execution. This can be compared to the arrival rates in our framework, which we use to proxy the immediacy dimension of liquidity.

We estimate the various parts of the framework using three models. The first model estimates imputed transaction costs (hereinafter '*cost model*') of the various transaction combinations. All possible combinations are SDB, SDD, SD, DD, DB and DDB[3]. We can only infer costs from transaction combinations that involve at least two bond transfers (SDB, SDD, DDB). The second model estimates the fraction of trades that dealers take into inventory for at least 15 minutes (hereinafter '*warehousing rate model*'). The last model estimates the arrival rate of buyers and sellers in the market (hereinafter '*arrival rate model*'). After development of the models, we complete this section by explaining how the models are combined into a probability of execution.

[1] There are three main reasons why dealers want to buy back the borrowed bonds as soon as possible: market risk (the bond price may rise), running costs (the cost of carry) and possible penalty costs (if the short-seller fails to deliver the bonds before settlement). Most dealers minimise potential costs by short offering liquid bonds only.

[2] Hendershott and Madhaven (2015) assume that investors choose the venue with the lowest expected cost. This means that the realised cost of transacting at a specific venue are observed conditional on the assumption that the expected cost at that venue is lower than at the other venue. This could cause a selection bias, which they account for by using inverse Mills ratios from the probit regression and including it in their cost model.

[3] Even though rare, if more than two dealers are involved in a transaction we also define them as SDB, SDD or DDB. For example, transactions that involve one buyer and three dealers are classified as DDB even though the correct representation would be DDDB. The same holds for SDB and SDD combinations.

2.1. The cost model

To estimate transaction costs, we employ a generalized linear model with appropriate distribution and link function. A generalized linear model is suitable for modelling transaction costs because costs are truncated at zero by definition. Additionally, the variance of transaction costs is approximately constant when measured on a logarithmic scale. Taking the logarithm of costs is not desirable because it would transform both the linearity and the variance of the data[4]. Given that the coefficient of variation is also approximately constant, we argue that a generalized linear model is an appropriate choice due to its flexibility and independence of transformations of the original data. We employ separate regressions for different trade sizes, similar to Edwards, Harris and Piwowar (2007). We group sizes as odd-lot [\$0–\$100k), round-lot [\$100k–\$1mm) and block sized [\$1mm, ∞). The trade types we can estimate are SDB, SDD and DDB because we need at least two prices to be able to impute a cost.

Let $\eta_{gp} = X_{gp}\beta_{gp}$, for design matrix X_{gp} and vector of coefficients β_{gp} for transactions of type p and size s belonging to group g. We can then estimate the expected cost as follows:

$$\mathbb{E}[C_{gp}|X_{gp}] = \mu_{gp} = g^{-1}(\eta_{gp}) \tag{1}$$

$$\mathbb{V}ar[C_{gp}|X_{gp}] = V(\mu_{gp}) = V(g^{-1}(\eta_{gp})) \tag{2}$$

$$\text{for} \quad g = \{s \in [0, 100\text{k})\},\ \{s \in [100\text{k}, 1\text{mm})\} \text{ or } \{s \in [1\text{mm}, \infty)\},$$
$$\text{and} \quad p = \text{SDB},\ \text{SDD or DDB}$$

Where $g^{-1}(\cdot)$ is the inverse link function and $V(\cdot)$ a function of the expected cost. The estimation error $C_{gp} - \mathbb{E}[C_{gp}|X_{gp}]$ is expected to follow a specific distribution. We illustrate the results of various model specifications of the cost model in Appendix I. We find that a log-link with gamma distribution is the most appropriate specification for transaction costs in the corporate bond market. This yields $\mathbb{E}[C_{gp}|X_{gp}] = \exp(\eta_{gp})$ and $\mathbb{V}ar[C_{gp}|X_{gp}] = V(\mu_{gp}) = \phi_{gp}\mu_{gp}^2$. The parameter ϕ_{gp} is the dispersion of the regression, which is estimated separately (Appendix B).

This model provides a broad representation of liquidity. The estimated parameters of the model can be used to uncover the parameters of the underlying distribution. The distribution can be interpreted as the probability of observing a specific cost, given that a transaction takes place. We can use this to proxy the liquidity dimensions width, depth and breadth (Section 1). The cost model is therefore not only useful for estimating expected costs, but can also be used to estimate the liquidity characteristics of a given market. This is further explained in Section 2.4.

[4] Jensen's inequality tells us that predictions from a regression with a transformed distribution of the dependent variable, e.g. lognormal, can be systematically biased. This because if the true distribution is not equal to the transformed distribution, the transformation will incur additional estimation error to the regression. This can be prevented by estimating the transformed expected value of the data, opposed to transforming the data itself.

2.2. The warehousing rate model

The fraction of trades that are either sold from, or absorbed in, dealer inventory is estimated with a probit model. The warehousing rates of buy and sell trades, γ_B and γ_S, are estimated by classifying whether individual trades involve dealer inventory. Let y_g^B be a vector of binary responses within size group g, where element y_{ig}^B is 1 if buy trade i was type DB or DBB and 0 otherwise. The same holds for y_g^S, for which element y_{ig}^S is 1 if sell trade i was type SD or SDD and 0 otherwise. Given a design matrix X_g for size group g, we estimate the following models:

$$\mathbb{P}[y_g^B = 1|X_g] = \Phi\left(X_g\delta_g^B\right) \qquad \textit{for buy trades} \tag{3}$$

$$\mathbb{P}[y_g^S = 1|X_g] = \Phi\left(X_g\delta_g^S\right) \qquad \textit{for sell trades} \tag{4}$$

$$\text{for} \quad g = \{s \in [0, 100\text{k})\},\ \{s \in [100\text{k}, 1\text{mm})\} \text{ or } \{s \in [1\text{mm}, \infty)\},$$

With δ_g^B and δ_g^S the vector of coefficients for buy and sell trades with size s in group g, respectively. We are mainly interested in the differences of the warehousing rates γ_B and γ_S in the cross-section. Therefore, we designed the study in such a way that the regressors explain cross-sectional variation between bonds, not between transactions on the same bond. By construction, the model therefore has low explanatory power to classify individual transaction types. Instead, we expect to find explanatory power in the cross-section as warehousing rates should vary with bond characteristics and market conditions.

To find the warehousing rates γ_B and γ_S, we simply use the expected probability that a trade involves dealer inventory. This because the probability of observing an outcome of a binary random variable is equal to the expected fraction of random draws with that outcome[5]. If we assume that the probability that any transaction goes through inventory is independent of other transactions, we find that the fraction of trades that involve dealer inventory equals the probability that any single transaction involves inventory:

$$\gamma_B(g, X_g) = \mathbb{P}[y_g^B = 1|X_g] \tag{5}$$

$$\gamma_S(g, X_g) = \mathbb{P}[y_g^S = 1|X_g] \tag{6}$$

We assume that dealers are rational and always execute instantaneous SDB roundtrips if possible. Because instantaneous roundtrips are effectively arbitrage opportunities for dealers, this is a feasible assumption (we explain the role of dealers in Appendix C). In our framework, the possibility of an instantaneous roundtrip thereby dictates what type of transaction we observe. Therefore, because the transaction combinations are mutually exclusive, we do not need to include a selection bias correction in our cost model as in Hendershott and Madhaven (2015).

[5]In essence, if we have n independent Bernoulli random variables X_i for $i = 1, ..., n$ such that $\mathbb{P}[X_i = 1] = p$ for all i, then $\mathbb{E}[\sum_{i=1}^n X_i] = np$. Clearly, the expected fraction of realisations equal to 1 is $np/n = p$. The fraction is therefore equal to the probability, which coincides with the intuition behind a binomial distribution.

2.3. The arrival rate model

To model the arrival flows λ_B and λ_S of buyers and sellers, we use a count model. The amount of incoming customers per fixed time period can also be interpreted as the 'rate' at which customers arrive. We estimate the amount of arriving buyers N^B_{btg} of size s or larger in group g for bond $b = 1, ..., K$ on day $t = 1, ..., T$ as the outcome of a Poisson distribution with conditional mean:

$$\mathbb{E}[N^B_{btg}|X_{btg}] = \lambda_B(X_{btg}) \tag{7}$$

$$\lambda_B(X_{btg}) = \exp\left(X_{btg}\kappa^B_g + \eta_{btg}\right) \tag{8}$$

$$\begin{aligned} \text{for bond} \quad & b = 1, ..., K, \\ \text{day} \quad & t = 1, ..., T, \\ \text{and group} \quad & g = \{s \in [0, \infty)\},\ \{s \in [100\text{k}, \infty)\} \text{ or } \{s \in [1\text{mm}, \infty)\} \end{aligned}$$

with $N^B_{btg} = 0, 1, 2, ...$, and κ^B_g the vector of coefficients for group g. The error η_{btg} captures individual variation in bonds. To allow for unobserved cross-sectional heterogeneity, we assume that η_{btg} follows the gamma distribution $Gamma(\theta_B, 1)$. This yields a negative binomial regression model with shape parameter θ_B and scale set to one. The negative binomial regression is therefore a generalisation of a poisson regression with support for overdispersion[6]. The parameter θ can be interpreted as the dispersion of the amount of arrivals. This count model also allows for zero outcomes, which is important as illiquid bonds have no trades on most days. We repeat the above estimation procedure to estimate the arrival rate of sellers as well. The corresponding notation is the same, except that parameters are denoted with 'S' instead of 'B'.

The probability of n_{btg} buyers arriving on day t for bond b, conditioned on the regressors X_{btg} of size group g can now be written as:

$$\mathbb{P}[N^B_{btg} = n_{btg}|X_{btg}] = \frac{\Gamma(n_{btg} + \theta_B)}{\Gamma(\theta_B)\Gamma(n_{btg} + 1)} \left(\frac{\theta_B}{\theta_B + \lambda_B(X_{btg})}\right)^{\theta_B} \left(\frac{\lambda_B(X_{btg})}{\theta_B + \lambda(X_{btg})}\right)^{n_{btg}}, \tag{9}$$

$$\begin{aligned} \text{for amount} \quad & n_{btg} = 0, 1, 2, ..., \\ \text{bond} \quad & b = 1, ..., K, \\ \text{day} \quad & t = 1, ..., T, \\ \text{group} \quad & g = \{s \in [0, \infty)\},\ \{s \in [100\text{k}, \infty)\} \text{ or } \{s \in [1\text{mm}, \infty)\} \end{aligned}$$

where $\Gamma(\cdot)$ is the gamma function, n_{btg} are a given number of arriving customers and θ_B is the shape parameter of the negative binomial distribution. Note that the size groups overlap as any group g also contains all higher size groups. The rationale for this is provided in the next section.

[6] In a poisson distribution, the variance equals the mean. Overdispersion in a poisson model occurs when the conditional variance is higher than the conditional mean. A negative binomial regression accounts for overdispersion by estimating the additional parameter θ in the variance term of the error distribution.

2.4. The probability of execution (PEX)

Now that we have established the three individual models, we can use them to make probabilistic inferences of bond-specific liquidity and construct the PEX measure. Before we develop these expressions, we first get a better grip on expected buy and sell costs. We combine the cost and warehousing rate model to calculate a weighted average of the different transaction types:

$$\textit{expected buy costs:} \qquad \mathbb{E}[C_B] = \widehat{\gamma}_B \mathbb{E}[C_{\text{DDB}}] + (1 - \widehat{\gamma}_B)\mathbb{E}[C_{\text{SDB}}] \tag{10}$$

$$\textit{expected sell costs:} \qquad \mathbb{E}[C_S] = \widehat{\gamma}_S \mathbb{E}[C_{\text{SDD}}] + (1 - \widehat{\gamma}_S)\mathbb{E}[C_{\text{SDB}}] \tag{11}$$

Next, we want to express the probability of observing the execution of a trade of size s for a given cost target or better. We do this by transforming the expected value from the cost model into the underlying cumulative probability function. The cost model regression yields an estimated dispersion parameter $\widehat{\phi}$ from which we can find α and β in $Gamma(\alpha, \beta)$. Specifically, for any set of bond characteristics and market conditions X_{gp}, we can use the model to find an estimated cost $\widehat{c}_{gp} = \mathbb{E}[C_{gp}|X_{gp}]$ for a transaction of type p in size group g. We then find the parameters of the gamma distribution as follows: $\widehat{\alpha}_{gp} = 1/\widehat{\phi}_{gp}$, $\widehat{\beta}_{gp} = \widehat{\alpha}_{gp}/\widehat{c}_{gp}$. We can now find the probability of observing the target cost c or better with the CDF of the gamma distribution:

$$\mathbb{P}[C_{gp} \leq c|X_{gp}] = \frac{\gamma(\widehat{\alpha}_{gp}, \widehat{\beta}_{gp}c)}{\Gamma(\widehat{\alpha}_{gp})} = \frac{\gamma(\widehat{\phi}_{gp}^{-1}, c\,\widehat{\phi}_{gp}^{-1}\widehat{c}_{gp}^{-1})}{\Gamma(\widehat{\phi}_{gp}^{-1})} \quad \text{with} \quad \gamma(\alpha, \beta x) = \int_0^{\beta x} t^{\alpha-1}e^{-t}\,\mathrm{d}t \tag{12}$$

Here $\Gamma(\cdot)$ is the gamma function and $\gamma(\widehat{\alpha}_{gp}, \widehat{\beta}_{gp}c)$ the lower incomplete gamma function evaluated at the target cost c. This expression is dependent on X_{gp}, because of the estimated $\widehat{c}_{gp}$.

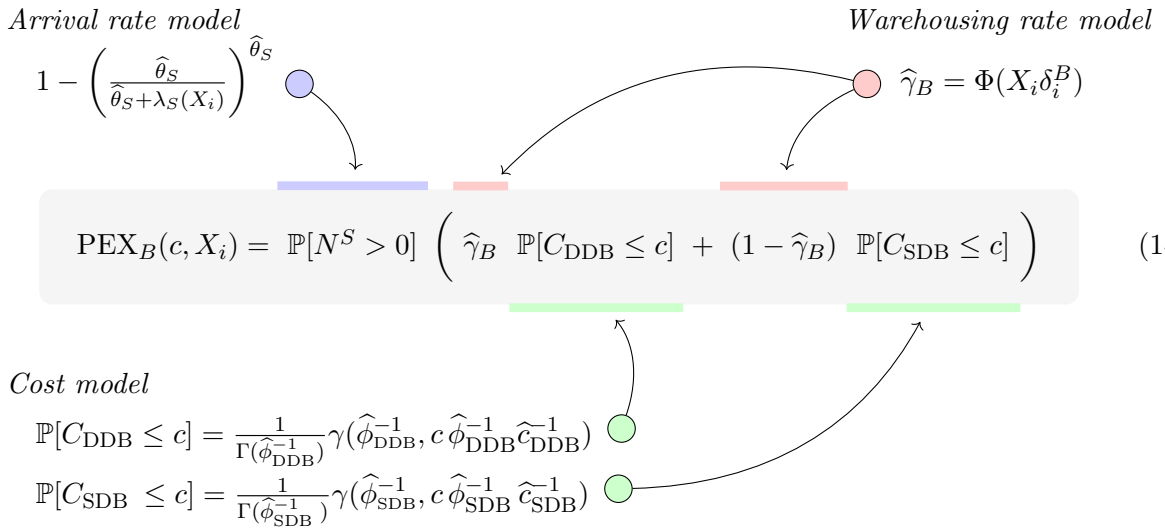
To measure the immediacy with which transactions can be executed, we assume that transactions can be executed with the same rate at which the opposite party arrives. For instantaneous transactions (SDB), this assumption is true. For the other transaction types we assume that dealers are willing to take the bonds into inventory with the same rate at which buyers arrive, given that they sell their inventory to buyers. Dealers are thereby believed to smooth market frictions with regard to transaction time and size. We argue the same for buyers, given that liquidity is provided by incoming sellers. For newly issued bonds, the rate at which buyers are accommodated can be larger than the rate at which sellers arrive because dealers hold a lot of inventory. We might therefore underestimate the buy-side liquidity circumstances of new issues. We also assume that any transaction in group g can be offset by an opposite transaction in the same group or higher. For example, any buy transaction with size $s_i \in [100\text{k}, 1\text{mm})$ is offset by any incoming sell transaction with size $s_j \in [100\text{k}, \infty)$, even if $s_i > s_j$. We assume that market frictions within size groups are absorbed by dealers. This assumption is consistent with the phenomenon that executing smaller transactions is easier than transacting larger ones. Nevertheless, there is no guarantee that large market frictions can be absorbed by the dealer. Our immediacy proxy can therefore be upward biased for very large transaction sizes.

To proxy the rate at which liquidity is available, we use the arrival rate model to find the probability that at least one opposite customer arrives. Let X_{gp} be the vector of regressors and $\widehat{\theta}$ the estimated shape parameter of the negative binomial distribution. For a buyer of bond b in size group g, the probability that a seller is available in group g or higher is then:

$$\mathbb{P}[N^S_{bg} > 0|X_{bg}] = 1 - \mathbb{P}[N^S_{bg} = 0|X_{bg}] = 1 - \left(\frac{\widehat{\theta}}{\widehat{\theta} + \lambda_S(X_{bg})}\right)^{\widehat{\theta}} \tag{13}$$

$$\text{with} \quad g = \{s \in [0, \infty)\},\ \{s \in [100\text{k}, \infty)\} \text{ or } \{s \in [1\text{mm}, \infty)\}$$

Finally, we combine the three models into a single probabilistic estimate. We do this by taking the CDF of equation (12) for the different transaction combinations and weighing them with the warehousing rate model. We multiply this with equation (13): the probability that an opposite party is available. For the set of bond characteristics X_i and target cost c, we now have our proposed liquidity measure, the **Probability of Execution**:



(14)

The PEX is a comprehensive measure that combines multiple dimensions of bond-specific liquidity. It estimates the *width*, *breadth* and *depth* of the market with the transaction cost term. The *immediacy* of being able to execute a transaction is proxied with the probability that an opposite party supplies liquidity. We assume that the probability of realising a cost of given type is independent of the probability that an opposite party is available. Additionally, we assume that dealer always prefer instantaneous roundtrips over other trades. The probability of observing a given transaction type is therefore independent of observing another type. For completeness, the PEX measure for both the buy and sell side are as follows:

$$\text{PEX}_B(c, X_i) = \mathbb{P}[N^S > 0]\,(\widehat{\gamma}_B \mathbb{P}[C_{\text{DDB}} \leq c] + (1 - \widehat{\gamma}_B)\mathbb{P}[C_{\text{SDB}} \leq c]) \qquad \textit{for buying}$$

$$\text{PEX}_S(c, X_i) = \mathbb{P}[N^B > 0](\widehat{\gamma}_S \mathbb{P}[C_{\text{SDD}} \leq c] + (1 - \widehat{\gamma}_S)\,\mathbb{P}[C_{\text{SDB}} \leq c]) \qquad \textit{for selling}$$

3. Methods

The three models that appear in this thesis can be estimated using traditional GLM estimation procedures (McCullagh & Nelder, 1989; Dobson & Barnett, 2008). This because all three models make use of a distribution from the exponential family. The cost model uses a gamma distribution, the warehousing rate model can be written as a GLM with binomial distribution and probit link function, and the arrival rate model is a GLM with a negative binomial distribution. The binomial and negative binomial distribution are part of the exponential family if the number of trials (or failures) is fixed. The estimation theory of generalized linear models is explained in Appendix B. In this section we explain how we estimate two-way clustered parameter variances and how we find the partial effects of variables on the individual models and the PEX.

3.1. Variance estimation and two-way error clustering

A disadvantage of using GLMs is that estimating the variance of the coefficients is nontrivial. There is no closed-form solution to the variance, such that we must resort to a Taylor approximation of the log-likelihood function around the estimated coefficients. This approach assumes that the model is correct and that there exists a true set of parameters θ_0. We can now use the Taylor approximation to find the distance between the set of estimated coefficients $\widehat{\theta}$ and the true coefficients θ_0 as $\widehat{\theta} - \theta_0$. The Taylor expansion approximates the likelihood function $l_n(\widehat{\theta})$ around θ_0 with sample size n as follows:

$$l_n(\widehat{\theta}) \approx l_n(\theta_0) + l_n'(\theta_0)(\widehat{\theta} - \theta_0) + \frac{1}{2}(\widehat{\theta} - \theta_0)^T l_n''(\theta_0)(\widehat{\theta} - \theta_0) \tag{15}$$

In order to maximise this likelihood, we set $l_n'(\theta) = 0$ and drop the last term. This yields:

$$l_n'(\widehat{\theta}) \approx l_n'(\theta_0) + (\widehat{\theta} - \theta_0)^T l_n''(\theta_0) = 0 \tag{16}$$

$$\widehat{\theta} - \theta_0 \approx \left(-l_n''(\theta_0)\right)^{-1} l_n'(\theta_0)^T \tag{17}$$

Now if we take the variance of $\widehat{\theta}$, we get $\mathbb{V}ar[\widehat{\theta}|\theta_0] = \mathbb{E}[(\widehat{\theta} - \theta_0)^2]$ because $\widehat{\theta}$ is unbiased. Hence:

$$\mathbb{V}ar[\widehat{\theta}|\theta_0] = \left[-l_n''(\theta_0)\right]^{-1} \left[\mathbb{V}ar[l_n'(\theta_0)|\theta_0]\right] \left[-l_n''(\theta_0)\right]^{-1} \tag{18}$$

where the covariance matrix is symmetric. By using asymptotic theory, for example that $\frac{1}{n} l_n'(\theta_0) \xrightarrow{P} 0$, and estimating the likelihood functions directly from the sample data by evaluating them at $\widehat{\theta}$, we yield the well-known asymptotic Huber sandwich estimator (Geyer, 2003):

$$\widehat{\theta} \overset{\mathcal{D}}{\sim} N(\theta_0, [-l_n''(\widehat{\theta})]^{-1} \, \widehat{\mathbb{V}}ar[l_n'(\widehat{\theta})] \, [-l_n''(\widehat{\theta})]^{-1}) \tag{19}$$

With $N(\mu, \sigma)$ the normal distribution and $\widehat{\mathbb{V}}ar[l_n'(\widehat{\theta})]$ the variance matrix of the first partial derivative of the likelihood function evaluated at $\widehat{\theta}$, where the sample size n is sufficiently large.

Now we can work out the first and second partial derivative of the log-likelihood as follows:

$$l'(\theta) = \sum_{i=1}^{n} g_i(y_i|\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(y_i|\theta_i, \phi) \tag{20}$$

$$l''(\theta) = \sum_{i=1}^{n} h_i(y_i|\theta) = \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log f(y_i|\theta_i, \phi) \tag{21}$$

where $f(y_i|\theta_i, \phi)$ is the conditional probability density of the data, further developed in equation (43) in Appendix B. If we substitute first and second partial derivates $l'(\theta)$ and $l''(\theta)$ back into the sandwich estimator, we can write the asymptotic variance of $\widehat{\theta}$ as:

$$\mathbb{V}ar[\widehat{\theta}] \overset{\mathcal{D}}{\approx} \left[-\sum_{i=1}^{n} h_i(y_i|\widehat{\theta}) \right]^{-1} \left[\sum_{i=1}^{n} g_i(y_i|\widehat{\theta})^T g_i(y_i|\widehat{\theta}) \right] \left[-\sum_{i=1}^{n} h_i(y_i|\widehat{\theta}) \right]^{-1} \tag{22}$$

For all models, we adjust the standard errors with two-way clustering on day and bond issue. This because the residuals can be clustered on individual days due to overall market conditions that the model cannot explain. Additionally, transactions that happen on the same day and on the same bond can have the same explanatory variables, such that their contribution to the variance of the estimators must be adjusted. We also cluster on bond issues to adjust for bond-specific biases. This because a bond that is classified as relatively liquid under the cost model may actually display unusually high transaction costs due to some external effect that is not captured by the regressors. Error clustering is therefore important because ignoring potential error correlations within clusters can lead to erroneous statistical inference.

