Price Formation in the Foreign Exchange Market*

Florent Gallien, Sergei Glebkin, Serge Kassibrakis, Semyon Malamud, and Alberto Teguia

August 23, 2023

Abstract

We study the joint price formation in the dealer-to-dealer (D2D) and dealer-to-customer (D2C) segments of the foreign exchange (FX) market, both theoretically and empirically. Our theory accounts for dealer heterogeneity, market power, and non-exclusive customer-dealer relationship and predicts that (i) D2D prices are negatively related to lagged cross-sectional covariance between D2C mid-quotes and bid-ask spreads and (ii) D2D spreads are negatively related to the lagged cross-sectional variance of D2C spreads. Our predictions are confirmed empirically using unique proprietary data. Model calibration implies that non-fundamental volatility due to liquidity shocks accounts for a third of overall short-term volatility.

Keywords: Liquidity, Foreign Exchange, OTC markets, Price Impact, Market Power
JEL Classification Numbers: F31, G12, G14, G21

*Florent Gallen is at Swissquote, Sergei Glebkin (glebkin@insead.edu) is at INSEAD, Serge Kassibrakis is at Swissquote, Semyon Malamud (semyon.malamud@epfl.ch) is at EPFL and the Swiss Finance Institute, and Alberto Teguia (alberto.mokakteguia@sauder.ubc.ca) is at UBC Sauder. This paper was previously circulated under the titles "Liquidity Provision in the Foreign Exchange Market" and "Dealer Heterogeneity and Exchange Rates." We thank Lukas Altermatt, Alain Chaboud, Thomas Gilbert, Pietro Gottardi, Vraham Kamara, Tomy Lee (discussant), Alexander Michaelides, Andreas Schrimpf, Vlad Sushko, Adrien Verdelhan, seminar participants at the University of Essex, conference participants at the 2023 Vienna Symposium on Foreign Exchange Markets, the 3rd Annual CEPR Symposium on Financial Economics, the 34th Annual Pacific Northwest Finance Conference and NES 30th Anniversary Conference for their helpful comments and remarks. Semyon Malamud gratefully acknowledges the financial support of the Swiss Finance Institute and the Swiss National Science Foundation, Grant 100018_192692. Sergei Glebkin, Semyon Malamud, and Alberto Teguia gratefully acknowledge the financial support of The Europlace Institute of Finance.

1 Introduction

Motivated by the failure of macroeconomic models to explain exchange rate dynamics, a growing literature emphasizes the role of dealers in price formation in foreign exchange (FX) markets.[1] The previous literature focused on the interdealer segment of the FX market and abstracted away its two-tiered structure, characterized by two principal segments: the dealer-to-customer (D2C) and the dealer-to-dealer (D2D) segments.[2] Such focus on the D2D market is likely due to lack of data availability: While high-quality D2D data has been available for a while (see e.g., Chaboud et al. (2007) for an overview of the widely-used EBS data), the D2C data has not yet been publicly available. Leveraging access to unique proprietary data on the cross-section of D2C quotes, this paper studies the joint price formation in the D2D and D2C segments of the FX market, both theoretically and empirically.

Our theory establishes and our empirical analysis estimates the following predictive relations between prices and bid-ask spreads in the D2D and D2C segments:

$$\begin{aligned}\text{Price}_{t+\ell}^{D2D} &= \beta_1^p E[\text{prices}_t^{D2C}] - \beta_2^p \operatorname{Cov}(\text{prices}_t^{D2C},\ \text{spreads}_t^{D2C}) + \text{const} + \epsilon, \\ \text{Spread}_{t+\ell}^{D2D} &= \beta_1^{BA} E[\text{spreads}_t^{D2C}] - \beta_2^{BA} \operatorname{Var}[\ \text{spreads}_t^{D2C}] + \text{const} + \epsilon.\end{aligned} \tag{1}$$

Here lag value ℓ is 10 sec in our baseline specification; the variables on the right-hand side are summary statistics of the cross-section of D2C quotes at time t: $E[\cdot]$, $\operatorname{Cov}(\cdot,\cdot)$ and $\operatorname{Var}[\cdot]$ stand for, respectively, cross-sectional mean, covariance, and variance; prices refer to mid-prices and spreads refer to bid-ask spreads.

We develop a pure inventory model that accounts for the key prominent features of the FX market: two-tiered market structure, dealer heterogeneity, dealer market power in both D2D and D2C market segments, and non-exclusive customer-dealer relationship.[3] Analyzing

[1] The fact that exchange rates are only weakly related to macroeconomic fundamentals is known as the Meese and Rogoff (1983) exchange rate disconnect puzzle.

[2] The literature review is in Section 7.

[3] See the overview of the FX market structure in Section 2. Non-exclusivity means that customers direct their orders to dealers offering the best prices instead of directing them to a preferred, exclusive dealer.

such a rich model is a challenge, which we overcome by studying the model in the limit when dealer heterogeneity is small and focusing on the first-order effects of dealer heterogeneity.[4] Thus, in our theoretical analysis, the terms ϵ in (1) represent the higher-order effects in the degree of dealer heterogeneity. Such an approach allows us to reduce the cross-section of D2C quotes to just a few summary statistics, facilitating both theoretical and empirical analysis. We believe it can be useful in other heterogenous-agents settings.

Our theory predicts that all β's in (1) are positive. Our empirical analysis offers complete confirmation for these predictions. In contrast, if the D2D market was competitive or if dealers were homogenous, $\beta_2^p = 0$. With exclusive customer-dealer relationship, $\beta_2^p < 0$. Thus, the positive sign of β_2^p is the unique prediction of our theory, distinguishing it from other theories of two-tiered markets featuring no heterogeneity (e.g., Vogler (1997)), perfectly competitive D2D market (e.g., Dunne et al. (2015)) or exclusive customer-dealer relationship (e.g., Babus and Parlatore (2022)).

Our theory also relates the magnitudes of β's to the parameters of the model, such as dealers' risk aversion. Using the estimated values of β's, we calibrate the elasticity of the D2D market: A typical liquidity shock originating from the D2C market moves mid-prices in the D2D market by 0.5 basis points. This is comparable to the average bid-ask spread of 0.44 basis points. The D2D market is illiquid. Coupled with our estimates of the arrival frequency of such shocks and their distribution, this further implies that the liquidity shocks account for around a third of overall short-term volatility in the FX market. The inelastic market hypothesis (Gabaix and Koijen, 2021) holds for the FX market.

In the rest of this section, we provide intuition for why relations (1) hold. While one naturally expects that average prices and spreads in D2C market and D2D markets should be related, there is no apparent reason as to why (i) the variance of spreads in the D2C market should be negatively related to D2D spreads (i.e., why $\beta_2^{BA} > 0$) or why (ii) the

[4]That is, we expand the equilibrium as the equilibrium in the homogeneous case, plus the corrections due to heterogeneity. We focus on the corrections that are first-order in the degree of heterogeneity.

covariance of prices and spreads in the D2C market should be negatively related to D2D prices (i.e., why $\beta_2^p > 0$).

To see why $\beta_2^{BA} > 0$, note that: (a) bid-ask spreads are positively related to dealers' risk-aversions[5] and (b) higher variance in risk-aversions across dealers is associated with greater liquidity of the D2D market. Part (a) is intuitive: More risk-averse dealers are less efficient at holding inventory and require higher compensation for doing so, resulting in wider spreads. To see (b), consider the following example. Imagine that we have three dealers with a risk aversion of 1 each. They each provide 1 unit of liquidity, 3 in total. Now consider what happens if dealers' risk aversions are 0.5, 1, and 1.5 (so that average risk aversion is the same, but the dispersion is higher). Because the liquidity (price elasticity) is inversely proportional to risk aversion, the dealers will provide 2, 1, and 1/1.5=0.66 units of liquidity, 3.66 in total, which is greater than 3.[6]

To see why $\beta_2^p > 0$, note that: (c) the covariance of D2C prices and spreads is positively related to the cross-sectional covariance of dealers' inventories and risk aversions and (d) the cross-sectional covariance of dealers' inventories and risk aversions is negatively related to D2D prices. Statement (c) holds because: (i) risk aversions and bid-ask spreads are positively related, and (ii) inventories and D2C prices are positively related. Part (i) is discussed above. To see part (ii), note that dealers posting the highest (lowest) prices will attract a disproportionate share of aggregate customer sell (buy) volume. Thus, dealers with the lowest (highest) prices will decrease (increase) their inventories, implying a positive cross-sectional relationship between inventories and D2C prices. Statement (d) holds because

[5] We interpret risk aversion broadly as a higher cost of holding more inventories for FX dealers. Heterogeneity in risk-aversions might also capture heterogeneity in customer bases of different dealers (order flow toxicity). The dealers are typically banks, and these costs usually come from banks' capital requirements, costs of obtaining funding, and regulatory constraints that may be binding at the bank level and transmitted into individual trading desk behavior. Moreover, the slackness of such constraints may change over time. Therefore, we see risk aversion as being time-varying. See Cenedese et al. (2021) for empirical evidence supporting this view.

[6] The price elasticity is inversely proportional to risk aversion when heterogeneity in risk aversions is small. In the general case, this relationship is more complex. However, price elasticity is still a convex function of risk aversion in the general case. Thus, the logic of our simple example still applies.

higher cross-sectional covariance of dealers' inventories and risk aversions is symptomatic of *liquidity mismatch*, whereby more risk-averse dealers hold more inventories. Under liquidity mismatch, the allocation of inventories across dealers is inefficient, resulting in higher effective risk aversion of the dealer sector and lower prices.

The rest of the paper is organized as follows. Section 2 gives an overview of the FX market structure. Section 3 presents the model. Our main theoretical predictions are derived in Section 4. They are tested in Section 5. Empirical estimates from Section 5 are then used to calibrate the model. The calibration is presented in Section 6. The literature review is in Section 7. Section 8 concludes. All proofs are presented in the Appendix.

2 Institutional Background

The real-world FX markets are fragmented. Neither retail nor institutional traders (e.g., hedge funds, corporates, or smaller regional banks) can trade directly. To trade in the FX market, they are usually subscribed to continuous quotes from a set of major dealer banks (and, potentially, a set of non-bank electronic market-makers) through these dealer banks' single-bank platforms (SBP).[7] Most of the D2C trading happens on such SBPs. Since a customer typically subscribes to multiple SBPs, the dealer banks compete in the D2C markets through quotes.[8]

Thus, the market is naturally fragmented into two segments with quite different natures of competition:

- In the D2C segment, dealer banks compete in prices, whereby prices are quoted

[7]On such an SBP, a dealer bank provides continuous quotes to its customers and usually uses these quotes to manage the direction of customers' trading and to internalize customer order flow. Some smaller regional banks also have SBP, whereby they offer quotes to their customers that aggregate multiple quotes that they receive from major dealers and market-makers. Many of these smaller banks act as retail aggregators, trying to internalize customer order flow and offload excess inventory by trading with major banks at their quoted prices.

[8]Anecdotal evidence suggests that smaller banks usually subscribe for quotes from all major dealers, while a typical hedge fund subscribes to approximately three to five quote streams.

continuously on both sides (bid and ask) as a function of order size. Separate quotes are provided for trades below USD 1 million and trades between USD 1 million and USD 5 million, with spreads naturally increasing in the trade size. Customers may split large orders (more than USD 1 million) across several dealers.

- In the D2D market segment, a few dealer banks trade and provide liquidity to each other to offload excess FX risk exposure (for example, inventory that they were not able to internalize through offsetting customer order flow).

- Given the relatively small number of major dealer banks, the competition is imperfect both in the D2C and the D2D segments, and dealers take their market power in both segments into account.

While this market structure is specific to FX markets, many other OTC markets have a similar two-tiered architecture. For example, in CDS markets, customers would often trade with dealers using a request for quote protocol, whereby they would simultaneously request quotes from multiple dealers (see, e.g., Collin-Dufresne et al. (2019) and Eisfeldt et al. (2023)). One unique feature of the FX market is the availability of both bid and ask quotes by multiple, non-anonymous dealers. This makes it different from (i) limit order markets, where one could see multiple quotes, but there is no way to know whether a given bid and a given ask come from the same dealer; (ii) one-sided RFQ markets, where it is possible to obtain quotes from multiple dealers, but only on one side of the market (either only the bid or only the ask).[9] Furthermore, we are unaware of any dataset containing such two-sided quotes. For our analysis, it is crucial to see the bid-ask spread quoted by every dealer because we use this information to back out important dealer characteristics. This is

[9]For example, this market structure is common in bond markets (see Bessembinder et al. (2020)) and CDS markets (see Collin-Dufresne et al. (2019)). While it is formally possible to request both bids and ask quotes from a given group of dealers, anecdotal evidence suggests that it is highly uncommon to do so because dealers in those markets want to know the actual direction of the customer's desired trade before providing the quote.

why the unique nature of our data (see Section 5 below) is ideally suited for our empirical analysis.[10]

3 The Model

There are three time periods, $t = 0,\ 1,\ 2$, and two tradable assets, a risk-free asset with a rate of return normalized to zero and a risky asset with a random payoff d at time $t = 2$. We assume that d is normally distributed with mean $\bar{d}$ and variance σ_d^2.

The D2D market operates at $t = 1$. The market is populated by $M > 2$ heterogeneous dealers, indexed by $l = 1, \cdots, M$, and having linear-quadratic utilities. Dealer l that begins $t = 1$ with χ_l units of risky asset and trades Q_l units at price $\mathcal{P}^{D2D}$ derives expected utility

$$U_l^{D2D} = \bar{d}\chi_l + (\bar{d} - \mathcal{P}^{D2D})Q_l - \frac{\Gamma_l}{2}(\chi_l + Q_l)^2.$$

We refer to the coefficient Γ_l as dealer l's risk aversion. We assume that the dealer knows his initial inventory χ_l but does not know that of other dealers. The D2D market is structured as a uniform-price double auction. Dealer l submits a (net) demand schedule $Q_l(\mathcal{P}^{D2D}) : \mathbb{R} \to \mathbb{R}$, which specifies demanded quantity of the asset given its price $\mathcal{P}^{D2D}$ in the interdealer market. The price $\mathcal{P}^{D2D}$ is such that the market clears, that is, $\sum_{l=1}^{M} Q_l(\mathcal{P}^{D2D}) = 0$. All dealers are strategic, and there are no noise traders.

The D2C market operates at $t = 0$. Dealers maximize a similar linear-quadratic objective. Dealer l that begins $t = 0$ with x_l units of risky asset and trades q_l units at price p_l derives expected utility

$$U_l^{D2C} = \bar{d}x_l + (\bar{d} - p_l)q_l - \frac{\Gamma_l}{2}(x_l + q_l)^2.$$

[10] Our empirical approach can be useful in markets with request-for-market (RFM) trading mechanisms, such as the interest rate swap market. In the RFM trading mechanism, a customer requests two-sided quotes from the dealers simultaneously. Such mechanisms are getting traction in the interest rate swaps market because they allow traders to hide their trading directions.

Note that the dealers do not account for the utility derived in the D2D trading round. It means that traders in the D2C trading desk trade independently from traders in the D2D desk. This is a simplifying assumption. Our extended model in Appendix IA.3 does not make it and delivers the same qualitative results.[11] Anecdotally, this is a realistic assumption, especially at high frequencies.[12]

We take a reduced-form approach to modeling customers. Dealers compete for customer order flow $\tilde{q}_c$ by providing bid and ask prices p_l^b and p_l^a, respectively. The prices p_l^b and p_l^a are provided to customers and are not observed by competitor dealers. We denote the mid-price by $\alpha_l \equiv \frac{p_l^b + p_l^a}{2}$ and the bid-ask spread by $b_l \equiv \frac{p_l^a - p_l^b}{2}$. We assume that $\tilde{q}_c$ is drawn from an arbitrary symmetric non-degenerate distribution with finite first two moments and that $\tilde{q}_c$ is independent of all other random variables in the model.[13] Positive (negative) realizations of $\tilde{q}_c$ correspond to customer sell (buy) volume. We denote $\tilde{q} = |\tilde{q}_c|$. In what follows, we normalize all quantities relative to the typical liquidity shock $\mathbb{E}[\tilde{q}]$. Thus, without loss of generality,

$$\mathbb{E}[\tilde{q}] = 1.$$

We assume dealers posting better prices are likelier to win the order flow. Consider a customer sell order $\tilde{q}$. We assume that for a dealer l who quotes bid_l, the probability of winning the order is given by

$$\psi(\text{bid}_l - \text{best bid}) \in [0, 1],$$

[11] That is, it yields the same signs of β's in (1).

[12] In our empirical application, one period of the model corresponds to a 10 seconds time interval. At such frequencies, D2D and D2C trading desks indeed act independently.

[13] The symmetry assumption can be relaxed. We only need the first two moments of $\tilde{q}_c|\tilde{q}_c > 0$ and $\tilde{q}_c|\tilde{q}_c < 0$ to be approximately the same in absolute value. This is true in our data: see Table 4, and compare the summary statistics for ProxyLiqShockBuy (our proxy for $\tilde{q}_c|\tilde{q}_c < 0$) and ProxyLiqShockSell (our proxy for $\tilde{q}_c|\tilde{q}_c > 0$).

where $\psi(\cdot)$ is an increasing and differentiable function. Similarly, for buy orders, the probability is $\psi(\text{best ask} - \text{ask}_l)$. Our formulation includes the following special cases:

- Exclusive customer-dealer relationship. In this case, a customer always trades with a preferred dealer, irrespective of that dealer's bid. In the symmetric case, $\psi(\cdot) = 1/M$: the order comes from the dealer l's "loyal" customer with probability $1/M$.

- Fully non-exclusive customer-dealer relationship: $\psi(x) = 1(x \geq 0)$. Note that in this case, $\psi(\cdot)$ is non-differentiable. However, we can approximate this case with, for example, a standard normal CDF $\Phi(\cdot)$, $\psi(x) = \Phi(x/\sigma)$, and σ sufficiently small.

In the intermediate cases, the idea is that a dealer who does not offer the best price might still win the order if his price is not too far from the best price. Such orders may come from loyal customers (who can still go to other dealers if the preferred dealer's price is too bad) or from customers splitting large orders across dealers (this is the mechanism at play in our extended model in Appendix IA.3).

Define the *reservation ask (bid) price* r_l^a (r_l^b) as the smallest ask price (largest bid price) acceptable to a dealer l. It is straightforward to derive[14]

$$r_l^a = \bar{d} - \Gamma_l \left(x_l - \frac{1}{2}\mathbb{E}[\tilde{q}^2] \right) \text{ and } r_l^b = \bar{d} - \Gamma_l \left(x_l + \frac{1}{2}\mathbb{E}[\tilde{q}^2] \right). \tag{2}$$

We assume that dealer's prices are affine functions of their reservation values, as follows:

$$p_l^a = k_0^a + k \cdot r_l^a \text{ and } p_l^b = k_0^b + k \cdot r_l^b, \tag{3}$$

where coefficients k, k_0^a and k_0^b are the same across dealers. The relation (3) holds exactly when the D2C market is structured as a second-price auction (in which case bidding reservation value is an equilibrium strategy). In the general case, (3) holds approximately

[14] The r_l^a solves $\bar{d}\,(x_l - \mathbb{E}\,[\tilde{q}]) + r_l^a \mathbb{E}\,[\tilde{q}] - \frac{\Gamma_l}{2}\mathbb{E}\left[(x_l - \tilde{q})^2\right] = \bar{d}x_l - \frac{\Gamma_l}{2}x_l^2$. A similar calculation applies to r_l^b.

when the gains from trade between customers and dealers are small (see Appendix F). Such approximate relation is sufficient for our main predictions.

Remark 1 (Simplifying assumptions and the extended model) In our model, we have assumed that dealers' prices in the D2C market cannot depend on the quantities demanded by customers. In other words, in our simple model, dealers have *no flexibility* in conditioning prices on quantities. As we discuss in Section 2, dealers have *partial flexibility* in the real FX market: they provide separate quotes for orders below \$1m and orders between \$1m and \$5m. In our extended model of Appendix IA.3, dealers can post linear price schedules; that is, they have *full flexibility* in conditioning prices on quantities. We believe that not capturing the partial flexibility of real FX markets is not crucial for our results because the results in the model with no flexibility (the model here) and full flexibility (the extended model) are the same. Second, we have assumed price-inelastic customer demand $\tilde{q}$. In our extended model, customer demand adjusts to prices posted by dealers. Because the implications in the two models are the same, this simplification of customer demand is not crucial for our results. Finally, unlike in the model here, customers can split their large orders across multiple dealers in the real world. The absence of such order splitting is not crucial for our results as our extended model allows for it and derives the same implications.

4 Theoretical Analysis

This section presents our main results about the interaction between prices and liquidity in the D2D and D2C market segments. We introduce some notation. For a vector $\{x_l\}_l$ we use $E[x_l] = \frac{1}{M}\sum_l x_l$ to denote its cross-sectional mean, and $\text{Var}[x_l] = \frac{1}{M}\sum_l (x_l - E[x_l])^2$ to denote the cross-sectional variance. For vectors $\{x_l\}_l$ and $\{y_l\}_l$ $\text{Cov}(x_l, y_l) = E[(x_l - E[x_l])(y_l - E[y_l])]$ denotes the cross-sectional covariance. $||x_l|| = \sqrt{\sum_l x_l^2}$ denotes the Euclidean norm of a vector $\{x_l\}_l$. The moments of random variables are denoted with blackboard letters, e.g., $\mathbb{E}[\tilde{q}]$ and $\mathbb{V}[\tilde{q}]$ denote the mean and the variance of a random variable $\tilde{q}$. See Appendix A for a summary of notation.