The two-way clustered asymptotic variance of $\widehat{\theta}$ can be found by summing over the clusters first when calculating the covariance matrix of the sandwich estimator. We assume that clusters are independent, but observations within clusters can have correlated errors. For the two way clustering, we cluster on both bond b and day t. Let c_b denote the set of observations of bond b and let c_t be the set of observations that occur on day t. Implementing the two-way clustering, the middle part of the sandwich estimator becomes the following:

$$\sum_{b} \sum_{t} \left[\sum_{i \in c_b \cup c_t} g_i(y_i|\widehat{\theta}) \right]^T \left[\sum_{i \in c_b \cup c_t} g_i(y_i|\widehat{\theta}) \right] \tag{23}$$

where the outer two sums go over all bonds b and days t. The two outside terms in equation (22) remain the same. As shown, we estimate the dependence between observations if they share either the same bond c_b or the same day c_t, such that both observations are in $c_b \cup c_t$. This adjusted sandwich estimator is robust under model misspecification and takes into account cluster-specific dependencies when estimating the covariance between observations.

3.2. Partial effects of variables

In order to find the effect sizes of individual variables on the three models and the PEX measure, we need to estimate partial effects. We calculate partial effects in three ways. First of all we can factor out the effect from the original model if the model specification allows for this. Second of all, we can take the partial derivative of a model with respect to a variable x_i, and use the derivative to linearise the effect of the variable for a given change Δx_i. The derivative is a linear approximation of the function near x_i and works relatively well for small changes Δx_i. The third method is to find the ratio between the function $f(x_1, ..., g(x_i), ..., x_k)$ and $f(x_1, ..., x_i, ..., x_k)$, where $g(x_i)$ is some projection of x_i and the model f takes k variables. This is exact for single observations but needs to be averaged for the partial effect over the full sample.

The partial effect of variables in a GLM with log-link function

Assume a GLM model with link function g such that $g(\mathbb{E}[Y]) = X\beta$. We can estimate this model as $Y = g^{-1}(X\beta) + \varepsilon$ where some distribution is assumed for ε. In order to interpret the estimated coefficients $\widehat{\beta}$, we need to understand the partial effects of the individual variables. For non-transformed variables, an additive change in the variable has a partial effect that is multiplicative under a log-link function. Given the vector of variables X_i of observation i, the estimated expected value of the dependent variable is:

$$\mathbb{E}[Y|X_i] = \exp\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_i x_i + \cdots + \widehat{\beta}_k x_k\right)$$

Given an additive change m in variable x, the partial effect under log-link is:

$$\begin{aligned}
\mathbb{E}[Y|x_1, x_2, ..., x_i + m, ..., x_k] &= \exp\left\{\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_i (x_i + m) + \cdots + \widehat{\beta}_k x_k\right\} \\
&= \exp\left\{\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_i x_i + \cdots + \widehat{\beta}_k x_k\right\} \exp\left\{\widehat{\beta}_i m\right\} \\
&= \mathbb{E}[Y|X_i] \exp\left\{\widehat{\beta}_i m\right\}
\end{aligned}$$

So for the additive effect m in $x_i + m$, we get the partial effect of the estimate as $\exp\left\{\widehat{\beta}_i m\right\}$. For log transformed independent variables, a multiplicative change in the variable has a partial effect that is multiplicative under a log-link function. Given variables $x_1, x_2, ..., \log(x_i), ..., x_k$, the partial effect under a log-link function is:

$$\begin{aligned}
\mathbb{E}[Y|x_1, x_2, ..., m \cdot x_i, ..., x_k] &= \exp\left\{\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_i \log(x_i \cdot m) + \cdots + \widehat{\beta}_k x_k\right\} \\
&= \exp\left\{\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_i \log(x_i) + \widehat{\beta}_i \log(m) + \cdots + \widehat{\beta}_k x_k\right\} \\
&= \mathbb{E}[Y|X_i] \exp\left\{\widehat{\beta}_i \log(m)\right\}
\end{aligned}$$

So for the multiplicative effect m in $\log(x_i \cdot m)$, we get the partial effect as $\exp\{\widehat{\beta}_i \log(m)\}$. To turn this into a percentage, we simply take $\exp\{\widehat{\beta}_i \log(m)\} - 1$.

The partial effect of variables in a probit model

We want to find the multiplicative partial effect of an independent variable x_i that has been transformed as $f(x_i)$. For a set of variables X_j and the estimated coefficients $\widehat{\delta}_g$ of trade j in size group g, we can write the estimated probit model as follows:

$$\widehat{\gamma}(g, X_j) = \mathbb{P}[y_g = 1|X_j] = \Phi(X_j\widehat{\delta}_g) = \Phi(\widehat{\delta}_0 + \widehat{\delta}_1 x_1 + \cdots + \widehat{\delta}_i f(x_i) + \cdots + \widehat{\delta}_k x_k) \tag{24}$$

We can now find the elasticity by taking the partial derivative of $\widehat{\gamma}(g, X_j)$ with respect to x_i:

$$\frac{\partial\widehat{\gamma}(g, X_j)}{\partial x_i} = \phi(X_j\widehat{\delta}_g) \cdot \widehat{\delta}_i \frac{\mathrm{d}f(x_i)}{\mathrm{d}x_i} \tag{25}$$

For the transformation $f(x) = x \cdot m$, we simply get $\widehat{\delta}_i m\phi(X_j\widehat{\delta}_g)$ as the partial effect. The corresponding semi-elasticity is $\widehat{\delta}_i m\phi(X_j\widehat{\delta}_g)/\Phi(X_j\widehat{\delta}_g)$. For the transformation $f(x) = \log(x)$, we get the following partial effect:

$$\frac{\partial\widehat{\gamma}(g, X_j)}{\partial x_i} = \phi(X_j\widehat{\delta}_g) \cdot \widehat{\delta}_i \frac{1}{x_i} \tag{26}$$

$$\xRightarrow{/\widehat{\gamma}(g,X_j)} \quad \frac{\partial\widehat{\gamma}(g, X_j)}{\partial x_i} \cdot \frac{x_i}{\widehat{\gamma}(g, X_j)} = \frac{\phi(X_j\widehat{\delta}_g) \cdot \widehat{\delta}_i}{\Phi(X_j\widehat{\delta}_g)} \tag{27}$$

This is the full elasticity E of the probability $\widehat{\gamma}$ for the unit change $\log(x) + 1 = \log(x \cdot e)$. For any other multiplier m, we need to make the adjustment: $\log(x \cdot e) + \log(m/e) = \log(x \cdot m)$. Alternatively, we can roughly adjust the coefficients with $(m/e) \cdot \widehat{\delta}_i$. This gives the following:

$$E_{\widehat{\gamma}(g,X_j)}(x_i, m) = \frac{\partial\widehat{\gamma}(g, X_j)}{\partial x_i} \cdot \frac{x_i m e^{-1}}{\widehat{\gamma}} = \widehat{\delta}_i m e^{-1} \frac{\phi(X_j\widehat{\delta}_g)}{\Phi(X_j\widehat{\delta}_g)} \tag{28}$$

Using method 3, we can also find the partial effect by increasing x_i with the multiplier m inside the function and dividing by the original, unadjusted probability:

$$\frac{\widehat{\gamma}(g, X_j + \widehat{\delta}_i \log(m))}{\widehat{\gamma}(g, X_j)} = \frac{\Phi(X_j\widehat{\delta}_g + \widehat{\delta}_i \log(m))}{\Phi(X_j\widehat{\delta}_g)} \tag{29}$$

Both methods yield approximately the same estimate of the partial effects. Because the estimate depends on the chosen variables X_j, we take the average of the individual partial effects over all observations. This is representative of the average partial effect of the variable on the sample.

The partial effect of variables on the CDF of the gamma distribution

As explained before, we relate the outcome of the cost model for group g and type p to the underlying gamma distribution using $\widehat{\alpha}_{gp} = \widehat{\phi}_{gp}^{-1}$ and $\widehat{\beta}_{gp} = \widehat{\phi}_{gp}^{-1}\widehat{c}_{gp}^{-1}$ for $Gamma(\widehat{\alpha}_{gp}, \widehat{\beta}_{gp})$ (original

explanation in Section 2.4 on page 13). For a cost target c and a set of k variables in vector X_j, we can calculate the corresponding cumulative distribution function as follows:

$$\mathbb{P}[C_{gp} \leq c|X_j] = \frac{\gamma(\widehat{\phi}_{gp}^{-1}, c\,\widehat{\phi}_{gp}^{-1}\widehat{c}_{gp}^{-1})}{\Gamma(\widehat{\phi}_{gp}^{-1})} = \frac{1}{\Gamma(\widehat{\phi}_{gp}^{-1})}\gamma\left(\widehat{\phi}_{gp}^{-1}, c\,\widehat{\phi}_{gp}^{-1}\exp\{-X_j\widehat{\beta}_{gp}\}\right) \tag{30}$$

where $\gamma(\cdot,\cdot)$ is the lower incomplete gamma function $\gamma(s,x) = \int_0^x t^{s-1}e^{-t}\,\mathrm{d}t$. To find the partial effect of a variable x_i in this function, we increase the log-transformed $\log(x_i)$ with the multiplier m inside the function and then divide by the original CDF:

$$\begin{aligned}\frac{\mathbb{P}[C_{gp} \leq c|x_1,...,\log(x_i\cdot m),...,x_k]}{\mathbb{P}[C_{gp} \leq c|X_j]} &= \frac{\gamma\left(\widehat{\phi}_{gp}^{-1}, c\,\widehat{\phi}_{gp}^{-1}\exp\{-X_j\widehat{\beta}_{gp} - \widehat{\beta}_i\log(m)\}\right)/\Gamma(\widehat{\phi}_{gp}^{-1})}{\gamma\left(\widehat{\phi}_{gp}^{-1}, c\,\widehat{\phi}_{gp}^{-1}\exp\{-X_j\widehat{\beta}_{gp}\}\right)/\Gamma(\widehat{\phi}_{gp}^{-1})} \\ &= \frac{\gamma\left(\widehat{\phi}_{gp}^{-1}, c\,\widehat{\phi}_{gp}^{-1}\exp\{-X_j\widehat{\beta}_{gp} - \widehat{\beta}_i\log(m)\}\right)}{\gamma\left(\widehat{\phi}_{gp}^{-1}, c\,\widehat{\phi}_{gp}^{-1}\exp\{-X_j\widehat{\beta}_{gp}\}\right)}\end{aligned} \tag{31}$$

Because the partial effect depends on the chosen variables X_j, we take the average of the individual partial effects over all observations to find the average partial effect.

The partial effect of the target cost on the CDF of the gamma distribution

Continuing with the notation from Section 3.2, we can find the partial effect of c in $\mathbb{P}[C_{gp} \leq c|X_j]$ by taking the partial derivative with respect to c:

$$\frac{\partial\mathbb{P}[C_{gp} \leq c|X_j]}{\partial c} = \frac{1}{\Gamma(\widehat{\phi}_{gp}^{-1})}\frac{\partial}{\partial c}\int_0^{c\,\widehat{\phi}_{gp}^{-1}\widehat{c}_{gp}^{-1}} t^{\widehat{\phi}_{gp}^{-1}-1}e^{-t}\,\mathrm{d}t \tag{32}$$

Using the product rule $\frac{\partial}{\partial x}h(g(x)) = h'(g(x))g'(x)$ with $h(x)$ the lower incomplete gamma function, $h(x) = \gamma(\widehat{\phi}_{gp}^{-1}, x) = \int_0^x t^{\widehat{\phi}_{gp}^{-1}-1}e^{-t}\,\mathrm{d}t$ and $g(x) = c\,\widehat{\phi}_{gp}^{-1}\widehat{c}_{gp}^{-1}$. Using the fundamental theorem of calculus, we see that $h'(x) = x^{\widehat{\phi}_{gp}^{-1}-1}e^{-x}$ and $g'(x) = \widehat{\phi}_{gp}^{-1}\widehat{c}_{gp}^{-1}$. Substituting terms, we then find the following:

$$\frac{\partial\mathbb{P}[C_{gp} \leq c|X_j]}{\partial c} = \frac{1}{\Gamma(\widehat{\phi}_{gp}^{-1})}\left(c\,\widehat{\phi}_{gp}^{-1}\widehat{c}_{gp}^{-1}\right)^{\widehat{\phi}_{gp}^{-1}-1}\left(e^{-c\,\widehat{\phi}_{gp}^{-1}\widehat{c}_{gp}^{-1}}\right)\left(\widehat{\phi}_{gp}^{-1}\widehat{c}_{gp}^{-1}\right) \tag{33}$$

$$= c^{-1}\frac{1}{\Gamma(\widehat{\phi}_{gp}^{-1})}\left(\widehat{\phi}_{gp}^{-1}\widehat{c}_{gp}^{-1}\right)^{\widehat{\phi}_{gp}^{-1}}\left(e^{-c\,\widehat{\phi}_{gp}^{-1}\widehat{c}_{gp}^{-1}}\right) \tag{34}$$

Which is the semi-elasticity of increasing the target cost c with one cent. To find the full elasticity, we can divide by the CDF again:

$$\frac{\partial\mathbb{P}[C_{gp} \leq c|X_j]}{\partial c}\cdot\frac{1}{\mathbb{P}[C_{gp} \leq c|X_j]} = c^{-1}\frac{\left(\widehat{\phi}_{gp}^{-1}\widehat{c}_{gp}^{-1}\right)^{\widehat{\phi}_{gp}^{-1}}\left(e^{-c\,\widehat{\phi}_{gp}^{-1}\widehat{c}_{gp}^{-1}}\right)}{\gamma(\widehat{\phi}_{gp}^{-1}, -c\,\widehat{\phi}_{gp}^{-1}\widehat{c}_{gp}^{-1})} \tag{35}$$

Additionally, we may also use method 3 and divide the function with the target cost plus one by the original function:

$$\frac{\mathbb{P}[C_{gp} \leq c+1|X_j]}{\mathbb{P}[C_{gp} \leq c|X_j]} = \frac{\gamma\left(\widehat{\phi}_{gp}^{-1}, (c+1)\,\widehat{\phi}_{gp}^{-1}\widehat{c}_{gp}^{-1}\right)}{\gamma\left(\widehat{\phi}_{gp}^{-1}, c\,\widehat{\phi}_{gp}^{-1}\widehat{c}_{gp}^{-1}\right)} \tag{36}$$

The partial effect of variables on the probability of at least one arrival

Using method 3, we find the partial effect of the effect $x_i \cdot m$ on the probability that at least one opposite party arrives by multiplying the log-transformed variable $\log(x_i)$ with the factor m in the original formula and dividing by the original probability function:

$$\frac{\mathbb{P}[N_g^S > 0|x_1, ..., \log(x_i \cdot m), ..., x_k]}{\mathbb{P}[N_g^S > 0|X_j]} = \frac{1 - \left(\frac{\widehat{\theta}_S}{\widehat{\theta}_S + \lambda_S(X_j)\exp\{\widehat{\kappa}_i \log(m)\}}\right)^{\widehat{\theta}_S}}{1 - \left(\frac{\widehat{\theta}_S}{\widehat{\theta}_S + \lambda_S(X_j)}\right)^{\widehat{\theta}_S}} \tag{37}$$

$$= \frac{1 - \widehat{\theta}_S^{\widehat{\theta}_S}\left(\widehat{\theta}_S + \exp\{X_j\widehat{\kappa}_g^S + \exp\{\widehat{\kappa}_i \log(m)\}\right)^{-\widehat{\theta}_S}}{1 - \widehat{\theta}_S^{\widehat{\theta}_S}\left(\widehat{\theta}_S + \exp\{X_j\widehat{\kappa}_g^S\}\right)^{-\widehat{\theta}_S}} \tag{38}$$

The partial effect depends on the chosen k variables in X_j, so we take the average of the individual partial effects over all observations.

The partial effect of variables on the PEX

Now that we have established the partial effects of the various components of PEX in the previous sections, we can compute the multiplicative partial effect Δ_i of a variable x_i on the PEX of a buy trade for a given set of independent variables X_j as follows:

$$\Delta_i \text{PEX}_B = \Delta_i \mathbb{P}[N^S > 0]\left(\widehat{\gamma}_B \Delta_i \widehat{\gamma}_B \Delta_i \mathbb{P}[C_{\text{DDB}} \leq c] + (1 - \widehat{\gamma}_B \Delta_i \widehat{\gamma}_B)\Delta_i \mathbb{P}[C_{\text{SDB}} \leq c]\right) \tag{39}$$

Here we have omitted the dependencies, but all terms are conditional on the given set of dependent variables X_j. In the same fashion, the partial effect on PEX_S may be calculated as follows:

$$\Delta_i \text{PEX}_S = \Delta_i \mathbb{P}[N^B > 0]\left(\widehat{\gamma}_S \Delta_i \widehat{\gamma}_S \Delta_i \mathbb{P}[C_{\text{SDD}} \leq c] + (1 - \widehat{\gamma}_S \Delta_i \widehat{\gamma}_S)\Delta_i \mathbb{P}[C_{\text{SDB}} \leq c]\right) \tag{40}$$

As before, we assume that the models in the PEX are independent such that the partial effects can simply be multiplied. Because the individual models rely on the given set of independent variables X_j, we estimate the partial effects separately per observation and take the average. Unlike the partial effects on the individual models, the partial effect of a variable on the PEX is not (log)linear. The average partial effects on the PEX are therefore not always representative of partial effects on individual observations. We illustrate this in Figure 3 in the results (Section 5).

4. Data

In order to find transaction combinations, we use all corporate bond trades reported to the Financial Industry Regulatory Authority's (FINRA) Trade Reporting and Compliance Engine (TRACE) between the years 2005 up and until 2013[7]. Specifically, we use the enhanced time-and-sales data which includes, among others, the original trade sizes and all reporting side indicators. This gives us an advantage over some of the older literature, where either trade sizes are downward biased or reporting sides have to be inferred from the data (Feldhütter, 2012). In total, we have 110,189,735 raw data records, including both trades and corrections to trades.

To filter the data, we first apply the filtering approach described by Dick-Nielsen (2009, 2014). We delete cancelled or reversed trades (2.61%)[8], apply corrections to erroneously reported records (1.13%) and delete records with odd or missing information (3.71%). We also delete all records of non-corporate bonds, records that modify other records, records with special prices[9], trades that are executed outside of market hours and trades that are reported long after they took place (6.63%). Like Dick-Nielsen, we also remove agency trades (11.57%) and double interdealer records (22.01%). Dick-Nielsen removes interdealer records using a price sequence filter. We find a more accurate way to remove interdealer records in the enhanced data: we delete all buy-side interdealer transactions that have not been disseminated[10].

On July 30, 2012, FINRA released a better version of TRACE with additional market reference data and improved process tracking capabilities. This greatly simplifies the cleaning procedure for data that was disseminated after this change. We find that these improvements, elaborately described by Dick-Nielsen (2014), do indeed improve the filtering accuracy. Because we are mainly interested in transaction costs of bonds that trade under normal conditions of the underlying company, we amend Dick-Nielsen's filter with several additional filters. We remove all records with extremely low prices, either when the corresponding bond ever reaches a price of \$1 or less, or if the transaction price is lower than \$50 (0.74%). This is done to prevent distressed debt from entering our sample because they give a large upward bias in the cost estimates. Additionally, dealers are unwilling to take distressed debt on their books, such that the fraction of trades that go through inventory is downward biased. The arrival rate of buyers

[7] FINRA releases the enhanced TRACE dataset with a lag of 18 months.

[8] Percentages denote the amount of deleted records from the original full dataset. The record types we refer to in the filter are the same as by the definitions of FINRA. The difference between cancels and reversals is that "firms report a 'cancellation' when trades are cancelled on the date of execution and a reversal when trades are cancelled on any day after the date of execution." (A311.3 of FINRA's FAQ).

[9] Harris (2015) explains the definition of 'special prices' in TRACE. In short, a transaction has a special price if its price is expected (by FINRA) to deviate from the normal market due to some irregularity.

[10] FINRA states the following on their website (FAQ 1.23): "For interdealer trades, TRACE disseminates only the sell side of the transaction. All Customer trades are disseminated." Therefore, we can safely delete all buy-side interdealer records that are also not disseminated. If a buy-side interdealer record has been disseminated (indicated by a flag), then it is not safe to delete this record because the corresponding sell-side record is either missing or contains an error. Because FINRA has access to dealer identifiers, they can perfectly match related interdealer records. Through the dissemination flag, we can therefore also make use of this information.

may also be influenced as hedge funds often step in to buy distressed debt. Our goal is to estimate transaction costs under normal company conditions and we are therefore not interested in distressed debt. We also remove bonds of which the price is larger than or equal to \$200 (0.75%). This is done to prevent convertible bonds from entering our sample as they behave differently than non-convertible bonds with respect to overall market conditions and company performance[11]. A full overview of the amount of deleted observations per year can be found in Table 7 on page 52. After filtering, we are left with 57,280,531 corrected trades.

Next, we apply detection logic to identify the different transaction combinations (described in the next section). This leaves us with 39,092,570 combinations (all types). We now only take the 30,690,032 observations of type SDB, SDD, DDB, SD or DB that are relevant for this study. After imputing transaction costs, we apply two more filters. First we delete all imputed roundtrip costs with zero or negative costs. We delete transactions with zero costs because dealers often transfer bonds between their subsidiaries without markup[12]. Because of the reporting obligation to TRACE, such transfers will show up as having zero cost. We also delete roundtrips with negative costs because they are probably caused by uncorrected records or a failure of the roundtrip discovery logic for ambiguous situations in which many trades happen in a short period of time[13]. Negative costs may also appear due to sudden market movements between the time that the dealer buys and sells the bonds, which hinders our assumption that costs are always positive. Deleting the negative costs, we end up with 29,161,855 combinations in total.

We acquire a wide variety of bond characteristics from constituent data of the Barclays U.S. Corporate Investment Grade index. Because not all bonds that are reported to TRACE are covered by the index, we only consider bonds that appear in both. We restrict our sample to investment grade bonds only because we find significantly different results with high yield bonds. High yield bonds seem to have different relations with respect to the explanatory variables. Nevertheless, the methodology in this thesis can be applied to high yield bonds too. At first glance, we find approximately similar results for high yield bonds, but with different coefficient magnitudes and significance. Further research is necessary to shed more light on these differences. After taking the intersection between the bonds available in TRACE and the bonds covered by the index, we are left with 13,277,112 observations. Finally, we split the sample into three different groups according to transaction size: \$0 to \$100k (odd-lot), \$100k to \$1mm (round-lot) and \$1mm or more (block sized[14]). We find that of the final trades in our sample, 62% has an

[11] Convertible bonds benefit from capital appreciation should the company do well: they can be interpreted as bonds that include a call option on the company's stock. This is reflected in the price of these bonds.

[12] For example, dealer subsidiary 'A' conducts a transaction with bonds from the inventory of subsidiary 'B'. Before subsidiary A can sell the bonds, subsidiary B must first transfer the bonds to A. This shows up as a seperate zero-cost transaction in TRACE. We delete such transfers because they concern the same dealer.

[13] Because we do not have dealer identifiers, it can happen that dealer records are accidentally intertwined. This can potentially result in the identification of negative costs.