4.1 D2D Prices and the Cross-Section of D2C Quotes

We start with the main result of this subsection. All proofs are in the Appendix.

Proposition 1 *Assume that heterogeneity in both* $\{\Gamma_l\}$ *and* $\{x_l\}$ *is small. Denote* $\epsilon = \max\{||x_l||, ||\Gamma_l - \Gamma^*||\}$. *Then we have*

$$\mathcal{P}^{D2D} = \beta_0^p + \beta_1^p E\left[\alpha_l\right] - \beta_2^p \operatorname{Cov}(\alpha_l, b_l) + o(\epsilon). \tag{4}$$

Where $\beta_0^p = -\frac{k_0^a + k_0^b}{k} \geq 0$, $\beta_1^p = \frac{1}{k} > 0$, $\beta_2^p = c_1(M)\frac{\phi_1}{k}\psi'(0) > 0$, $\phi_1 = \frac{2}{\mathbb{E}[q^2]k}$ *and* $c_1(M) = \frac{M}{(M-2)M+2}$. *In a model without heterogeneity,* $\mathcal{P}^{D2D} = \beta_0^p + \beta_1^p E\left[\alpha_l\right]$.[15]

To understand the economic mechanism behind Proposition 1, it is worth taking an intermediate step of first relating prices to the cross-section of dealers' inventories and risk-aversions and then relating the latter to the cross-section of D2C mid quotes and bid-ask spreads. We do such analysis in the Appendix IA.1 and only present the intuition here.

The fact that D2D prices and average mid-quotes are positively related is intuitive. In our model, this happens because (1) average D2C mid-quote $E[\alpha_l]$ is negatively related to dealers' average inventory $E[x_l]$ and (2) average inventory is, in turn, negatively related to D2D price $\mathcal{P}^{D2D}$. Part (1) is intuitive and reflects that dealers' marginal cost of holding more inventory is increasing in the amount they already have. Part (2) is also intuitive and is at the heart of Evans and Lyons (2002b) seminal findings about the predictability of FX rates with D2D order flow.

We now turn to the central finding of Proposition 1: that prices in the D2D market are negatively related to $\operatorname{Cov}(\alpha_l, b_l)$. As we explain below, this happens because higher

[15] Note that in a model without heterogeneity, all dealers quote the same prices $\alpha_l = E[\alpha_l]$ for any l.

$\text{Cov}(\alpha_l, b_l)$ is symptomatic of a larger *liquidity mismatch*: a situation whereby more risk-averse dealers hold more inventories.

Our first measure of mismatch mirrors the definition:

$$Y_{mismatch} \equiv \text{Cov}(\chi_l, \Gamma_l),$$

Which is a cross-sectional covariance between dealers' post-D2C inventories χ_l and risk-aversions Γ_l.

That prices should be lower when $Y_{mismatch}$ is larger is intuitive. When $Y_{mismatch} > 0$, the allocation of inventories across dealers is inefficient: it could be improved by reallocating the inventory from more to less risk-averse dealers. Due to market power, such inefficiency cannot be resolved in one round of D2D trade. As a result, under liquidity mismatch, the dealer sector as a whole is more risk-averse and will require higher compensation for holding the inventories, resulting in lower prices.[16]

Our final step shows that higher $Y_{mismatch}$ is associated with higher $\text{Cov}(\alpha_l, b_l)$. Thus, the observable quantity $\text{Cov}(\alpha_l, b_l)$ can be used as a proxy for unobservable $Y_{mismatch}$ allowing us to define

$$\mathcal{A}_{mismatch} \equiv \text{Cov}(\alpha_l, b_l),$$

as our second (price-based) mismatch measure. The relationship between the two mismatch measures holds because: (i) risk aversions and bid-ask spreads are positively related, and (ii) inventories (with which dealers start the D2D trading round) and D2C prices are positively related. Part (i) is intuitive. More risk-averse dealers are less efficient at holding inventory and require higher compensation, resulting in wider spreads. To see part (ii), note that dealers posting the highest (lowest) prices will attract a disproportionate share of aggregate

[16]See Proposition IA.7 for a formal analysis underlying the intuition here.

customer sell (buy) volume. Thus dealers with the lowest (highest) prices will decrease (increase) their inventories, implying a positive cross-sectional relationship between inventories and D2C prices.[17]

4.2 D2D liquidity and the Cross-Section of D2C Quotes

We start by introducing our theoretical measure of illiquidity in the D2D market. Denote $Q_l(\mathcal{P}^{D2D})$ the demand schedule of dealer l. Our illiquidity measure is Λ, where

$$\Lambda^{-1} = 2 \sum_l \frac{\partial Q_l}{\partial \mathcal{P}^{D2D}}.$$

Here 2 is just a convenient normalizing factor. Notably, Λ is proportional to the reciprocal of the slope of the aggregate dealer's demand, which is a standard measure of liquidity in the theoretical literature (see Vayanos and Wang (2013)). For our empirical work, we relate Λ to percentage bid-ask spreads (see Section 5.1). Our key result is as follows.

Proposition 2 *Assume that cross-sectional mean of dealer's risk aversions is close to* Γ^* *and that cross-sectional variance of dealers' risk aversions is small. Denote* $E[\Gamma_l] - \Gamma^* = \epsilon_1$, $Var[\Gamma_l] = \epsilon_2$ *and* $\epsilon = \max(\epsilon_1, \epsilon_2)$. *The following is true*

$$\Lambda = \beta_0^\Lambda + \beta_1^\Lambda E[b_l] - \beta_2^\Lambda Var[b_l] + o(\epsilon), \tag{5}$$

where $\beta_0^\Lambda = \phi_0 \frac{\Lambda_0}{\Gamma^*} > 0$, $\beta_1^\Lambda = \phi_1 \frac{\Lambda_0}{\Gamma^*} > 0$, $\beta_2^\Lambda = \frac{1}{2}\phi_1^2 \frac{\Lambda_0^3\left(1-\sqrt{\frac{\Gamma_*^2}{\Lambda_0^2}+1}\right)+2\Gamma_*^2\Lambda_0}{\Gamma_*^2\left(\Gamma_*^2+\Lambda_0^2\right)} > 0$, $\Lambda_0 = \frac{2\Gamma_*(M-1)}{(M-2)M}$, $\phi_0 = -\frac{(k_0^a - k_0^b)}{\mathbb{E}[q^2]k}$ *and* $\phi_1 = \frac{2}{\mathbb{E}[q^2]k}$. *In a model without heterogeneity,* $\Lambda = \beta_0^\Lambda + \beta_1^\Lambda E[b_l]$.

It makes sense that illiquidity in the D2D market is positively related to the average bid-ask spreads $E[b_l]$. In our model, $E[b_l]$ reflects the dealers' average risk-aversion $E[\Gamma_l]$ (see (3)), which is positively related to illiquidity.

[17] The formal result underlying the intuition here is Proposition IA.10.

The central finding of Proposition 2 is a negative relationship between the illiquidity of the D2D market and the cross-sectional variance of D2D spreads. It arises because spreads are linearly related to risk aversions (see (3)), and illiquidity is smaller when cross-sectional variance or dealers' risk aversion is smaller. To understand the intuition behind the last statement, consider an example. Imagine we have three dealers with a risk aversion of 1 each. They each provide 1 unit of liquidity, 3 in total. Now consider what happens if dealers' average risk aversion is 0.5, 1, and 1.5 (so that average risk aversion is the same, but we increased the variance). Since liquidity provided (price elasticity) is inversely related to risk aversion, the dealers will provide 2, 1, and 1/1.5=0.66 units of liquidity, 3.66 in total. Thus, greater dispersion in risk aversions results in more liquidity.[18]

4.3 Relation to Other Theoretical Work

The negative relationship between prices in the D2D market and $\mathcal{A}_{mismatch}$, ($\beta_2^p > 0$) is the unique prediction of our theory. Non-exclusivity, dealer heterogeneity, and imperfect competition in the D2D market are all important ingredients underlying it. Indeed, without heterogeneity, the only possibility for $\mathcal{A}_{mismatch}$ to be different from zero is due to noise in the data (in theory, homogenous dealers should post the same mid-quotes and bid-ask spreads). Even in that case $\mathcal{A}_{mismatch}$ should not be related to $\mathcal{P}^{D2D}$ and so $\beta_2^p = 0$.

In the competitive case, standard aggregation results imply that all dealers can be replaced with a representative dealer holding the dealer sector's aggregate inventory $\sum_l \chi_l$. In that case, the aggregate inventory, but not its distribution across dealers, should matter for asset prices. Consistent with that, we demonstrate in Appendix E that the competitive version of our model yields $\beta_2^p = 0$.

[18]The price elasticity is inversely proportional to risk aversion when heterogeneity in risk aversions is small. In the general case, this relationship is more complex. However, price elasticity is still a convex function of risk aversion in the general case. Thus the logic of our simple example still applies. We also note that Proposition IA.9 establishes negative relationship between Λ and $\mathrm{Var}[b_l]$ without relying on small heterogeneity approximation.

Finally, the pure exclusive version of our model implies $\psi'(\cdot) = 0$ and so $\beta_2^p = 0$. In Appendix IA.6, we consider a model of Babus and Parlatore (2022) that, compared to our model, additionally features price-elastic customer demand. We show that $\beta_2^p < 0$ there. Intuitively, customers and dealers trade toward an efficient allocation, where their risks are perfectly shared, but they only do so imperfectly due to dealers' market power. Thus, dealers who start with relatively high pre-D2C inventory x_l will end up with relatively high post-D2C inventories $\tilde{\chi}_l$. These dealers also tend to post lower D2C prices, creating a *negative* relationship between prices and inventories. (In contrast, we discussed in Section 4.1 that such a relationship is positive in our model.) Overall, we can conclude that $\beta_2^p \leq 0$ with exclusivity.

To conclude, the positive sign of β_2^p is the unique prediction of our theory, distinguishing it from models with no heterogeneity (e.g., Vogler (1997)), perfectly competitive D2D market (e.g., Dunne et al. (2015)) or exclusive customer-dealer relationship (e.g., Babus and Parlatore (2022)).

5 Empirical Analysis

In this section, we present and test our empirical hypotheses concerning the signs of the coefficients in the Propositions 1 and 2. The implications of the magnitudes of these coefficients are explored in Section 6, where we use the theoretical values of these coefficients to calibrate the elasticity of the FX market.

5.1 Identifying Assumptions

Our theoretical predictions in Proposition 1 are regarding price levels. Price time series are non-stationary. To facilitate our empirical analysis, we derive implications for returns. We assume that prices in the D2D market are consistent with Proposition 1 at any time. Then, one can write:

$$\mathcal{P}_{t+\ell}^{D2D} = \beta_0^p + \beta_1^p E\left[\alpha_{l,t}\right] - \beta_2^p cov(\alpha_{l,t}, b_{l,t}) + o(\epsilon).$$

This is as given by equation (4) where we additionally assume that one period in the model corresponds to ℓ seconds in the data. Taking differences and dividing by $\mathcal{P}_t^{D2D}$ we can rewrite the above equation as follows:

$$\mathcal{R}_{t+\ell}^{D2D} \equiv \frac{\Delta\mathcal{P}_{t+\ell}^{D2D}}{\mathcal{P}_t^{D2D}} = \beta_1^p \frac{\Delta E\left[\alpha_{l,t}\right]}{\mathcal{P}_t^{D2D}} - \beta_2^p \frac{\Delta cov(\alpha_{l,t}, b_{l,t})}{\mathcal{P}_t^{D2D}} + o(\epsilon), \tag{6}$$

where we use the notation $\Delta X_t = X_t - X_{t-\ell}$. To proceed, we note that in our data, $\mathcal{P}_t^{D2D} \approx E[\alpha_{l,t}]$ (see Table 4) and, for ℓ small, $E[\alpha_{l,t+\ell}] \approx E[\alpha_{l,t}]$.[19] This allows us to rewrite (6) in terms of returns and percentage spreads as follows

$$\mathcal{R}_{t+\ell}^{D2D} = \beta_1^p \mathcal{R}_t^{D2C} - \beta_2^p \Delta\widehat{\mathcal{A}}_{mismatch,t} + \epsilon_{t+\ell}, \tag{7}$$

where we have denoted

$$\mathcal{R}_t^{D2C} \equiv \frac{\Delta E\left[\alpha_{l,t}\right]}{E\left[\alpha_{l,t}\right]},\ BA_{l,t}^{D2C} \equiv \frac{b_{l,t}}{E\left[\alpha_{l,t}\right]} \text{ and } \widehat{\mathcal{A}}_{mismatch,t} \equiv cov(\alpha_{l,t}, BA_{l,t}^{D2C}),$$

and replaced $o(\epsilon)$ with the noise term $\epsilon_{t+\ell}$. We formulate our first identifying assumption.

Assumption 1 *The data-generating process for D2D returns is given by (7), where β_1^p and β_2^p are constants, and the noise $\epsilon_{t+\ell}$ is orthogonal to $\mathcal{R}_{t+\ell}^{D2D}$.*

Similarly, our predictions in Proposition 2 are in terms of illiquidity Λ, which we do not directly observe in the data. In our second identifying assumption, we relate Λ to the percentage bid-ask spreads

$$BA^{D2D} = \frac{\text{Ask}^{D2D} - \text{Bid}^{D2D}}{\mathcal{P}^{D2D}}.$$

[19]For example, for $\ell = 10s$, corr$(E[\alpha_{l,t+\ell}], E[\alpha_{l,t}]) = 0.999977$ in our sample.

Assumption 2 *There is an affine relationship between percentage price impact* $\Lambda_t^{\%} = \Lambda_t/\mathcal{P}_t^{D2D}$ *and percentage bid ask spreads* BA_{D2D}*:* $h_0 + h_1\Lambda_t^{\%} = BA_t^{D2D}$, *where* h_0 *and* h_1 *are constants.*

One microfoundation for Assumption 2 is as follows. In a linear model, one can write that $(\text{Ask} - \text{Bid})/(2 \cdot \text{minimum lot size}) = \Lambda$. Therefore, the above holds with $h_1 = 2 \cdot \text{minimum lot size}$ and $h_0 = 0$. However, we don't take a stand on a particular microfoundation and let the data decide the values of both h_0 and h_1.

Next, we clarify our assumption (3). It says that coefficients k_0^a, k_0^b and k are the same across dealers. However, we also need to state how these coefficients change over time. We clarify this in our next assumption.

Assumption 3 *There is an affine relationship between* $p_{l,t}^a/\mathcal{P}_t^{D2D}$ *and* $r_{l,t}^a/\mathcal{P}_t^{D2D}$, *as well as* $p_{l,t}^b/\mathcal{P}_t^{D2D}$ *and* $r_{l,t}^b/\mathcal{P}_t^{D2D}$*:* $p_{l,t}^a/\mathcal{P}_t^{D2D} = \varkappa_0^a + \varkappa \cdot r_{l,t}^a/\mathcal{P}_t^{D2D}$ *and* $p_{l,t}^b/\mathcal{P}_t^{D2D} = \varkappa_0^b + \varkappa \cdot r_{l,t}^b/\mathcal{P}_t^{D2D}$, *where* $\varkappa_0^a$, $\varkappa_0^b$ *and* $\varkappa$ *are constants.*

The idea behind this assumption is that the time variation in D2D prices absorbs the time variation in the cross-section of D2C prices and reservation values. This is a reasonable assumption, given the tight link between D2D and D2C prices (see Table 4).[20]

Using Assumptions 2 and 3, and proceeding similarly to the way equation (7) was derived, we rewrite (5) as

$$BA_{t+\ell}^{D2D} = \beta_0^{BA} + \beta_1^{BA} E[BA_{l,t}^{D2C}] - \beta_2^{BA} \mathrm{Var}[BA_{l,t}^{D2C}] + \epsilon_{t+l}. \tag{8}$$

We relegate the derivation to Appendix B.2, where we also present the closed-form expressions for β_0^{BA}, β_1^{BA} and β_2^{BA}. We present our final identifying assumption.

Assumption 4 *The data-generating process for D2D percentage spreads is given by* (8), *where* β_0^{BA}, β_1^{BA} *and* β_2^{BA} *are constants, and the noise* $\epsilon_{t+\ell}$ *is orthogonal to* $BA_{t+\ell}^{D2D}$.

[20] We also note that this assumption is not crucial for neither our estimation nor calibration. If we proceed assuming that k_0^a, k_0^b, and k are constant over time, which amounts to using bid-ask spreads instead of percentage spreads, our results are very similar.

5.2 Data Description

We test the model's predictions using three high-frequency datasets time-stamped with 1-second precision. The first two datasets are provided to us by a large Swiss retail aggregator (the retail aggregator hereafter). The first dataset contains price quotes for the currency pair "EUR/USD" for orders submitted to the retail aggregator by the largest FOREX dealers.[21] Separate quotes are provided for trades below USD 1 million and between USD 1 million and USD 5 million. This dataset covers the period from June 1, 2016, to September 30, 2016. The retail aggregator cannot access the D2D market and can manage its inventory only by trading with the dealers. Thus, from the point of view of our model, the retail aggregator is a customer trading with dealers in the D2C market.

The second dataset contains the orders submitted for execution by its clients for the same currency pair and period. The geographical composition of the retail aggregator's clients who trade the EUR/USD currency pair is highly heterogeneous and is dominated by large groups of clients from Italy, Switzerland, China, and Spain. 95% of clients in this dataset are retail, while the rest are corporate. For 99% of the observations, the order size does not exceed USD 1 mln. Thus, for our empirical analysis, we filter the data to keep the quotes and the orders with an order size not exceeding 1 million. We refer to the merged first two datasets as the D2C dataset.

The third dataset is the EBS dataset provided to us by NEX Data. We refer to it as D2D dataset going forward. EBS is one of the largest D2D FX platforms in the world. This dataset is a comprehensive account of FX's best bids and asks aggregated within each second. It also indicates when a transaction occurs and its price and size. Mancini et al. (2013) compared the EBS dataset with other FX datasets and concluded that "EBS is effectively the only current data source for intraday data." For the sake of consistency with the D2C dataset, we restrain our D2D dataset to the subset of transactions with an order

[21] According to the Triennial Central Bank Survey (April 2016), this currency pair accounts for the most significant turnover on the FOREX market.

size of 1 mln or less.[22] We follow Mancini et al. (2013) to clean our quote datasets. In addition, for the D2C datasets, we filter the data sample on quotes to retain price quotes from a subsample of dealers simultaneously providing quotes during a long continuous time interval within each trading day. After applying such a filtering procedure, we retain quotes from 10 large dealers.

In our D2C dataset, we observe the bid $p^b_{l,t}$ and the ask $p^a_{l,t}$ for each dealer l at each time instant t, which we then use to directly back out the mid-prices $\alpha_l = \frac{p^a_{l,t}+p^b_{l,t}}{2}$ and spreads $b_l = \frac{p^a_{l,t}-p^b_{l,t}}{2}$; the average mid-prices $E[\alpha_{l,t}] = \frac{1}{M}\sum_{l=1}^{M} \alpha_{l,t}$, average spreads $E[b_{l,t}] = \frac{1}{M}\sum_{l=1}^{M} b_{l,t}$ and variance of spreads $\mathrm{Var}[b_{l,t}] = \frac{1}{M}\sum(b_{l,t} - E[b_{l,t}])^2$; percentage spreads $BA^{D2C}_{l,t} = b_{l,t}/E[\alpha_{l,t}]$, average percentage spreads $E[BA^{D2C}_{l,t}] = \frac{1}{M}\sum_{l=1}^{M} BA^{D2C}_{l,t}$ and mismatch $\widehat{\mathcal{A}}_{mismatch,t} = \mathrm{Cov}(\alpha_{l,t}, BA^{D2C}_{l,t}) = \frac{1}{M}\sum_l (\alpha_{l,t} - E[\alpha_{l,t}])\left(BA^{D2C}_{l,t} - E[BA^{D2C}_{l,t}]\right)$. In our D2D dataset, we observe the bid Bid^{D2D}_t and ask Ask^{D2D}_t, from which we compute $\mathcal{P}^{D2D}_t = \frac{Bid^{D2D}_t+Ask_t}{2}$ and $BA^{D2D}_t = \frac{Ask^{D2D}_t-Bid^{D2D}_t}{2\mathcal{P}^{D2D}_t}$. For values of lag $\ell = 5, 10, 30, 60$, and 90 seconds we also compute $\mathcal{R}^{D2D}_t = \frac{\mathcal{P}^{D2D}_t-\mathcal{P}^{D2D}_{t-\ell}}{\mathcal{P}^{D2D}_t}$ and $\mathcal{R}^{D2C}_t = \frac{E\left[\alpha_{l,t}\right]-E\left[\alpha_{l,t-\ell}\right]}{E\left[\alpha_{l,t}\right]}$. All of these quantities are exactly as in our theoretical analysis.