[14] Industry conventions sometimes dictate another split for transactions sized between \$1mm and \$5mm and sizes of \$5mm or higher. We found no special differences between the two groups so we decided to aggregate them.

odd-lot transaction size, 23% is a round-lot trade and 15% is block sized. When performing the regressions, we further delete observations for which we have missing bond characteristics. To show how many observations we end up using, we display the amount of included observations separately per regression. For the cost model, we additionally delete the 0.5% of highest transaction costs per size group. These records have extremely high costs, likely due to uncorrected errors by inaccuracies in the filtering procedure[15] or fat-finger errors[16].

4.1. Identifying and imputing transaction costs

Before we estimate the models, we first explain how we identify the different transaction combinations. For this, we build on the methodology of Feldhütter (2012). Feldhütter calculates imputed roundtrip trades (IRTs) based on the phenomenon that bonds often do not trade for hours, or even days, and then two or more transactions are reported to TRACE within a very short time span. We can assume that these trades are part of a 'pre-matched arrangement' where a dealer has matched both a buying and selling customer. Once the dealer has found such a match, two transactions take place. One between the seller and the dealer and one between the dealer and the buyer. If any other dealers are involved in the pre-matching, there can also be additional trades that are part of the same roundtrip. Feldhütter classifies trades as part of an IRT if two or more consecutive trades take place within 15 minutes of each other, on the same bond (same CUSIP) and with equal size (par value volume in dollars). For any such event, Feldhütter takes the lowest price to be from the seller and the highest price to be from the buyer. The roundtrip cost is then taken as the buyer's purchasing price minus the price that the seller received[17]. Under Feldhütter's definition of IRTs, a roundtrip is detected when at least one customer is present and two transactions with the same size happen at approximately the same time. This causes the transaction combinations SDB, SDD and DDB to all be classified as the same IRT. Feldhütter admits that SDD and DDB transactions are in fact representative of just half (or approximately half) of the effective bid-ask spread. Because he does not differentiate between the three types, it causes the IRTs to underestimate actual transaction costs.

Because we have access to the enhanced TRACE data, we can amend Feldhütter's (2012) approach to avoid misclassifications. First of all, we have access to uncapped transaction sizes[18]. This improves the accuracy of accurately detecting roundtrips. Secondly, we also have buy and sell flags for every trade. This means that we can see exactly which trades involve buyers or sellers. As a consequence, we can observe the buy and sell prices directly from the trades

[15] Especially before the 2012 improvement, it is sometimes not possible to find the record that should be corrected (this is quite rare). We always delete all records if a correction or deletion happens to match multiple.

[16] Sometimes we observe a roundtrip with, for example, buy price \$110.50 and sell price \$101.25. Likely, the sell price should have been \$110.25 because the last 0 and 1 were switched by accident and never corrected. These type of errors are extremely rare, but are sometimes observed when dealers report information manually. This is more common in the beginning of the sample; nowadays almost all reporting is automated.

[17] Transaction costs are taken as a fixed dollar amount with respect to price, not with respect to yield spread.

[18] In the non-enhanced data, transaction sizes in TRACE are capped at \$5mm for investment grade bonds and \$1mm for high yield bonds. This can lead to erroneously matching trades that actually have different sizes.

instead of inferring them from the imputed spread. This gives us the opportunity to classify the different transaction combinations exactly as they occur, thus making the distinction between SDB and SDD or DDB types. Because costs can not be imputed for DB and SD combinations, we cannot use them. We therefore assume that the imputed transaction costs from SDD and DDB combinations are also approximately representative of the SD and DB types. This means that we might overestimate costs of transactions that involve dealer inventory, given that an additional dealer must be involved for costs to be identifiable. Because this is a common phenomenon in the corporate bond market, it can be argued that this is a decent assumption[19].

We define transaction costs as being half of the realised bid-ask spread. Assuming that the theoretical value of a bond is the midpoint of the bid-ask spread, we argue that the cost of trading is the deviation from the true value, the midpoint. Given that SDB combinations represent the full bid-ask spread, we divide the bid-ask spread by two to get the implied transaction costs. Because SDD and DDB combinations are representative of approximately half the bid-ask spread, we can directly infer transaction costs from them[20]. Mathematically, the cost C is calculated as $C = (p^B - p^S)/2$ for SDB combinations, and $C = (p^B - p^S)$ for SDD and DDB combinations. Here, p^B is the buyer's price and p^S is the seller's price. We denominate costs in dollar cents because we observe that dealers generally do not adjust their markups for different prices. Instead, dealers seem to use fixed markup amounts (10¢, 25¢, etc.) and we therefore argue against denominating costs in basis points ex ante. Doing so can result in the accidental inclusion of a price effect[21]. We have demonstrated this effect in Figure 7 on page 51, where the cost distribution for costs in basis points shows a relation to price. As a consequence, we find that denominating costs in dollar cents, opposed to basis points, removes most of the explanatory power of price observed by Harris (2015).

For the full period from 2005 up and until 2013, we find 17.10% of all transaction combinations to be DD interdealer trades, 2.73% to be SDB instantaneous roundtrips, 12.50% are SDD halftrips and 18.94% are DDB halftrips. Additionally, 19.96% are single SD trades and 24.71% are single DB trades. Finally, we are not able to classify 4.08% of record combinations because they are ambiguous[22]. Descriptive statistics of the filtered sample can be found in Table 1. It must be noted that the sample statistics are primarily dominated by liquid bonds. This is caused by the fact that liquid bonds have many transactions, thereby contributing a large portion of transaction to the total sample. This causes liquid bonds to dominate aggregate

[19] Feldhütter (2012) also makes this assumption, arguing that SDD and DDB transactions are still representative of the then-prevailing bid-ask spread. Nevertheless, if one prefers to remove this assumption, then the SDD and DDB results in this thesis can be interpreted as unrepresentative for SD and DB transactions. This would limit the scope of inference for the SDD and DDB types, but retains interpretation for SDB combinations.

[20] The definition of transaction costs can differ according to preferences. If the full bid-ask spread is preferred as a measure of transaction costs, one can also retain the SDB costs and double the SDD or DDB spreads.

[21] Nota bene: given a transaction cost of 30 cents for an asset with price \$90 and for an asset that costs \$110, the relative transaction costs will be higher for the cheaper asset by construction (33 bps versus 27 bps).

[22] The combination is flagged as ambiguous if the identification logic finds that more than one buying or selling customer with the same transaction size appears before the roundtrip has ended (= opposite party found).

statistics. Table 1 is therefore only representative of an average transaction, but not of an average transaction on an average bond.

Table 1: ***Descriptive statistics of transaction costs***. This table gives an overview of the sample statistics related to the imputed transaction costs. As discussed in Section 1, the statistics in the table capture different liquidity dimensions. Specifically, the first percentile is related to the 'width', the mean and median to the 'depth' and the skewness and kurtosis to the 'breadth' of the aggregate bond market. Costs are denoted in dollar cents and are taken as half of the imputed bid-ask spread. SDB stands for Seller-Dealer-Buyer imputed roundtrip costs. SDD and DDB are imputed Seller-Dealer-Dealer and Dealer-Dealer-Buyer transaction costs, respectively. Statistics are calculated over the separate size groups after cleaning the data. Transactions for which we do not have bond characteristics are deleted. The metrics in this table are representative of the same sample that is used for our cost model estimates. The notation 'p.b.' denotes 'per bond'. The number of trades are listed in thousands, indicated by (k). The mean and median trade sizes are denoted in thousands of dollars, denoted by ($k).

	$0–$100k			$100k–$1mm			$1mm+		
	SDB	SDD	DDB	SDB	SDD	DDB	SDB	SDD	DDB
Mean cost	67	57	88	23	39	60	13	19	22
1st percentile	1	2	3	0	1	1	0	0	1
Median cost	50	50	88	10	25	39	9	8	12
99th percentile	213	225	259	150	200	244	75	125	150
Std	58	49	67	31	42	59	15	26	30
Skewness	0.8	1.4	0.6	2.5	1.9	1.2	3.4	3.3	3.1
Kurtosis	2.7	5.3	2.6	10.6	7.6	3.9	24.0	19	16
Mean trades p.b.	19	135	277	7	23	48	7	7	10
Median trades p.b.	5	31	74	3	8	17	4	3	5
Mean size ($k)	24	22	24	284	227	214	5100	3200	2900
Median size ($k)	20	15	20	200	150	150	3000	2000	2000
Bonds	5183	6852	7315	5616	6430	7038	6354	5414	6422
Trades (k)	97	922	2023	40	148	336	42	36	62

Analysing Table 1, we observe several patterns. First of all, average costs decrease for higher transaction sizes. The SDB cost for a transaction between $0–$100k averages at 67 cents. Likewise, an SDB transaction of $1mm+ has an average cost of just 13 cents. Costs are higher when a trade goes through dealer inventory (SDD, DDB) than if it is transacted instantaneously (SDB). We also observe that DDB transactions are more expensive than SDD transactions. A reason for this might be that SDD transactions are relatively simple for a dealer: he buys the bonds and charges the expected inventory costs. For DDB transactions, it might be that the dealer needs to borrow the bonds from the interdealer repo market. This entails more costs and could therefore lead to higher markups. We see that the distribution of costs shifts more towards the midpoint

as transaction size increases, as can be judged from the first percentiles and median costs. This would imply that the width of the market is smaller for larger transaction sizes. On the other hand, we also observe that, as transaction size increases, the market becomes less liquid in other dimensions. The growing skewness and kurtosis convey that markets thin out in terms of breadth for higher sizes. Even though median costs become cheaper, it also becomes more likely to observe a cost in the tail of the distribution. In general, we observe that the average cost is always higher than the median. The same holds for the average amount of trades per bond, where the average amount of trades per bond is always higher than the median. This gives evidence that liquidity is unevenly distributed across bonds, with outliers both for very liquid bonds and very illiquid bonds.

When looking at the different transaction combinations, we also observe some interesting properties. SDD trades have both a higher skewness and kurtosis than DDB trades, but not higher average or median costs. This implies that the distribution of realised costs for SDD trades is more dispersed, suggesting that sellers are willing to occasionally accept high costs. We imagine this happens in times of distressed selling or weakened overall market conditions. The prevalence of the different transaction combinations also changes with transaction size. As transaction size increases, dealers become increasingly unwilling to absorb bonds into inventory. This causes the SDB transaction combinations to appear more often than the SDD combinations in the \$1mm+ size group. This motivates weighing transaction costs with respect to the probability of observing a given combination. Specifically, because SDD and DDB costs are higher than costs for the immediate SDB roundtrip trades, we argue that transaction costs should indeed be expressed as an average cost that is weighted with the prevalence of the different combinations.

4.2. Variable selection

Now that we have developed the models, we need to select variables that we expect to explain cross-sectional differences in corporate bond liquidity. Ex ante, the variable selection poses several problems. First of all, we may not use statistical significance as a 'decisive' factor for selecting variables. As elaborately discussed by Ziliak and McCloskey (2008), statistical significance is useless if the effect has no practical significance or if the size of the effect is small. In our case, this cautionary note holds especially true: we are bound to find statistical significance of regression coefficients, regardless of practical significance or economic intuition. This is due to the extremely large sample sizes that we work with. This is also known as the p-value problem: p-values based on consistent estimators approach zero in the limit if the estimated parameter is not *exactly* equal to the null hypothesis (Lin et al., 2013). Therefore, it makes no sense to use statistical significance as a selection criterion. Instead, we investigate the size of effects and try to understand the interaction between variables. This coincides with the investigation of multicollinearity. Specifically, multicollinearity affects the estimated parameters such that the interpretation of effects can become deceptive. We expect to find multicollinearity between variables that are related to the same risk factor of corporate bonds. Examples of such risk

factors are interest rate risk, credit risk or overall market risk. More obvious relations, such as that between 'age' and 'maturity', should also be taken into account. Age and maturity have a one-to-one relation, where maturity decreases with one year if the bond becomes one year older. In the same way, there is also a relation between the duration of the bond and its age. But because 'duration' also depends on other characteristics, such as the yield to maturity and the coupon rate of the bond, this relation is much less pronounced.

For the selection, we use a bottom-up approach, starting with the variables for which we have evidence of explanatory power by accompanying literature. For each variable, we take the following steps to determine whether we want to add it to our framework:

1. First we create an ex ante hypothesis regarding the sign of the effect of the variable.
2. We add the variable to the existing regression and inspect the resulting estimated coefficients.
3. If the coefficients of previously added variables change sign, the regression has become instable. In this case, we investigate the reason for the instability. If the coefficient of the newly added variable is not as expected, we proceed with caution to the next steps.
4. We inspect multicollinearity by calculating the 'variance inflation factors' (VIF) of the different variables (Kutner, 2005). If the VIF of a variable is higher than 2, we investigate its dependency structure with existing variables in the regression.
5. If a variable is believed to cause instability in the regression, we compare the new variable to the variables that are highly correlated. We run the regression again, each time including one of the correlated variables. We record the size of the effect and the variance of the corresponding estimated coefficient. Of this set of correlated variables, we choose the one that (1) has a clear economic interpretation, (2) exhibits the largest effect, (3) is supported by previous literature and (4) has satisfactory precision of the estimated coefficient. These aspects are considered in order of appearance, with (1) being the most important and (4) posing as a final check. If the sign of the variable is not as expected, we reason ex post why this may be the case. We discard the variable if we cannot explain the observed sign.

In practice, this procedure is difficult to implement due to the three different size groups and different types of transaction combinations that we consider. To maintain comparability between the different size groups and regressions, the selection procedure is performed jointly over the various size groups and combinations. We use cluster-robust standard errors to get a more objective picture of the variation in regression coefficients (see Section 3.1). We include the variable 'price' in both the cost and warehousing rate model because we want to show its effect and discuss its low explanatory power.

4.3. Variable overview

After performing the selection procedure, we end up with the following selection of variables:

Price	The price of the transaction. Taken as the buy-side price whenever possible.
Duration	The (remaining) duration of the bond in years.
Spread	The yield spread of the bond in basis points. Calculated as the difference in yield between the corporate bond and the U.S. Treasury yield curve for given maturity.
Age	The age of the bond, taken as the number of years since the date of issuance.
AmtOut	The amount of outstanding debt in thousands, consisting of the issue size of the bond plus any additional debt redemption or reissuance. Because the amount outstanding represents the actual amount of tradable debt in the market, it is believed to be a better proxy of liquidity than just the issue size (Hom 2004).
Size	The transaction size, taken as the par value volume in dollars.
AvgSize	The average transaction size of the bond, calculated separately per size group by using all transactions in that group and in higher groups.
Volume	The past monthly transaction volume of the bond, calculated using TRACE.
VIX	The level of the CBOE Volatility Index, which is a measure of implied volatility of S&P 500 index options. The VIX proxies general risk aversion in the markets.
Senior	A dummy variable indicating whether the bond is senior or subordinated debt.
Callable	A dummy variable indicating whether the bond has an embedded option.

The variables *Duration*, *Spread*, *Age* and *AmtOut* are recalculated monthly. The other variables are available on a daily basis. All calculations are performed without look-ahead bias, by excluding the information on the day of the transaction. For example, *Volume* is calculated over the period $t-23$ until $t-1$ for a transaction that takes place on day t. For the variables that are calculated on a monthly basis, we use the information of the previous month. A description of the other variables that we investigated can be found in Appendix G on page 53. An overview of the variables that have been investigated in other studies can be found in Table 8 on page 55.

For the cost and arrival rate models, we find that the relation between the regressors and the dependent variable is best represented in multiplicative form, by using a log-link function. We have included Q-Q plots for various cost model specifications in Appendix I. Secondly, plotting regression residuals versus individual covariates conveys heteroscedasticity, indicating that the regressors should be transformed. We present formal deviance tests for \$1mm+ transactions in Table 9 in Appendix I. From this table, it can be observed how the log transformation yields the highest improvement in deviance for almost all regressors in the models. For *Age*, we find that the square-root transform also works well, but we use the log transform for consistency and ease of interpretation. We do not transform *VIX* because its effect is already linear.

5. Results

We regress the models from Section 2 on the cleaned data from Section 4. From the estimated coefficients, we calculate partial effects as derived in Section 3. In this section we show all results and discuss the most interesting ones. We begin with the results of the PEX measure.

5.1. Results of the PEX measure

Table 2: ***Probability of Execution partial effects***. This table shows the partial effects on the PEX measure for buy and sell transaction for sizes between \$0–\$100k, \$100k–\$1mm and \$1mm+. The first row denotes the mean PEX of the observations. The partial effects are denoted in percentages and calculated as explained in Section 3.2. The effect multiplier m_k we use for variable x_k appears in square brackets after the name. To give an example of interpretation: given a \$1mm+ buy trade, if the *AmtOut* doubles [×2], the PEX_B increases by 42.96%. For the buy side we use all DDB and SDB combinations and for the sell side the SDD and SDB types. Standard deviations appear in brackets below the corresponding estimate. The mean PEX and the partial effect of the cost target are calculated by setting the cost target at 25¢.

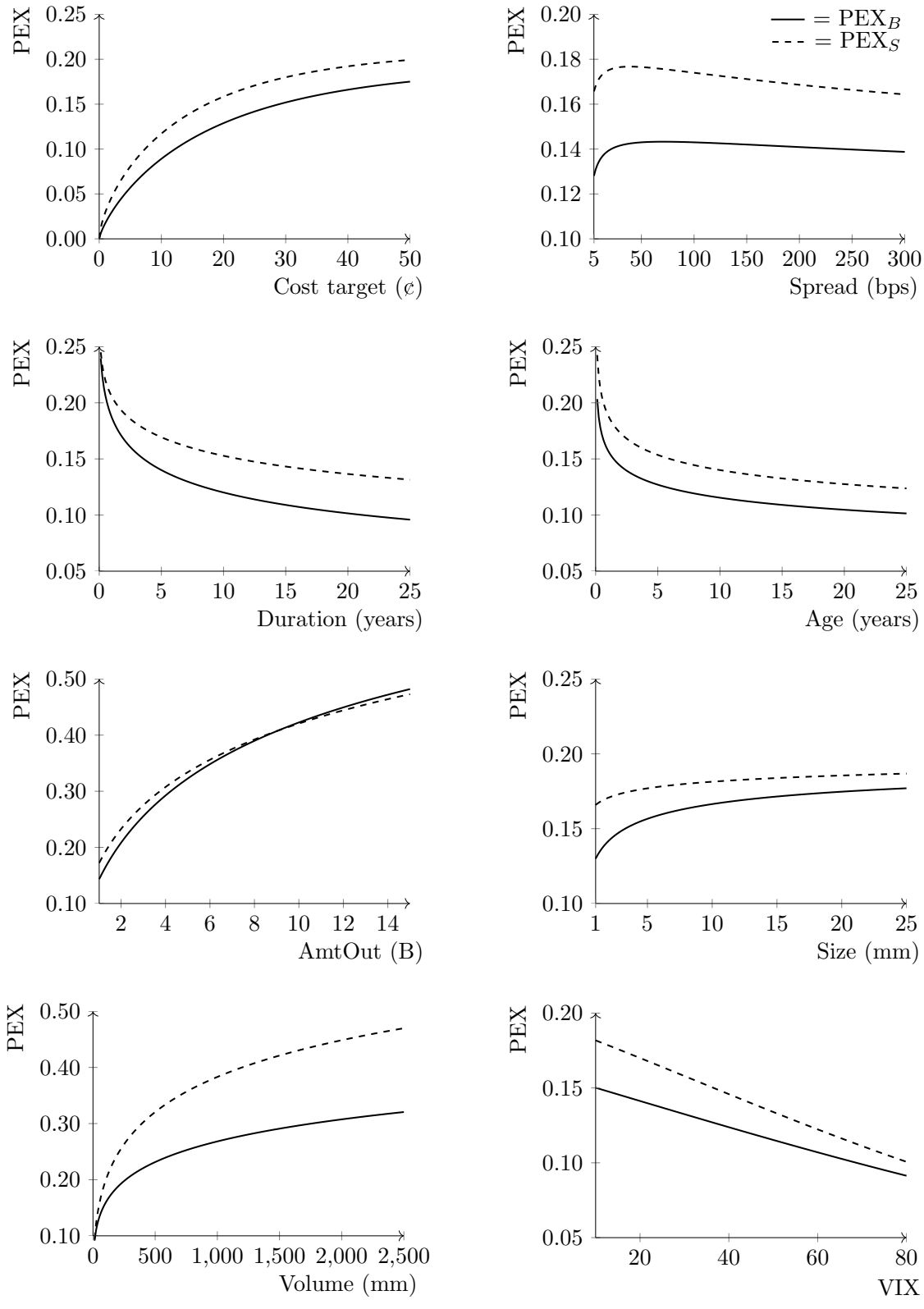
		\$0–\$100k		\$100k–\$1mm		\$1mm+	
		PEX_B	PEX_S	PEX_B	PEX_S	PEX_B	PEX_S
Mean		0.11	0.22	0.17	0.28	0.20	0.25
		(0.11)	(0.09)	(0.11)	(0.15)	(0.14)	(0.18)
Cost target	[+1]	4.94	4.61	2.09	1.91	1.68	1.42
		(0.70)	(0.41)	(0.35)	(0.34)	(0.67)	(0.67)
log(Price)	[×10%]	−8.27	−3.07	0.29	−2.24	0.19	−0.88
		(1.27)	(0.28)	(0.06)	(0.43)	(0.09)	(0.49)
log(Spread)	[×2]	−23.39	−1.29	−15.31	−8.17	−4.46	−4.19
		(3.43)	(6.87)	(4.04)	(3.62)	(5.45)	(3.28)
log(Duration)	[×2]	−52.87	−22.67	−31.15	−15.90	−16.17	−10.03
		(5.23)	(2.07)	(4.14)	(2.58)	(6.85)	(3.11)
log(Age)	[×2]	−3.64	−11.23	−5.77	−10.55	−7.80	−8.53
		(0.93)	(3.29)	(1.73)	(2.93)	(2.12)	(2.22)
log(AmtOut)	[×2]	31.17	13.37	40.20	24.47	42.96	32.25
		(19.40)	(10.68)	(16.66)	(13.45)	(13.18)	(11.51)
log(Size)	[×2]	−2.87	6.75	21.30	15.66	9.08	4.44
		(0.59)	(0.67)	(5.29)	(3.74)	(5.18)	(3.71)
log(AvgSize)	[×2]	−16.71	−12.96	−5.77	−6.57	0.52	0.64
		(9.56)	(9.90)	(2.80)	(4.24)	(4.20)	(5.20)
log(Volume)	[×2]	2.15	7.01	9.33	14.61	16.19	22.06
		(3.62)	(5.85)	(3.88)	(6.48)	(5.55)	(7.80)
VIX	[+1]	−0.20	0.20	−0.23	−0.25	−0.50	−0.56
		(0.06)	(0.07)	(0.07)	(0.04)	(0.21)	(0.28)
Senior	[=1]	19.71	2.30	19.68	3.66	13.99	7.68
		(4.27)	(0.83)	(3.98)	(0.82)	(3.81)	(3.48)
Callable	[=1]	−32.24	−20.44	−7.69	−11.59	−0.68	−3.83
		(4.69)	(1.51)	(3.02)	(1.89)	(3.23)	(3.27)
Observations		3,041,914		523,936		140,167	

The effect of a bond characteristic or market condition on the PEX is a multiplicative combination of the individual partial effects on the cost, warehousing rate and arrival rate models (Section 3.2). Because of this nonlinearity, we find that the relation between a variable and the PEX gives a rich representation of how a variable influences a bond's liquidity. These relationships have been visualised in Figure 3 for the set of most influential variables. Because the partial effects depend on the characteristics of the chosen bond and the size of a trade, we also provide the average partial effects of the buy and sell transactions in our sample in Table 2.