In addition, we construct proxies for customers' liquidity demand and D2D order flow. Our proxy for the customers' liquidity demand is $ProxyLiqShock$, the net of all seller-initiated orders and buyer-initiated orders from the retail aggregator's clients, that is, the net selling pressure. This definition follows our assumption that any online broker acts as an intermediary between retail clients and large dealers, and such brokers are competitive. The net selling pressure can then be seen as *a scaled proxy* of the total customer shocks $\tilde{q}_c$. In addition, we will use the summary statistics for the positive and negative parts of $ProxyLiqShock$, which we denote $ProxyLiqShockSell$ and $ProxyLiqShockBuy$, respectively. The last two variables are the scaled proxies of $\tilde{q}$. We construct a proxy for order flow in the D2D market using the EBS dataset, aggregated at the second level, which provides

[22] Transactions of size one mln or less account for more than 50% of deals in the EBS data. We ran our analysis with all transaction data in an earlier paper version and obtained the same qualitative results.

volume and buy/sell information for each executed transaction. We use completed orders of size 1 million and construct the following proxy for order flow in the D2D market: OF^{D2D}, the aggregate (signed) net order flow (in mln) in the D2D market computed within a trailing window of 24 hours.[23]

We report the summary statistics and correlation table for the final merged dataset in Tables 4 and 5. Mid-price distributions are essentially the same in the D2C and the D2D markets, while the spreads differ. Differences in prices are small because FX markets are very liquid and have very narrow bid-ask spreads. As one can see from Table 1, a typical bid-ask spread is about 0.5 basis points. Thus, one can only see the difference in the fifth digit after the comma. We also note that even though $\mathcal{A}_{mismatch}$ is constructed using the cross-section of both mid prices α_l and percentage bid-ask spreads BA_l^{D2C} in the D2C market, it is correlated to $E[BA_l^{D2C}]$ and std$[BA_l^{D2C}]$, but not $E[\alpha_l]$. The correlation of $\mathcal{A}_{mismatch}$ with both $E[BA_l^{D2C}]$ and std$[BA_l^{D2C}]$ is approximately 0.6 and so is far from 1. Thus, our measure contains additional information relative to both $E[BA_l^{D2C}]$ and std$[BA_l^{D2C}]$.

[Include Tables 4 and 5 here.]

5.3 Empirical Hypotheses

We split our predictions into two parts: returns and spreads. Assumption 1 implies that (7) can be estimated using OLS regression. Our regression specification is

$$\mathcal{R}_{t+\ell}^{D2D} = \beta_1^p \mathcal{R}_t^{D2C} - \beta_2^p \Delta \widehat{\mathcal{A}}_{mismatch,t} + \text{Constant}. \tag{9}$$

Our theory implies the following signs of regression coefficients: $\beta_1^p > 0$, $\beta_2^p > 0$, and Constant $= 0$.In contrast, as discussed in Section 4 if the D2D market was competitive or

[23]In the previous version of the paper, we have obtained similar results to here with the following alternative proxies for the D2D order flow: (a) aggregate daily (signed) net order flow in the D2D market, and (b) aggregate (signed) net order flow (in mln) in the D2D market computed within a trailing window of 1 hour.

if dealers were homogenous, $\beta_2^p = 0$ and with exclusive customer-dealer realtionship $\beta_2^p < 0$. Thus, the positive sign of β_2^p is the unique prediction of our theory, distinguishing it from models with no heterogeneity (e.g., Vogler (1997)), perfectly competitive D2D market (e.g., Dunne et al. (2015)) or exclusive customer-dealer relationship (e.g., Babus and Parlatore (2022)). Table 1 summarizes our theoretical predictions and that of other theories (in parenthesis).

Table 1: Theoretical predictions, returns. Summarizes our predictions and that of other theories (in parenthesis).

Regression: $\mathcal{R}_{t+\ell}^{D2D} = \beta_1^p \mathcal{R}_t^{D2C} - \beta_2^p \Delta\widehat{\mathcal{A}}_{mismatch,t} + \text{Constant}\,,$	
Regression coefficient	Predicted sign
$\mathcal{R}_t^{D2C}$	> 0
$\Delta\widehat{\mathcal{A}}_{mismatch,t}$	> 0, non-exclusive customer-dealer relationship ($= 0$, competitive benchmark) ($= 0$, no heterogeneity) (< 0, exclusive customer-dealer relationship)
Constant	$= 0$

Similarly, assumption 4 implies that (8) can be estimated using OLS regression

$$BA_{t+\ell}^{D2D} = \beta_1^{BA} E[BA_{l,t}^{D2C}] - \beta_2^{BA} \text{Var}[BA_{l,t}^{D2C}] + \text{Constant}. \tag{10}$$

Our theory predicts $\beta_1^{BA} > 0$ and $\beta_2^{BA} > 0$, whereas without heterogeneity $\beta_2^{BA} = 0$. We summarise the second part of our predictions in Table 2.

In the next section, we estimate regressions (9) and (10) to test our implications about the

Table 2: Theoretical predictions, spreads. Summarizes our predictions and that of other theories (in parenthesis).

Regression: $BA^{D2D}_{t+\ell} = \beta^{BA}_1 E[BA^{D2C}_{l,t}] - \beta^{BA}_2 \mathrm{Var}[BA^{D2C}_{l,t}] + \text{Constant}$	
Regression coefficient	Predicted sign
$E[BA^{D2C}_{l,t}]$	> 0
$\mathrm{Var}[BA^{D2C}_{l,t}]$	$= 0$, no heterogeneity > 0, the model, competitive benchmark

signs of the regression coefficients. Our predictions about the magnitudes of these coefficients are derived and discussed in Section 6.

5.4 Results

To test predictions summarized in Table 1, we estimate (9), with lag value $\ell = 10$ seconds. Table 6 presents the results. Specifications (3) and (4) additionally include $ProxyLiqShock$ and OF^{D2D} as controls. The results perfectly align with the predictions: the regression coefficient on $\mathcal{R}^{D2C}$ and $\Delta\widehat{\mathcal{A}}_{mismatch}$ is positive, while the intercept is zero.

[Include Table 6 here.]

The positive relationship between D2D returns and lagged D2C returns echoes the findings of Evans and Lyons (2002b), who were the first to document the predictive relationship between D2D order flow and exchange rates. Indeed, both the D2C returns and order flow reflect changes in inventory, which, in turn, are associated with returns going forward. However, in our sample at least, the predictive power of lagged D2C returns seems to dominate: while the coefficient on OF^{D2D} has the expected sign, it is not statistically significant (see panel (4) of Table 6).

In contrast, Column (2) shows that customers' order flow in the D2C market ($ProxyLiqShock$) negatively forecasts price changes in the D2D market. The negative sign means that an increase in the customers' liquidity demand results in lower future prices. Given that a significant fraction of the retail aggregator's customers are based in Switzerland, it is astonishing that their order flow predicts changes in the EUR/USD exchange rates quoted on EBS, one of the largest international inter-dealer platforms in the world. Note, however, that the coefficient is quite small. These results align with our interpretation of $ProxyLiqShock$ as a scaled proxy of $\tilde{q}_c$, with a small scaling coefficient. The scaling should not affect the statistical significance but makes the regression coefficient smaller.

The negative relationship between D2D returns and lagged changes in $\widehat{\mathcal{A}}_{mismatch}$ confirms the unique prediction of our theory, distinguishing it from theories with no heterogeneity (e.g., Vogler (1997)), perfectly competitive D2D market (e.g., Dunne et al. (2015)) or exclusive customer-dealer relationship (e.g., Babus and Parlatore (2022)). This finding also highlights that dealer heterogeneity, imperfect competition, and non-exclusivity are all jointly important features underlying the high-frequency dynamics of exchange rates.

We repeat the same empirical test for lags $\ell = 5, 30, 60$, and 90 seconds and present the results in Section D.1. The main observation is that the impact of liquidity mismatch on future D2D prices is significant for small lags ℓ but loses significance as we increase ℓ. This result is intuitive: Liquidity mismatch matters because price impact precludes heterogeneously risk-averse dealers from immediately trading toward the efficient allocation of risk. Instead, they are forced to trade toward their risk target over multiple rounds. In an ultra-fast foreign exchange market, it takes dealers several seconds to rebalance their portfolios.

We turn to predictions about spreads (Table 2). To this end, we estimate (10). Table 7 presents the results. Again, the results perfectly align with the predictions: the regression coefficients on $E[BA^{D2C}]$ and $\mathrm{Var}[BA^{D2C}]$ are both positive. The negative association between the variance of D2C spreads and D2D spreads going forward underlines the importance of heterogeneity in explaining the high-frequency of FX market liquidity. Controlling for D2D order flow and customers' liquidity shocks does not impact the results. We repeat this empirical test for lags $\ell = 5, 20, 30, 60$, and 90 and present the results in Section D.2. As one can see, regression results strongly support our empirical predictions at all horizons. The predictive power of the variance of D2C spreads remains statistically significant for longer lags.

[Include Table 7 here.]

6 Calibration

In this section, we will use our closed-form solutions for the coefficients in the regressions (9) and (10) to calibrate the elasticity of FX market. The reciprocal of elasticity (referred to as inverse-elasticity going forward) is defined as price change (expressed in basis points) induced by a typical liquidity shock $\mathbb{E}[\tilde{q}]$, normalized by the size of the shock:

$$\mathcal{E}^{-1} \equiv \frac{\mathbb{E}[\tilde{q}]}{\mathcal{P}^{D2D}} \frac{d\mathcal{P}^{D2D}}{dQ} \cdot 10^4 = \frac{\Lambda}{2\mathcal{P}^{D2D}} \cdot 10^4. \tag{11}$$

To obtain the last expression we substituted $\Lambda/2 = \frac{d\mathcal{P}^{D2D}}{dQ}$ and accounted for the fact that we normalized $\mathbb{E}[\tilde{q}] = 1$. Further, we approximate $\mathcal{E}^{-1} \approx \mathcal{E}_0^{-1} = \frac{\Lambda_0}{2\mathcal{P}^{D2D}} \cdot 10^4$.[24]

[24] Λ_0 is the leading term in the "small heterogeneity" approximation to Λ (see Proof of Proposition 2, equation (19)): $\Lambda_0 = \lim_{\epsilon \to 0} \Lambda(\epsilon)$, where ϵ is defined in the Proposition 2. Using higher-order terms has little effect on our numerical results, but complicates the algebra.

6.1 Identifying Assumptions

Proceeding with calibration requires knowing the number of dealers M and estimating the moments of liquidity shock $\tilde{q}$. For the first requirement, we proceed with

$$M = 10,$$

the number of dealers in our dataset as our baseline scenario.[25] As for the first one, the needed quantity is *coefficient of variation* of $\tilde{q}$,

$$\mathbb{CV}[\tilde{q}] \equiv \frac{\sqrt{\mathbb{V}[\tilde{q}]}}{\mathbb{E}[\tilde{q}]}.$$

As discussed in Section 5.2 we view the variable $ProxyLiqShock$ as a scaled proxy for $\tilde{q}_c$, which we now formulate as an assumption.

Assumption 5 *The liquidity shock $\tilde{q}_c$ is proportional to the variable $ProxyLiqShock$, $\tilde{q}_c \propto ProxyLiqShock$.*

Given Assuption 5 we can compute $\mathbb{CV}[\tilde{q}]$ using either positive or negative parts of $ProxyLiqShock$ ($ProxyLiqShockSell$ and $ProxyLiqShockBuy$, respectively). The two estimates are quite similar (see Table 4), and we take the average of the two:

$$\check{\mathbb{CV}}[\tilde{q}] = \frac{\mathbb{CV}[ProxyLiqShockSell] + \mathbb{CV}[ProxyLiqShockBuy]}{2} = 5.324.$$

Here we introduced the notation $\check{X}$ for an empirical estimate of X. Invoking Assumption 5, $\mathcal{E}_0$ can be uniquely identified.

[25] Our closed-form solution for elasticity makes it straightforward to do the sensitivity analysis with respect to M.

Proposition 3 *The elasticity $\mathcal{E}_0$ is uniquely identified as follows*

$$\mathcal{E}_0 = \frac{\check{\beta}_2^{BA}}{\check{\beta}_1^{BA}\check{\beta}_1^{p}} \cdot (1 + \check{\mathbb{CV}}[\tilde{q}]^2) \cdot c_{\mathcal{E}}(M), \tag{12}$$

where $c_{\mathcal{E}}(M) = \frac{M((M-2)M+2)^2}{2(M-2)(M-1)((M-1)M+1)} \cdot 10^{-4}$.

We note that heterogeneity is key to identifying elasticity. Without $\check{\beta}_2^{BA}$ the elasticity $\mathcal{E}_0$ cannot be identified.

6.2 Results

Using our closed-form expression (12), summary statistics in Table 4, and our baseline empirical results for $\ell = 10s$ in Tables 6 and 7 we obtain the inverse-elasticity of

$$\mathcal{E}_0^{-1} = 0.4958.$$

Our estimate implies that a typical liquidity shock moves mid-prices in the D2D market by 0.5 basis points. This is comparable to the average bid-ask spread of 0.44 basis points (see Table 4). The D2D market is illiquid.

In contrast, had the D2D market been perfectly competitive the inverse-elasticity would have been (see the derivation in Section E.3)

$$\mathcal{E}_0^{-1} = 0.4407.$$

This suggests that bid-ask spreads in the D2D market would have been $12.5\% = 0.4958/0.4407 - 1$ smaller with competitive dealers. The D2D market is non-competitive.

6.2.1 Inelastic Market Hypothesis in FX market

Here we ask the following question: can the illiquidity of the D2D market and the shocks originating in the D2C market explain the volatility of high-frequency FX returns? That is, does the inelastic market hypothesis (Gabaix and Koijen, 2021) hold for the FX market? To address this question, we compute the volatility of the return component generated by the liquidity shocks and compare it to the total return volatility.

In addition to Assumption 5 specifying the distribution of liquidity shocks, we need to specify how these shocks arrive over time.

Assumption 6 *Liquidity shocks $\tilde{q}_{c,t}$ are i.i.d., satisfy the Assumption 5, and arrive at a Poisson intensity ν. In addition, we assume that ν can be identified as the frequency of changes in the D2D order flow in our sample, $\nu = 0.0647\ \mathrm{sec}^{-1}$.*

Given Assumption 6, the sum of liquidity shocks between time t and $t+\ell$ is given by

$$\widetilde{Q}_c^{t\to t+\ell} = \sum_{i}^{N(\ell)} \tilde{q}_{c,i},$$

where $N(\tau)$ denotes Poisson counting process with intensity ν. The change in price between t and $t+\ell$ is given by

$$\mathcal{P}_{t+\ell}^{D2D} - \mathcal{P}_t^{D2D} = \text{change in fundamentals} + \frac{\mathcal{P}_t^{D2D}\mathcal{E}^{-1}}{\mathbb{E}[\tilde{q}]} \cdot \widetilde{Q}_c^{t\to t+\ell}.$$

The second component is the price reaction to the liquidity shock.[26] Dividing by $\mathcal{P}_t^{D2D}$ we decompose the return into two parts: $\mathcal{R}_{t+\ell}^{D2D} = \mathcal{R}_{t+\ell}^{f,D2D} + \mathcal{R}_{t+\ell}^{liq,D2D}$, where the fundamental part is by definition $\mathcal{R}_{t+\ell}^{f,D2D} \equiv$ change in fundamentals$/\mathcal{P}_t^{D2D}$ and non-fundamental part is

[26]The price reaction to the liquidity shock is the price impact of the liquidity shock, $\left(\frac{\mathcal{P}_t^{D2D}\mathcal{E}^{-1}}{\mathbb{E}[\tilde{q}]}\right)$ (see (11)), times the sum of liquidity shocks ($\widetilde{Q}_c^{t\to t+\ell}$).

caused by liquidity shocks and is given by

$$\mathcal{R}_{t+\ell}^{liq,D2D} = \frac{\mathcal{E}^{-1}}{\mathbb{E}[\tilde{q}]}\widetilde{Q}_c^{t\to t+\ell}.$$

This gives the volatility of the non-fundamental component

$$\begin{aligned}
\sigma\left[\mathcal{R}_{t+\ell}^{liq,D2D}\right] &= \frac{\mathcal{E}^{-1}}{\mathbb{E}[\tilde{q}]}\cdot\sigma\left[\widetilde{Q}_c^{t\to t+\ell}\right] \\
&= \frac{\mathcal{E}^{-1}}{\mathbb{E}[\tilde{q}]}\cdot\nu\sqrt{\ell}\sigma\left[\tilde{q}_c\right] \\
&= \nu\sqrt{\ell}\mathcal{E}^{-1}\mathbb{CV}[\tilde{q}]\frac{\sigma\left[\tilde{q}_c\right]}{\sigma\left[\tilde{q}\right]} \\
&= 0.1812 \text{ bps.}
\end{aligned}$$

In the second line, we used the standard properties of a Poisson process to compute $\sigma\left[\widetilde{Q}_c^{t\to t+\ell}\right] = \nu\sqrt{\ell}\sigma\left[\tilde{q}_c\right]$. In the third line, we multiplied and divided by $\sigma[\tilde{q}]$ and substituted $\mathbb{CV}[\tilde{q}] = \sigma[\tilde{q}]/E[\tilde{q}]$. In the fourth line, we use Assumption 5 to substitute

$$\frac{\sigma[\tilde{q}_c]}{\sigma[\tilde{q}]} = \frac{\text{std}[ProxyLiqShock]}{0.5(\text{std}[ProxyLiqShockBuy] + \text{std}[ProxyLiqShockSell])} = 0.3355,$$

along with the values of $l = 10$, $\nu = 0.0647$, and $\mathbb{CV}[\tilde{q}] = 5.324$. The volatility of 10-second returns in our sample is $\sigma\left[\mathcal{R}_{t+\ell}^{D2D}\right] = 0.5539$. This means

$$\sigma\left[\mathcal{R}_{t+\ell}^{liq,D2D}\right]/\sigma\left[\mathcal{R}_{t+\ell}^{D2D}\right] = 32.71\%$$

The non-fundamental volatility accounts for around a third of overall short-term volatility in the FX market. The inelastic market hypothesis holds.

7 Literature Review

A growing body of literature shows how order flow and dealer inventories serve as essential determinants of exchange rates. Lyons (1995) was one of the first to provide strong empirical evidence that dealers actively control their inventories and study how this inventory control creates a link between order flow and exchange rates. In particular, consistent with classical theories (see, e.g., Ho and Stoll (1981)), Lyons (1995) shows that dealers "shade prices"–that is, they shift prices in the direction opposite to their inventory. While such price shading has been documented in other markets,[27] recent studies, such as Bjønnes and Rime (2005) and Osler et al. (2011), have not found evidence of price shading in the FX markets. In contrast to these papers, and in agreement with Lyons (1995), our empirical results provide strong evidence of price shading in the D2C market.

In a seminal contribution, Evans and Lyons (2002b) develop the first theoretical model to account for the two-tier structure of the FX market and derive an endogenous link between order flow and exchange rates. Consistent with the constraints that the real-world FX dealers face, Evans and Lyons (2002b) assumes that dealers need to hold zero inventory overnight. As a result, dealers optimally shade their prices to achieve the zero inventory target. This price shading leads to a contemporaneous relationship between order flow and exchange rates.

Previous research has mostly focused on interdealer trading. An incomplete list of papers focusing on inter-dealer trading includes Reiss and Werner (1998), Reiss and Werner (2005) and, more recently, Li and Schürhoff (2019). A recent paper of Eisfeldt, Herskovic and Liu (2022) (circulated after this one), considers a corporate bond market and shows that when interdealer price dispersion is high, bond prices are low. Similarly, we show that exchange rates are low when liquidity mismatch is high. Our analysis (available upon request) found that D2C price dispersion is not statistically significant in predicting D2D prices. At least in

[27] See, e.g., (Madhavan and Smidt, 1993) and (Dunne et al., 2010).

the FX market, the liquidity mismatch appears to be a better measure than price dispersion. We note, however, that liquidity mismatch cannot be computed in all markets as one needs to see both bid and ask prices at all times.

Although many papers study (both theoretically and empirically) prices and liquidity in the D2D segment of the FX market, to the best of our knowledge, there are no papers that study the joint price formation in the D2C and D2D segments. This is an important gap in our understanding of FX markets, especially given that the D2C trading volume is higher than the D2D volume (see, Moore et al. (2016)). Our paper seeks to fill this gap. Similar to Evans and Lyons (2002b), ours is a pure inventory-theoretic model.[28] However, it differs from that of Evans and Lyons (2002b) in several important dimensions. First, we assume that dealers are heterogeneous in their risk-bearing capacities. Second, we introduce strategic competition between dealers for order flow in the D2C market. Third, we assume that dealers are also strategic when trading in the D2D market. This assumption is crucial for our main results: It implies that (heterogeneous) dealers have (heterogeneous) price impact and, hence, the joint distribution of inventories and price impacts (as captured by the liquidity mismatch) matters for equilibrium prices and allocations. Finally, because we apply our model to very short horizons (up to ten seconds), we do not need to impose the zero inventory constraint. We believe that all of these new ingredients are important for correctly modeling real-world FX markets and allow us to capture new effects that have not previously been studied in the literature.

Our theoretical model is related to several existing models of market fragmentation. Dunne et al. (2015) develop a dynamic model of dealer intermediation between a monopolistic customer-dealer market with homogeneous dealers and a competitive inter-dealer market. Vogler (1997) develops a fragmented market model with homogeneous dealers. The inter-

[28] See, Bacchetta and Van Wincoop (2006), Evans et al. (2011), Evans and Lyons (2005c); Cao et al. (2006), Frankel et al. (2009), Lyons et al. (2001), King et al. (2010), Michaelides et al. (2018), and Gargano et al. (2018) for FX microstructure models of exchange rates that rely on private, heterogeneous information. Incorporating such informational asymmetries into our model is an important direction for future research.

dealer market in Vogler (1997) operates through the double auction protocol, which is similar to ours, but absent heterogeneity, the notion of liquidity mismatch does not arise. Most importantly, Vogler (1997) assumes Bertrand competition between dealers in the D2C market: Dealers quote a single price, and customers can both buy and sell unlimited amounts at this price.[29] Thus, counterfactually, there is no bid-ask spread in the D2C market. Summarizing, neither Vogler (1997) nor Dunne et al. (2015) can account for the key features of our model: (1) dealer heterogeneity; (2) imperfect competition (price impact) in the inter-dealer market; and (3) illiquid D2C markets. All of these features are crucial for our ability to compute the price-based liquidity mismatch from the observed cross-section of D2C prices and spreads.[30]

Colliard et al. (2018) study the effects of market fragmentation and dealer market power in which dealers differ in their connectivity. However, they deliberately keep the matching process between dealers and their clients simple to focus on the D2D market. While their market mechanisms are well-suited for many OTC markets, such as the bond market, ours could be better suited for modeling forex trading since the market structure is very different.

The closest to ours is the paper by Babus and Parlatore (2022), who developed a fragmented market model with identical (homogeneous) dealers and a D2D double auction market protocol as in our paper. In addition, they assume that each dealer runs a local D2C market and has a customer base that can only trade in this local market with one particular dealer. They show how market fragmentation can arise endogenously when customers endogenously decide upon the local market they want to join. Thus, in Babus and Parlatore (2022), conditional on the chosen market participation structure, competition between dealers is nonexistent. Although their model might apply to some real-world fragmented markets, the assumed market structure differs from the structure we observe in real-world FX markets. Yet, introducing dealer heterogeneity into their model would lead

[29] Dunne et al. (2015) also consider an extension of their model with the same trading protocol.

[30] Similarly, the models of Ho and Stoll (1983) and Biais (1993), related to our simple model abstract from dealer heterogeneity.

to dispersion in dealer-specific prices and spreads (price impacts) in the D2C market. Hence, their model might produce the joint behavior of D2D and D2C prices that we observe in our data. In Appendix IA.6, we extend the model of Babus and Parlatore (2022) to allow for dealer heterogeneity and then derive the equilibrium relationship between the price-based liquidity mismatch in the D2C market and the price level in the D2D market. Surprisingly, we find that this model implies the sign of a relationship that is opposite to the one that we observe in the data. By contrast, our model does indeed produce the empirically observed positive relationship between the liquidity mismatch and D2D prices. In Section IA.1.1, we explain how the nature of competition between dealers drives this difference in model predictions. Namely, in our model (as in the real world), dealers in the D2C market compete in mechanisms (price schedules; see Biais et al. (2000, 2013) and Back and Baruch (2013)). This competition implies that they use D2C markets to aggressively manage their inventories to the extent that their post-D2C inventories are negatively related to pre-D2C inventories, reverting the sign of the link between D2C and D2D markets.

Our notion of liquidity mismatch is indirectly related to the liquidity mismatch defined in Brunnermeier and Krishnamurthy (2012) (see also Bai et al. (2018)) as the mismatch between the market liquidity of assets and the funding liquidity of liabilities. It is natural to expect that the inventory cost in our model is closely related to the dealers' funding liquidity. Hence, the total inventory cost for a given dealer is linked to the liquidity mismatch of Brunnermeier and Krishnamurthy (2012), defined on a single bank level. However, in stark contrast to Brunnermeier and Krishnamurthy (2012), our liquidity mismatch measure is cross-sectional and captures a mismatch in the distribution of assets across different dealers with different liquidity needs and different price impacts. The fact that one can identify the mismatch directly from prices in the D2C segment is a surprising and novel prediction of our model.

Babus and Parlatore (2022) belongs to a larger stream of literature on fragmented markets (see, e.g., Babus and Kondor (2018), Malamud and Rostek (2017), and Babus and Parlatore (2022)) that assumes identical, auction-like trading protocols in the two market segments.[31] Under such protocols, effectively, dealers do not play any unique role, resulting in customers and dealers ending up equally providing liquidity to each other. We show that, contrary to our model, this behavior leads to an opposite (negative) sign of the relationship between liquidity mismatch and price level. Empirically, we find strong evidence for the positive relationship, suggesting that our modeling of D2C trading is crucial for matching the data.

Numerous papers provide evidence that order flow contains information about contemporaneous (Evans and Lyons (2002b), Evans and Lyons (2002a), Hau et al. (2002), Fan and Lyons (2003), Froot and Ramadorai (2005), Bjønnes et al. (2005), Danielsson and Love (2006), Killeen et al. (2006), Berger et al. (2008), Brunnermeier et al. (2008), King et al. (2010), Rime et al. (2010), Breedon and Vitale (2010), and Bjønnes et al. (2011)) and future (Evans and Lyons (2005b), Danielsson et al. (2012), and Evans and Rime (2016), Collin-Dufresne et al. (2019)) prices. In this paper, our focus is on predictive relationships at horizons that are short enough so that efficient allocation cannot be achieved due to market imperfections. The key insight from our model is that price dispersion in the D2C market can be used to recover information about the distribution of inventories and the liquidity mismatch. This novel, endogenous, purely price-based object is unique to our model.

Our model assumes that dealer inventories constitute an important driving force behind price dynamics. Recent empirical evidence supports this assumption. For instance, Friewald et al. (2019), Anderson and Liu (2019), and Randall (2015) show that inventory costs explain a significant fraction of yield spread changes in corporate bonds. Hendershott and Menkveld (2014) provide similar evidence for equity markets. Numerous papers have also shown the importance of the D2D market as a key venue for inventory management. See, for example, Schultz (2017) Schürhoff and Li (2019), Collin-Dufresne et al. (2019), Hollifield et al. (2017),

[31] Some papers (see, e.g., Liu et al. (2017)) also assume a competitive D2D market, implying that the allocation of risk among dealers is irrelevant.

and Anderson and Liu (2019). The most closely related to ours is the paper by Collin-Dufresne et al. (2019), who study two-tired CDS markets and show that bid-ask spreads are wider in the D2C market, and D2C transactions are largely institutional trades that have a large permanent price impact. By contrast, the FX market is relatively less concentrated and has a significant retail segment, which is captured in our data. We find that bid-ask spreads are narrower in the D2C market and have only a small, transitory price impact, consistent with our inventory-driven model.

Our main predictions stem from dealers' heterogeneity. Data strongly support the presence of such heterogeneity. For example, Evans (2002) found that most short-term volatility in exchange rates is a result of dealers' heterogeneous trading decisions. Similarly, Bjønnes and Rime (2005) document significant differences in dealers' trading styles, especially related to how they control their inventories. Randall (2015) provides evidence of heterogeneous and time-varying costs of holding inventory for dealers in the US corporate bond market. Our data on dealer quotes in the D2C market also suggests the presence of significant heterogeneity: Both the bid-ask spreads and the sensitivity of quotes to shocks are highly heterogeneous across dealers. Han et al. (2022) argue that dispersed beliefs represent another important source of dealer heterogeneity. While dealer disagreement is undoubtedly important, in our model, heterogeneous inventory generates a distinctive prediction about the link between prices in the two market segments that heterogeneous beliefs cannot generate. Investigating the joint role of inventories and beliefs in FX markets is an important direction for future research.

Most of our key predictions depend crucially on the fact that the FX market is not perfectly liquid, and dealers have a price impact. As a result, markets cannot efficiently allocate risk at short horizons, and the distribution of inventories across dealers impacts price dynamics. However, the idea that price impact is linked to order flow is not new (see, for example, Evans and Lyons (2005a)), to the best of our knowledge, our model is the first to micro-found this price impact in a model that accounts for the two-tier structure of the FX market. In particular, our model can be used to recover market liquidity from dealer quotes, providing an explicit, micro-founded measure of liquidity risk and shedding new light on the findings of Banti et al. (2012) and Mancini et al. (2013).

8 Conclusion

Our findings highlight the importance of a two-tiered market structure, dealer heterogeneity, market power, and non-exclusive customer-dealer relationships as important features of the FX market affecting its high-frequency dynamics. We have related D2D prices and spreads to the two first moments of the joint cross-sectional distribution of D2C prices and spreads. The following research can look at higher moments. For example, is the cross-sectional skewness of D2D spreads (or prices) an important determinant of high-frequency FX returns?

On the theory side, going deeper into a dealer's problem dynamics can be fruitful. Shall dealers pass on the liquidity shocks from customers to the D2D market immediately? Or shall they accumulate inventories, partially netting the shocks that arrived at different times? Our static model cannot answer such questions, so one must derive a dynamic model instead.

References

Anderson, Chris and Weiling Liu, "Intermediary Trading and Risk Constraints," Working paper 2019.

Babus, Ana and Cecilia Parlatore, "Strategic fragmented markets," *Journal of Financial Economics*, 2022, *145* (3), 876–908.

___ **and Peter Kondor**, "Trading and Information Diffusion in Over-the-Counter markets," *Econometrica*, 2018, *86* (5), 1727–1769.

Bacchetta, Philippe and Eric Van Wincoop, "Can information heterogeneity explain the exchange rate determination puzzle?," *The American Economic Review*, 2006, *96* (3), 552–576.

Back, Kerry and Shmuel Baruch, "Strategic liquidity provision in limit order markets," *Econometrica*, 2013, *81* (1), 363–392.

Bai, Jenny, Arvind Krishnamurthy, and Charles-Henri Weymuller, "Measuring Liquidity Mismatch in the Banking Sector," *Journal of Finance*, 2018, *73* (1), 51–93.

Banti, Chiara, Kate Phylaktis, and Lucio Sarno, "Global liquidity risk in the foreign exchange market," *Journal of International Money and Finance*, 2012, *31* (2), 267–291.

Berger, David W, Alain P Chaboud, Sergey V Chernenko, Edward Howorka, and Jonathan H Wright, "Order flow and exchange rate dynamics in electronic brokerage system data," *Journal of international Economics*, 2008, *75* (1), 93–109.

Bessembinder, Hendrik, Chester Spatt, and Kumar Venkataraman, "A survey of the microstructure of fixed-income markets," *Journal of Financial and Quantitative Analysis*, 2020, *55* (1), 1–45.

Biais, Bruno, "Price Formation and Equilibrium Liquidity in Fragmented and Centralized Markets," *The Journal of Finance*, 1993, *48*, 157–185.

___ **, David Martimort, and Jean-Charles Rochet**, "Competing mechanisms in a common value environment," *Econometrica*, 2000, *68* (4), 799–837.

___ **,** ___ **, and** ___ , "Competing mechanisms in a common value environment: A Corrigendum," *Econometrica*, 2013, *81* (4), 393–406.

Bjønnes, Geir H, Carol L Osler, Dagfinn Rime, and Norges Bank, "Sources of information advantage in the foreign exchange market," *Norges Bank*, 2011, p. 46.

Bjønnes, Geir Høidal and Dagfinn Rime, "Dealer behavior and trading systems in foreign exchange markets," *Journal of Financial Economics*, 2005, *75* (3), 571 – 605.

___ **,** ___ **, and Haakon O Aa Solheim**, "Liquidity provision in the overnight foreign exchange market," *Journal of International Money and Finance*, 2005, *24* (2), 175–196.

Breedon, Francis and Paolo Vitale, "An empirical study of portfolio-balance and information effects of order flow on exchange rates," *Journal of International Money and Finance*, 2010, *29* (3), 504–524.

Brunnermeier, Markus K, Stefan Nagel, and Lasse H Pedersen, "Carry trades and currency crashes," *NBER macroeconomics annual*, 2008, *23* (1), 313–348.

Cao, H Henry, Martin D Evans, Richard K Lyons et al., "Inventory Information," *The Journal of Business*, 2006, *79* (1), 325–364.

Cenedese, Gino, Pasquale Della Corte, and Tianyu Wang, "Currency mispricing and dealer balance sheets," *The Journal of Finance*, 2021, *76* (6), 2763–2803.

Chaboud, Alain, Sergey Chernenko, and Jonathan H Wright, "Trading activity and exchange rates in high-frequency EBS data," *FRB International Finance Discussion Paper*, 2007, (903).

Colliard, Jean-Edouard, Thierry Foucault, and Peter Hoffmann, "Inventory Management, Dealers' Connections and Prices in OTC Markets," *Working Paper*, 2018.

Collin-Dufresne, Pierre, Benjamin Junge, and Trolle anders B, "Market Structure and Transaction Costs of Index CDSs," *Journal of Finance*, 2019.

Danielsson, Jon and Ryan Love, "Feedback trading," *International Journal of Finance & Economics*, 2006, *11* (1), 35–53.

___ , Jinhui Luo, and Richard Payne, "Exchange rate determination and inter-market order flow effects," *The European Journal of Finance*, 2012, *18* (9), 823–840.

Duffie, Darrell, Nicolae Gârleanu, and Lasse Heje Pedersen, "Over-the-counter markets," *Econometrica*, 2005, *73* (6), 1815–1847.

Dunne, Peter G, Harald Hau, and Michael Moore, "A tale of two platforms: dealer intermediation in the European sovereign bond market," *Journal of the European Economic Association*, 2010.

Dunne, Peter, Harald Hau, and Michael Moore, "Dealer Intermediation Between Markets," *Journal of the European Economic Association*, 2015, *13* (5), 770–804.

Eisfeldt, Andrea L, Bernard Herskovic, and Shuo Liu, "Interdealer Price Dispersion," *Available at SSRN*, 2022.

___ , ___ , Sriram Rajan, and Emil Siriwardane, "OTC intermediaries," *The Review of Financial Studies*, 2023, *36* (2), 615–677.

Evans, Martin DD, "FX trading and exchange rate dynamics," *The Journal of Finance*, 2002, *57* (6), 2405–2447.

___ and Dagfinn Rime, "Order flow information and spot rate dynamics," *Journal of International Money and Finance*, 2016, *69*, 45–68.

___ and Richard K Lyons, "Informational integration and FX trading," *Journal of International Money and Finance*, 2002, *21* (6), 807–831.

___ and ___ , "Order flow and exchange rate dynamics," *Journal of political economy*, 2002, *110* (1), 170–180.

___ and ___ , "Do currency markets absorb news quickly?," *Journal of International Money and Finance*, 2005, *24* (2), 197–217.

___ and ___ , "Meese-Rogoff Redux: Micro-Based Exchange-Rate Forecasting," *American Economic Review*, 2005, pp. 405–414.

Evans, Martin D.D. and Richard K. Lyons, "A New Micro Model of Exchange Rate Dynamics," Working paper, Georgetown University 2005.

Evans, Martin DD et al., *Exchange-Rate Dynamics*, Princeton University Press, 2011.

Fan, Mintao and Richard K Lyons, "Customer trades and extreme events in foreign exchange," *Monetary History, Exchange Rates and Financial Markets: Essays in Honour of Charles Goodhart*, 2003, pp. 160–179.

Frankel, Jeffrey A, Giampaolo Galli, and Alberto Giovannini, *The microstructure of foreign exchange markets*, University of Chicago Press, 2009.

Friewald, Nils, , and Florian Nagler, "Over-the-Counter Market Frictions and Yield Spread Changes," *The Journal of Finance*, 2019, *74*, 3217–3257.

Froot, Kenneth A and Tarun Ramadorai, "Currency Returns, Intrinsic Value and Institutional-Investor Flows," *The Journal of Finance*, 2005, *60* (3), 1535–1566.

Gabaix, Xavier and Ralph SJ Koijen, "In search of the origins of financial fluctuations: The inelastic markets hypothesis," Technical Report, National Bureau of Economic Research 2021.

Gargano, Antonio, Steven J. Riddiough, and Lucio Sarno, "The Value of Volume in Foreign Exchange," Working Paper, University of Melbourne 2018.

Han, Brandon Yueyang, Sophia Zhengzi Li, and Zhaogang Song, "Dealer Disagreement and Asset Prices in FX Markets," *Available at SSRN*, 2022.

Hau, Harald, William Killeen, and Michael Moore, "The euro as an international currency: explaining puzzling first evidence from the foreign exchange markets," *Journal of International Money and Finance*, 2002, *21* (3), 351–383.

Hendershott, Terrence and Albert J. Menkveld, "Price pressures," *Journal of Financial Economics*, 2014, *114*, 405–423.

Ho, Thomas and Hans R Stoll, "Optimal dealer pricing under transactions and return uncertainty," *Journal of Financial economics*, 1981, *9* (1), 47–73.

Ho, Thomas SY and Hans R Stoll, "The dynamics of dealer markets under competition," *The Journal of finance*, 1983, *38* (4), 1053–1074.

Hollifield, Burton, Artem Neklyudov, and Chester Spatt, "Bid-ask spreads, trading networks and the pricing of securitizations," *Review of Financial Studies*, 2017, *30*, 3048–3085.

K., Gary Gorton Brunnermeier Markus and Arvind Krishnamurthy, "Risk topography," *NBER Macroeconomics Annual*, 2012, *26*, 149–176.

Killeen, William P, Richard K Lyons, and Michael J Moore, "Fixed versus flexible: Lessons from EMS order flow," *Journal of International Money and Finance*, 2006, *25* (4), 551–579.

King, Michael, Lucio Sarno, and Elvira Sojli, "Timing exchange rates using order flow: The case of the Loonie," *Journal of Banking & Finance*, 2010, *34* (12), 2917–2928.

Kyle, Albert S, "Continuous auctions and insider trading," *Econometrica: Journal of the Econometric Society*, 1985, pp. 1315–1335.

___ , "Informed speculation with imperfect competition," *The Review of Economic Studies*, 1989, *56* (3), 317–355.

Li, Dan and Norman Schürhoff, "Dealer networks," *The Journal of Finance*, 2019, *74* (1), 91–144.

Liu, Ying, Sebastian Vogel, and Yuan Zhang, "Electronic Trading in OTC Markets vs. Centralized Exchange," Technical Report, SFI 2017.

Lyons, Richard K, "Tests of microstructural hypotheses in the foreign exchange market," *Journal of Financial Economics*, 1995, *39* (2), 321–351.

___ **et al.**, *The Microstructure Approach to Exchange Rates*, Vol. 1, The MIT Press, 2001.

Madhavan, Ananth and Seymour Smidt, "An analysis of changes in specialist inventories and quotations," *The Journal of Finance*, 1993, *48* (5), 1595–1628.

Malamud, Semyon and Marzena Rostek, "Decentralizaed Exchange," *American Economic Review*, 2017, *107* (11), 3320–3362.

Mancini, Loriano, Angelo Ranaldo, and Jan Wrampelmeyer, "Liquidity in the foreign exchange market: Measurement, commonality and risk premiums," *The Journal of Finance*, 2013, *68* (5), 1805–1841.

Meese, Richard A and Kenneth Rogoff, "Empirical exchange rate models of the seventies: Do they fit out of sample?," *Journal of international economics*, 1983, *14* (1-2), 3–24.

Michaelides, Alexander, Andreas Milidonis, and George Nishiotis, "Private Information in Currency Markets," Technical Report, Journal of Financial Economics (forthcoiming) 2018.

Moore, Michael, Andreas Schrimpf, and Vladyslav Sushko, "Downsized FX markets: causes and implications," *BIS Quaterly Review*, December 2016.

Osler, Carol L, Alexander Mende, and Lukas Menkhoff, "Price discovery in currency markets," *Journal of International Money and Finance*, 2011, *30* (8), 1696–1718.

Randall, Oliver, "How Do Inventory Costs Affect Dealer Behavior in the US Corporate Bond Market?," *Working Paper*, 2015.

Reiss, Peter C and Ingrid M Werner, "Does risk sharing motivate interdealer trading?," *The Journal of Finance*, 1998, *53* (5), 1657–1703.

___ **and** ___ , "Anonymity, adverse selection, and the sorting of interdealer trades," *The Review of Financial Studies*, 2005, *18* (2), 599–636.

Rime, Dagfinn, Lucio Sarno, and Elvira Sojli, "Exchange rate forecasting, order flow and macroeconomic information," *Journal of International Economics*, 2010, *80* (1), 72–88.

Rostek, Marzena and Marek Weretka, "Dynamic thin markets," *The Review of Financial Studies*, 2015, *28* (10), 2946–2992.

Schultz, Paul, "Inventory management by corporate bond dealers," Working paper 2017.