The biggest effect on the probability of execution is a bond's total amount outstanding and the transaction volume of the previous month. Both these effects yield an increase in the PEX for both buyers and sellers of any amount of bonds. We find that the effect increases with transaction size: a bond with twice the amount outstanding than a similar counterpart, yields an average 13% increase in the probability of selling a trade below \$100k. In the same way, the effect gives an increase in probability of almost 43% for transacting above one million. Other variables that positively influence the PEX of buyers and sellers is the size of the transaction and whether the debt is senior or subordinated. The effect of transaction size, more accurately described in Figure 3 for a transaction size of \$2mm, has become a stylised fact of corporate bond liquidity since Schultz (2001) described transaction costs as decreasing logarithmically with size. Akin to corresponding literature, we confirm the same relation between size and liquidity, with the strongest effect for trades between \$100k and \$1mm. However, we find that the effect largely declines, and can even be negative, for small odd-lot transactions. Moving on to seniority, we argue that the increased liquidity of senior debt can be explained by its direct relation to the risk profile of a bond: it has a higher payment priority than subordinated debt in case of default. In turn, this decreases risk for dealers and thereby lowers expected transaction costs. The observed positive effect of seniority on the PEX thereby coheres with our expectations.

On the other side of the liquidity spectrum, we observe that the duration of a bond has the largest negative effect on the PEX, with a considerably larger effect for smaller trades. In addition, duration seems to affect buyers approximately twice as much as sellers. From the graph in Figure 3, we see that the impact on PEX is especially large for bonds with low duration. Other variables that show a negative relation to the PEX are the age of the bond, the level of the VIX and the yield spread. The effect of the latter is not as strong for all sizes as we would expect. The probability of transacting at least one million decreases by only approximately 4% if the yield spread of a bond doubles. By further investigating the individual models later in this section, we find that yield spread has a relatively large effect on arrival rates. Apparently bonds of high yield spread are more popular, thereby generating more market activity and thus liquidity. The total negative effect of the yield spread on the PEX is therefore small. Lastly, we find that callable bonds also result in a lower PEX, especially for odd-lot trades. We argue that dealers charge a premium for callable bonds due to their complexity and inherent call risk. In turn, this negatively influences the PEX when the target cost remains fixed.

Figure 3: ***Partial effects of the PEX for a median $1mm+ transaction.*** These graphs show the partial effects of the variables with the largest influence on the PEX for buying (——) and selling (- - -). The median $1mm+ transaction has a *Size* of $2mm, *Price* of $103, *Spread* of 158 bps, *Duration* of 4.71y, *Age* of 2.2y, *AmtOut* of $1B, *AvgSize* of $1.05mm, past monthly *Volume* of $65mm, *VIX* of 19.4, is *Senior* and does not have an embedded option.



5.2. Results of the individual models

In this section we analyse the results of the three models by interpreting the effects of the various liquidity determinants and providing possible explanations for the observed relations. We are mainly interested in the sizes of the effects and their sign, and therefore only discuss statistical significance if unsatisfactory. The goodness of fit measures used in this section, the deviance mean squared error, Brier score, McFadden R^2 and deviance tests, are explained in Appendix H.

Figure 4: ***Cost distributions.*** These graphs show probability density plots of the cost distributions of the various size groups for buy (DDB, SDB) and sell (SDD, SDB) transactions. The distributions show liquidity circumstances for the bid and ask side, as explained in Figure 1. They are constructed by taking the first percentile (P1 = ·····), median (P50 = ——) and last percentile (P99 = - - -) of bond characteristics, where for each variable we take P99 as the most liquid percentile of the variable. For example, the P99 distribution is calculated using the lowest percentile of durations in the sample, because the effect of duration on costs is positive (Table 3). The expected cost of the median transaction is indicated by a bullet (•).

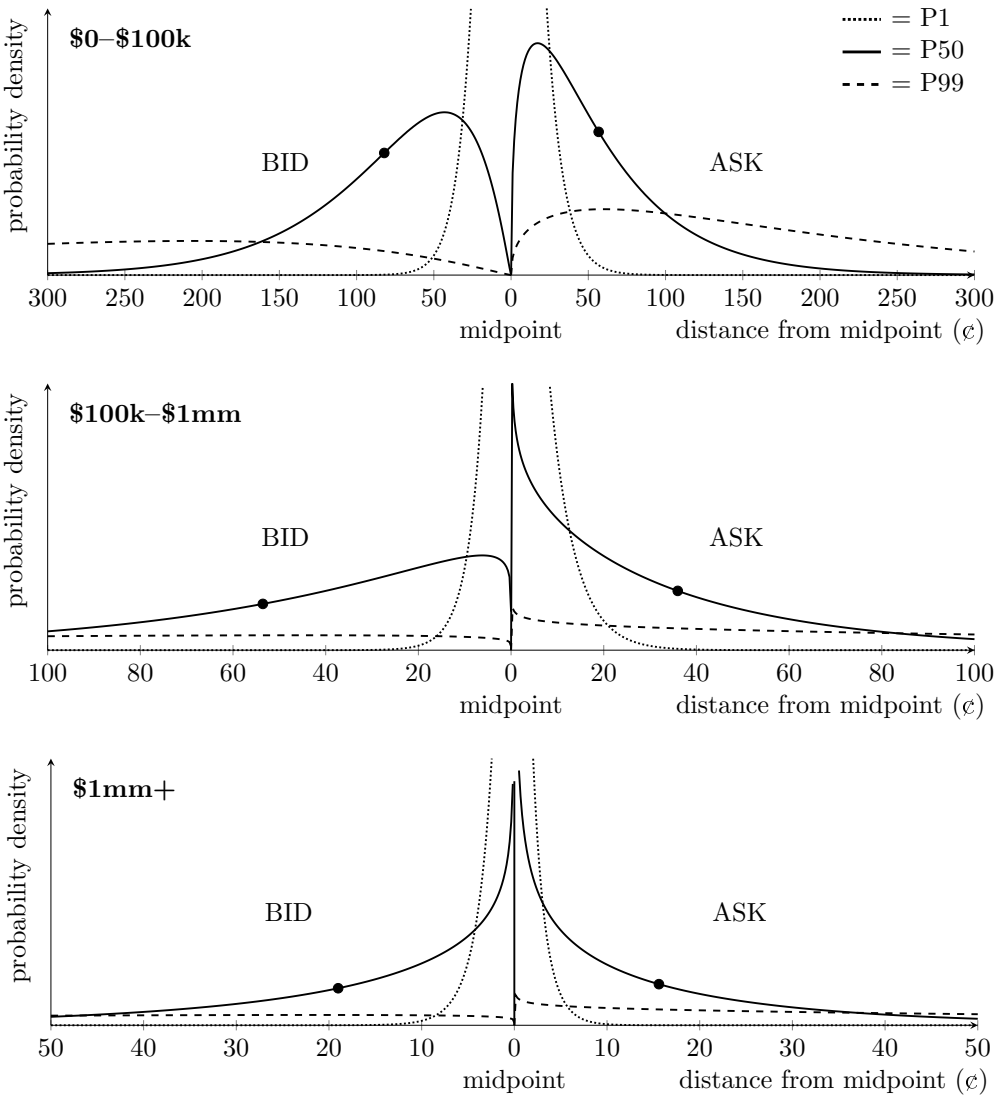


Table 3: ***Cost model estimates***. This table shows the partial effects of a GLM regression with log-link and gamma distribution of imputed transaction costs on bond characteristics and market conditions for trade combinations of sizes between \$0–\$100k, \$100k–\$1mm and \$1mm+. Continuous regressors have been demeaned before estimation. Costs are taken as half the bid-ask spread and are denominated in dollar cents. The listed partial effects are denoted in percentage terms, except for the intercept. The model coefficient $\widehat{\beta}_k$ is transformed into its multiplicative partial effect using either $\exp(\widehat{\beta}_k m_k)$ for the effect $x_k + m_k$ on non-transformed variables or $\exp(\widehat{\beta}_k \ln(m_k))$ for the effect $x_k \cdot m_k$ on log transformed variables (see Section 3.2). The effect multiplier m_k we use for variable x_k appears in square brackets after the name. To give an example of interpretation: given a \$1mm+ SDB trade, if yield spread doubles [×2] the expected cost becomes 11.02% more expensive. SDB stands for Seller-Dealer-Buyer imputed roundtrip costs. SDD and DDB are imputed Seller-Dealer-Dealer and Dealer-Dealer-Buyer transaction costs, respectively. The symbol ** indicates significance at the 1% level and * at the 5% level. Standard errors are two-way clustered on day and bond issue. The corresponding cluster-robust t-statistics appear in brackets below the corresponding coefficients. Also reported are the deviance MSE, dispersion parameter $\widehat{\phi}$, McFadden pseudo-R^2 (in percentages), and deviance F-statistics.

		\$0–\$100k			\$100k–\$1mm			\$1mm+		
		SDB	SDD	DDB	SDB	SDD	DDB	SDB	SDD	DDB
Intercept		62.53**	56.10**	85.48**	23.33**	36.15**	60.03**	11.78**	17.51**	23.73**
		(184.14)	(326.00)	(390.07)	(89.85)	(188.61)	(261.25)	(77.22)	(91.74)	(106.16)
log(Price)	[×10%]	−0.77	1.36**	2.39**	−0.44	1.79**	−0.14	−0.79	1.15	−0.16
		(−1.76)	(7.36)	(9.42)	(−0.91)	(3.23)	(−0.49)	(−0.92)	(1.26)	(−0.36)
log(Spread)	[×2]	11.59**	3.61**	7.78**	17.47**	10.47**	11.46**	11.02**	12.17**	10.64**
		(15.77)	(13.64)	(24.50)	(22.98)	(25.64)	(24.77)	(23.12)	(20.46)	(18.17)
log(Duration)	[×2]	15.14**	9.08**	21.07**	10.48**	8.21**	17.13**	14.92**	8.16**	10.34**
		(28.30)	(46.51)	(76.49)	(15.50)	(24.97)	(18.62)	(31.61)	(15.18)	(22.29)
log(Age)	[×2]	1.50**	2.86**	0.57**	0.58	2.30**	0.10	1.05**	0.99**	−0.74**
		(5.44)	(15.48)	(6.48)	(1.55)	(10.34)	(0.39)	(4.69)	(3.48)	(−3.10)
log(AmtOut)	[×2]	0.79	−0.45*	−1.70**	−0.75	1.56**	−0.93**	−2.39**	−1.74**	−1.95**
		(1.90)	(−2.25)	(−9.36)	(−1.13)	(4.32)	(−3.63)	(−3.13)	(−2.85)	(−4.62)
log(Size)	[×2]	−6.09**	−2.41**	0.94**	−13.59**	−10.44**	−9.90**	−2.79**	−4.04**	−8.69**
		(−27.11)	(−17.92)	(12.07)	(−33.86)	(−37.84)	(−45.63)	(−8.59)	(−10.26)	(−29.16)
log(AvgSize)	[×2]	−5.53**	−0.96**	−0.50**	−3.98**	−3.04**	−1.47**	−1.83**	−3.28**	−2.95**
		(−14.69)	(−5.40)	(−3.04)	(−7.82)	(−10.66)	(−5.68)	(−5.28)	(−7.96)	(−9.07)
log(Volume)	[×2]	1.47**	0.00	0.91**	1.07**	−0.53**	0.70**	−0.62**	−1.61**	−0.26
		(8.40)	(−0.05)	(11.89)	(3.96)	(−3.89)	(5.54)	(−3.18)	(−6.75)	(−1.20)
VIX	[+1]	0.45**	−0.14**	0.06	0.57**	0.31**	0.01	1.37**	1.72**	1.12**
		(4.25)	(−3.12)	(1.34)	(3.66)	(3.04)	(0.12)	(10.00)	(10.42)	(9.22)
Senior	[=1]	−15.33**	−1.90	−7.88**	−25.85**	−3.66	−16.51**	−4.58	−9.27**	−20.39**
		(−6.81)	(−1.49)	(−6.59)	(−7.92)	(−1.81)	(−11.18)	(−1.09)	(−2.84)	(−7.28)
Callable	[=1]	44.64**	19.86**	29.34**	−2.79	22.69**	17.28**	2.74	19.66**	10.24**
		(7.96)	(4.99)	(6.17)	(−0.58)	(4.08)	(4.01)	(0.83)	(4.96)	(2.64)
Deviance MSE		0.70	0.82	0.73	1.16	1.18	1.04	0.84	1.17	1.09
Dispersion ($\widehat{\phi}$)		0.64	0.70	0.47	1.47	1.12	0.85	1.04	1.60	1.34
McFadden R^2		35.12	8.05	20.05	30.82	13.35	21.82	29.04	19.04	20.75
F-statistic		6,337	10,494	86,988	1,560	2,675	12,732	1,554	689	1,459
Observations		96,468	922,001	2,023,445	39,702	148,159	336,075	42,246	36,156	61,765

The cost model is arguably one of the most important parts of the PEX measure, such that most of the effects on transaction costs are closely related to what we observe in the PEX. The distinctive way in which we use the cost model to estimate liquidity is pronounced in Figure 4. From this figure we can see how the shape of the cost distribution changes on the bid and ask sides, both for different transaction sizes and liquidity circumstances. The latter can be observed by comparing the shapes of the distributions when taking the most liquid (P99) versus the most illiquid values (P1) for each variable. The ability of the variables to completely change the shape of the underlying distribution demonstrates the flexible character of the cost model. As expected, the cost distributions shrink as transaction size increases. Where the expected cost of the median transaction is more than 50¢ for a size below \$100k, the cost decreases to just below 20¢ for sizes above \$1mm. One explanation for this is that the size of transactions is related to the size of the institution behind it. We expect a bigger investor to have a better network of dealers and generate more business, resulting in more bargaining power when negotiating prices.

The partial effects of the cost model are shown in Table 3. Performing a likelihood ratio test with the reported F-values, we find that the cost model significantly improves the deviance compared to a model with just an intercept ($\alpha = 0.001$). In addition, the McFadden R^2 shows that the variables also yield a good improvement in explanatory power. Interestingly, SDD combinations prove to be more difficult to estimate than the DDB types. The SDB combinations yield the highest R^2 and the lowest mean squared error. The differences in the estimates of the dispersion $\widehat{\phi}$ for the different groups, give a good indication of the amount of variance in the corresponding samples. Even though the dispersion increases for higher transaction sizes, the explanatory power of the model remains satisfactory and even improves for SDD flows.

The yield spread and duration of a bond appear as the most important determinants for transaction costs. Both their effect size and statistical significance are strong and robust over all possible model specifications. Together with a bond's age, callability and the level of the VIX, these variables give the strongest increase in expected transaction costs. Surprisingly, a higher age seems to improve liquidity in the \$1mm+ DDB sample. A possible explanation is that dealers are giving buyers a discount for removing illiquid bonds from their inventory. Going over the negative effects, we observe that costs decrease when the amount outstanding is higher, the transaction size is larger, the bond is senior, or the average transaction size for that bond is above-average. The latter implies that dealers for that bond are used to transacting big lots, making the average trade easier to handle and thereby cheaper. We expected that costs also decrease if the monthly volume is higher, but this effect is not so clear-cut for all groups. Specifically, a high volume seems to adversely affect the SDB and DDB types for sizes below \$1mm, although the effect size is small. We also included the price of transactions to demonstrate its effect. Controlling for spread and duration, we find that price loses its explanatory power, thereby rendering most coefficients insignificant. Interestingly, the effect of price becomes significant again when denominating costs in basis points. This is because of the inclusion of a 'price effect', as shown in Appendix D.

Table 4: ***Warehousing rate model, average partial effects***. This table shows the average partial effects of a probit regression on the event that trades are taken into inventory by dealers ($y = 1$) or are instantaneous roundtrips ($y = 0$). We perform separate regression per size group and buy or sell side. Continuous regressors have been demeaned before estimation. The listed numbers are partial effects and are denoted in percentage terms, except for the intercept. The intercept can be interpreted as the average warehousing rate in the corresponding sample because it is transformed with the standard normal CDF, $\Phi(\cdot)$. The average partial effect of variable x_k is calculated and transformed as described in Section 3.2. The effect multiplier m_k we use for variable x_k appears in square brackets after the name. The symbol ** indicates significance at the 1% level and * at the 5% level. Standard errors are two-way clustered on day and bond issue. The corresponding cluster-robust t-statistics appear in brackets below the corresponding coefficients. Also reported are the Brier score, McFadden pseudo-R^2 (in percentages) and sum of squared Pearson residuals (Pearson SSR). The Brier score can be interpreted as a MSE (Appendix H). A chi-squared test of the Pearson SSR is highly insignificant ($p \approx 1$) for all samples.

		\$0–\$100k		\$100k–\$1mm		\$1mm+	
		buying	selling	buying	selling	buying	selling
Φ(Intercept)		0.98**	0.96**	0.96**	0.93**	0.92**	0.91**
		(85.80)	(53.37)	(58.22)	(49.16)	(74.90)	(72.11)
log(Price)	[×10%]	0.01	0.72**	−0.05	−0.10	0.68**	0.37**
		(0.18)	(6.81)	(−0.29)	(−0.56)	(5.78)	(2.78)
log(Spread)	[×2]	0.51**	−0.33**	−0.20	−0.80**	−1.41**	−1.25**
		(11.92)	(−3.86)	(−1.33)	(−3.89)	(−7.75)	(−6.72)
log(Duration)	[×2]	0.21**	0.29**	0.50**	0.79**	1.25**	1.28**
		(3.39)	(3.70)	(3.50)	(4.64)	(9.86)	(9.63)
log(Age)	[×2]	−0.30**	0.48**	−0.29**	0.62**	0.06	−0.06
		(−8.26)	(7.58)	(−3.15)	(4.95)	(0.81)	(−0.73)
log(AmtOut)	[×2]	−1.07**	−0.13	−0.58**	0.22	−0.81**	0.10
		(−14.32)	(−1.42)	(−3.21)	(0.90)	(−4.68)	(0.58)
log(Size)	[×2]	−0.08*	−0.64**	−0.84**	−0.30**	−1.23**	0.18*
		(−2.16)	(−13.03)	(−16.00)	(−4.01)	(−19.76)	(2.54)
log(AvgSize)	[×2]	0.25**	0.41**	−0.88**	−0.74**	−0.10	−0.46**
		(4.06)	(3.62)	(−6.69)	(−4.25)	(−0.77)	(−3.24)
log(Volume)	[×2]	0.23**	−0.18**	0.27**	−0.30**	0.77**	0.05
		(11.02)	(−5.01)	(5.38)	(−3.80)	(12.64)	(0.82)
VIX	[+1]	0.01	−0.03**	0.04**	0.03	−0.02	−0.02
		(1.78)	(−3.74)	(3.45)	(1.75)	(−1.58)	(−1.44)
Senior	[=1]	−0.05*	0.01	−0.12**	−0.11*	−0.18**	−0.18**
		(−2.16)	(0.24)	(−3.51)	(−2.53)	(−6.09)	(−5.95)
Callable	[=1]	−1.47**	−0.95	−0.13	1.00	0.52	0.79
		(−3.89)	(−1.62)	(−0.30)	(1.85)	(1.29)	(1.85)
Brier score		0.03	0.04	0.05	0.07	0.09	0.10
McFadden R^2		2.66	2.26	0.85	0.83	1.07	0.44
Pearson SSR		5,150,238	3,355,462	1,812,272	1,285,795	1,018,426	924,749
Observations		5,109,938	3,401,398	1,811,855	1,285,281	1,016,708	924,726

Table 5: ***Arrival rate model estimates***. This table shows the coefficient estimates of a negative binomial regression with log-link of the amount of incoming buyers and sellers per day. Regressions are repeated for different size groups. Continuous regressors have been demeaned before estimation. Listed numbers are transformed incidence rate ratios, expressed in percentages. The intercept can be interpreted as the average expected amount of arrivals per day. The model coefficient $\widehat{\beta}_k$ is transformed into its incidence rate ratio using either $\exp(\widehat{\beta}_k m_k)$ for the effect $x_k + m_k$ on non-transformed variables or $\exp(\widehat{\beta}_k \ln(m_k))$ for the effect $x_k \cdot m_k$ on log transformed variables (Section 3.2). The effect multiplier m_k we use for variable x_k appears in square brackets after the name. To give an example of interpretation: given a \$1mm+ trade, if bond age doubles [×2] the expected amount of incoming buyers per day decreases with 4.37%. The symbol ** indicates significance at the 1% level and * at the 5% level. Standard errors are two-way clustered on day and bond issue. The corresponding cluster-robust t-statistics appear in brackets below the corresponding coefficients. Also reported are the deviance MSE, the estimated shape parameter $\widehat{\theta}$ of the negative binomial distribution, the McFadden pseudo-R^2 (in percentages) and the sum of squared Pearson residuals (Pearson SSR). A chi-squared test of the Pearson SSR is highly insignificant ($p \approx 1$) for all samples. The listed metrics are further explained in Appendix H.

		\$0–\$100k		\$100k–\$1mm		\$1mm+	
		buyers	sellers	buyers	sellers	buyers	sellers
Intercept		0.73**	0.52**	0.25**	0.21**	0.08**	0.08**
		(−10.70)	(−21.19)	(−65.53)	(−75.37)	(−165.93)	(−166.77)
log(Spread)	[×2]	14.75**	2.73**	5.01**	3.62**	2.52**	3.97**
		(38.82)	(9.19)	(17.84)	(11.90)	(9.94)	(13.58)
log(Duration)	[×2]	−4.40**	−1.27**	−4.55**	−4.10**	−1.86**	−2.50**
		(−13.97)	(−4.65)	(−22.24)	(−20.13)	(−10.50)	(−13.10)
log(Age)	[×2]	−7.10**	−1.81**	−6.06**	−3.89**	−4.37**	−5.07**
		(−38.14)	(−9.61)	(−53.58)	(−31.82)	(−46.65)	(−46.04)
log(AmtOut)	[×2]	21.31**	34.86**	22.40**	27.72**	18.40**	22.55**
		(49.72)	(86.23)	(75.03)	(91.90)	(76.72)	(81.98)
log(AvgSize)	[×2]	−20.42**	−20.41**	−8.07**	−5.92**	0.61**	0.11
		(−77.79)	(−78.16)	(−38.78)	(−28.88)	(4.03)	(0.67)
log(Volume)	[×2]	11.78**	7.37**	11.24**	7.66**	13.07**	9.89**
		(73.71)	(60.69)	(100.19)	(73.43)	(120.05)	(79.16)
VIX	[+1]	0.29**	−0.27**	−0.11	−0.35**	−0.23**	−0.22**
		(3.26)	(−3.66)	(−1.49)	(−4.18)	(−3.14)	(−2.58)
Senior	[=1]	−2.08	14.13**	0.87	5.44**	8.66**	7.38**
		(−0.69)	(4.18)	(0.41)	(2.60)	(5.56)	(4.83)
Callable	[=1]	−5.83*	13.69**	−0.83	8.56**	3.87**	5.56**
		(−2.17)	(4.42)	(−0.46)	(4.15)	(2.71)	(3.41)
Deviance MSE		10.31	4.74	10.86	8.22	43.91	29.28
Dispersion ($\widehat{\theta}$)		1.13	2.55	1.88	2.10	1.11	1.16
McFadden R^2		46.92	48.72	35.57	28.80	26.76	23.94
Pearson SSR		13,579,059	7,721,764	7,257,008	6,927,541	7,540,585	7,185,467
Observations		6,519,328	6,519,328	6,512,396	6,512,396	6,488,779	6,488,779

Apart from the cost model, the warehousing and arrival rate models can also largely influence the final estimate of the probability of execution. The explanatory power of the warehousing rate model, as shown in Table 4, might appear surprisingly low on first sight. However, the listed McFadden R^2 is low by construction: the probit regression aims to classify whether individual trades involve dealer inventory or not. Our goal for the warehousing rate model, on the other hand, is to estimate the average warehousing rate of transactions for a bond in general. The regressors we employ are therefore not aimed at correctly classifying the warehousing rate of individual trades, which keeps the explanatory power artificially low. Nevertheless, the model achieves satisfactory Brier scores and generates significant coefficients for most variables.