Schürhoff, Norman and Dan Li, "Dealer Networks," *forthcming in the Journal of Finance*, 2019.

Vayanos, Dimitri, "Strategic trading and welfare in a dynamic market," *The Review of Economic Studies*, 1999, *66* (2), 219–254.

—— **and Jiang Wang**, "Market liquidity—theory and empirical evidence," in "Handbook of the Economics of Finance," Vol. 2, Elsevier, 2013, pp. 1289–1361.

—— **and Paul Woolley**, "An institutional theory of momentum and reversal," *The Review of Financial Studies*, 2013, *26* (5), 1087–1145.

Vives, Xavier, "Strategic supply function competition with private information," *Econometrica*, 2011, *79* (6), 1919–1966.

Vogler, Karl-Hubert, "Risk allocation and inter-dealer trading," *European Economic Review*, 1997, *41*, 1615–1634.

A Summary of notation

Notation	Explanation
	General mathematical notation
$E[x_l] = \frac{1}{M}\sum_l x_l$	Cross-sectional mean of a variable x_l
$Var[x_l] = \frac{1}{M}\sum_l (x_l - E[x_l])^2$	Cross-sectional variance of a variable x_l
$cov(x_l, y_l) = \frac{1}{M}\sum_l (x_l - E[x_l])(y_l - E[y_l])$	Cross-sectional covariance of a variables x_l and y_l
$\mathbb{E}[\tilde{q}]$	Expectation of a random variable $\tilde{q}$
$\mathbb{V}[\tilde{q}]$	Variance of a random variable $\tilde{q}$
$\mathbb{CV}[\tilde{q}] = \sqrt{\mathbb{V}[\tilde{q}]}/\mathbb{E}[\tilde{q}]$	Coefficient of variation of a random variable $\tilde{q}$
$cov(x_l, y_l) = \frac{1}{M}\sum_l (x_l - E[x_l])(y_l - E[y_l])$	Cross-sectional covariance of a variables x_l and y_l
	Model variables
$\tilde{q}_c$	customer order flow (signed)
$\tilde{q} = \lvert\tilde{q}_c\rvert$	customer order flow (non-signed)
Γ_l	Dealer l's risk aversion
M	Number of dealers
α_l	Dealer l's D2C miq-quote
b_l	Dealer l's D2C bid-ask spread
$\mathcal{P}$	D2D mid-price spread
$\mathcal{R}_t^{D2C} = \Delta E\left[\alpha_{l,t}\right]/E\left[\alpha_{l,t}\right]$	D2C return
$\mathcal{R}_{t+\ell}^{D2D} = \Delta\mathcal{P}_{t+\ell}^{D2D}/\mathcal{P}_t^{D2D}$	D2D return
$BA_{l,t}^{D2C} = b_{l,t}/E\left[\alpha_{l,t}\right]$	Percentage bid-ask spread in D2C
$BA_t^{D2D} = (\mathrm{Ask_t}^{D2D} - \mathrm{Bid_t}^{D2D})/\mathcal{P}_t^{D2D}$	Percentage bid-ask spread in D2D
$\widehat{\mathcal{A}}_{mismatch,t} \equiv cov(\alpha_{l,t}, BA_{l,t}^{D2C})$	(Price-based) liquidity mismatch
l and ℓ	l indexes dealers; ℓ denotes the lag value.

B Derivations and Proofs

B.1 Equilibrium in the D2D Market

This section uses the results in Malamud and Rostek (2017) to characterize the unique, robust linear Nash equilibrium in the D2D double auction game. Our D2D game is the same in the model and the extended models.

Given his asset holdings χ_l, dealer l's objective is to choose the trade size Q_l in the D2D market that maximizes his utility by choosing the optimal trade size Q_l. The equilibrium demand schedule $Q_l = Q_l(\chi_l, \mathcal{P}^{D2D})$ of dealer l equalizes his marginal utility with his marginal payment for each price,

$$d - \Gamma_l\left(\chi_l + Q_l\right) \; = \; \mathcal{P}^{D2D} + \beta_l Q_l, \tag{13}$$

where β_l measures the *price impact of dealer l in the D2D market* ("Kyle's lambda"; see Kyle (1985)). Formally, β_l is the derivative of the inverse residual supply of dealer l, which is defined by aggregation through the market clearing of the schedules submitted by other traders, $\{Q_\ell(\chi_\ell, \mathcal{P}^{D2D})\}_{\ell \neq l}$. It follows that the price elasticity of delaer l's demand is $1/(\beta_l + \Gamma_l)$ and the illiquiidty of D2D market is given by

$$\Lambda = 2/\sum_l \frac{1}{\beta_l + \Gamma_l}.$$

To pin down equilibrium price impacts, we note that the market clearing condition requires that the price impact assumed by dealer l be equal to the slope of his inverse residual supply, which results from the aggregation of the other dealers' submitted schedules. Proposition 4 (Proposition 1 in Malamud and Rostek (2017)) shows that the system for equilibrium price impacts can be solved explicitly.[32] Additionally, the proposition below derives some comparative statics.

Proposition 4 *There exists a unique D2D market equilibrium. The equilibrium price is given by*

$$\mathcal{P}^{D2D} = \bar{d} - \mathbf{Q}^* \textit{ with } \mathbf{Q}^* \; \equiv \; \mathcal{B}^{-1}\sum_{l=1}^{M}(\Gamma_l + \beta_l)^{-1}\Gamma_l\chi_l\,. \tag{14}$$

Trader l's price impact β_l is given by

$$\beta_l = \frac{2\Gamma_l}{\Gamma_l\mathcal{B} - 2 + \sqrt{(\Gamma_l\mathcal{B})^2 + 4}}\,,$$

[32] For symmetric risk aversions, the equilibrium of Proposition 4 coincides with the equilibrium in Rostek and Weretka (2011), which in turn coincides with Kyle (1989), without nonstrategic traders and assuming independent values). The case of symmetric risk aversions has also been studied in Vayanos (1999), and Vives (2011).

where $\mathcal{B} \in \mathbb{R}_+$ is the unique positive solution to $\sum_l (\Gamma_l \mathcal{B} \ + 2 + \sqrt{(\Gamma_l \mathcal{B})^2 + 4})^{-1} \ = \ 1/2$. The illiquidity is given by $\Lambda = 2/\mathcal{B}$. Moreover: (i) the price impact β_l is cross-sectionally monotone decreasing in Γ_l, i.e., if $\Gamma_{l_1} > \Gamma_{l_2}$, then $\beta_{l_1} < \beta_{l_2}$; and (ii) β_l is monotone increasing Γ_ℓ for any $\ell \neq l$: That is, an increase in risk aversion of any trader worsens liquidity for all other traders. The equilibrium post-D2D trade inventory of dealer l is

$$\tilde{\chi}_l \ = \ (\Gamma_l + \beta_l)^{-1} \mathbf{Q}^* + (\Gamma_l + \beta_l)^{-1} \beta_l \chi_l \,. \tag{15}$$

Dealers' equilibrium indirect utility is given by

$$\mathcal{U}_l \ = \ E\left[\chi_l d - 0.5 \Gamma_l \chi_l^2 \ + \ (0.5\Gamma_l + \beta_l)(\Gamma_l + \beta_l)^{-2} (\mathbf{Q}^* - \Gamma_l \chi_l)^2\right] . \tag{16}$$

Proof of Proposition 4. The results regarding the dealers' price impact (β_ℓ) and $\mathcal{B}$, except (ii), are given in Malamud and Rostek (2017). We now prove (ii) : β_l is monotone increasing Γ_ℓ for any $\ell \neq l$. Fix l. It follows from differentiation that

$$\beta_l(\mathcal{B}) \ = \ \frac{2\Gamma_l}{\Gamma_l \mathcal{B} - 2 + \sqrt{(\Gamma_l \mathcal{B})^2 + 4}}$$

is increasing in $\mathcal{B}$. Let $\mathbf{\Gamma} \ = \ \{\Gamma_\ell\}_{\ell=1}^M$ be a set of dealers' risk-aversion and suppose that $\tilde{\mathbf{\Gamma}} = \left\{\tilde{\Gamma}_\ell\right\}_{\ell=1}^M$ is such that

$$\tilde{\Gamma}_l \ > \ \Gamma_l \quad \text{and} \quad \tilde{\Gamma}_\ell \ = \ \Gamma_\ell \quad \forall \ \ell \ \neq \ l \,.$$

Since $\beta_l(\mathcal{B})$ is increasing, it is enough to show that the corresponding $\mathcal{B}$ and $\tilde{\mathcal{B}}$ satisfy

$$\tilde{\mathcal{B}} \ < \ \mathcal{B}$$

to prove the statement. By definition,

$$\begin{aligned} \frac{1}{2} &= \ \sum_\ell \frac{1}{\Gamma_\ell \mathcal{B} \ + 2 + \sqrt{(\Gamma_\ell \mathcal{B})^2 + 4}} \ + \ \frac{1}{\Gamma_l \mathcal{B} \ + 2 + \sqrt{(\Gamma_l \mathcal{B})^2 + 4}} \\ &> \ \sum_\ell \frac{1}{\tilde{\Gamma}_\ell \tilde{\mathcal{B}} \ + 2 + \sqrt{(\tilde{\Gamma}_\ell \tilde{\mathcal{B}})^2 + 4}} \,. \end{aligned}$$

It follows that $\tilde{\mathcal{B}} \ < \ \mathcal{B}$ since

$$\sum_\ell \frac{1}{\tilde{\Gamma}_\ell \tilde{\mathcal{B}} \ + 2 + \sqrt{(\tilde{\Gamma}_\ell \tilde{\mathcal{B}})^2 + 4}} \ = \ \frac{1}{2} \,.$$

We finish by proving the remaining parts of the proposition. First, we prove that $\mathcal{B}$ is the aggregate liquidity, that is

$$\mathcal{B} \;=\; \sum_{\ell} \frac{1}{\Gamma_\ell + \beta_\ell}$$

We have

$$\begin{aligned}(\beta_\ell + \Gamma_\ell)^{-1} = \; & \frac{1}{\Gamma_\ell}\left[1 - \frac{2}{\Gamma_\ell \mathcal{B} + \sqrt{(\Gamma_\ell \mathcal{B})^2 + 4}}\right] \;=\; \frac{1}{2\Gamma_\ell}\left[2 + \Gamma_\ell \mathcal{B} - \sqrt{(\Gamma_\ell \mathcal{B})^2 + 4}\right] \\ = \; & 2\left[\frac{\mathcal{B}}{2 + \Gamma_\ell \mathcal{B} - \sqrt{(\Gamma_\ell \mathcal{B})^2 + 4}}\right].\end{aligned}$$

Summing over ℓ then yields the result.

Next, we multiply Equation (13) by $(\Gamma_l + \beta_l)^{-1}$ and rearrange the terms to obtain

$$(\Gamma_l + \beta_l)^{-1}(\mathcal{P}^{D2D} - d) \;=\; -(\Gamma_l + \beta_l)^{-1}\Gamma_l \chi_l - Q_l,$$

We can then sum over ℓ and use Equation B.1 (which shows that $\mathcal{B}$ is equal to aggregate liquidity) to obtain

$$\mathcal{P}^{D2D} = \bar{d} - \mathbf{Q}^* - \sum_{l=1}^{M} Q_\ell\,.$$

Equation 14 then follows from market clearing. We finish the proof by plugging

$$Q_\ell \;=\; (\Gamma_\ell + \beta_\ell)^{-1}(d - \mathcal{P}^{D2D} - \Gamma_\ell \chi_\ell) \;=\; (\Gamma_\ell + \beta_\ell)^{-1}\mathbf{Q}^* - (\Gamma_\ell + \beta_\ell)^{-1}\Gamma_\ell \chi_\ell$$

into $\tilde{\chi}_\ell = \chi_\ell + Q_\ell$ to obtain Equations (15) and (16). Q.E.D.

B.2 Derivation of equation (8)

As with the derivation of (7), we assume that (i) illiquidity in D2D market is consistent with Proposition 2 at any time and (ii) a period of the model corresponds to ℓ seconds in the data. Then,

$$\Lambda_{t+\ell} = \beta_0^\Lambda + \beta_1^\Lambda E[b_{t,l}] - \beta_2^\Lambda \mathrm{Var}[b_{t,l}] + o(\epsilon).$$

Dividing by $\mathcal{P}_t^{D2D}$ and noting that $\mathcal{P}_{t+\ell}^{D2D} \approx \mathcal{P}_t^{D2D} \approx E[\alpha_{l_t}]$ (see the discussion in Section 5.1) we rewrite the last expression as

$$\Lambda_{t+\ell}^{\%} = \beta_0^{\Lambda^{\%}} + \beta_1^\Lambda E[BA_{t,l}^{D2C}] - \beta_2^{\Lambda^{\%}} \mathrm{Var}[BA_{t,l}^{D2C}] + o(\epsilon).$$

Here

$$\beta_0^{\Lambda\%} \equiv \beta_0^{\Lambda}/\mathcal{P}_t^{D2D}$$
$$= \frac{(\varkappa_0^b - \varkappa_0^a)}{\mathbb{E}[q^2]\varkappa}\frac{2(M-1)}{(M-2)M}.$$

We used the fact that $\varkappa_0^a = \mathcal{P}_t^{D2D}k_0^a$, $\varkappa_0^b = \mathcal{P}_t^{D2D}k_0^b$, and $\varkappa = k$.

Similarly,

$$\beta_1^{\Lambda} = \phi_1\frac{\Lambda_0}{\Gamma^*} > 0$$
$$= \phi_1\frac{2(M-1)}{(M-2)M}.$$

Finally,

$$\beta_2^{\Lambda\%} \equiv \beta_2^{\Lambda}\cdot\mathcal{P}_t^{D2D}$$
$$= \mathcal{E}_0\phi_1^2\frac{2(M-1)^2((M-1)M+1)}{M^2((M-2)M+2)^2}.$$

Here we denoted the leading term of elasticity of the D2D market by $\mathcal{E}_0$. Indeed, the elasticity can be written as

$$\mathcal{E} = \frac{\mathcal{P}_t^{D2D}}{\mathbb{E}[\tilde{q}]}\frac{dQ_t^{D2D}}{d\mathcal{P}_t^{D2D}}$$
$$= \frac{2\mathcal{P}_t^{D2D}}{\Lambda_t}$$
$$= \frac{2\mathcal{P}_t^{D2D}}{\Lambda_0} + O(\epsilon)$$
$$= \mathcal{E}_0 + O(\epsilon).$$

In the second line, we accounted for the fact that $\mathbb{E}[\tilde{q}]$ is normalized to one. We approximated Λ with Λ_0 in the third line.

Invoking Assumption 2 we obtain

$$\beta_0^{BA} = h_0 + h_1\frac{(\varkappa_0^b - \varkappa_0^a)}{\mathbb{E}[q^2]\varkappa}\frac{2(M-1)}{(M-2)M},$$
$$\beta_1^{BA} = h_1\phi_1\frac{2(M-1)}{(M-2)M} > 0, \text{ and} \tag{17}$$
$$\beta_2^{BA} = h_1\phi_1^2\mathcal{E}_0\frac{2(M-1)^2((M-1)M+1)}{M^2((M-2)M+2)^2} > 0. \tag{18}$$

B.3 Proof of Proposition 1

Proof of Proposition 1. The closed-form expression for $\mathcal{P}^{D2D}$ can be written as (see Section B.1)

$$\mathcal{P}^{D2D} = \bar{d} - \sum_l \frac{\frac{\Lambda}{2}}{\Gamma_l + \beta_l} \Gamma_l \chi_l.$$

Substituting the expressions for β_l it can be further rewritten as

$$\mathcal{P}^{D2D} = \bar{d} - \sum_l \underbrace{\frac{\Lambda}{2} \frac{\Gamma_l - \Lambda + \sqrt{\Gamma_l^2 + \Lambda^2}}{\Gamma_l + \sqrt{\Gamma_l^2 + \Lambda^2}}}_{\equiv f_l(\overrightarrow{\Gamma})} \chi_l$$

Define f_l as follows

$$f_l \;=\; \frac{\Lambda}{2} \frac{\Gamma_l - \Lambda + \sqrt{\Gamma_l^2 + \Lambda^2}}{\Gamma_l + \sqrt{\Gamma_l^2 + \Lambda^2}}.$$

Now Taylor expand $f_l(\overrightarrow{\Gamma})$ around $\overrightarrow{\Gamma} = \Gamma_* \overrightarrow{1}$. We keep fixed χ_l.

$$f_l(\overrightarrow{\Gamma}) = f_l(\Gamma_* \overrightarrow{1}) + \frac{\partial f_l}{\partial \Gamma_l} \left(\Gamma_* \overrightarrow{1} \right) (\Gamma_l - \Gamma_*) + \sum_{k \neq l} \frac{\partial f_l}{\partial \Gamma_k} \left(\Gamma_* \overrightarrow{1} \right) (\Gamma_k - \Gamma_*) + O\left(||\Gamma_l - \Gamma^*||^2 \right)$$

In equilibrium $\overrightarrow{\chi}$ depends on $\overrightarrow{\Gamma}$ and $\overrightarrow{x}, \overrightarrow{\chi} = \overrightarrow{\chi}_{eq}\left(\overrightarrow{x}, \overrightarrow{\Gamma} \right)$, but the expression above is still true for $\overrightarrow{\chi} = \overrightarrow{\chi}_{eq}\left(\overrightarrow{x}, \overrightarrow{\Gamma} \right)$. In our next step we'll augment the asymptotics just described by also expanding $\overrightarrow{\chi}_{eq}\left(\overrightarrow{x}, \overrightarrow{\Gamma} \right)$.

Zero-order term:

$$\begin{aligned}
\sum_l f_l(\Gamma_* \overrightarrow{1}) \chi_l &= \sum_l \frac{\Gamma_*}{m} \chi_l \\
&= \Gamma_* E[\chi_l] \\
&= \Gamma_* E[x_l] + \Gamma_* E[q_l]
\end{aligned}$$

where q_l denotes the equilibrium trade of the l-th dealer in the D2C market.

In our model, the inventory of one dealer will increase by $\tilde{q}$ and the inventory of another one will decrease by (another realization of) $\tilde{q}$. Due to this symmetry, we get $E[q_l] = 0$ after averaging across the realizations of $\tilde{q}$. Proceeding further, we write all expressions after computing $\mathbb{E}_{\tilde{q}}[\cdot]$. We have

$$\sum_l f_l(\Gamma_* \overrightarrow{1})\chi_l = \Gamma_* E[x_l]$$

Denote

$$\epsilon = \max\{||x_l||, ||\Gamma_l - \Gamma^*||\}.$$

Proceed to first-order term

$$\frac{\partial f_l}{\partial \Gamma_l}\left(\Gamma_* \overrightarrow{1}\right) = \frac{M(2M-3)+2}{M^2((M-2)M+2)} \equiv \tilde{c}_1(M)$$

$$\frac{\partial f_l}{\partial \Gamma_k}\left(\Gamma_* \overrightarrow{1}\right) = \frac{(M-2)(M-1)}{M^2((M-2)M+2)} \equiv \tilde{c}_2(M)$$

$$\frac{\partial f_l}{\partial \Gamma_l}\left(\Gamma_* \overrightarrow{1}\right)(\Gamma_l - \Gamma_*) + \sum_{k \neq l} \frac{\partial f_l}{\partial \Gamma_k}\left(\Gamma_* \overrightarrow{1}\right)(\Gamma_k - \Gamma_*) =$$
$$\left(\tilde{c}_1(M) - \tilde{c}_2(M)\right)(\Gamma_l - \Gamma_*) + \tilde{c}_2(M)\sum_k (\Gamma_k - \Gamma_*)$$

Denote

$$c_1(M) \equiv M\left(\tilde{c}_1(M) - \tilde{c}_2(M)\right) = \frac{M}{(M-2)M+2}$$

Note that

$$\sum_k (\Gamma_k - \Gamma_*) = ME[\Gamma_l - \Gamma_*].$$

The first-order term becomes

$$c_1 \frac{1}{M}\sum_l (\Gamma_l - \Gamma_*)\chi_l + \tilde{c}_2(M)ME[\Gamma_l - \Gamma_*]\sum_l \chi_l =$$
$$\overbrace{c_1 \frac{1}{M}\sum_l (\Gamma_l - \Gamma_*)x_l}^{O(\epsilon^2)} + c_1 \frac{1}{M}\sum_l (\Gamma_l - \Gamma_*)q_l + \overbrace{\tilde{c}_2(M)ME[\Gamma_l - \Gamma_*]\sum_l \chi_l}^{O(\epsilon^2)} =$$
$$c_1(M)cov(\Gamma_l, q_l) + O(\epsilon^2).$$

Combining everything together, we obtain

$$\mathcal{P}^{D2D} = \bar{d} - \Gamma_* E[x_l] - c_1(M)cov(q_l, \Gamma_l) + O(\epsilon^2).$$

Now note that

$$\begin{aligned} E[\frac{r_l^a + r_l^b}{2}] &= \bar{d} - E[\Gamma_l x_l] \\ &= \bar{d} - \underbrace{E[(\Gamma_l - \Gamma_*)\, x_l]}_{O(\epsilon^2)} - \Gamma_* E[x_l] \\ &= \bar{d} - \Gamma_* E[x_l] + O(\epsilon^2) \end{aligned}$$

Thus we have

$$\begin{aligned} \bar{d} - \Gamma_* E[x_l] &= E[\frac{r_l^a + r_l^b}{2}] + O(\epsilon^2) \\ &= \frac{1}{k} E\left[\alpha_l\right] - \frac{k_0^a + k_0^b}{k} + O(\epsilon^2). \end{aligned}$$

And so the expression for $\mathcal{P}^{D2D}$ becomes:

$$\mathcal{P}^{D2D} = -\frac{k_0^a + k_0^b}{k} + \frac{1}{k} E\left[\alpha_l\right] - c_1(M)cov(q_l, \Gamma_l) + O(\epsilon^2).$$

Recall that $\Gamma_l = \phi_o + \phi_1 b_l$, Then,

$$cov(q_l, \Gamma_l) = \phi_1 cov(q_l, b_l).$$

Finally, we Taylor expand q_l:

$$\begin{aligned} q_l &= \mathbb{E}_{\tilde{q}}\left[\tilde{q}\psi(\text{bid}_l - \text{best bid}) - \tilde{q}\psi\left(\text{best ask} - \text{ask}_l\right)\right] \\ &= 1/2\mathbb{E}_{\tilde{q}}[\tilde{q}]\psi(r_l - \bar{r}) - 1/2\mathbb{E}_{\tilde{q}}[\tilde{q}]\psi\left(\underline{r} - r_l\right) \\ &= \frac{1}{2}\mathbb{E}_{\tilde{q}}[\tilde{q}]\psi'(0)\left(2r_l - \underline{r} - \bar{r}\right) + O(\epsilon^2) \end{aligned}$$

where $r_l = \frac{r_l^a + r_l^b}{2}$ and $\bar{r} = \max_l r_l$. Then,

$$\begin{aligned} cov(q_l, b_l) &= \mathbb{E}_{\tilde{q}}[\tilde{q}]\psi'(0)cov(r_l, b_l) + O(\epsilon^2) \\ &= \frac{1}{k}\mathbb{E}_{\tilde{q}}[\tilde{q}]\psi'(0)cov(\alpha_l, b_l) + O(\epsilon^2) \end{aligned}$$

The last transition is true because $\alpha_l = \frac{k_0^a + k_0^b}{2} + k\frac{r_l^a + r_l^b}{2}$. We therefore obtain

$$\mathcal{P}^{D2D} = -\frac{k_0^a + k_0^b}{k} + \frac{1}{k} E\left[\alpha_l\right] - c_1(M)\frac{\phi_1}{k}\mathbb{E}[\tilde{q}]cov(\alpha_l, b_l) + o(\epsilon).$$

Q.E.D.

B.4 Proof of Proposition 2

Proof. Recall our expression for $\mathcal{B}$ (see Section B.1):

$$\sum_l \left(\Gamma_l \mathcal{B} + 2 + \sqrt{(\Gamma_l \mathcal{B})^2 + 4}\right)^{-1} = 1/2.$$

Multiply by $\mathcal{B}$ and denote $\Lambda = 2/\mathcal{B}$ (our current definition of bid-ask spread in D2D).

$$\sum_l \frac{1}{\Gamma_l/\Lambda + 1 + \sqrt{(\Gamma_l/\Lambda)^2 + 1}} = 1.$$

The last equation defines implicitly $\Lambda(\overrightarrow{\Gamma})$. Denote

$$f(x) = \frac{1}{x + 1 + \sqrt{x^2 + 1}}.$$

Denote $\Lambda_{kl} = \partial^2 \Lambda / (\partial \Gamma_k \partial \Gamma_l)$ and similarly for the first derivatives. Write second-order Taylor expansion for the last equation around $\overrightarrow{\Gamma} = \Gamma^* \overrightarrow{1}$:

$$\Lambda(\overrightarrow{\Gamma}) = \Lambda(\Gamma^* \overrightarrow{1}) + \sum_l \Lambda_l(\Gamma^* \overrightarrow{1})\,(\Gamma_l - \Gamma^*) + \\ + \frac{1}{2} \sum_{k,l} \Lambda_{k,l}(\Gamma^* \overrightarrow{1})\,(\Gamma_l - \Gamma^*)\,(\Gamma_k - \Gamma^*) + o\left(||\Gamma_l - \Gamma^*||^2\right).$$

Now compute $\Lambda(\Gamma^* \overrightarrow{1})$, $\Lambda_l(\Gamma^* \overrightarrow{1})$ and $\Lambda_{k,l}(\Gamma^* \overrightarrow{1})$.
$\Lambda(\Gamma^* \overrightarrow{1})$ solves $\frac{M}{\Gamma^*/\Lambda + 1 + \sqrt{(\Gamma^*/\Lambda)^2 + 1}} = 1$. We obtain

$$\Lambda(\Gamma^* \overrightarrow{1}) \equiv \Lambda_0 = \frac{2\Gamma_*(M-1)}{(M-2)M}. \tag{19}$$

To compute $\Lambda_l(\Gamma^* \overrightarrow{1})$ we differentiate implicitly $\frac{\partial}{\partial \Gamma_l} \sum_k f(\Gamma_k / \Lambda\left(\overrightarrow{\Gamma}\right)) = 0$. We get

$$\Lambda_l(\Gamma^* \overrightarrow{1}) = \frac{\Lambda_0}{\Gamma^* M}.$$

Similarly, we obtain the second derivatives via implicit differentiation.

$$\Lambda_{ll}(\Gamma^* \overrightarrow{1}) \equiv \Lambda_2^d = \frac{\Lambda_0(M-1)\left(\Lambda_0^2\left(\sqrt{\frac{\Gamma_*^2}{\Lambda_0^2} + 1} - 1\right) - 2\Gamma_*^2\right)}{\Gamma_*^2 M^2 \left(\Gamma_*^2 + \Lambda_0^2\right)},$$

$$\Lambda_{lk} \equiv \Lambda_2^{od} = \frac{\Lambda_0^3\left(1-\sqrt{\frac{\Gamma_*^2}{\Lambda_0^2}+1}\right)+2\Gamma^2\Lambda_0}{\Gamma_*^2 M^2\left(\Gamma_*^2+\Lambda_0^2\right)}.$$

Now plug eveything back to the Taylor. Start with linear term

$$\begin{aligned}
\Lambda(\Gamma^*\overrightarrow{1})+\sum_l \Lambda_l(\Gamma^*\overrightarrow{1})\left(\Gamma_l-\Gamma^*\right) &= \Lambda_0+\frac{\Lambda_0}{\Gamma^* M}\sum_l\left(\Gamma_l-\Gamma^*\right)\\
&= \frac{\Lambda_0}{\Gamma^*}\frac{1}{M}\sum_l \Gamma_l\\
&= \frac{\Lambda_0}{\Gamma^*}E[\Gamma_l].
\end{aligned}$$

Carry on to quadratic

$$\sum_{k,l}\Lambda_{k,l}(\Gamma^*\overrightarrow{1})\left(\Gamma_l-\Gamma^*\right)\left(\Gamma_k-\Gamma^*\right)=\left(\Lambda_2^d-\Lambda_2^{od}\right)\sum_l\left(\Gamma_l-\Gamma^*\right)^2+\Lambda_2^{od}\sum_{k,l}\left(\Gamma_l-\Gamma^*\right)\left(\Gamma_k-\Gamma^*\right).$$

Now note

$$\begin{aligned}
\sum_l\left(\Gamma_l-\Gamma^*\right)^2 &= ME\left[\left(\Gamma_l-E[\Gamma_l]+\underbrace{E[\Gamma_l]-\Gamma^*}_{\equiv\epsilon_1}\right)^2\right]\\
&= M\left(Var[\Gamma_l]+2\epsilon_1\underbrace{E\left[\Gamma_l-E[\Gamma_l]\right]}_{=0}+\epsilon_1^2\right)\\
&= MVar[\Gamma_l]+M\epsilon_1^2
\end{aligned}$$

$$\begin{aligned}
\sum_{k,l}\left(\Gamma_l-\Gamma^*\right)\left(\Gamma_k-\Gamma^*\right) &= \sum_k\left(\Gamma_k-\Gamma^*\right)\sum_l\left(\Gamma_l-E[\Gamma_l]+\epsilon_1\right)\\
&= M\epsilon_1\sum_k\left(\Gamma_k-\Gamma^*\right)\\
&= M^2\epsilon_1^2
\end{aligned}$$

Combining everything we get

$$\Lambda(\overrightarrow{\Gamma})=\frac{\Lambda_0}{\Gamma^*}E[\Gamma_l]+\frac{1}{2}\left(\Lambda_2^d-\Lambda_2^{od}\right)MVar[\Gamma_l]+o\left(||\Gamma_l-\Gamma^*||^2\right)+\epsilon_1^2\cdot const$$

where

$$\Lambda_0=\frac{2\Gamma_*(M-1)}{(M-2)M},$$

$$\left(\Lambda_2^d - \Lambda_2^{od}\right) M = -\frac{\Lambda_0^3\left(1-\sqrt{\frac{\Gamma_*^2}{\Lambda_0^2}+1}\right)+2\Gamma_*^2\Lambda_0}{\Gamma_*^2\left(\Gamma_*^2+\Lambda_0^2\right)},$$

And the last two terms can be combined into $o(\epsilon)$. The final step is to note that from (3) we have: $\Gamma_l = \phi_0 + \phi_1 b_l$ with $\phi_0 = -\frac{\mathbb{E}[q](k_0^a - k_0^b)}{\mathbb{E}[q^2]k}$ and $\phi_1 = \frac{2\mathbb{E}[q]}{\mathbb{E}[q^2]k}$. Therefore, we get

$$\Lambda(\overrightarrow{\Gamma}) = \phi_0 \frac{\Lambda_0}{\Gamma^*} + \phi_1 \frac{\Lambda_0}{\Gamma^*} E[b_l] + \frac{1}{2}\left(\Lambda_2^d - \Lambda_2^{od}\right) M \phi_1^2 Var[b_l] + o(\epsilon).$$

Q.E.D.

B.5 Proof of Proposition 3

Proof of Proposition 3. Dividing (18) by (17) we obtain

$$\phi_1 \mathcal{E}_0 \frac{(M-1)(M-2)(M(M-1)+1)}{M((M-2)M+2)^2} = \frac{\check{\beta}_2^{BA}}{\check{\beta}_1^{BA}}.$$

Substituting $\phi_1 = \frac{2}{\mathbb{E}[q^2]\varkappa} = 2/((\check{\mathbb{C}\mathbb{V}}[\tilde{q}]^2 + 1)\varkappa)$ and $\check{\beta}_1^{p} = 1/\varkappa$ we obtain the desired result. Q.E.D.

C Summary Statistics and Regression Tables

C.1 Summary statistics

Table 4: Summary statistics

Variable	mean	sd	min	q05	q50	q95	max
$E[\alpha_l]$	1.1187	0.0096	1.0952	1.1012	1.1186	1.1349	1.1416
$\mathcal{P}^{D2D}$	1.1187	0.0096	1.0952	1.1012	1.1186	1.1349	1.1416
$10^4 \cdot BA^{D2D}$	0.4919	0.3640	0.0091	0.2500	0.5000	1	15
$10^4 \cdot E[b_l]$	0.4785	0.6753	0.1875	0.2938	0.4111	0.8357	66.70
$10^4 \cdot \text{std}[b_l]$	0.2824	1.783	0.0378	0.1050	0.2138	0.4899	178.7
$10^8 \cdot \mathcal{A}_{mismatch}$	0.6718	96.59	-5,605	-0.0671	0.0003	0.0721	20,711
$ProxyLiqShock$ (\$mln)	0.0001	0.1638	-60.00	-0.0010	0	0.0010	42.58
$ProxyLiqShockSell$ (\$mln)	0.0896	0.4788	0.000001	0.0010	0.0090	0.3440	42.58
$ProxyLiqShockBuy$ (\$mln)	0.0938	0.4976	0.000001	0.0010	0.0100	0.3860	60.00
OF^{D2D} (\$mln)	54.01	138.0	-464	-185	40	299	388

Table 5: Correlation Table

$E[\alpha_l]$	$\mathcal{P}^{D2D}$	BA^{D2D}	$E[b_l]$	$\text{std}[b_l]$	$\mathcal{A}_{mismatch}$	$ProxyLiqShock$	OF^{D2D}
1.000	1.000***	-0.051***	0.057***	-0.000	-0.001	0.008***	0.054***
1.000***	1.000	-0.051***	0.056***	-0.002***	-0.003***	0.008***	0.054***
-0.051***	-0.051***	1.000	0.340***	0.101***	0.004***	-0.000	-0.031***
0.057***	0.056***	0.340***	1.000	0.910***	0.539***	-0.000	-0.005***
-0.000	-0.002***	0.101***	0.910***	1.000	0.651***	-0.000	-0.003***
-0.001	-0.003***	0.004***	0.539***	0.651***	1.000	-0.000	0.002***
0.008***	0.008***	-0.000	-0.000	-0.000	-0.000	1.000	-0.005***
0.054***	0.054***	-0.031***	-0.005***	-0.003***	0.002***	-0.005***	1.000

C.2 Returns

Table 6: Forecasting FX rates: Ten Seconds Ahead.— The table reports estimates for the regression

$$\mathcal{R}^{D2D}_{t+\ell} = \beta^p_1 \mathcal{R}^{D2C}_t - \beta^p_2 \Delta\widehat{\mathcal{A}}_{mismatch,t} + \text{controls}_t + \text{Constant}\,,$$

where $\Delta X_t = X_t - X_{t-\ell}$ and $\ell = 10s$. Heteroscedasticity and autocorrelation robust standard errors are shown in parenthesis.

	Dependent variable:			
	$\mathcal{R}^{D2D}_{t+\ell}$			
	(1)	(2)	(3)	(4)
$\mathcal{R}^{D2C}_t$	0.1847***	0.1849***	0.1847***	0.1849***
	(0.0326)	(0.0326)	(0.0322)	(0.0322)
$\Delta\widehat{\mathcal{A}}_{mismatch,t}$	4.4862***	4.4909***	4.4863***	4.4909***
	(0.8055)	(0.8060)	(0.7963)	(0.7968)
$ProxyLiqShock_t$		−0.000003***		−0.000003***
		(0.000000)		(0.000000)
OF^{D2D}_t			−0.0000	−0.0000
			(0.0000)	(0.0000)
Constant	0.0000	0.0000	0.0000	0.000000
	(0.000000)	(0.000000)	(0.000000)	(0.000000)
Observations	6,061,553	6,061,553	6,061,553	6,061,553
R^2	0.0320	0.0321	0.0320	0.0321
Adjusted R^2	0.0320	0.0321	0.0320	0.0321
Residual Std. Error	0.0001	0.0001	0.0001	0.0001
	(df = 6061550)	(df = 6061549)	(df = 6061549)	(df = 6061548)
F Statistic	100,167.2000***	66,919.5300***	66,778.4700***	50,189.9100***
	(df = 2; 6061550)	(df = 3; 6061549)	(df = 3; 6061549)	(df = 4; 6061548)
Note:				*p<0.1; **p<0.05; ***p<0.01

C.3 Spreads

Table 7: Forecasting FX Spreads: Ten Seconds Ahead.— The table reports estimates for the regression

$$BA^{D2D}_{t+\ell} = \beta_1^{BA} E[BA^{D2C}_{l,t}] - \beta_2^{BA}\mathrm{Var}[BA^{D2C}_{l,t}] + \mathrm{controls}_t + \mathrm{Constant}\,,$$

where $\Delta X_t = X_t - X_{t-\ell}$ and $\ell = 10s$. Heteroscedasticity and autocorrelation robust standard errors are shown in parenthesis.

	Dependent variable:			
	$BA^{D2D}_{t+\ell}$			
	(1)	(2)	(3)	(4)
$E[BA^{D2C}_{l,t}]$	0.5445***	0.5445***	0.5442***	0.5442***
	(0.0126)	(0.0126)	(0.0126)	(0.0127)
$\mathrm{Var}[BA^{D2C}_{l,t}]$	13.4702***	13.4702***	13.4601***	13.4601***
	(0.3227)	(0.3247)	(0.3240)	(0.3260)
$ProxyLiqShock_t$		−0.000000		−0.000000
		(0.000000)		(0.000000)
OF^{D2D}_t			−0.0000***	−0.0000***
			(0.0000)	(0.0000)
Constant	0.00001***	0.00001***	0.00001***	0.00001***
	(0.000001)	(0.000001)	(0.000001)	(0.000001)
Observations	6,149,128	6,149,128	6,149,128	6,149,128
R^2	0.3060	0.3060	0.3067	0.3067
Adjusted R^2	0.3060	0.3060	0.3067	0.3067
Residual Std. Error	0.00003	0.00003	0.00003	0.00003
	(df = 6149125)	(df = 6149124)	(df = 6149124)	(df = 6149123)
F Statistic	1,355,530.0000***	903,687.0000***	906,710.8000***	680,033.2000***
	(df = 2; 6149125)	(df = 3; 6149124)	(df = 3; 6149124)	(df = 4; 6149123)
Note:				*p<0.1; **p<0.05; ***p<0.01

D Additional Regression Tables

D.1 Returns

Table 8: Forecasting FX rates: Five Seconds Ahead.— The table reports estimates for the regression

$$\mathcal{R}^{D2D}_{t+\ell} = \beta^p_1 \mathcal{R}^{D2C}_t - \beta^p_2 \Delta\widehat{\mathcal{A}}_{mismatch,t} + \text{controls}_t + \text{Constant}\,,$$

where $\Delta X_t = X_t - X_{t-\ell}$ and $\ell = 5s$. Heteroscedasticity and autocorrelation robust standard errors are shown in parenthesis.

	Dependent variable:			
	$\mathcal{R}^{D2D}_{t+\ell}$			
	(1)	(2)	(3)	(4)
$\mathcal{R}^{D2C}_t$	0.1830***	0.1831***	0.1830***	0.1831***
	(0.0319)	(0.0319)	(0.0315)	(0.0315)
$\Delta\widehat{\mathcal{A}}_{mismatch,t}$	4.2254***	4.2293***	4.2254***	4.2293***
	(0.8106)	(0.8097)	(0.8022)	(0.8012)
$ProxyLiqShock_t$		−0.000002***		−0.000002***
		(0.000000)		(0.000000)
OF^{D2D}_t			−0.0000	−0.0000
			(0.0000)	(0.0000)
Constant	0.0000	0.0000	0.0000	0.0000
	(0.000000)	(0.000000)	(0.000000)	(0.000000)
Observations	6,061,573	6,061,573	6,061,573	6,061,573
R^2	0.0309	0.0310	0.0309	0.0310
Adjusted R^2	0.0309	0.0310	0.0309	0.0310
Residual Std. Error	0.00004	0.00004	0.00004	0.00004
	(df = 6061570)	(df = 6061569)	(df = 6061569)	(df = 6061568)
F Statistic	96,646.9200***	64,567.7000***	64,431.4600***	48,425.9100***
	(df = 2; 6061570)	(df = 3; 6061569)	(df = 3; 6061569)	(df = 4; 6061568)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 9: Forecasting FX rates: Thirty Seconds Ahead.— The table reports estimates for the regression

$$\mathcal{R}_{t+\ell}^{D2D} = \beta_1^p \mathcal{R}_t^{D2C} - \beta_2^p \Delta\widehat{\mathcal{A}}_{mismatch,t} + \text{controls}_t + \text{Constant}\,,$$

where $\Delta X_t = X_t - X_{t-\ell}$ and $\ell = 30s$. Heteroscedasticity and autocorrelation robust standard errors are shown in parenthesis.

	Dependent variable:			
	$\mathcal{R}_{t+\ell}^{D2D}$			
	(1)	(2)	(3)	(4)
$\mathcal{R}_t^{D2C}$	0.1254***	0.1255***	0.1254***	0.1255***
	(0.0202)	(0.0202)	(0.0199)	(0.0199)
$\Delta\widehat{\mathcal{A}}_{mismatch,t}$	2.9001***	2.9036***	2.9001***	2.9036***
	(0.4921)	(0.4928)	(0.4861)	(0.4868)
$ProxyLiqShock_t$		−0.000003***		−0.000003***
		(0.000001)		(0.000001)
OF_t^{D2D}			−0.0000	−0.0000
			(0.0000)	(0.0000)
Constant	0.000000	0.000000	0.000000	0.000000
	(0.000000)	(0.000000)	(0.000000)	(0.000000)
Observations	6,061,473	6,061,473	6,061,473	6,061,473
R^2	0.0151	0.0151	0.0151	0.0151
Adjusted R^2	0.0151	0.0151	0.0151	0.0151
Residual Std. Error	0.0001	0.0001	0.0001	0.0001
	(df = 6061470)	(df = 6061469)	(df = 6061469)	(df = 6061468)
F Statistic	46,471.3300***	31,045.1200***	30,981.5300***	23,284.3300***
	(df = 2; 6061470)	(df = 3; 6061469)	(df = 3; 6061469)	(df = 4; 6061468)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 10: Forecasting FX rates: Sixty Seconds Ahead.— The table reports estimates for the regression

$$\mathcal{R}^{D2D}_{t+\ell} = \beta^p_1 \mathcal{R}^{D2C}_t - \beta^p_2 \Delta\widehat{\mathcal{A}}_{mismatch,t} + \text{controls}_t + \text{Constant}\,,$$

where $\Delta X_t = X_t - X_{t-\ell}$ and $\ell = 60s$. Heteroscedasticity and autocorrelation robust standard errors are shown in parenthesis.