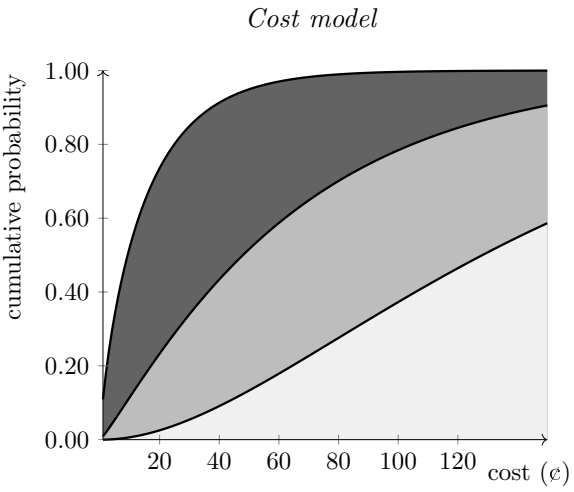
The most influential variables on the warehousing rate are a bond's yield spread, duration, amount outstanding and monthly volume. Of these, the yield spread and duration are the most stable, but show opposite effects. The effect of yield spread has an obvious interpretation: higher spreads lead to increased inventory risk, which makes dealers increasingly wary to take bonds into inventory. As a consequence, we observe more instantaneous roundtrips. For duration, the effect is the opposite: a possible explanation is that a higher duration leads to a less liquid market. Dealers are therefore forced to use their inventory in order to make a market, simply because pre-arranging transactions is more difficult. Moving on to the size and average size of transactions, we observe an influence on the warehousing rate that is mostly negative. This can be argued by the theory that dealers increasingly prefer to pre-arrange roundtrips for large transfers. Lastly, we find that seniority instills confidence in dealers to use their inventory more.

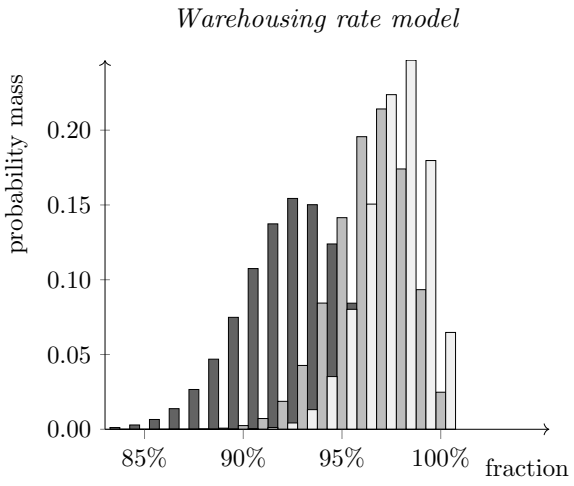
The results for the arrival rate model, displayed in Table 5, advocate its importance within the PEX measure. It achieves a relatively high explanatory power, despite the samples showing large dispersion. Driven by outliers, bonds that are very liquid and have a lot of transactions per day, the estimated Pearson residuals are higher than would be ideal. Nevertheless, almost all variables are highly significant and have a large influence on the arrival rate. Of specific interest are the amount outstanding and the bond's volume of the previous month. Both variables increase the arrival rates of both buyers and sellers with considerable effect. Interestingly, bonds with a higher yield spread also generate more market activity. We hypothesise that the higher yield either attracts more investors to a bond, or results in more transactions due to a higher turnover. On the contrary, we find that market activity mostly decreases for old bonds, which can be explained by the on-the-run effect: the most recently issued bonds of given tenor attract the most trading volume. Finally, we observe that senior bonds generate more activity than their subordinated counterparts. Likewise, callable bonds generate more arrivals than normal bonds.

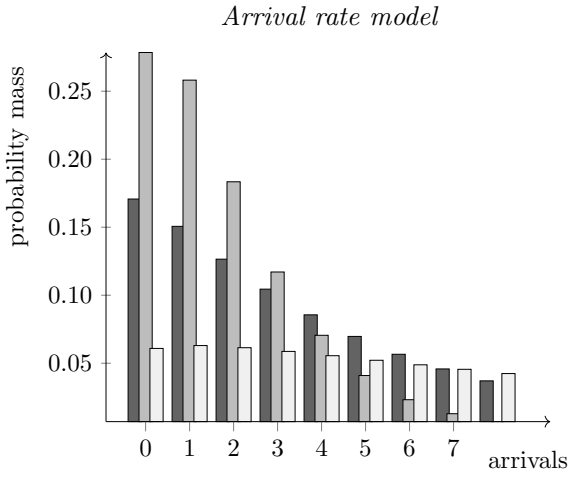
To give an impression of the results of the models and their contribution to the PEX, we have compiled three examples of DDB transfers in Figure 5. The imputed costs of the three examples can be compared to the estimated buy cost in our framework, denoted as $\mathbb{E}[C_B]$. The distributions in the plots are those used in the resulting PEX calculations.

Figure 5: ***Example transactions***. This figure shows all model results for three DDB trades of General Motors 8% 1/11/31 (example A), Bank of America 7.4% 15/1/11 (B) and Apple 2.4% 3/5/23 (C). Costs are in dollar cents and the PEX is calculated for a target cost of 25¢.

Examples of buy transactions			
Example	A	B	C
Bond ID	GM-31	BAC-11	AAPL-23
Transaction date	28/7/06	14/11/08	11/6/13
Transaction time	09:26:00	13:51:09	09:46:01
Type	DDB	DDB	DDB
Imputed costs	150	61	20
Price ($)	99.8	102.3	93.9
Spread (bps)	329	625	72.5
Duration (years)	10.36	2.02	9.06
Age (years)	4.68	7.79	0.09
AmtOut ($mm)	4000	1629	5500
Size ($k)	50	190	2000
Average size ($k)	1575	377	2684
Volume ($mm)	1265	121	2254
VIX	14.9	59.8	15.4
Senior	yes	no	yes
Callable	no	no	no
Model results			
$\mathbb{E}[C_{\text{SDB}}]$	81	62	7
$\mathbb{E}[C_{\text{SDD}}]$	69	60	8
$\mathbb{E}[C_{\text{DDB}}]$	153	65	16
$\widehat{\gamma}_B$	0.97	0.96	0.93
$\widehat{\gamma}_S$	0.95	0.93	0.90
$\mathbb{E}[C_B]$	151	66	15
$\mathbb{E}[C_S]$	70	60	7
$\mathbb{P}[C_B \leq 25]$	0.04	0.29	0.80
$\mathbb{P}[C_S \leq 25]$	0.23	0.37	0.94
$\mathbb{P}[N^B > 0]$	0.94	0.72	0.83
$\mathbb{P}[N^S > 0]$	0.92	0.59	0.79
PEX buy (c=25)	0.04	0.17	0.63
PEX sell (c=25)	0.21	0.27	0.78
Distribution parameters for buy transactions			
Cost model			
$\widehat{\alpha}$ in $\Gamma(\widehat{\alpha}, \widehat{\beta})$	2.11	1.14	0.76
$\widehat{\beta}$ in $\Gamma(\widehat{\alpha}, \widehat{\beta})$	0.01	0.02	0.05
Warehousing rate			
$\widehat{p}$ in Bin$(100, \widehat{p})$	0.97	0.96	0.93
Arrival rate			
$\widehat{m}$ in NB$(\widehat{m}, \widehat{r})$	12.34	1.83	4.36
$\widehat{r}$ in NB$(\widehat{m}, \widehat{r})$	1.13	1.88	1.11







Now that we have presented the main results for the three models and the PEX measure, we shortly discuss the decisions and assumptions behind the presented models. For the cost model, we use a log-link GLM with gamma distribution, because it gives the best fit. We demonstrate the fit for $1mm+ SDB transactions for both normal linear models and GLMs in Appendix I. As can be seen from the various model specifications, a normal OLS model is not appropriate: the normal distribution does not fit the cost distribution well, even if costs are log-transformed. We do not consider limited dependent variable models, such as censored or truncated regressions, because transaction costs are neither censored nor truncated: costs are simply positive by definition. We find that the log-link GLM with gamma distribution is the only specification that is able to accurately capture low transactions costs, even though it is not able to fit extremely high outliers. The gamma distribution is also flexible enough to model the various shapes of the cost distribution for different types of market participants (visualised in Figure 4).

For the warehousing rate model, we use a probit regression to stay consistent with Hendershott and Madhaven (2015) and to stay close to the probabilistic setting of the PEX. We have no doubt that a logit regression would give the same results. Nevertheless, the estimated warehousing rates are largely dominated by positive outcomes, and it might therefore be worth investigating whether zero-inflated models, or complementary-log-log models, provide a better fit. That being said, our sample sizes are large enough such that we did not run into finite sample problems when estimating the probit regressions. For the arrival rate model, we use the negative binomial regression because the data suffers from significant overdispersion. We provide formal results of overdispersion in Table 10 in Appendix I, where we use the test of Cameron and Trivedi (1990). Investigating the accuracy of the arrival rate model, we find that the negative binomial regression gives a good fit for arrivals under $1mm. For the highest size group, it might be worth looking into zero-inflated negative binomial models to achieve a better fit.

When analysing the deviance and Pearson mean squared errors, we find that the three models are able to estimate their respective liquidity dimensions with satisfactory precision. Additionally, the regressors generally have significant coefficients. When comparing the performance of the models to naive moving averages, we find that they also mostly outperform them. An overview of the performance of naive moving averages is provided in Appendix K. Interestingly, the naive moving average approach works quite well for estimating transaction costs and arrival rates for smaller transaction sizes. Nevertheless, the performance results depend heavily on the chosen lag. In the same Appendix K, we also replicate Roll's measure as in Bao et al. (2008) and show that half of the implied bid-ask spread by Roll's measure largely underestimates imputed transaction costs. A major weakness of naive measures is that they rely heavily on data of recent transactions. Especially for illiquid bonds, where an accurate liquidity proxy is most useful, it can therefore be difficult to do inferences. For example, a bond may not trade for a prolonged period of time, such that the proxy is outdated for both market conditions and bond characteristics may have changed. Additionally, naive liquidity proxies lack a connection to

transaction size, thereby generalising the same estimate for all market participants. We observe that liquidity is fragmented between different size groups: some bonds may be very liquid for small transaction sizes but can simultaneously be illiquid for large sized block trades. This makes such naive measures less practical for institutional investors.

The moving averages do bring up the importance of time variation, given that the corporate bond market has changed a lot over the past couple of years. The 2010 Volcker Rule and the Third Basel Accord prevent banks from making speculative investments, which, in turn, decreases the potential of banks to provide liquidity for many assets. As a consequence, the liquidity landscape of corporate bonds is believed to be in a deteriorating state: there seem to be too many bonds and too little dealers. Nevertheless, the amount of issued debt is ever increasing and the total amount of trading volume is not disappointing (SIFMA, 2016). We therefore propose to extend our research by further investigating how the performance of the PEX measure changes throughout time. Even though we have included regressors that incorporate market conditions, a comparison of parameter estimates in recent years compared to periods such as the 2008 financial crisis should provide extra insight into parameter stability.

To investigate the sensitivity of the PEX measure, we perform a robustness check on the detection period of the imputed transaction types in Appendix L. We find that increasing this window leads to the detection of more combinations, and slightly higher transaction costs on average. The latter is probably caused by intraday interest rate movements. Nevertheless, the costs increase only slightly because the warehousing rate scales the types according to their prevalence. If more SDB transactions are found, and DDB transactions become more expensive, the weighted expected buy costs can still remain relatively constant. Other detection periods are therefore also viable options, and lead to slightly different cost estimates.

Apart from the detection period, the PEX also depends on other assumptions. We assume that SDD and DDB transfers are representative of approximately half of the effective bid-ask spread, which might not always be true: it can be argued that the second dealer is in fact a customer too. This because the second dealer is willing to take the bonds into inventory, or sell from inventory, thereby posing as the opposite liquidity provider to the transaction. From that point of view, the SDD and DDB transactions are representative of more than half of the bid-ask spread. Additionally, we assume that buyers can transact with the same rate at which sellers arrive at the market and vice versa. Although we take the differences in the arrival rates to be smoothed out by dealers, this approach is still quite rough. The arrival rate model could be improved to better estimate the immediacy dimension of liquidity, for example by estimating the actual volume that either arrives at the market or temporarily resides in the interdealer circuit. We also assume independence of the three models within the PEX measure, which might not hold. In order to investigate this assumption, we bootstrap residuals of the various components in the PEX and plot them against each other. Because the three models use different samples,

we only use those observations that appear in all three. For the warehousing and arrival rate probability components, we calculate residuals by taking the average rates over the month in which the transaction occurred. Although true residuals cannot be calculated, given that the liquidity aspects are unobserved, this approximation gives an idea of how the residuals look like. We bootstrap the resulting sample using 100,000 draws. In order to calculate 95% confidence regions, we employ two-dimensional kernel density estimation using a Gaussian kernel and the rule-of-thumb bandwidth of Venables and Ripley (2013). The resulting plots are displayed in Appendix J. We find that the various PEX components do not suffer from obvious correlations that are persistent throughout the size groups. Nevertheless, the plots show that the residuals can covary, such that the PEX may suffer from biases for certain bonds. Further research is needed to investigate whether the independence assumption is indeed valid. From the plots, we also observe that transforming the arrival rates into a probability makes the result slightly upward biased. As a result, the PEX might overestimate liquidity. Because true residuals do not exist, we are not able to formally verify this. Another issue, is that the causality between the liquidity dimensions is still unclear: do cheaper transaction costs attract more activity, or does more activity result in cheaper transaction costs? The same goes for our limited understanding of the warehousing rates. FINRA's proposal for constructing a dataset with masked dealer identifiers might change this (SEC, 2016). Using this new data, future research can study the behaviour of individual dealers and the circumstances under which they change their inventory policy.

Lastly, we would like to make the recommendation to institutional investors to employ a flexible cost target, not a fixed one. From the various examples presented in this thesis, it can be observed how the PEX reacts heavily to deteriorating liquidity conditions. By adjusting the cost target accordingly, investors can stabilise their chances of getting an execution while maintaining a competitive position in the bilateral bargaining process with dealers.

6. Conclusion

The proposed PEX measure, as developed in this thesis, is the result of three individual models. We observe that each model is able to capture a different aspect of corporate bond liquidity which sufficient accuracy, such that we believe the resulting PEX measure to be a good proxy for the probability of executing a transaction. Empirical evidence suggests that the cost model can accurately estimate imputed transactions, thereby outperforming naive estimators. The warehousing rate model has low explanatory power per construction, but nevertheless accomplishes its task and yields satisfactory statistical significance for most regressors. The arrival rate model proves to be a great addition to any liquidity proxy, and explains the immediacy dimension of liquidity with good explanatory power. Although we acknowledge that more research is needed to further establish the validity and implications of the PEX measure, this thesis gives first evidence that the PEX measure is a stepping stone towards successfully capturing the multiple liquidity dimensions of transacting in the corporate bond market.

Bibliography

Amihud, Y. (2002). Illiquidity and stock returns: cross-section and time-series effects. *Journal of financial markets*, 5(1):31–56.

Augustin, N. H., Sauleau, E.-A., and Wood, S. N. (2012). On quantile quantile plots for generalized linear models. *Computational Statistics & Data Analysis*, 56(8):2404–2409.

Bao, J., Pan, J., and Wang, J. (2011). The illiquidity of corporate bonds. *The Journal of Finance*, 66(3):911–946.

Ben Dor, A., Dynkin, L., Hyman, J., Houweling, P., Leeuwen, E. V., and Penninga, O. (2007). DTS (duration times spread). *Journal of Portfolio Management, Winter*.

Ben Dor, A., Dynkin, L., Hyman, J., and Phelps, B. D. (2012). *Quantitative Credit Portfolio Management: Practical Innovations for Measuring and Controlling Liquidity, Spread, and Issuer Concentration Risk*. John Wiley & Sons, Inc.

Bessembinder, H., Maxwell, W., and Venkataraman, K. (2006). Market transparency, liquidity externalities, and institutional trading costs in corporate bonds. *Journal of Financial Economics*, 82(2):251–288.

Biswas, G., Nikolova, S., and Stahel, C. W. (2014). The transaction costs of trading corporate credit. *Available at SSRN 2532805*.

Black, F. (1970). Fundamentals of liquidity. *Mimeograph, Associates in Finance, June*.

Breslow, N. E. (1984). Extra-poisson variation in log-linear models. *Applied statistics*, pages 38–44.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2012). Robust inference with multiway clustering. *Journal of Business & Economic Statistics*.

Cameron, A. C. and Trivedi, P. K. (1990). Regression-based tests for overdispersion in the poisson model. *Journal of econometrics*, 46(3):347–364.

CGFS (2016). Fixed income market liquidity. *CGFS Papers No 55, January 2016, Committee on the Global Financial System.* Available at http://www.bis.org/publ/cgfs55.pdf.

Davidian, M. (2005). Lecture notes in applied longitudinal data analysis: Generalized linear models for nonnormal response (chapter 11).

Dick-Nielsen, J. (2009). Liquidity biases in TRACE. *Journal of Fixed Income*, 19(2).

Dick-Nielsen, J. (2014). How to clean enhanced TRACE data. *Available at SSRN 2337908*.

Dick-Nielsen, J. and Rossi, M. (2016, forthcoming). The cost of immediacy for corporate bonds.

Dobson, A. J. and Barnett, A. (2008). *An introduction to generalized linear models*. New York: Chapman & Hall.

Dunson, D. (2005). Generalized linear models [lecture notes]. *Department of Statistical Science, Duke University*. Accessed at https://www2.stat.duke.edu/courses/Fall05/sta216/.

Edwards, A. K., Harris, L., and Piwowar, M. S. (2007). Corporate bond market transaction costs and transparency. *The Journal of Finance*, 62(3):1421–1451.

European Banking Authority (February 2013). On defining liquid assets in the LCR under the draft CRR. Technical report. https://www.eba.europa.eu.

Feldhütter, P. (2012). The same bond at different prices: identifying search frictions and selling pressures. *Review of Financial Studies*, 25(4):1155–1206.

Friewald, N., Jankowitsch, R., and Subrahmanyam, M. G. (2012a). Liquidity, transparency and disclosure in the securitized product market. Available at SSRN 2139310.

Friewald, N., Jankowitsch, R., and Subrahmanyam, M. G. (2012b). Illiquidity or credit deterioration: A study of liquidity in the US corporate bond market during financial crises. *Journal of Financial Economics*, 105(1):18–36.

Geyer, C. J. (2003). 5601 Notes: The Sandwich Estimator. *School of Statistics, University of Minnesota*. Available from http://www.stat.umn.edu/geyer/5601/notes/sand.pdf.

Goldstein, M. A., Hotchkiss, E. S., and Sirri, E. R. (2007). Transparency and liquidity: A controlled experiment on corporate bonds. *Review of Financial Studies*, 20(2):235–273.

Green, R. C., Hollifield, B., and Schürhoff, N. (2007). Financial intermediation and the costs of trading in an opaque market. *Review of Financial Studies*, 20(2):275–314.

Harris, L. (2015, forthcoming). Transaction costs, trade throughs, and riskless principal trading in corporate bond markets. Working paper, version 1.03, October 22, 2015.

Harris, L. et al. (1990). Liquidity, trading rules and electronic trading systems. Technical report.

Harris, L. and Piwowar, M. S. (2006). Secondary trading costs in the municipal bond market. *The Journal of Finance*, 61(3):1361–1397.

Hendershott, T. and Madhavan, A. (2015). Click or call? auction versus search in the over-the-counter market. *The Journal of Finance*, 70(1):419–447.

Hinde, J. and Demétrio, C. G. (1998). Overdispersion: models and estimation. *Computational Statistics & Data Analysis*, 27(2):151–170.

Ho, T. S. and Lee, S. B. (2004). *The Oxford Guide to Financial Modeling: Applications for Capital Markets, Corporate Finance, Risk Management and Financial Institutions*. Oxford university press.

Houweling, P., Mentink, A., and Vorst, T. (2005). Comparing possible proxies of corporate bond liquidity. *Journal of Banking & Finance*, 29(6):1331–1358.

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2005). *Applied linear statistical models*, volume 5. McGraw-Hill/Irwin, New York.

Lin, M., Lucas Jr, H. C., and Shmueli, G. (2013). Too big to fail: large samples and the p-value problem. *Information Systems Research*, 24(4):906–917.

Lybek, T. and Sarr, A. (2002). *Measuring liquidity in financial markets*. Number 2-232. International Monetary Fund.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, volume 37. New York: Chapman & Hall.

Mizrach, B. (2015). Analysis of corporate bond liquidity. Technical report, FINRA Office of the Chief Economist.

Qualtieri, A. J. and McCusker, P. (2014). Evolving liquidity dynamics of fixed income and cash markets. Technical report, State Street Corporation, Global Advisors.

Rindi, B. (2008). Informed traders as liquidity providers: Anonymity, liquidity and price formation. *Review of Finance*, 12(3):497–532.

Roll, R. (1984). A simple implicit measure of the effective bid-ask spread in an efficient market. *The Journal of Finance*, 39(4):1127–1139.

Schultz, P. (2001). Corporate bond trading costs: A peek behind the curtain. *The Journal of Finance*, 56(2):677–698.

SEC, Securities and Exchange Commission (2016). Order Granting Approval of Proposed Rule Change to Create an Academic Corporate Bond TRACE Data Product. Available from the SEC at https://www.sec.gov/rules/sro/finra/2016/34-78759.pdf.

SIFMA, Securities Industry and Financial Markets Association (2016). September 2016, US bond market issuance and outstanding [dataset], US bond market trading volume [dataset]. Available from SIFMA at http://www.sifma.org/research/statistics.aspx.

Sommer, P. and Pasquali, S. (2016). Liquidity - how to capture a multidimensional beast. *The Journal of Trading*, 11(2):21–39.

Venables, W. N. and Ripley, B. D. (2013). *Modern applied statistics with S-PLUS*. Springer Science & Business Media.

Ziliak, S. T. and McCloskey, D. N. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. University of Michigan Press.

Appendices

A. Liquidity dimensions for limit order markets

One of the first comprehensive definitions of liquidity was proposed by Harris (1990). According to Harris, liquidity can be measured in four different dimensions: *width* (or tightness), *depth*, *immediacy* and *resiliency*. This definition set the stage for much of the following literature and is still in use today (European Banking Authority, 2013). It can be argued that there exists a fifth dimension, as done by Lybek and Sarr in IMF's working paper: 'Measuring Liquidity in Financial Markets' (2002). They show that market liquidity can also be measured by its *breadth*. Sommer and Pasquali (2016) provide a short explanation for each of the first four dimensions. We describe the last dimension, breadth, by following the definition of Lybek and Sarr.

Width measures the cost of consuming liquidity immediately, without regard for tradable quantity. The width of a market is often measured by the minimum spread between bid and ask prices. *Depth* is the total quantity of assets available for consumption. This can be measured by the quoted quantities both above or below the current tradable prices of the asset, signifying the existence of abundant orders. *Immediacy* is the speed with which orders can be transacted, which is especially important for larger trades. *Resiliency* is the time it takes for the price to return to value (or the 'pre-trade equilibrium') after a trade has taken place. Lastly, *breadth* is the size and frequency of orders. It can be argued that a market is more liquid if its average orders at the bid and ask sides are larger and appear with higher frequency.