	Dependent variable:			
	$\mathcal{R}^{D2D}_{t+\ell}$			
	(1)	(2)	(3)	(4)
$\mathcal{R}^{D2C}_t$	0.0766***	0.0767***	0.0766***	0.0767***
	(0.0126)	(0.0126)	(0.0125)	(0.0125)
$\Delta\widehat{\mathcal{A}}_{mismatch,t}$	1.4709***	1.4737***	1.4710***	1.4737***
	(0.3571)	(0.3577)	(0.3539)	(0.3545)
$ProxyLiqShock_t$		−0.000004***		−0.000004***
		(0.000001)		(0.000001)
OF^{D2D}_t			−0.0000	−0.0000
			(0.0000)	(0.0000)
Constant	0.000000	0.000000	0.000000	0.000000
	(0.000001)	(0.000001)	(0.000001)	(0.000001)
Observations	6,061,353	6,061,353	6,061,353	6,061,353
R^2	0.0057	0.0057	0.0057	0.0057
Adjusted R^2	0.0057	0.0057	0.0057	0.0057
Residual Std. Error	0.0001	0.0001	0.0001	0.0001
	(df = 6061350)	(df = 6061349)	(df = 6061349)	(df = 6061348)
F Statistic	17,402.6300***	11,641.5600***	11,602.5600***	8,731.7870***
	(df = 2; 6061350)	(df = 3; 6061349)	(df = 3; 6061349)	(df = 4; 6061348)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 11: Forecasting FX rates: Ninety Seconds Ahead.— The table reports estimates for the regression

$$\mathcal{R}^{D2D}_{t+\ell} = \beta^p_1 \mathcal{R}^{D2C}_t - \beta^p_2 \Delta\widehat{\mathcal{A}}_{mismatch,t} + \text{controls}_t + \text{Constant}\,,$$

where $\Delta X_t = X_t - X_{t-\ell}$ and $\ell = 90s$. Heteroscedasticity and autocorrelation robust standard errors are shown in parenthesis.

	Dependent variable:			
	$\mathcal{R}^{D2D}_{t+\ell}$			
	(1)	(2)	(3)	(4)
$\mathcal{R}^{D2C}_t$	0.0463***	0.0463***	0.0463***	0.0463***
	(0.0090)	(0.0090)	(0.0088)	(0.0089)
$\Delta\widehat{\mathcal{A}}_{mismatch,t}$	0.7132	0.7156	0.7133	0.7156
	(0.4520)	(0.4541)	(0.4500)	(0.4520)
$ProxyLiqShock_t$		−0.000004***		−0.000004***
		(0.000001)		(0.000001)
OF^{D2D}_t			−0.0000	−0.0000
			(0.0000)	(0.0000)
Constant	0.000000	0.000000	0.000000	0.000000
	(0.000001)	(0.000001)	(0.000001)	(0.000001)
Observations	6,061,233	6,061,233	6,061,233	6,061,233
R^2	0.0021	0.0021	0.0021	0.0021
Adjusted R^2	0.0021	0.0021	0.0021	0.0021
Residual Std. Error	0.0002	0.0002	0.0002	0.0002
	(df = 6061230)	(df = 6061229)	(df = 6061229)	(df = 6061228)
F Statistic	6,364.3250***	4,270.8460***	4,243.7040***	3,203.7530***
	(df = 2; 6061230)	(df = 3; 6061229)	(df = 3; 6061229)	(df = 4; 6061228)
Note:				*p<0.1; **p<0.05; ***p<0.01

D.2 Spreads

Table 12: Forecasting FX Spreads: Five Seconds Ahead.— The table reports estimates for the regression

$$BA^{D2D}_{t+\ell} = \beta_1^{BA} E[BA^{D2C}_{l,t}] - \beta_2^{BA} \text{Var}[BA^{D2C}_{l,t}] + \text{controls}_t + \text{Constant}\,,$$

where $\Delta X_t = X_t - X_{t-\ell}$ and $\ell = 5s$. Heteroscedasticity and autocorrelation robust standard errors are shown in parenthesis.

	Dependent variable:			
	$BA^{D2D}_{t+\ell}$			
	(1)	(2)	(3)	(4)
$E[BA^{D2C}_{l,t}]$	0.5480***	0.5480***	0.5477***	0.5477***
	(0.0119)	(0.0119)	(0.0120)	(0.0120)
$\text{Var}[BA^{D2C}_{l,t}]$	13.5613***	13.5613***	13.5512***	13.5512***
	(0.3058)	(0.3070)	(0.3075)	(0.3087)
$ProxyLiqShock_t$		0.000000*		0.000000
		(0.000000)		(0.000000)
OF^{D2D}_t			−0.0000***	−0.0000***
			(0.0000)	(0.0000)
Constant	0.00001***	0.00001***	0.00001***	0.00001***
	(0.000001)	(0.000001)	(0.000001)	(0.000001)
Observations	6,149,158	6,149,158	6,149,158	6,149,158
R^2	0.3102	0.3102	0.3109	0.3109
Adjusted R^2	0.3102	0.3102	0.3109	0.3109
Residual Std. Error	0.00003	0.00003	0.00003	0.00003
	(df = 6149155)	(df = 6149154)	(df = 6149154)	(df = 6149153)
F Statistic	1,382,450.0000***	921,635.1000***	924,684.1000***	693,514.4000***
	(df = 2; 6149155)	(df = 3; 6149154)	(df = 3; 6149154)	(df = 4; 6149153)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 13: Forecasting FX Spreads: Thirty Seconds Ahead.— The table reports estimates for the regression

$$BA^{D2D}_{t+\ell} = \beta^{BA}_1 E[BA^{D2C}_{l,t}] - \beta^{BA}_2 \mathrm{Var}[BA^{D2C}_{l,t}] + \mathrm{controls}_t + \mathrm{Constant}\,,$$

where $\Delta X_t = X_t - X_{t-\ell}$ and $\ell = 30s$. Heteroscedasticity and autocorrelation robust standard errors are shown in parenthesis.

	Dependent variable:			
	$BA^{D2D}_{t+\ell}$			
	(1)	(2)	(3)	(4)
$E[BA^{D2C}_{l,t}]$	0.5339***	0.5336***	0.5336***	
	(0.0128)	(0.0129)	(0.0129)	(0.0129)
$\mathrm{Var}[BA^{D2C}_{l,t}]$	13.1912***	13.1912***	13.1817***	13.1818***
	(0.3269)	(0.3280)	(0.3286)	(0.3298)
$ProxyLiqShock_t$		−0.000000		−0.000000
		(0.000000)		(0.000000)
OF^{D2D}_t			−0.0000***	−0.0000***
			(0.0000)	(0.0000)
Constant	0.00001***	0.00001***	0.00001***	0.00001***
	(0.000001)	(0.000001)	(0.000001)	(0.000001)
Observations	6,147,812	6,147,812	6,147,812	6,147,812
R^2	0.2927	0.2927	0.2934	0.2934
Adjusted R^2	0.2927	0.2927	0.2934	0.2934
Residual Std. Error	0.00003	0.00003	0.00003	0.00003
	(df = 6147809)	(df = 6147808)	(df = 6147808)	(df = 6147807)
F Statistic	1,271,888.0000***	847,932.4000***	850,870.2000***	638,158.8000***
	(df = 2; 6147809)	(df = 3; 6147808)	(df = 3; 6147808)	(df = 4; 6147807)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 14: Forecasting FX Spreads: Sixty Seconds Ahead.— The table reports estimates for the regression

$$BA^{D2D}_{t+\ell} = \beta^{BA}_1 E[BA^{D2C}_{l,t}] - \beta^{BA}_2 \mathrm{Var}[BA^{D2C}_{l,t}] + \mathrm{controls}_t + \mathrm{Constant}\,,$$

where $\Delta X_t = X_t - X_{t-\ell}$ and $\ell = 60s$. Heteroscedasticity and autocorrelation robust standard errors are shown in parenthesis.

	Dependent variable:			
	$BA^{D2D}_{t+\ell}$			
	(1)	(2)	(3)	(4)
$E[BA^{D2C}_{l,t}]$	0.5221***	0.5221***	0.5218***	0.5218***
	(0.0133)	(0.0134)	(0.0134)	(0.0134)
$\mathrm{Var}[BA^{D2C}_{l,t}]$	12.8970***	12.8971***	12.8866***	12.8867***
	(0.3405)	(0.3415)	(0.3420)	(0.3429)
$ProxyLiqShock_t$		−0.000000*		−0.000000*
		(0.000000)		(0.000000)
OF^{D2D}_t			−0.0000***	−0.0000***
			(0.0000)	(0.0000)
Constant	0.00002***	0.00002***	0.00002***	0.00002***
	(0.000001)	(0.000001)	(0.000001)	(0.000001)
Observations	6,148,828	6,148,828	6,148,828	6,148,828
R^2	0.2792	0.2792	0.2799	0.2799
Adjusted R^2	0.2792	0.2792	0.2799	0.2799
Residual Std. Error	0.00003	0.00003	0.00003	0.00003
	(df = 6148825)	(df = 6148824)	(df = 6148824)	(df = 6148823)
F Statistic	1,190,598.0000***	793,734.0000***	796,609.7000***	597,458.8000***
	(df = 2; 6148825)	(df = 3; 6148824)	(df = 3; 6148824)	(df = 4; 6148823)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 15: Forecasting FX Spreads: Ninety Seconds Ahead.— The table reports estimates for the regression

$$BA^{D2D}_{t+\ell} = \beta_1^{BA} E[BA^{D2C}_{l,t}] - \beta_2^{BA} \mathrm{Var}[BA^{D2C}_{l,t}] + \mathrm{controls}_t + \mathrm{Constant}\,,$$

where $\Delta X_t = X_t - X_{t-\ell}$ and $\ell = 90s$. Heteroscedasticity and autocorrelation robust standard errors are shown in parenthesis.

	Dependent variable:			
	$BA^{D2D}_{t+\ell}$			
	(1)	(2)	(3)	(4)
$E[BA^{D2C}_{l,t}]$	0.5124***	0.5124***	0.5120***	0.5120***
	(0.0129)	(0.0130)	(0.0130)	(0.0130)
$\mathrm{Var}[BA^{D2C}_{l,t}]$	12.6485***	12.6486***	12.6380***	12.6380***
	(0.3327)	(0.3336)	(0.3338)	(0.3348)
$ProxyLiqShock_t$		−0.000000		−0.000000
		(0.000000)		(0.000000)
OF^{D2D}_t			−0.0000***	−0.0000***
			(0.0000)	(0.0000)
Constant	0.00002***	0.00002***	0.00002***	0.00002***
	(0.000001)	(0.000001)	(0.000001)	(0.000001)
Observations	6,148,648	6,148,648	6,148,648	6,148,648
R^2	0.2676	0.2676	0.2684	0.2684
Adjusted R^2	0.2676	0.2676	0.2684	0.2684
Residual Std. Error	0.00003	0.00003	0.00003	0.00003
	(df = 6147449)	(df = 6147448)	(df = 6147448)	(df = 6147447)
F Statistic	0.00003	0.00003	0.00003	0.00003
	(df = 6148645)	(df = 6148644)	(df = 6148644)	(df = 6148643)
Note:				*p<0.1; **p<0.05; ***p<0.01

E Competitive benchmark

In this appendix, we consider the same model except that the D2D market is assumed to be competitive.

E.1 D2D Prices and the Cross-Section of D2C Quotes in the Competitive Model

Proposition 5 *Assume that cross-sectional mean of dealer's risk aversions is close to* Γ^* *and that cross-sectional variance of dealers' risk aversions is small. Denote* $E[\Gamma_l] - \Gamma^* = \epsilon_1$, $Var[\Gamma_l] = \epsilon_2$ *and* $\epsilon = \max(\epsilon_1, \epsilon_2, ||x_l||\}$. *Then we have*

$$\mathcal{P}^{D2D} = \beta_0^{p_c} + \beta_1^{p_c} E\left[\alpha_l\right] + O(\epsilon)\,,$$

where $\beta_0^{p_c} = -\frac{k_0^a + k_0^b}{k} > 0$ *and* $\beta_1^{p_c} = \frac{1}{k} > 0$.

Proof of Proposition 5. Dealers' FOC is

$$d - \Gamma_l\left(\chi_l + Q_l\right) \ = \ \mathcal{P}^{D2D}, \tag{20}$$

when the D2D market is competitive. Dividing by Γ_l, summing over all dealers, using market clearing and rearranging expression implies that

$$\mathcal{P}^{D2D} \ = \ d - \left(\sum_l \frac{1}{\Gamma_l}\right)^{-1} \sum \chi_l\,.$$

We now derive the Taylor series expansion of $\mathcal{P}^{D2D}$ as a function of dealers' risk aversion around $\mathbf{\Gamma}^* = \Gamma^* \times \mathbf{1}$. We have

$$\mathcal{P}^{D2D} = \ \mathcal{P}^{D2D}|_{\mathbf{\Gamma}^*} + \sum_j d_{\Gamma_j} \mathcal{P}^{D2D}|_{\mathbf{\Gamma}^*}(\Gamma_j - \Gamma^*)\,,$$

with

$$\begin{aligned} \mathcal{P}^{D2D}|_{\mathbf{\Gamma}^*} &= \ d - \Gamma^* E[\tilde{\chi}] \\ d_{\Gamma_j}\mathcal{P}^{D2D}|_{\mathbf{\Gamma}^*} &= \ -\frac{1}{M}E[\tilde{\chi}] \end{aligned}$$

Thus, to first-order, we have

$$\begin{aligned}\mathcal{P}^{D2D} &= d - \Gamma^* E[\tilde{\chi}] - \sum_j \frac{1}{M} E[\tilde{\chi}](\Gamma_j - \Gamma^*) \\ &= d - \Gamma^* E[\tilde{\chi}] - \underbrace{E[\tilde{\chi}]E[\Gamma_j - \Gamma^*]}_{o(\epsilon)} \\ &= d - \Gamma^* E[\tilde{x}] - \underbrace{E[\tilde{\chi}]E[\Gamma_j - \Gamma^*]}_{o(\epsilon)},\end{aligned}$$

where the last equality follows from the fact that $E[\tilde{q}] = 0$. Recall that

$$d - \Gamma^* E[\tilde{x}] = \frac{1}{k} E[\alpha_l] - \frac{k_0^a + k_0^b}{k} + O(\epsilon^2).$$

Hence, $\mathcal{P}^{D2D}$ becomes:

$$\mathcal{P}^{D2D} = -\frac{k_0^a + k_0^b}{k} + \frac{1}{k} E[\alpha_l] + O(\epsilon).$$

Q.E.D.

E.2 D2D Spreads and the Cross-Section of D2C Quotes in the Competitive Model

Proposition 6 *Assume that cross-sectional mean of dealer's risk aversions is close to Γ^* and that cross-sectional variance of dealers' risk aversions is small. Denote $E[\Gamma_l] - \Gamma^* = \epsilon_1$, $Var[\Gamma_l] = \epsilon_2$ and $\epsilon = \max(\epsilon_1, \epsilon_2)$. The following is true*

$$\Lambda = \beta_0^\Lambda + \beta_1^\Lambda E[b_l] - \beta_2^\Lambda Var[b_l] + o\left(\epsilon^2\right),$$

where $\beta_0^\Lambda = \frac{2\phi_0}{M} > 0$, $\beta_1^\Lambda = \frac{2\phi_1}{M} > 0$, $\beta_2^\Lambda = \frac{2\phi_1^2}{M\Gamma^*} > 0$, $\phi_0 = -\frac{(k_0^a - k_0^b)}{\mathbb{E}[q^2]k}$ and $\phi_1 = \frac{2}{\mathbb{E}[q^2]k}$.

Proof of Proposition 6. The BA spread Λ is given by

$$\Lambda = 2\left(\sum_l \frac{1}{\Gamma_l}\right)^{-1}$$

under the assumption of a competitive D2D market. The Taylor Series expansion of Λ (as a function of $\{\Gamma_l\}$) around $\mathbf{\Gamma}^* = \Gamma^* \times \mathbf{1}$ gives

$$\begin{aligned}
\frac{1}{2}\Lambda = & \frac{\Gamma^*}{M} + \sum_j \Lambda_j|_{\mathbf{\Gamma}^*}(\Gamma_j - \Gamma^*) + \frac{1}{2}\sum_j \sum_i \Lambda_{ij}|_{\mathbf{\Gamma}^*}(\Gamma_j - \Gamma^*)(\Gamma_i - \Gamma^*) \\
= & \frac{\Gamma^*}{M} + \sum_j \frac{1}{M^2}(\Gamma_j - \Gamma^*) - \frac{1}{M^2\Gamma^*}\sum_i (\Gamma_i - \Gamma^*)^2 + \frac{1}{M^3\Gamma^*}\sum_j \sum_i (\Gamma_j - \Gamma^*)(\Gamma_i - \Gamma^*) \\
= & \frac{\Gamma^*}{M} + \frac{1}{M^2}\sum_j (\Gamma_j - \Gamma) - \frac{1}{M^2\Gamma^*}\left(MVar[\tilde{\Gamma}] + M\epsilon_1^2\right) + \frac{1}{M^3\Gamma^*}\left(M^2\epsilon_1^2\right) \\
= & \frac{1}{M}E[\tilde{\Gamma}] - \frac{1}{M\Gamma^*}Var[\tilde{\Gamma}] + o\left(\epsilon^2\right)
\end{aligned}$$

Recall that $\Gamma_l = \phi_0 + \phi_1 b_l$. The result then follows.

Q.E.D.

E.3 Derivation of the Ratio of Elasticities in the Competitive and non-competitive Model

It follows from (20) that the leading term for the illiquidity Λ^c in the competitive model is

$$\Lambda_0^c = \frac{2\Gamma^*}{M}.$$

Comparing it to the expression for Λ_0 in Proposition 1 we find that $\Lambda_0/\Lambda_0^c = (M-1)/(M-2)$ and thus, for the ratio of elasticities we have

$$\frac{\mathcal{E}_0^c}{\mathcal{E}_0} = \frac{M-1}{M-2}.$$

F A microfoundation for (3)

In this section, we characterize dealers' equilibrium quoting strategies and demonstrate that (3) holds approximately when gains from trade between customers and dealers are small.

We start with the analysis of dealers' bids. Analysis for asks is analogous and is omitted for brevity. We assume that all dealers face the same uncertainty about the reservation value of others. In particular, each dealer believes that the distribution of any other dealer k's reservation value is

$$Pr(r_k^b < x) = F(x) \text{ with support } [\underline{r}, \bar{r}].$$

We assume that customers have reservation value r_c and that the gains from trade between customers and dealers are small: We will characterize the equilibrium asymptotically, when $\bar{r} - r_c$ is small. Consistent with the idea customers will not choose a dealer posting the price r_c or less we assume that $\psi(x - \bar{r}) = 0$, $x \leq r_c$ and $\psi'(r_c - \bar{r}) = 0$.

We consider a symmetric equilibrium in which a dealer with a reservation value r bids according to a strictly increasing strategy $\beta(r)$. We employ the revelation principle. Each dealer l reports a reservation value $\hat{r}_l$, which is then transformed into his bid $\beta(\hat{r}_l)$. Equilibrium $\beta(\cdot)$ must be such that truth-telling is optimal: $\hat{r}_l = r_l$ in equilibrium.

Denote

$$G(x) = F(x)^{M-1}.$$

$G(\cdot)$ is the CDF of the maximum of the reservation values of $M-1$ "other" dealers.

Dealer's profit given reported reservation value $\hat{r}$ is

$$(r - \beta(\hat{r})) \int \psi\left(\beta(\hat{r}) - \beta\left(x\right)\right) dG(x)$$

Differentiating with respect to $\hat{r}$ and substituting $\hat{r} = r$ yields the necessary condition for symmetric equilibrium

$$-\beta'(r) \int \psi\left(\beta(\hat{r}) - \beta\left(x\right)\right) dG(x) + \left(r - \beta(r)\right) \beta'(r) \int \psi'\left(\beta(r) - \beta\left(x\right)\right) dG(x) = 0.$$

This can be rewritten as

$$r\beta'(r) \int \psi'\left(\beta(r) - \beta\left(x\right)\right) dG(x) = \left(\beta(r) \int \psi\left(\beta(r) - \beta\left(x\right)\right) dG(x)\right)' \tag{21}$$

Denote

$$H(r) = \int_{r_c}^{\bar{r}} \psi\left(\beta(r) - \beta\left(x\right)\right) dG(x)$$

Then

$$h(r) = H'(r) = \beta'(r) \int \psi'\left(\beta(r) - \beta\left(x\right)\right) dG(x).$$

Then we can rewrite (21) as

$$rh(r) = \left(\beta(r)H(r)\right)'$$

And so the ODE (21) is equivalent to an integral equation

$$\beta(r) = r - \int_{r_c}^{r} \frac{H(t)}{H(r)} dt,$$

where we accounted for the boundary condition $\beta(r_c) = r_c$.