It is possible to describe the five liquidity dimensions in terms of cumulatively available volume at the bid and ask side of market. We have visualised this in Figure 6. This definition of liquidity is based on exchange-traded limit order markets like the equity markets. These markets guarantee that transactions take place against the best prevailing price on an exchange. Here, the breadth of the market is visualised as the slope of available volume over the distance from the midpoint. This because breadth captures the frequency and size of orders at the bid and ask sides.

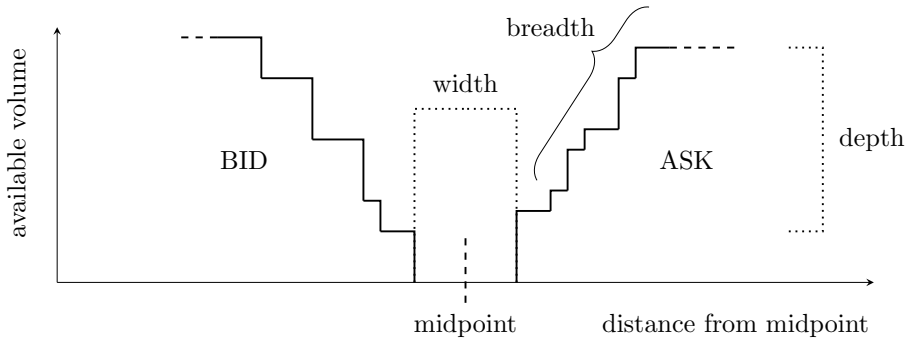


Figure 6: Visualisation of the liquidity dimensions for a limit order market

B. Estimation theory for generalized linear models

We assume that observations y_i are conditionally-independent given a vector of regressors x_i for $i = 1, ..., n$. We also assume that the conditional distribution of $y_i|x_i$ belongs to the exponential family. Let μ_i be the expected value of y_i and let $g(\cdot)$ be a one-to-one continuously differentiable transformation. We define the linear predictor η_i as:

$$g(\mu_i) = \eta_i = x_i'\beta \tag{41}$$

$$\text{s.t.} \quad \mathbb{E}[y_i] = \mu_i = g^{-1}(\eta_i) = g^{-1}(x_i'\beta) \tag{42}$$

where β is a vector of unknown coefficients. We can express the probability density of any probability distribution in the exponential family as:

$$f(y_i|\theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\} \tag{43}$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known functions. Functions $a(\cdot)$ and $c(\cdot)$ identify the chosen distribution and $b(\cdot)$ depends on the chosen link function. The relation between $g(\cdot)$ and $b(\cdot)$ is $b'(\theta_i) = g^{-1}(x_i'\beta)$ (Dobson & Barnett, 2008). If $\eta_i = \theta_i$ because $b'(\cdot)$ is $g^{-1}(\cdot)$, the link function is called *canonical*. We use non-canonical link functions for all three models. The mean and variance are $\mathbb{E}[y_i] = \mu_i = b'(\theta_i)$ and $\mathbb{V}ar[y_i] = \sigma_i^2(\theta_i, \phi) = b''(\theta_i)a(\phi)$. We take the dispersion of the regression $a(\phi)$ to be constant: $a(\phi) = \phi$. The variance function then boils down to $\sigma_i^2(\theta_i, \phi) = b''(\theta_i)\phi$, where θ_i is a function of η_i. The corresponding likelihood function becomes:

$$L(\beta|y, \phi, x) = \prod_{i=1}^{n} \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right\} \tag{44}$$

Where y, β and x are vectors and θ a function of the vector η. We can find the maximum likelihood estimate of β for a fixed dispersion parameter ϕ as follows:

$$\widehat{\beta} = \arg\max_{\beta} L(\beta|y, \phi, x) \tag{45}$$

As explained in Dobson and Barnett (2008), we can now use a Fisher scoring method to maximise the likelihood. First, note that the log-likelihood can be written as follows:

$$l(\beta|y, \phi, x) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \tag{46}$$

To maximise this log-likelihood, we simply apply Fermat's theorem by taking $\partial l/\partial \beta_j = 0$ for parameter β_j. Using the chain rule and working out terms, this yields:

$$\frac{\partial l(\beta|y,\phi,x)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} \tag{47}$$

$$= \sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi}\frac{1}{\sigma_i^2}\frac{\partial \mu_i}{\partial \eta_i}x_{ij} \tag{48}$$

$$= \sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi}W_i\frac{\partial \eta_i}{\partial \mu_i}x_{ij} \tag{49}$$

which is repeated for all coefficients. The weight term W_i is defined as follows:

$$W_i = \frac{1}{\sigma_i^2}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \tag{50}$$

Putting (49) equal to zero and multiplying with ϕ yields the following scoring equation for β_j:

$$U_j = \sum_{i=1}^{n} W_i(y_i - \mu_i)\frac{\partial \eta_i}{\partial \mu_i}x_{ij} \tag{51}$$

We can now derive the variance-covariance matrix $\mathfrak{J}_{jk}$ between coefficient j and k to be:

$$\mathfrak{J}_{jk} = \mathbb{E}[U_j U_k] = \sum_{i=1}^{n} \frac{x_{ij}x_{ik}}{\sigma_i^2}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \tag{52}$$

where we assumed independence between the random variables Y_i and Y_j for $i \neq j$. In order to find the estimates b of the coefficients β, we can use Fisher's iterative scoring equation:

$$b^{(m)} = b^{(m-1)} + \left[\mathfrak{J}^{(m-1)}\right]^{-1} U^{(m-1)} \tag{53}$$

where $\mathfrak{J}$ is the information matrix with elements $\mathfrak{J}_{jk}$ for all j and k. The term $\left[\mathfrak{J}^{(m-1)}\right]^{-1}$ denotes its inverse. The U matrix contains the coefficient scores. The vector $b^{(m)}$ denotes the parameter estimates after the mth iteration.

Rewriting this scoring algorithm, we find that equation (53) yields (Dobson & Barnett, 2008):

$$X^T W X b^{(m)} = X^T W z \tag{54}$$

with W the matrix of weights and z a vector with the adjusted dependent variables:

$$z_i = \widehat{\eta}_i + (y_i - \widehat{\mu}_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right) \qquad \text{for } i = 1, ..., n \tag{55}$$

where $\widehat{\eta}_i^{(m)} = x_i' b^{(m)}$, $\widehat{\mu}_i^{(m)} = g^{-1}(\widehat{\eta}_i^{(m)})$ and the derivative $\partial\eta_i/\partial\mu_i$ is evaluated at $\widehat{\mu}_i^{(m)}$. This cannot be solved at once because b depends on W and z, and in turn W and z depend on b. However, this representation lends itself to iteratively reweighted least squares estimation:

1. Given an estimate $b^{(m)}$, calculate $\widehat{\eta}_i^{(m)}$ and $\widehat{\mu}_i^{(m)}$ for all $i = 1, ..., n$. Now calculate the new adjusted dependent variables from equation (55), which yields a new vector $z^{(m+1)}$.

2. Calculate the new weights from equation (50):

$$W_i^{(m+1)} = \frac{1}{\sigma_i^2\left(\widehat{\mu}_i^{(m)}, \phi\right)} \left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 \tag{56}$$

3. We can now solve equation 54 to yield the new estimate $b^{(m+1)}$. One way of doing this is by regressing $z^{(m+1)}$ on X with weight $W^{(m+1)}$ (a.k.a. weighted least squares). Now continue with step 1 until consecutive improvements in the estimates of b are sufficiently small.

This procedure is initialised with an initial approximation $b^{(0)}$ and weights W set to one.

For the cost model, we estimate the dispersion parameter $\widehat{\phi}$ as explained in the next section. For the negative binomial distribution, the dispersion parameter $\widehat{\phi}$ coincides with the shape parameter of the gamma distribution (written as $\widehat{\theta}$ in Section 2.3). In order to estimate $\widehat{\theta}$, we need to use the Gauss-Seidel method. Specifically, we initialise $\widehat{\theta}$ by fitting a standard Poisson model and relate the variance of the linear predictor to the mean of the regression (Breslow, 1984). Given this estimate of dispersion, we now simply estimate the coefficients as before. Fixing the estimated coefficients, we can find the dispersion parameter using a Newton-Raphson iterative scheme. This process of fixing either the estimated coefficients or the estimated dispersion parameter is iterated until convergence of both is achieved (Hinde et al., 1998).

Estimating the dispersion in a GLM with gamma distribution

First, we repeat that the density of an exponential family distribution can be expressed as:

$$f(y_i|\theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\} \tag{57}$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known functions. Functions $a(\cdot)$ and $c(\cdot)$ identify the chosen distribution and $b(\cdot)$ depends on the chosen link function. The relation between $g(\cdot)$ and $b(\cdot)$ is $b'(\theta_i) = g^{-1}(x_i'\beta)$ (Dobson & Barnett, 2008). For the gamma density function with mean parametrisation $Gamma(\mu, \nu)$, we can write the density in the exponential family form as follows:

$$f(y|\mu, \nu) = \frac{1}{\Gamma(\nu)y}\left(\frac{\nu y}{\mu}\right)^\nu \exp\left(\frac{\nu y}{\mu}\right) = \exp\left\{\frac{-\frac{1}{\mu}y + \ln(\frac{1}{\mu})}{\frac{1}{\nu}} + \ln\left(\frac{\nu^\nu y^{\nu-1}}{\Gamma(\nu)}\right)\right\} \tag{58}$$

From which we can deduce that $\theta = -1/\mu$ and $\mu = -1/\theta$. Also $a(\phi) = \frac{1}{\nu}$. Now in order to derive the first two moments, we do:

$$b(\theta_i) = -\ln(-\theta_i) \tag{59}$$

$$\frac{\mathrm{d}b(\theta_i)}{\mathrm{d}\theta_i} = b'(\theta_i) = \frac{-1}{\theta_i} = \frac{-1}{\frac{-1}{\mu_i}} = \mu_i \tag{60}$$

$$\frac{\mathrm{d}b'(\theta_i)}{\mathrm{d}\theta_i} = b''(\theta_i) = \frac{1}{\theta_i^2} = \mu_i^2 \tag{61}$$

It can now be derived that $\mathbb{E}[y_i|\theta_i] = b'(\theta_i)$ and $\mathbb{V}ar[y_i|\theta_i, \phi] = b''(\theta_i)a(\phi)$ (McGullagh & Nelder, 1989). Taking a constant dispersion parameter $a(\phi) = \phi$, we find that:

$$\mathbb{E}[y_i|\theta_i] = b'(\theta_i) = \mu_i \tag{62}$$

$$\mathbb{V}ar[y_i|\theta_i, \phi] = b''(\theta_i)a(\phi) = \phi\mu_i^2 \tag{63}$$

Let there be n observations and p parameters. The dispersion parameter can now be estimated:

$$\widehat{\phi} = \frac{\mathbb{V}ar[y|\theta, \phi]}{\widehat{\mu}^2} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i^2} \tag{64}$$

We find that the mean squared pearson residual is approximately equal to the dispersion:

$$MSE_p = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(y_i - \widehat{\mu}_i)}{\sqrt{V(\widehat{\mu}_i)}}\right]^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i^2} \tag{65}$$

$$\frac{MSE_p}{\widehat{\phi}} = \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i^2}}{\frac{1}{n-p}\sum_{i=1}^{n}\frac{(y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i^2}} = \frac{n-p}{n} \tag{66}$$

$$\text{s.t.} \quad MSE_p = \widehat{\phi}(n-p)/n \quad \text{and} \quad MSE_p \approx \widehat{\phi} \text{ for large } n \tag{67}$$

C. The role of dealers

The first and foremost understanding about liquidity is that it relies heavily on dealers. Even though this is a basic notion about liquidity, it is important to understand the fundamental motives and risks of dealers to understand their influence on liquidity and transaction costs.

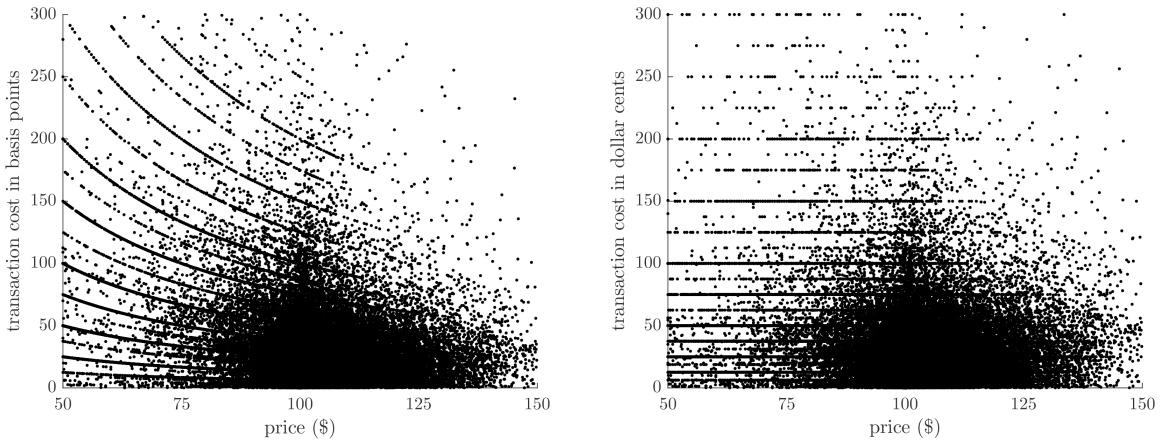
For example, consider a market without dealers. Transactions in this market can only take place if (1) opposite parties agree to transact at the same price, the same time (2) and with the same quantity (3). These are both necessary and sufficient conditions for a transaction to take place. Therefore, a certain level of heterogeneity between market participants is a prerequisite for market activity. The more homogeneous the market is, the less likely it is that you will find a counterparty that is willing to make a transaction. Specifically, this is due to the 'fair

value' of the asset. Homogeneous participants will have the same thoughts about this fair value, whereas heterogeneous participants will have differing views. In essence, if everyone knows an asset is worth exactly a specific amount p_x and all market participants expect the asset to be worth $p_y > p_x$ tomorrow, then all participants will be buyers and no transactions can take place.

So now that we have established that the liquidity of a market is directly related to the heterogeneity of its participants, we can also understand the role of the dealer. Given that a liquid market requires heterogeneity, a dealer's role is to bridge the heterogeneity between market participants with opposing views. This is accomplished by continuously acting as an opposite party to whoever is willing to transact. This bridges heterogeneity in at least one of the three necessary conditions for a transaction: price, time and quantity. But there's no such thing as a free lunch and dealers want to be compensated for taking risk on their books. This is accomplished through the bid-ask spread in which markets are quoted. The dealer offers to buy at a bid price and sell at an ask price, thus earning a small spread. Bagehot (1971; Bagehot is a pseudonym of J.L. Treynor) was one of the first to point out the biggest risk of market making: trading against informed traders that possess 'special information' about the true value of an asset. Following this line of thought, the true objective of dealers is to accommodate as many transactions as possible, while minimising risk and maximising the earned bid-ask spread. Instantaneous roundtrips, as mentioned in this thesis, are the type of transactions that are closest to a free lunch: the dealer will only need to take the corresponding bonds on his books for a negligibly small period of time (which is why we call them 'instantaneous'). This implies that the dealer is exposed to very little inventory risk and is not vulnerable to informed traders. Especially for larger transaction sizes, such pre-arranged instantaneous roundtrips are a good tool for dealers to protect against undesirable market movements and potential inventory risk.

D. The 'price effect' when costs are denominated in basis points

Harris (2015) finds that transaction prices, specifically the inverse of the price, to be statistically significant, and to explain a large part of the relative cost spreads that he uses as the dependent variable. Additionally, Harris finds that the inverse of the par value size is also significant, but negative. Harris did not anticipate this result and explains that the estimated coefficient suggests that there exists a fixed cost of \$−0.89 per trade. He reasons that this result might be wrong, and simply be due to multicollinearity. We argue that the observed fixed dollar amount cost is an artefact of the accidental inclusion of a 'price effect' when transaction costs are denominated in relative spread terms. We observe from our imputed transaction costs that dealers prefer to use fixed dollar amount markups. Subsequently, an effect between price and the resulting costs appears when transforming these fixed dollar markups relatively to price. We display this effect in Figure 7. From these graphs, it can be seen how the fixed markups begin to display an effect with respect to price when denominated in basis points instead.



(a) Costs denominated in basis points **(b)** Costs denominated in dollar cents

Figure 7: Distribution of costs versus price, for SDB costs of size $1mm+

E. Characteristics of bonds in the sample

Table 6: ***Bond characteristics***. This table gives descriptive statistics of some characteristics of the bonds in our sample (2005–2013, investment grade bonds only). The metrics in this table are dominated by liquid bonds and are therefore only representative of an average transaction in the sample, not of an average transaction on an average bond. Units appear after the variable name. Variables that are denoted in thousands of dollars have a trailing '($k)' behind the variable name. Millions of dollars are denoted by '($mm)'. P1 and P99 denote the first and 99th percentile, respectively. Variable definitions are in Section 4.3 and Appendix G.

	Mean	P1	Median	P99	Std	Min	Max
Duration (years)	6.15	0.97	4.66	17.55	4.45	0.07	28.21
Spread (bps)	162	25	137	638	142	4	3100
Age (years)	3.13	0.05	2.04	15.19	3.18	0.01	27.18
Maturity (years)	9.40	1.03	5.43	29.97	10.12	0.07	99.74
AmtOut ($mm)	757	250	500	3186	677	215	15000
Other AmtOut ($mm)	17705	0	7464	116021	26097	0	341335
Average size ($k)	1644	116	1319	7412	1616	0	39827
Monthly volume ($mm)	98.7	1.4	44.2	792.8	234.5	0	12567
Monthly turnover	9.0%	0.2%	6.2%	55.1%	10.6%	0%	163.6%
Equity listed	94.2%						
Senior	93.5%						
Callable	14.0%						

F. Filtered observations per year

Table 7: ***Filtered observations per year***. This table shows the amount of observations that are either deleted or changed as a result of each step in the filtering process. Filter conditions are described in Section 4, page 21.

	2005	2006	2007	2008	2009	2010	2011	2012	2013	Total	%
Cancels	126,988	121,933	105,745	130,517	202,684	192,329	168,385	210,525	200,365	1,459,471	1.32%
Deleted	126,549	121,803	105,652	130,402	202,514	192,165	168,249	210,695	200,583	1,458,612	1.32%
Corrections	114,606	107,420	87,546	94,888	129,863	143,316	121,302	254,409	240,874	1,294,224	1.17%
Corrected	107,883	102,130	82,986	90,422	123,982	137,157	116,350	246,963	233,080	1,240,953	1.13%
Reversals	128,877	105,926	78,645	90,646	151,186	228,596	121,535	12,165	2,765	920,341	0.84%
Deleted	172,626	149,842	114,556	126,346	214,358	407,377	212,455	18,526	2,834	1,418,851	1.29%
High price outliers	25,471	64,938	38,194	67,196	184,595	98,451	96,123	126,505	122,416	823,889	0.75%
Low price outliers	51,969	30,326	31,175	103,449	89,622	292,583	120,859	79,540	15,540	815,063	0.74%
Other outliers	622,410	363,482	266,224	615,950	1,051,960	1,241,073	1,105,435	1,078,924	960,896	7,306,354	6.63%
Double inter-dealer	1,572,434	1,513,815	1,441,356	1,923,909	3,526,709	3,540,847	3,359,465	3,696,332	3,676,370	24,251,237	22.01%
Ambiguous records	363,750	329,991	267,380	311,625	477,987	558,635	407,823	705,706	666,456	4,089,353	3.71%
Agency records	818,530	768,366	738,592	1,072,766	1,814,748	1,893,642	1,777,765	1,936,649	1,924,718	12,745,776	11.57%
Before filter	8,106,864	7,300,388	6,700,012	8,982,733	15,509,598	16,196,366	14,815,442	16,323,772	16,254,560	110,189,735	
After filter	4,353,125	3,957,825	3,696,883	4,631,090	7,947,105	7,971,593	7,567,268	8,470,895	8,684,747	57,280,531	51.98%

G. Other investigated variables of interest

Apart from the variables listed on page 28, we also tested a range of other variables. These variables are not selected for the final models because they were either (1) highly correlated to a more powerful variable that was already present in the regression, (2) did not have a clear economic interpretation for the effect on the dependent variable, (3) did not have a large effect on the dependent variable or (4) did not achieve satisfactory statistical significance.

Maturity	The remaining time in years until the maturity date of the bond.
Issue size	The number of issued bonds multiplied by the par value.
Coupon rate	The annual coupon payments of the bond, relative to its par value.
Duration×*Spread*	The interaction between the *Duration* of the bond and its *Spread*.
Other issue size	The sum of the issue sizes of the other U.S. bonds of the issuer.
Other AmtOut	The sum of the amount outstanding of the other U.S. bonds of the issuer.
Only issue	A dummy indicating whether the bond is the only U.S. issue of the issuer.
No. of other issues	The amount of other outstanding U.S. issues of the issuer.
Equity listed	A dummy indicating whether the issuer also has listed equity.
Rating	The credit rating of the bond, taken as the average between S&P and Moody's rating and converted to a 1–22 numerical scale of decreasing credit quality.
Total return	Monthly return on the bond, calculated as the change in yield to maturity.
Volatility	The volatility of the issuer's equity, calculated over 1- and 6-month windows.
Turnover	The past monthly *Volume* divided by the amount outstanding (*AmtOut*).
Missing prices	A dummy variable indicating whether the bond traded on the previous day.
Fraction missing	The fraction of days with no transactions, calculated over the last 100 days.
On-the-run	A dummy variable indicating whether a bond is both *Senior* and the youngest bond of the issuer (lowest *Age* of all outstanding bonds of the issuer).
Inventory change	The sum of aggregate dealer inventory, calculated as a running sum of the volume of bonds added to the aggregate dealer inventory (SD, SDD) and sold from dealer inventory (DB, DDB). We tested both a window of 100 days and an expanding window of the full available history of the bond.
CDX	The daily CDX investment grade credit default swap index (by Markit Group). We investigated both the level of the CDX and the daily percentage change.
Swap spread (5y)	The difference between the fixed rate component of an interest rate swap and the on-the-run U.S. Treasury yield, both for a maturity of 5 years.
TED spread (3m)	The difference between the three-month LIBOR rate (for interbank loans) and the three-month Treasury bill interest rate (government debt).

Selection procedure results for other variables of interest

Below are individual explanations of why we do not use a specific variable. For the reasons provided in Section 4.2, we decided to exclude these variables from this study. Nevertheless, most variables did show a relation to liquidity and might therefore still be interesting for other studies.

Maturity	Correlated to *Age* and *Duration*, where *Duration* is more powerful.
Issue size	Correlated to *AmtOut*, where *AmtOut* has more power and better economic interpretation.
Coupon rate	Controlling for the final set of variables, *Coupon rate* does not have a large effect on the dependent variables.
Duration×*Spread*	High explanatory power, but provides no added value to the regression if *Spread* and *Duration* are already included. Also creates instability in the regression when the other two variables are already included.
Other issue size	Correlated to *Other AmtOut*, where *Other AmtOut* has a better economic interpretation.
Other AmtOut	Decent effect, but gives some unexplainable coefficient signs. Also coefficient signs are not robust over different size groups or model specifications.
Only issue	Explanatory power is negligible and has low statistical significance.
No. of other issues	Small explanatory power, low statistical significance.
Equity listed	Small explanatory power, low statistical significance.
Rating	Because we use investment grade bonds only, *Rating* does not provide a lot of extra information. Also some coefficient signs could not be explained and coefficient signs are not robust over different size groups or model specifications.
Total return	Small explanatory power, low statistical significance.
Volatility	Small explanatory power when controlling for *Volume* and *VIX*.
Turnover	Correlated to *Volume*, where *Volume* has a larger effect.
Missing prices	Small explanatory power, low statistical significance and not robust.
Fraction missing	Small explanatory power, some coefficient signs could not be explained.
On-the-run	Small explanatory power when controlling for *Senior* and *Age*.
Inventory change	Decent effect on its own, but negligible power when controlling for *Volume*, *VIX* and *AmtOut*. Still we think this variable could provide additional information, but more research is needed to confirm this.
CDX	Correlated to *VIX*, where *VIX* has the most explanatory power.
Swap spread (5y)	Correlated to *VIX*, where *VIX* has the most explanatory power.
TED spread (3m)	Correlated to *VIX*, where *VIX* has the most explanatory power.

Table 8: ***Variables investigated in other studies.*** This table shows the variables that studies about liquidity in dealer markets have used. Both the application and the definition of the variables may differ between studies. The table shows whether a study investigates a specific variable *at all*, using a tick (✓). Note that because the studies have differing goals, they also employ different selection criteria.

		This study (2016)	Schultz (2001)	Houweling et al. (2004)	Harris et al. (2006)	Goldstein et al. (2007)	Green et al. (2007)	Edwards et al. (2007)	Bao et al. (2008)	Friewald et al. (a,b) (2012)	Hendershott et al. (2015)	Harris (2015)	Dick-Nielsen et al. (2016)
included in final framework	Price	✓			✓			✓	✓			✓	
	Duration	✓											
	Spread	✓											
	Age	✓		✓	✓	✓	✓	✓	✓		✓		
	AmtOut	✓											
	Size	✓	✓			✓	✓	✓	✓	✓	✓	✓	
	AvgSize	✓										✓	
	Volume	✓				✓	✓	✓		✓			
	VIX	✓											✓
	Senior	✓											
	Callable	✓			✓			✓					
investigated but excluded from our final framework	Maturity	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Issue size	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Coupon rate	✓					✓	✓	✓	✓		✓	
	Duration×Spread	✓											
	Other issue size	✓			✓			✓			✓		
	Other AmtOut	✓											
	Only issue	✓											
	No. of other issues	✓											
	Equity listed	✓		✓	✓								
	Rating	✓	✓	✓	✓			✓	✓		✓	✓	✓
	Total return	✓							✓		✓	✓	
	Volatility	✓							✓				
	Turnover	✓						✓	✓				
	Missing prices	✓		✓		✓				✓			
	Fraction missing	✓						✓					
	On-the-run	✓		✓									
	Inventory change	✓					✓						✓
	CDX	✓											
	Swap spread (5y)	✓											
	TED spread (3m)	✓											✓

H. Definitions of goodness of fit measures and tests

Mean Squared Error (MSE)

The Mean Squared Error (MSE), as used in Appendix K, is defined as follows:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\widehat{y}_i - y_i)^2 \tag{68}$$

where $\widehat{y}_i$ and y_i are the estimated value and actual value of observation i, respectively. The number of observations in the sample is denoted by N.

Deviance MSE

The 'deviance residual' of a GLM model is the difference between the log likelihood estimate of the 'saturated' model $l_S(\widehat{\psi}|y)$ minus the log likelihood estimate of the proposed model $l(\widehat{\beta}|y)$:

$$D = 2\left(l_S(\widehat{\psi}|y) - l(\widehat{\beta}|y)\right) \tag{69}$$

where the saturated model is the model with the 'most general possible mean structure': the model has as many parameters (collected in vector $\widehat{\psi}$) as observations y (Kutner, 2005). The deviance mean squared error (MSE) for data y_i, where $i = 1, ..., N$, is then:

$$\text{Deviance MSE} = \frac{1}{N}\sum_{i=1}^{N} 2\left(l_S(\widehat{\psi}|y_i) - l(\widehat{\beta}|y_i)\right) \tag{70}$$

Brier score

The Brier score is a strictly proper scoring rule and has a similar interpretation as the mean squared error for continuous models. The Brier score is bounded between 0 and 1, with 0 being the best score. If $\widehat{p}_i$ is the estimated probability for observation $i = 1, ..., N$, and b_i is the actual binary outcome of the event (0 or 1), then the Brier score is defined as:

$$\text{Brier score} = \frac{1}{N}\sum_{i=1}^{N}(\widehat{p}_i - b_i)^2 \tag{71}$$

McFadden pseudo-R^2

The McFadden pseudo-R^2 is defined as one minus the log likelihood of a model $l(\widehat{\beta}|y)$, with k estimated parameters in the vector $\widehat{\beta}$, divided by the log likelihood of a model that only contains an intercept $\widehat{\alpha}$ with log likelihood $l(\widehat{\alpha}|y)$. For given data y, this yields:

$$\text{McFadden pseudo-R}^2 = 1 - \frac{l(\widehat{\beta}|y)}{l(\alpha|y)} \tag{72}$$

Likelihood ratio test (Chi-squared test)

Assume there is a 'full' model with p parameters and a nested restricted model with k less parameters than the full model. We can test if the fit of the small model is significantly worse than the fit of the full model using a likelihood ratio test:

$$-2\ln\left[\frac{L(\text{restricted})}{L(\text{full})}\right] \sim \chi^2_{p-k} \tag{73}$$

where $L(\text{restricted})$ and $L(\text{full})$ are the likelihoods of the restricted and full models, respectively. This follows a chi-squared distribution asymptotically, with $p-k$ degrees of freedom. The test can also be formulated in terms of deviances. Specifically, the deviance of a model is twice the difference between the log likelihood of the model $l(\text{model})$ and the log likelihood of the saturated version of the model $l(\text{saturated})$ (as shown in equation (69) on page 56). If we substract the deviance of the full model D_{full} from the deviance of the restricted model $D_{\text{restricted}}$, we get the likelihood ratio test again:

$$\begin{aligned} D_{\text{restricted}} - D_{\text{full}} &= 2\,l(\text{saturated}) - 2\,l(\text{restricted}) - \big[2\,l(\text{saturated}) - 2\,l(\text{full})\big] && (74)\\ &= 2\,l(\text{full}) - 2\,l(\text{restricted}) && (75)\end{aligned}$$

The likelihood ratio test can therefore be used to test nested GLMs using their likelihoods or deviances. The test is known to be incorrect for GLMs that include a separate estimation of the dispersion parameter through quasi-likelihood estimation (as is the case with our cost model).

Approximate likelihood ratio test (F-test)

Following Venables and Ripley (2013), we test whether omitting k parameters from a model with p parameters in total yields a significantly higher scaled deviance with an approximate F-test:

$$\frac{D_{\text{restricted}} - D_{\text{full}}}{\widehat{\phi}(p-k)} \sim F_{p-k,N-k} \tag{76}$$

where D is the deviance, N is the number of observations and $\widehat{\phi}$ the estimated dispersion parameter. We use this test both to find the significance of a decrease in deviance when adding individual variables (or different transformations of individual variables) and to compare full models to models with just an intercept (i.e. 'null model'). The degrees of freedom $p-k$ is equal to the difference in parameters between the two specifications. This test is more appropriate than the simple likelihood ratio test for models that use quasi-likelihood estimation: when a separate estimate of the dispersion parameter is included. We therefore use this test for our cost model.

Pearson's chi-squared test

To measure the goodness-of-fit (or rather the "badness of fit" in this case) for categorical variables (the warehousing rate model and the arrival rate model), we calculate the sum of squared errors

of Pearson residuals. These follow a χ^2 distribution asymptotically, that can be used to test whether fitted values are independent of the actual observations. If this hypothesis is rejected for some significance level, it can be concluded that the proposed model might not be a good fit for the data. In essence, we want the distribution of the fitted values to be dependent on the distribution of the actual values. The value of the chi-squared test for Pearson residuals is an indication of how likely it is that the fitted values do not coincide with the corresponding observed values. Rejecting the test thereby suggests that the model does not fit the data well.

Using the notation from Appendix B, we calculate the Pearson residuals as follows:

$$p_i = \frac{y_i - \widehat{\mu}_i}{\sqrt{V(\widehat{\mu}_i)}} \tag{77}$$

where y_i is the actual observation i and $\widehat{\mu}_i$ the estimated expected value of the observation. The variance function $V(\widehat{\mu}_i)$ differs per distribution. We take the dispersion in the variance function to be $a(\phi) = \phi$ and estimate the dispersion as explained in Appendix B.

The Pearson χ^2 test can now be written as follows (Venables & Ripley, 2013):

$$\sum_{i=1}^{N} p_i^2 \overset{\mathcal{D}}{\sim} \chi^2_{N-p} \tag{78}$$

where p is the number of parameters in the full model and N is the amount of observations. Note that this result holds asymptotically: it is approximately correct for large sample sizes.

Testing for overdispersion

To test for overdispersion in the arrival rates, we use the test of Cameron and Trivedi (1990). Specifically, we estimate a Poisson model to find the conditional expected value $\mathbb{E}[y_i]$ of dependent variable y_i on the vector of explanatory variables X. Let $\mu_i \equiv \mathbb{E}[y_i|X]$. Using this expected value, we then test whether $\mathbb{V}ar[Y_i] = \mu_i$ with the following hypotheses:

$$\mathrm{H}_0: \quad \mathbb{V}ar[y_i] = \mu_i \tag{79}$$

$$\mathrm{H}_1: \quad \mathbb{V}ar[y_i] = \mu_i + \alpha \cdot g(\mu_i) \tag{80}$$

where $g(\cdot)$ is some monotonic function (often $g(x) = x$ or $g(x) = x^2$). To test for overdispersion, we can simply confirm whether $\alpha = 0$ in (80). We do this by running (80) as a regression and then checking whether α is statistically significant from zero with a t-test. A two-sided t-test clarifies whether a Poisson model is sufficient, where a one-sided t-test can confirm whether the data exhibits either significant under- or overdispersion. We are interested in testing overdispersion in the samples, so we test the one-sided alternative $\mathrm{H}_1 : \alpha > 0$.

I. Model specification results

In this section, we provide statistical evidence of our various model specifications. In Figures 8 and 9, we show residuals for alternative model specifications of the cost model, using both normal linear models (LM) and a GLM with gamma distribution for different link functions. The Q-Q plots for the GLMs are based on Augustin et al. (2011). Investigating the plots, we find that normal LMs are not suitable for transaction costs. A GLM with log-link and gamma distribution appears to have the best fit, even though it slightly underestimates large outliers. In Table 9 we present formal tests for different transformations of the regressors. We find that the log transform gives the highest improvement in deviance for almost all regressors in the models. For VIX, we find that it should not be transformed. For Age, the square-root transform also works well, but we stick to the log transform for ease of interpretation. Lastly, we find significant overdispersion in the arrival rates as observed from Table 10. This motivates the necessity of modelling the arrival rates with the negative binomial regression instead of a Poisson model.

Figure 8: ***'fitted vs residual' and 'Q-Q plots' for linear models of $1mm+ SDB costs.*** These plots show the fit of two different linear model (LM) specifications: both with normal and log transformed costs. These models assume a normal distribution for the error terms.

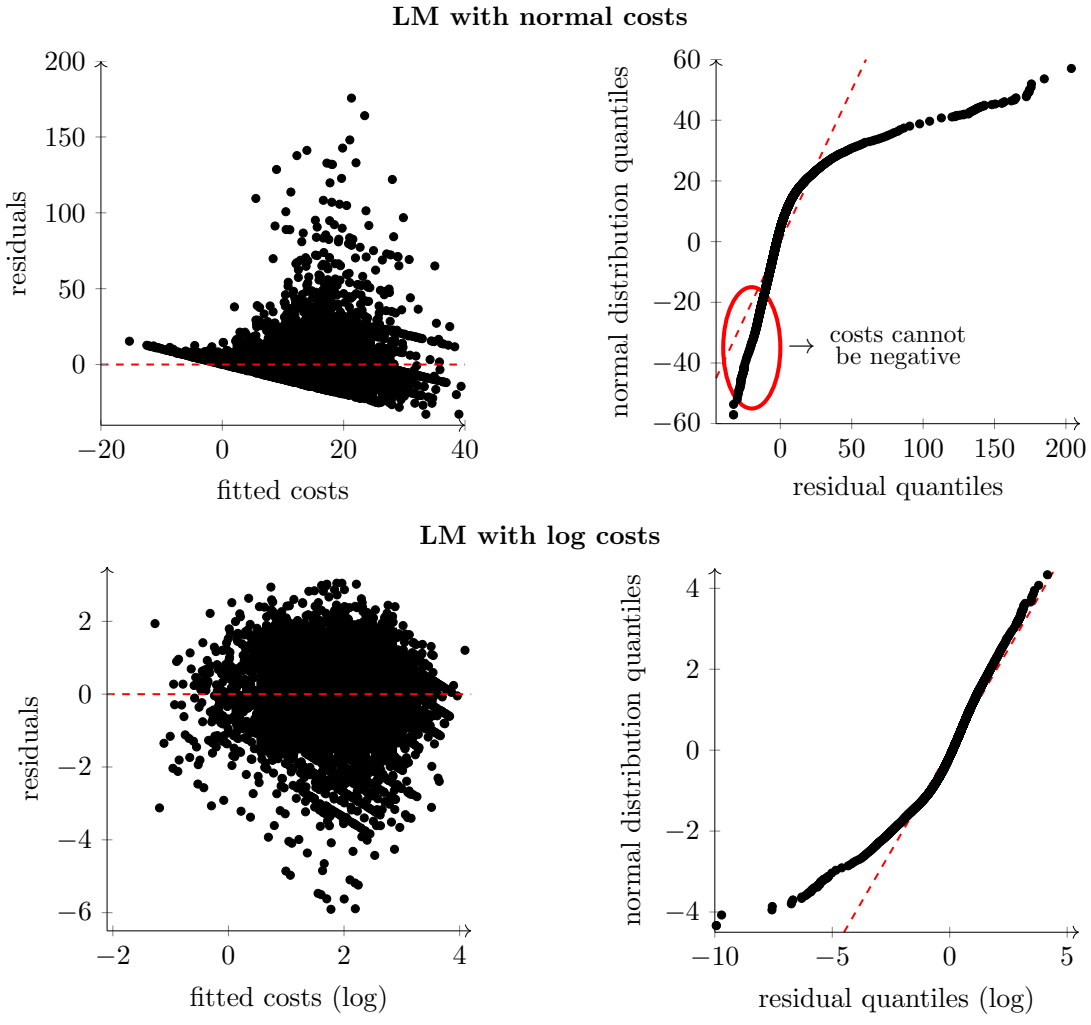


Figure 9: ***'fitted vs residual' and 'Q-Q plots' for GLMs of $1mm+ SDB costs.*** These plots show the fits for different GLM specifications of the gamma distribution using the deviance Q-Q plots of Augustin et al. (2012). The first two specifications use the identity link function, both with and without log transformed costs. The last specification is the one used in this thesis: a log-link with gamma distribution and no transformation of the dependent variable.

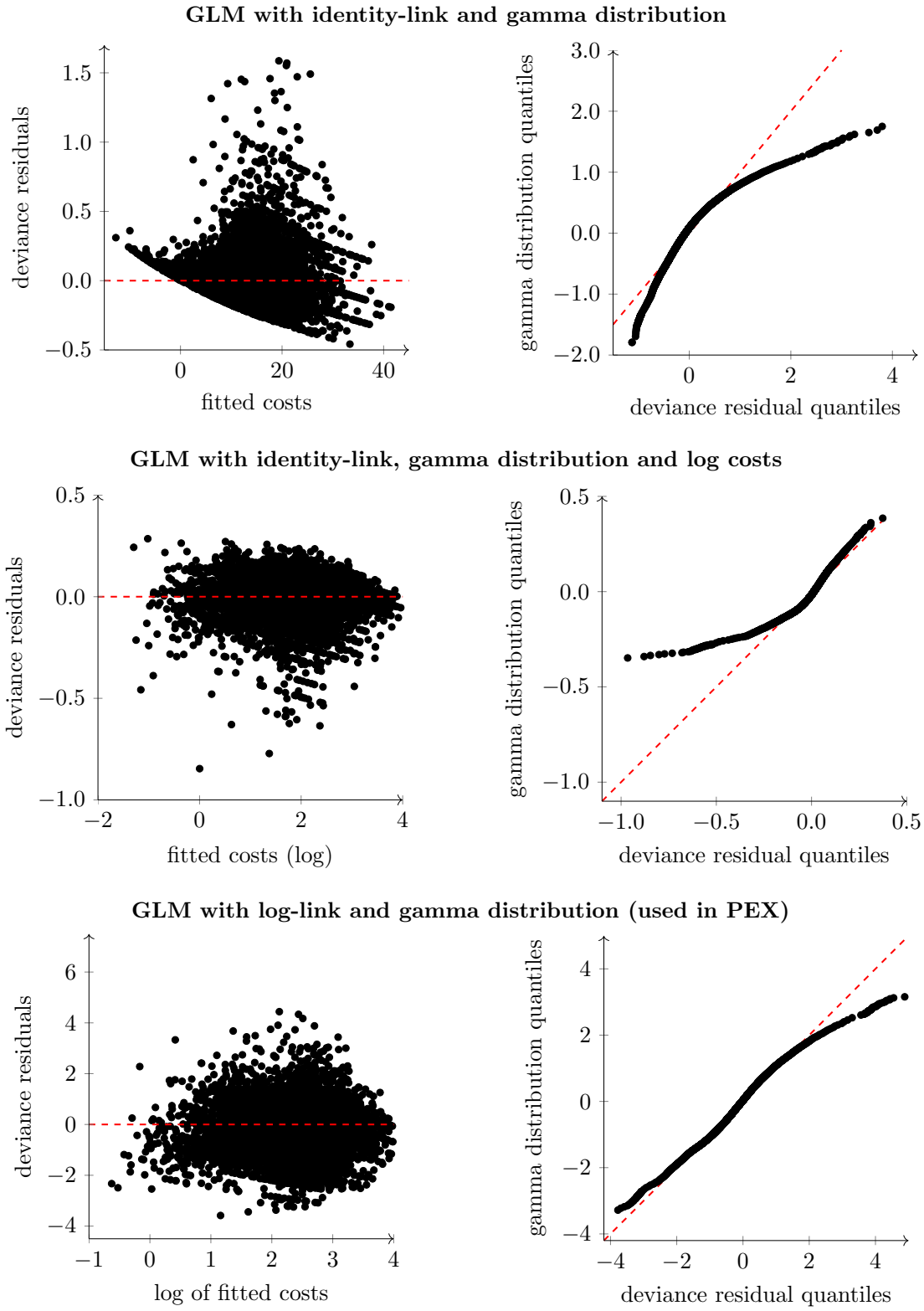


Table 9: ***Variable specification results***. This table shows the deviance improvements when including either variables without transformation (x), with a square-root transformation ($\sqrt{x}$), or with a log transformation ($\log x$) for specific subsamples (in brackets below model name). The results are slightly different for other sample groups (different types, sizes), but the same pattern emerges. The significance of the deviance improvements for the cost model is tested with an F-test. The F-values appear in brackets below the corresponding deviance improvements. For the warehousing rate and arrival rate models, we use a likelihood ratio test (LRT) that is based on the chi-squared distribution. These tests are more elaborately described in Appendix H. The symbol * indicates a significant improvement in the deviance (or likelihood) at the 1% level.

	Cost model ($1mm+, SDB, F-test)			**Warehousing rate model** ($1mm+, buying, LRT)			**Arrival rate model** ($1mm+, buyers, LRT)		
	x	$\sqrt{x}$	$\log(x)$	x	$\sqrt{x}$	$\log(x)$	x	$\sqrt{x}$	$\log(x)$
Price	18.38* (18)	25.77* (25)	32.30* (31)	306.55*	371.07*	435.88*			
Spread	698.39* (674)	1614.38* (1558)	2210.23* (2134)	793.11*	1440.34*	1084.38*	123.73*	172.18*	140.26*
Duration	4160.52* (4016)	4473.42* (4318)	4572.66* (4414)	484.75*	782.38*	1294.27*	1.97	10.46*	84.26*
Age	121.27* (117)	102.09* (99)	62.12* (60)	9.77*	0.18	8.20*	939.29*	1467.45*	1770.31*
AmtOut	36.06* (35)	59.32* (57)	79.97* (77)	116.41*	218.88*	296.50*	3429.41*	4425.42*	4941.99*
Size	34.39* (33)	256.37* (247)	277.16* (268)	17.62*	1064.42*	1865.64*			
AvgSize	12.35* (12)	52.24* (50)	87.86* (85)	61.85*	41.42*	9.48*			
Volume	3.96 (4)	5.47 (5)	28.09* (27)	23.20*	53.74*	1138.38*	1154.25*	5344.80*	10552.48*
VIX	1755.72* (730)	881.95* (851)	969.23* (936)	128.66*	43.94*	59.13*	124.38*	22.23*	19.42*
Observations		42,246			1,016,708			6,488,779	

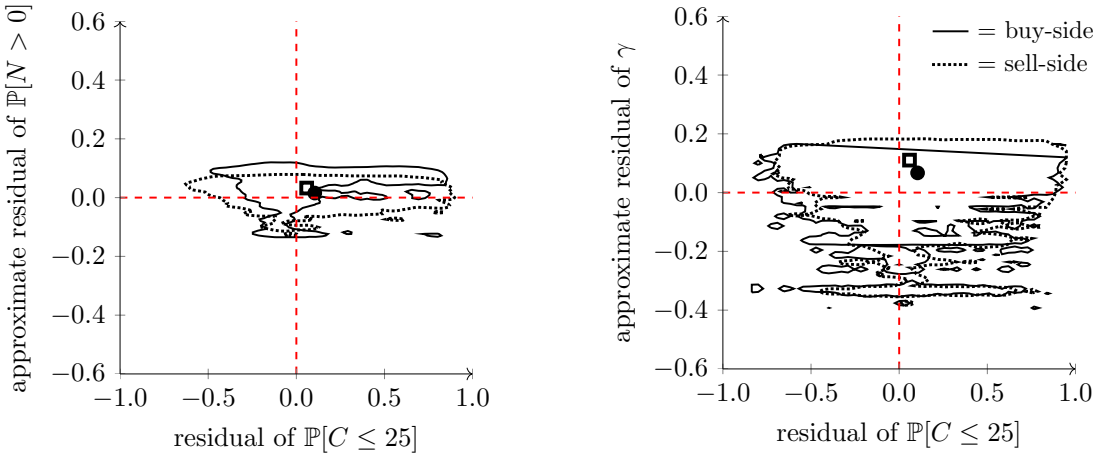
Table 10: ***Overdispersion results of the arrival rates***. This table shows results of the Cameron and Trivedi (1990) test for overdispersion. The test is further explained in Appendix H. The symbol * indicates significant overdispersion ($\delta > 1, \alpha > 0$) at the 1% level.

	$0–$100k		$100k–$1mm		$1mm+	
	buyers	sellers	buyers	sellers	buyers	sellers
$\mathbb{V}ar[y_i] = \delta\mu_i$						
dispersion ($\widehat{\delta}$)	5.04*	2.07*	1.40*	1.26*	1.27*	1.20*
(t-value)	(12.47)	(53.10)	(95.79)	(97.43)	(55.87)	(73.90)
$\mathbb{V}ar[y_i] = \mu_i + \alpha\mu_i^2$						
alpha ($\widehat{\alpha}$)	1.81*	0.66*	0.58*	0.53*	0.90*	0.95*
(t-value)	(14.03)	(56.96)	(106.59)	(110.73)	(52.92)	(84.67)
Observations	6,519,328	6,519,328	6,512,396	6,512,396	6,488,779	6,488,779

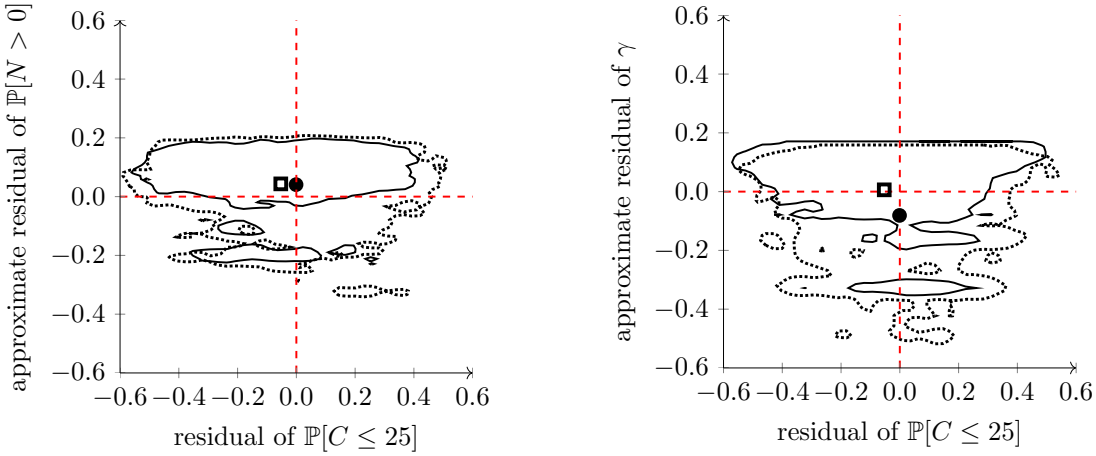
J. Residual covariance analysis

The figures below show 95% confidence regions (2D Gaussian kernel density estimation with bandwidth $\widehat{b} = 4.24 \min(\widehat{\sigma}, \text{IQR}/1.34) n^{-1/5}$ from Venables and Ripley (2013), with n the sample size, $\widehat{\sigma}$ the standard deviation and IQR the interquartile range) of bootstrapped residuals for the different components in the PEX. The residuals of the warehousing and arrival rate probabilities are approximated by taking the average rates over the month in which the transaction occurred. Buy-side regions appear as —— with mean □ and sell-side regions appear as ····· with mean ●.

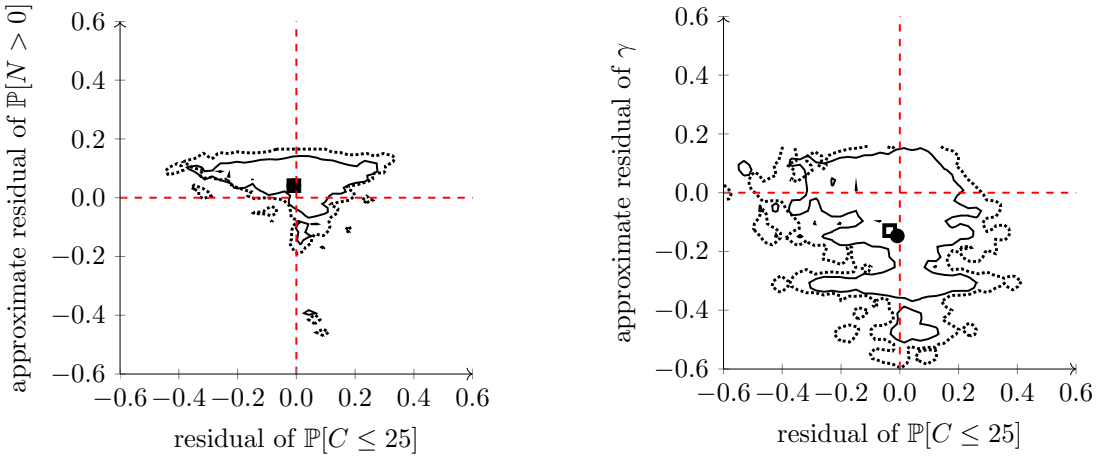
Results for $0–$100k transactions



Results for $100k–$1mm transactions



Results for $1mm+ transactions



K. Performance comparison

Roll's model to calculate the bid-ask spread is based on the premise that trading costs cause negative serial dependence in consecutive changes in transaction prices (Roll, 1984). If we follow the notation of Bao et al. (2011), we denote the price at time t as P_t. The implied bid-ask spread by Roll's model can then be written as two times the square root of the negative covariance between consecutive price changes:

$$\text{bid-ask spread} = 2\sqrt{-\mathbb{C}ov[P_t - P_{t-1}, P_{t+1} - P_t]} = 2\sqrt{-\mathbb{C}ov[\Delta P_t, \Delta P_{t+1}]} \tag{81}$$

Albeit simple, Roll's measure has proven to be a robust estimator of the bid-ask spread. Additionally, the advantage of Roll's model is that only transaction prices are needed to calculate it. The big disadvantage of the Roll model is its inability to generalise across bonds. Plenty of recent transaction data is needed to accurately estimate the measure, such that the measure is less useful for less liquid bonds. Naturally, a liquidity measure is most useful for such illiquid bonds and this demonstrates why measures that use longitudinal, opposed to cross-sectional, data have limited applicability. In Table 11 we show the performance results when taking half of the implied bid-ask spread by Roll's model for estimating imputed transaction costs.

Table 11: ***Roll model cost estimation performance***. This table shows the performance of the Roll model's implied bid-ask spread to estimate realised transaction costs. Listed are the mean error (ME), mean squared error (MSE), R^2 (percentages) and the sample size (Obs.).

		\$0–\$100k			\$100k–\$1mm			\$1mm+		
Lag		SDB	SDD	DDB	SDB	SDD	DDB	SDB	SDD	DDB
10	ME	−67.2	−56.6	−86.0	−21.6	−37.4	−58.4	−10.3	−17.4	−20.8
	MSE	84.7	63.7	120.7	15.8	34.1	71.4	4.0	12.9	15.8
	R^2 (%)	−8.4	−39.0	−53.0	−15.5	−29.2	−28.0	−16.6	−12.3	−15.8
	Obs.	97,038	928,668	2,021,319	36,472	140,047	314,502	32,225	28,730	47,905
50	ME	−63.8	−52.0	−81.4	−21.0	−35.8	−56.8	−6.4	−14.5	−17.6
	MSE	78.9	57.8	112.0	16.7	32.6	69.1	3.0	10.5	16.5
	R^2 (%)	12.8	−19.6	−25.6	1.7	−15.1	−7.7	−9.0	−14.6	−11.4
	Obs.	91,692	896,846	1,901,288	26,690	111,703	230,531	11,160	12,198	18,165
100	ME	−61.5	−48.6	−78.0	−20.8	−34.8	−55.8	−4.8	−11.9	−14.3
	MSE	75.3	53.8	105.8	17.6	32.0	68.2	3.4	9.6	11.1
	R^2 (%)	17.3	−15.0	−19.4	7.7	−9.2	−2.0	−6.1	−11.5	−10.2
	Obs.	86,208	859,631	1,771,796	20,123	89,350	170,054	4,571	5,506	7,487
250	ME	−57.6	−41.9	−70.8	−20.2	−32.4	−52.5	−2.3	−8.2	−9.0
	MSE	69.3	46.8	94.2	19.4	31.1	64.9	4.9	12.0	7.8
	R^2 (%)	19.8	−10.7	−14.0	9.4	−3.0	5.6	−4.5	1.6	2.5
	Obs.	74,142	766,373	1,486,643	11,323	54,461	86,466	1,102	1,256	1,410
all	ME	0.9	3.8	−29.7	−8.0	−21.0	−43.7	−6.9	−12.5	−16.9
	MSE	43.8	47.5	65.6	10.0	24.2	54.0	3.4	11.0	14.0
	R^2 (%)	2.0	−26.9	−37.4	10.5	−9.3	−9.1	−16.2	−12.5	−17.0
	Obs.	98,570	936,049	2,046,721	39,726	148,325	337,393	41,493	35,313	60,600

Additionally, we also employ naive moving averages to estimate transactions costs and the warehousing and arrival rates. For the naive moving average, we employ the simplest form of an historical average to forecast a timeseries. Let y_t for $t = 1, ..., T$ denote the dependent variable at hand and let $\mathcal{F}_t$ denote the filtration with all information up to and including time t. The forecast we estimate for the value of y_{t+1} given the information up to time t, is then:

$$\widehat{y}_{t+1}|\mathcal{F}_t = \frac{1}{l}\sum_{i=0}^{l} y_{t-i} \tag{82}$$

Where the window l is taken to use the information of the previous l values. We deliberately keep this model as simple as possible to give a rough overview of its performance without resorting to the estimation of the appropriate number of lags and the elaborate process of assessing parameter stability. Naturally, this model therefore only poses as a naive average of past observations.

Table 12: ***Moving average cost estimation performance***. This table shows the performance of a naive moving average to estimate realised transaction costs. Listed are the mean error (ME), mean squared error (MSE), R^2 in percentages and the number of available observations (Obs.).

		\$0–\$100k			\$100k–\$1mm			\$1mm+		
Lag		SDB	SDD	DDB	SDB	SDD	DDB	SDB	SDD	DDB
10	ME	5.5	7.3	−9.0	11.3	−0.8	−19.7	0.8	−6.2	−10.3
	MSE	30.2	36.6	36.0	13.5	22.4	36.6	3.4	9.4	11.7
	R^2 (%)	28.1	−3.8	24.9	10.2	−4.8	8.1	−2.0	4.3	1.8
	Obs.	97,655	932,260	2,031,645	37,702	144,651	327,188	35,591	33,096	55,742
50	ME	6.1	9.9	−8.1	12.7	1.1	−18.7	0.2	−6.6	−10.8
	MSE	27.9	33.8	33.2	12.9	20.4	34.5	3.2	8.8	13.0
	R^2 (%)	33.1	1.7	31.1	16.8	2.1	17.5	3.9	5.8	6.1
	Obs.	93,268	904,527	1950,884	28,774	120,640	264,244	18,020	21,459	33,190
100	ME	6.2	11.4	−7.1	13.5	2.4	−18.2	0.4	−6.5	−10.9
	MSE	28.0	33.4	32.5	13.7	20.4	34.5	3.3	8.7	14.0
	R^2 (%)	33.4	2.5	31.8	18.3	3.2	19.9	3.7	5.2	5.0
	Obs.	87,860	865,207	1,840,656	22,247	98,558	207,205	9,686	13,260	19,761
250	ME	5.5	14.3	−5.7	14.4	4.4	−17.6	1.8	−5.9	−9.8
	MSE	29.0	33.1	32.1	16.2	21.6	35.9	3.5	9.4	11.3
	R^2 (%)	31.7	3.0	31.5	18.6	3.5	21.8	4.7	5.7	7.6
	Obs.	74,291	761,319	1,577,748	11,725	62,435	117,646	2,602	4,024	5,439
all	ME	16.9	25.2	−1.9	18.9	8.6	−12.9	0.1	−6.7	−10.4
	MSE	31.6	38.2	34.0	14.2	20.7	32.3	3.2	9.6	11.7
	R^2 (%)	28.8	6.9	28.9	14.1	7.2	19.5	−3.2	2.1	1.4
	Obs.	98,565	936,414	2,043,504	39,847	149,029	337,643	41,849	35,759	61,671

Table 13: ***Moving average warehousing rate performance.*** This table shows the performance of a naive moving average to estimate the warehousing rate. Listed are the mean error (ME), Brier score (Brier) and the number of available observations (Obs.).

		$0–$100k		$100k–$1mm		$1mm+	
Lag		buying	selling	buying	selling	buying	selling
10	ME	−0.02	−0.04	−0.05	−0.11	−0.23	−0.31
	Brier	0.05	0.05	0.08	0.10	0.20	0.25
	Obs.	2,748,745	3,181,667	1,110,650	856,681	484,083	414,477
50	ME	−0.01	−0.04	−0.03	−0.09	−0.19	−0.25
	Brier	0.05	0.04	0.07	0.09	0.18	0.21
	Obs.	2,292,677	2,627,864	619,682	379,948	92,753	67,514
100	ME	−0.01	−0.03	−0.03	−0.09	−0.17	−0.23
	Brier	0.05	0.04	0.07	0.09	0.20	0.24
	Obs.	1,959,124	2,182,452	384,502	206,635	19,869	15,058
250	ME	0.00	−0.03	−0.01	−0.10	−0.12	−0.21
	Brier	0.05	0.04	0.08	0.10	0.24	0.24
	Obs.	1,387,635	1,446,088	127,476	52,565	660	708
all	ME	−0.03	−0.06	−0.05	−0.14	−0.23	−0.33
	Brier	0.05	0.05	0.08	0.13	0.22	0.29
	Obs.	2,918,789	3,373,030	1,370,103	1,202,433	846,698	819,503

Table 14: ***Moving average arrival rate performance.*** This table shows the performance of a naive moving average to estimate the arrival rate of buyers and sellers. Listed are the mean error (ME), mean squared error (MSE), R^2 (percentages) and the number of observations (Obs.).

		$0–$100k		$100k–$1mm		$1mm+	
Lag		buyers	sellers	buyers	sellers	buyers	sellers
10	ME	0.00	0.00	0.00	0.00	0.00	0.00
	MSE	3.42	1.02	0.53	0.36	0.21	0.16
	R^2 (%)	60.76	62.81	33.71	25.98	14.87	15.36
	Obs.	17,873,130	17,873,130	17,832,540	17,832,540	17,758,125	17,758,125
50	ME	0.00	0.00	0.00	0.00	0.00	0.00
	MSE	4.55	1.05	0.55	0.36	0.20	0.16
	R^2 (%)	48.19	61.77	29.80	25.57	13.53	13.86
	Obs.	17,556,090	17,556,090	17,516,220	17,516,220	17,443,125	17,443,125
100	ME	−0.01	−0.01	0.00	0.00	0.00	0.00
	MSE	5.24	1.10	0.58	0.37	0.21	0.17
	R^2 (%)	41.17	60.20	26.69	24.15	11.98	12.41
	Obs.	17,159,790	17,159,790	17,120,820	17,120,820	17,049,375	17,049,375
250	ME	−0.03	−0.03	−0.01	−0.01	0.00	0.00
	MSE	6.54	1.28	0.66	0.40	0.23	0.18
	R^2 (%)	30.59	55.26	21.20	20.73	9.14	9.68
	Obs.	15,970,890	15,970,890	15,934,620	15,934,620	15,868,125	15,868,125
all	ME	−0.17	−0.14	−0.06	−0.05	−0.02	−0.02
	MSE	7.22	1.86	0.70	0.43	0.22	0.18
	R^2 (%)	15.78	31.70	10.13	10.18	4.20	4.52
	Obs.	17,952,390	17,952,390	17,911,620	17,911,620	17,836,875	17,836,875

L. Robustness checks

Table 15: ***Robustness check $0–$100k.*** Robustness check of the transaction cost imputation results for different window sizes and trades between $0 and $100k. These transactions are for investment grade bonds only, where the top 0.5% of costs are deleted per roundtrip type (SDB, SDD or DDB). In order to make a fair comparison, we do not delete transaction combinations if they have missing bond characteristics. The sample used in this table is therefore slightly different from the samples that we use in the regressions. Costs are representative of half of the realised bid-ask spread and are denoted in dollar cents. We examine imputed costs for window sizes ranging from 15 minutes (m), to hours (h), days (d) and a full workweek (w).

		odd-lot transactions ($0–$100k)					
	Detection period	15m	1h	3h	1d	2d	1w
SDB	Mean cost	68	80	83	82	83	82
	1st percentile	1	2	2	2	2	1
	Median	50	67	70	67	66	63
	99th percentile	234	256	271	286	298	318
	Std	60	63	65	67	68	72
	Skewness	0.9	0.8	0.9	1.0	1.1	1.3
	Kurtosis	0.1	0.1	0.4	0.9	1.3	1.9
SDD	Mean cost	58	59	60	63	65	68
	1st percentile	1	1	1	1	1	1
	Median	50	50	50	50	50	50
	99th percentile	228	242	250	267	288	304
	Std	50	52	54	57	60	64
	Skewness	1.5	1.5	1.6	1.6	1.7	1.8
	Kurtosis	2.5	2.8	3.0	3.3	3.7	4.2
DDB	Mean cost	88	87	87	87	89	89
	1st percentile	3	3	3	3	3	3
	Median	87	84	80	80	80	78
	99th percentile	260	266	274	284	296	300
	Std	67	68	69	71	73	75
	Skewness	0.6	0.7	0.7	0.8	0.8	0.9
	Kurtosis	−0.4	−0.3	−0.2	−0.1	0.0	0.1
observed order types	SDB	1.0%	2.0%	3.0%	3.6%	4.2%	2.6%
	SDD	9.9%	10.1%	9.8%	7.2%	6.8%	3.5%
	DDB	21.6%	19.9%	17.8%	12.2%	11.6%	5.7%
	SD	22.7%	22.6%	22.3%	19.8%	20.0%	17.9%
	DB	30.1%	30.4%	30.5%	32.4%	31.8%	33.7%
	DD	14.7%	15.0%	16.5%	24.9%	25.5%	36.6%
	Ambiguous	429,491	580,074	684,978	678,801	678,642	530,373
	Observations	9,453,376	8,733,976	8,061,004	8,690,128	8,049,736	11,322,109
inventory	Avg. buy fraction γ_B	98.0%	96.3%	94.1%	92.6%	91.2%	93.9%
	Avg. sell fraction γ_S	96.9%	94.4%	91.3%	88.3%	86.5%	89.3%
	Weighted buy costs	87	87	87	87	89	89
	Weighted sell costs	58	60	62	65	68	69

Table 16: ***Robustness check \$100k–\$1mm.*** Robustness check of the transaction cost imputation results for trades between \$100k and \$1mm. These transactions are for investment grade bonds only, where the top 0.5% of costs are deleted per roundtrip type (SDB, SDD or DDB). In order to make a fair comparison, we do not delete transaction combinations if they have missing bond characteristics. The sample used in this table is therefore slightly different from the samples that we use in the regressions. Costs are representative of half of the realised bid-ask spread and are denoted in dollar cents. We examine imputed costs for windows ranging from 15 minutes (m), to hours (h), days (d) and a full workweek (w).

		round-lot transactions (\$100k–\$1mm)					
	Detection period	15m	1h	3h	1d	2d	1w
SDB	Mean cost	21	27	31	33	35	35
	1st percentile	0	0	0	0	0	0
	Median	10	13	15	18	19	19
	99th percentile	122	147	161	173	181	189
	Std	27	33	36	38	40	41
	Skewness	2.0	1.9	1.8	1.8	1.9	2.0
	Kurtosis	3.9	3.4	3.0	3.4	3.7	4.5
SDD	Mean cost	38	39	39	40	42	43
	1st percentile	1	1	1	1	1	0
	Median	25	25	25	25	25	25
	99th percentile	188	200	200	209	224	243
	Std	40	41	42	45	47	50
	Skewness	1.7	1.8	1.9	2.0	2.1	2.2
	Kurtosis	3.4	3.8	4.0	4.9	5.3	6.2
DDB	Mean cost	59	58	56	55	55	54
	1st percentile	1	1	1	1	1	1
	Median	38	36	33	31	30	29
	99th percentile	238	241	244	250	253	260
	Std	58	58	58	58	59	60
	Skewness	1.2	1.3	1.4	1.5	1.5	1.6
	Kurtosis	0.7	1.1	1.4	1.8	2.1	2.6
observed order types	SDB	1.0%	1.5%	2.1%	3.0%	3.7%	3.8%
	SDD	3.9%	4.4%	4.6%	4.5%	4.5%	3.8%
	DDB	8.9%	9.6%	9.6%	8.7%	8.4%	6.7%
	SD	26.6%	27.1%	27.2%	26.5%	26.6%	25.0%
	DB	35.2%	35.6%	35.5%	35.1%	34.9%	34.5%
	DD	24.3%	21.8%	21.0%	22.0%	22.0%	26.1%
	Ambiguous	48,231	71,706	98,644	134,210	157,098	164,171
	Observations	3,818,196	3,590,066	3,427,003	3,298,275	3,151,754	3,319,430
inventory	Avg. buy fraction γ_B	97.7%	96.8%	95.6%	93.5%	92.1%	91.6%
	Avg. sell fraction γ_S	96.7%	95.4%	93.9%	91.1%	89.3%	88.4%
	Weighted buy costs	58	57	55	53	54	53
	Weighted sell costs	37	38	39	40	41	42

Table 17: ***Robustness check $1mm+***. Robustness check of the transaction cost imputation results for different window sizes and trades of $1mm or larger. These transactions are for investment grade bonds only, where the top 0.5% of costs are deleted per roundtrip type (SDB, SDD or DDB). In order to make a fair comparison, we do not delete transaction combinations if they have missing bond characteristics. The sample used in this table is therefore slightly different from the samples that we use in the regressions. Costs are representative of half of the realised bid-ask spread and are denoted in dollar cents. We examine imputed costs for window sizes ranging from 15 minutes (m), to hours (h), days (d) and a full workweek (w).

		block sized transactions ($1mm+)					
	Detection period	15m	1h	3h	1d	2d	1w
SDB	Mean cost	12	13	14	16	17	19
	1st percentile	0	0	0	0	0	0
	Median	9	10	10	11	11	12
	99th percentile	51	61	68	83	95	107
	Std	12	13	14	17	19	21
	Skewness	1.6	1.8	1.9	2.1	2.3	2.4
	Kurtosis	2.6	3.7	4.3	5.6	6.5	7.4
SDD	Mean cost	18	20	22	25	28	31
	1st percentile	0	0	0	0	0	0
	Median	8	10	12	13	15	16
	99th percentile	107	117	126	150	170	202
	Std	22	24	27	31	35	41
	Skewness	2.4	2.3	2.3	2.4	2.5	2.6
	Kurtosis	6.7	6.4	6.5	7.0	7.8	8.7
DDB	Mean cost	21	23	24	26	27	29
	1st percentile	0	1	1	1	1	1
	Median	11	13	13	14	15	15
	99th percentile	125	134	141	150	169	185
	Std	26	27	29	31	34	37
	Skewness	2.4	2.4	2.4	2.4	2.6	2.6
	Kurtosis	6.6	6.5	6.8	7.0	8.1	8.5
observed order types	SDB	1.7%	2.7%	3.5%	4.7%	5.2%	5.5%
	SDD	1.5%	1.9%	2.2%	2.4%	2.5%	2.4%
	DDB	2.5%	3.3%	3.7%	4.0%	4.0%	3.7%
	SD	31.3%	32.2%	32.3%	31.6%	31.5%	30.5%
	DB	33.9%	34.6%	34.5%	33.7%	33.5%	32.4%
	DD	29.0%	25.2%	23.8%	23.7%	23.2%	25.6%
	Ambiguous	18,057	29,509	41,096	59,414	71,985	79,696
	Observations	2,463,986	2,266,249	2,157,060	2,053,788	1,976,843	2,004,364
inventory	Avg. buy fraction γ_B	95.5%	93.4%	91.5%	89.0%	87.8%	86.8%
	Avg. sell fraction γ_S	95.0%	92.7%	90.7%	87.9%	86.7%	85.8%
	Weighted buy costs	21	22	23	25	26	28
	Weighted sell costs	17	19	21	24	26	29