Now note that

$$\int_{r_c}^{r} H(t)dt = \int_{r_c}^{r} dt \left\{ h(r_c)(t-r_c) + O\left((t-r_c)^2\right) \right\} = \frac{1}{2} h(r_c)\left(r-r_c\right)^2 + o\left(\bar{r}-r_c\right)$$

$$H(r) = h(r_c)(r-r_c) + o\left(\bar{r}-r_c\right).$$

Combining the two we get

$$\beta(r) = r - \frac{1}{2}(r-r_c) + o\left(\bar{r}-r_c\right).$$

Thus, (3) holds approximately when $\bar{r}-r_c$ is small.

Internet Appendix for "Dealer Heterogeneity and Exchange Rates"

IA.1 Additional Theoretical Results

In this section, we derive additional results about the signs of the coefficients β_2^p and β_2^{BA}, that help to shed some light on the economic mechanisms behind (4) and (5).

IA.1.1 Inventory-based Liquidity Mismatch

We introduce some notation. We say that $\{\hat{x}_l\}_l$ is a *mean-preserving spread* of $\{x_l\}_l$ if $\hat{x}_l = x_l + \epsilon_l$ and $E[\epsilon_l] = 0$. We use the shortcut x for a vector $\{x_l\}_l$ and we denote $\|x\| = (\sum_l x_l^2)^{1/2}$ the Euclidean norm of x. We also write $a \stackrel{s}{=} b$ whenever $\text{sign}(a) = \text{sign}(b)$.

We say that a *liquidity mismatch* occurs when high-risk aversion dealers have large inventories after the D2C trading round. Our inventory-based measure of liquidity mismatch $Y_{mismatch}$, mirrors this definition:

$$Y_{mismatch} \equiv \text{Cov}(\Gamma_l, \chi_l).$$

Recall our price-based measure $\mathcal{A}_{mismatch}$, defined as follows:

$$\mathcal{A}_{mismatch} \equiv \text{Cov}(b_l, \alpha_l).$$

We show below that $\mathcal{A}_{mismatch} \stackrel{s}{=} Y_{mismatch}$, that is, when there is a positive (negative) liquidity mismatch, both measures are positive (negative). Thus, $\mathcal{A}_{mismatch}$ can be vieved as a proxy for $Y_{mismatch}$.

IA.1.2 D2D Pricing and Dealer Characteristics

When there is a liquidity mismatch, the allocation of inventories across dealers is inefficient. Due to market power, such inefficiency will not be resolved after the D2D trade. As a result, dealers (on average) will require higher compensation for holding the inventories, and the prices will be lower. Consistent with this intuition, we show below that positive (negative) mismatch $Y_{mismatch}$ introduces a downward (upward) distortion in the equilibrium D2D price.

Indeed, one can rewrite (14) as follows:

$$\mathcal{P}^{D2D} \;=\; \bar{d} - \mathcal{B}^{-1} M\Big(E[(\Gamma_l + \beta_l)^{-1}\Gamma_l]E[\chi_l] \;+\; \text{Cov}((\Gamma_l + \beta_l)^{-1}\Gamma_l, \chi_l)\Big). \tag{IA.1}$$

The key term in (IA.1) is $\text{Cov}((\Gamma_l+\beta_l)^{-1}\Gamma_l, \chi_l)$. We show below that $\text{Cov}((\Gamma_l+\beta_l)^{-1}\Gamma_l, \chi_l)) \stackrel{s}{=} Y_{mismatch}$, and so, indeed, positive (negative) mismatch $Y_{mismatch}$ introduces a downward (upward) distortion in the equilibrium D2D price.

To see that $\mathrm{Cov}((\Gamma_l+\beta_l)^{-1}\Gamma_l, \chi_l)) \stackrel{s}{=} Y_{mismatch}$, note that by Proposition 4, *less risk averse dealers have greater price impact*: If $\Gamma_1 < ... < \Gamma_M$, then dealers' price impacts in the D2D market satisfy $\beta_1 > \cdots > \beta_M$. This is intuitive: Less risk-averse dealers face a more risk-averse rest of the market and, therefore, a less elastic residual supply. By direct calculation, the weights $(\Gamma_l + \beta_l)^{-1}\Gamma_l$ in Equation (14) are monotone increasing in Γ_l. Therefore, we can write

$$\mathrm{Cov}((\Gamma_l + \beta_l)^{-1}\Gamma_l, \chi_l)) \stackrel{s}{=} \mathrm{Cov}(\Gamma_l, \chi_l) = Y_{mismatch}.$$

The following proposition follows immediately.

Proposition IA.7 *Fix* $\{\Gamma_l\}_l$. *Consider three possibilities for distribution of inventories* $\{\chi_l\}_l$ *across dealers:* $\{\chi_l\}_l$ *is such that (a)* $\mathrm{Cov}(\chi_l, \Gamma_l) > 0$, *(b)* $\mathrm{Cov}(\chi_l, \Gamma_l) = 0$, *and (c)* $\mathrm{Cov}(\chi_l, \Gamma_l) < 0$. *Suppose that* $E[\chi_l]$ *is the same in all three cases. Then,* $\mathcal{P}^{D2D}_{(a)} < \mathcal{P}^{D2D}_{(b)} < \mathcal{P}^{D2D}_{(c)}$.

We also derive implications of changes in the dispersion of risk aversions.

Proposition IA.8 *An increase in the dispersion of risk aversions* Γ_l *(defined as a mean-preserving spread) leads to a decrease in bid-ask spread* BA *in D2D market.*

To understand the intuition behind the proposition above, consider the following example. Imagine we have three dealers with a risk aversion of 1 each. They each provide 1 unit of liquidity, 3 in total. Now consider what happens if dealers' average risk aversion is 0.5, 1, and 1.5 (so that average risk aversion is the same, but we increased the dispersion). Since liquidity provided (price elasticity) is inversely related to risk aversion, the dealers will provide 2, 1, and 1/1.5=0.66 units of liquidity, 3.66 in total.[33] Thus, greater dispersion in risk aversions results in more liquidity.

Propositions IA.7 and IA.8 establish the link between prices and liquidity in the D2D market and some statistics of unobservable dealer characteristics. The next section demonstrates how to identify these statistics from prices and spreads in the D2C market.

IA.2 Linking Dealer Characteristics to D2C Prices and Spreads

Because neither risk aversions Γ_l nor inventories χ_l are observable, we cannot directly take the predictions of Propositions IA.7 and IA.8 to the data. However, mid-prices α_l and bid-ask spreads b_l in the D2C market are informative about unobservable dealer characteristics Γ_l and χ_l. Here, we use our model to map $Y_{mismatch}$ and cross-sectional dispersion of risk aversions to observable statistics of prices and spreads in the D2C market. Section IA.3.2 shows that the same mapping is applicable in the full model.

[33] The price elasticity is inversely proportional to risk aversion when heterogeneity in risk aversions is small. In the general case, this relationship is more complex. However, price elasticity is still a convex function of risk aversion in the general case. Thus the logic of our simple example still applies.

First, we note that D2C bid-ask spreads are informative about dealers' risk aversions. Indeed, it follows directly from (2) and (3) that

$$b_l = k_0^a - k_0^b + \frac{1}{2}k\Gamma_l \left(\frac{E[\tilde{q}^2|\tilde{q} > 0]}{E\left[\tilde{q}|\tilde{q} > 0\right]} - \frac{E[\tilde{q}^2|\tilde{q} < 0]}{E\left[\tilde{q}|\tilde{q} < 0\right]} \right).$$

Thus, changes in Γ_l are proportional to changes in b_l. This is intuitive: More risk-averse dealers are less efficient at holding inventory and require higher compensation for doing so, resulting in wider spreads. Directly from Proposition IA.8, we obtain the following.

Proposition IA.9 *An increase in the dispersion of D2C bid-ask spreads b_l (defined as a mean-preserving spread) is associated with a decrease in bid-ask spread BA in D2D market.*

We now show that $Y_{mismatch} \overset{s}{=} \mathcal{A}_{mismatch}$. The equilibrium relationship between prices in the D2C market and dealers' inventories and risk aversions is generally complex. However, it is possible to derive analytical approximations when dealer heterogeneity is small. Such an approximation allows us to capture the first-order effects of heterogeneity on equilibrium quantities while preserving analytical tractability. The case of small heterogeneity corresponds to small values of $\|\Gamma - \Gamma^*\|$, that is, when risk aversions Γ_l are close to some average level Γ^*. Under such an approximation, we can write

$$p_l^a \approx const^a - k\Gamma^* x_l,\; p_l^b \approx const^b - k\Gamma^* x_l, \text{ and } \alpha_l \approx const - k\Gamma^* x_l, \tag{IA.2}$$

where $const^a$, $const^b$, and $const$ are some constants that are the same across dealers and $\approx$ denotes approximate equality, up to terms of order $O(\|\Gamma - \Gamma^*\|)$.

It follows from (IA.2) that the dealer posting the lowest ask (highest bid) price is also the dealer with the lowest (highest) mid-price. Because the customer order flow is directed to the dealer with the best price, the dealer with the smallest (highest) mid-price will decrease (increase) his inventory, while other dealers' inventories will be unchanged. Thus, we get a positive cross-sectional relationship between changes in inventories $\chi_l - x_l$ and mid-prices α_l. If total customer orderflow $\tilde{q}$ is large compared to dealers' initial inventories x_l, dealers' final inventory χ_l and change in inventories $\chi_l - x_l$ have the same sign. Thus, we have a positive cross-sectional relationship between χ_l and α_l. Then we have $\mathrm{Cov}(\chi_l, \Gamma_l) \overset{s}{=} \mathrm{Cov}(\alpha_l, \Gamma_l) \overset{s}{=} \mathrm{Cov}(\alpha_l, b_l)$, where the last equality follows because b_l is related positively to Γ_l. This results in $Y_{mismatch} \overset{s}{=} \mathcal{A}_{mismatch}$, and the following proposition follows directly from Proposition IA.7.

Proposition IA.10 *Suppose that $\|\Gamma - \Gamma^*\|^2$ is sufficiently small and that the realization $|\tilde{q}|$ is sufficiently large. Then $Y_{mismatch} \overset{s}{=} \mathcal{A}_{mismatch}$. Holding $\{\Gamma_l\}_l$ fixed, consider three possibilities for the joint distribution of mid-prices and bid-ask spreads in the D2C market: (a) $\mathcal{A}_{mismatch} > 0$, (b) $\mathcal{A}_{mismatch} > 0 = 0$, and (c) $\mathcal{A}_{mismatch} < 0$. Suppose that in all three cases, $E[\alpha_l]$ is the same. Then, $\mathcal{P}^{D2D}_{(a)} < \mathcal{P}^{D2D}_{(b)} < \mathcal{P}^{D2D}_{(c)}$.*

IA.3 Extended Model

IA.3.1 The Model

This section aims to develop a theoretical model that captures all aspects of the real-world market structure described in Section 2. First, each dealer bank must be able to quote prices to its customers through its own SBP, accounting for the imperfect quote competition with other dealers. Second, each customer must be able to choose a dealer with whom to trade. Third, a customer shall be able to split orders across several dealers. Fourth, in the case of a customer order imbalance, each dealer must be able to offload some of its inventory in the inter-dealer market. Fifth, we need a way of modeling imperfect competition and market power in both segments. Finally, the model should feature dealer heterogeneity to speak to heterogeneity in prices and bid-ask spreads in the data. Below, we outline a model that accounts for all these frictions.

There are five time periods, $t = -1,\ 0-\ ,\ 0,\ 1,\ 2$, and two tradable assets, a risk-free asset with a rate of return normalized to zero and a risky asset with a random payoff d at time $t = 2$. We assume that d is normally distributed with mean $\bar{d}$ and variance σ_d^2. The market is populated by M heterogeneous dealers, indexed by $l = 1, \cdots, M$, and n ex-ante identical customers, indexed by $c = 1, \cdots, n$. All agents start with zero asset inventories. The timeline is as follows:

- At time $t = -1$, dealers' inventory shocks $x_l,\ l = 1, \cdots, M$ are realized and are public information.[34] Making such information private is irrelevant for the D2D trading round because of the ex-post nature of the D2D market mechanism. As for the D2C round, the absence of uncertainty about other dealers' holdings is a simplifying assumption. In the real world, dealers have many sources of price-based information, allowing them to make inferences about other dealers' inventory constraints. For example, dealers may observe other dealers' quotes on various platforms. However, modeling price-based inference would drastically complicate the analysis, introducing signaling and belief manipulation aspects into the equilibrium behavior. For this reason, almost all existing models of OTC markets make the same simplifying assumption of publicly observed private types. (See, for example, Duffie et al. (2005), Schürhoff and Li (2019), Babus and Parlatore (2022).) We abstract from these effects and leave them for future research. These shocks may originate from previous trading rounds on the dealer-specific SBP.

- At time $t = 0-$, customers' *endowment shocks* vector $\Theta = \{\theta_c\}_{c=1}^n$ is realized; each shock is customers' private information. We assume that these shocks are independent and identically distributed across customers and are drawn from an arbitrary non-degenerate distribution with finite first two moments. Parameters $\bar{\theta} = E[\theta]$ and $\sigma_\theta^2 = \mathrm{Var}[\theta]$ are public information. For simplicity, we assume that customers are ex-ante homogeneous and have identical parameters $\bar{\theta},\ \sigma_\theta$. In the real world, customers, might

[34]This assumption is made for simplicity.

be heterogeneous across dealers, which may be an important aspect of heterogeneity in dealer behavior. We leave this aspect for future research. As for the case of dealers, we could also interpret θ as a private demand/taste shock unrelated to actual asset holdings.

- At the time $t = 0$, each dealer trades with customers in the D2C market using the bank's own SBP. Namely, the dealer l publishes a (dealer-specific) binding price schedule $p_l(q),\ l = 1, \cdots, M$ to each customer on the SBP, describing the per-unit price at which he is willing to sell q units of the risky asset. When $q < 0$, then $-p_l(q)$ is the per-unit bid price for many $-q$ units. Although in real life the price schedules are restricted to step functions (with different prices quoted for orders below USD 1 million and between USD 1 million and USD 5 million), we restrict dealers to use linear schedules instead. That is, we assume $p_l(q) = \alpha_l + b_l q$. Such a restriction captures the ability of real-world dealers to quote different prices for different order sizes while preserving analytical tractability.[35]

- We assume that each customer has access to the SBPs of all dealers.[36] Given the quoted price schedules on all the SBPs, each customer c optimally chooses the vector of quantities $q_c = (q_l(\theta_c))_{l=1}^M$, where $q_l(\theta_c)$ specifies the quantity of the asset acquired from dealer l by a customer with endowment shock θ_c. The total amount paid by customer c to the dealers is then given by

$$\pi(q_c) \ = \ \sum_{l=1}^{M} q_l(\theta_c)\, p_l(q_l(\theta_c)).$$

  A customer c ends up holding a total of

$$\bar{q}_c \ = \ \theta_c \ + \ \sum_{l=1}^{M} q_l(\theta_c)$$

[35] Allowing dealers to optimize in a more general class of functions complicates the analysis significantly. Each dealer selects a mechanism (the price schedule $p_l(q)$), while at the same time competing with other dealers' mechanisms and simultaneously taking into account the fact that all dealers serve as liquidity providers to each other during the second stage of the game (the inter-dealer trading). Such games in competing mechanisms are known to be extremely complex; even with homogeneous traders and a single trading round, optimal price schedules are highly nonlinear, and a symmetric equilibrium often fails to exist. See, for example, Biais et al. (2000, 2013), and Back and Baruch (2013). In our paper, the problem is much more involved because dealers are asymmetric, and there is a second rebalancing stage, introducing another dimension to the strategic interaction. In particular, when posting a price schedule, a dealer l has to account for the fact that his schedule affects customers' trades with other dealers, which in turn affects other dealers' inventories, thereby affecting the ability of dealer l to trade with other dealers in the subsequent D2D trading round.

[36] It is a simplifying assumption. If the customer is a regional bank, it may have a subscription to quotes from all major dealers; by contrast, if the customer is a hedge fund, it may decide to subscribe to only a subset of those dealers. Investigating the architecture of the D2C trading network and the endogenous decision of each customer with which dealers connect is an important direction for future research.

units of the asset. Thus, our model features order-splitting by customers. As discussed above, in the real world, order-splitting occurs often for orders above USD 1 million.

- After this D2C trading round on the SBPs, dealer l receives the vector of orders $Q_l = (q_l(\theta_c))_{c=1}^n$ and a total cash transfer of

$$\Pi_l(Q_l) \;\equiv\; \sum_{c=1}^{n} p_l(q_l(\theta_c))q_l(\theta_c) \tag{IA.3}$$

from the customers and ends up holding

$$\chi_l \;=\; x_l \;-\; \sum_{c=1}^{n} q_l(\theta_c) \tag{IA.4}$$

units of the asset. The aggregation of the D2C order flow Q_l into the sum $\sum_{c=1}^{n} q_l(\theta_c)$ represents the process of *internalization of order flow* by dealer l.

- At the time $t = 1$, dealers trade in the centralized inter-dealer market to rebalance their inventories. In the real world, D2D platforms function closely to a centralized limit-order market. We capture this fact by assuming that the D2D market operates as the standard uniform-price double auction (see, e.g., Kyle (1989), Vives (2011), Rostek and Weretka (2015), and Malamud and Rostek (2017)). Dealer l submits a (net) demand schedule $Q_l(\mathcal{P}^{D2D}) : \mathbb{R} \;\to\; \mathbb{R}$, which specifies demanded quantity of the asset given its price $\mathcal{P}^{D2D}$ in the inter-dealer market. All dealers are strategic; in particular, there are no noise traders. As is standard in strategic centralized market models for divisible goods or assets, we study the Nash equilibrium in linear bid schedules (hereafter, *equilibrium*). With divisible goods, equilibrium is invariant to the distribution of independent private uncertainty.[37] We denote by $Q_l(\chi_l, \mathcal{P}^{D2D})$ the D2D net trade of dealer l with inventory χ_l (see (IA.4)). The latter is given by the total initial inventory plus the total non-internalized customer order flow in the D2C market. Post-D2D trade, the dealer ends up with an inventory of

$$\tilde{\chi}_l \;=\; \chi_l \;+\; Q_l(\chi_l, \mathcal{P}^{D2D})\,. \tag{IA.5}$$

[37] That is, the linear Bayesian Nash Equilibrium with independent private endowment values has an *ex post* property and coincides with the linear equilibrium that is robust to adding noise in trade (robust Nash Equilibrium; e.g., Vayanos (1999) and Rostek and Weretka (2015)). Equilibrium is *linear* if schedules have the functional form of $q_l(\cdot) = \alpha_0 + \alpha_{l,q}q_l^0 + \alpha_{l,p}p$. Strategies are not restricted to linear schedules; rather, it is optimal for a trader to submit a linear demand, given that others do. The approach of analyzing the symmetric linear equilibrium is common in centralized market models (e.g., Kyle (1989), Vayanos (1999), and Vives (2011)). Our analysis does not assume equilibrium symmetry. As equilibrium schedules are optimal even if traders learn the independent value endowments q_l^0 (or equivalently, stochastic marginal utility intercepts, $\tilde{d} = d - \alpha\Sigma\tilde{q}_l^0$) of all other agents, equilibrium is *ex post* Bayesian Nash. The key to the *ex post* property is that permitting pointwise optimization – for each price – equilibrium demand schedules are optimal for any distribution of independent private information and are independent of agents' expectations about others' endowments.

- At time $t = 2$, the asset pays off.

We assume that all agents (dealers and customers) incur quadratic costs for holding inventories, equivalent to linearly decreasing marginal values. Importantly, these inventory holding costs are heterogeneous across dealers: Although customers are assumed to be homogeneous, all having the same cost γ, dealers are heterogeneous, with dealer l incurring cost Γ_l.[38] Thus, customers' total expected utility is given by

$$u(q_c) \;=\; -\pi(q_c) \;+\; E\left[d \cdot \bar{q}_c \;-\; 0.5\gamma\, \bar{q}_c^2\right] . \tag{IA.6}$$

Dealers' total expected utility has two components:

$$U_l \;=\; E\left[\Pi_l(Q_l) \;+\; \mathcal{U}_l(\tilde{\chi}_l, \mathcal{P}^{D2D})\right] , \tag{IA.7}$$

where $\Pi_l(Q_l)$ is the total transfer (IA.3) received from customers, while

$$\mathcal{U}_l(\tilde{\chi}_l, \mathcal{P}^{D2D}) \;=\; d \cdot \tilde{\chi}_l - 0.5\Gamma_l \tilde{\chi}_l^2 - \mathcal{P}^{D2D}\, Q_l(\chi_l, \mathcal{P}^{D2D})$$

is dealers' quadratic utility of their post-D2D trade inventory (IA.5) net of the total price $\mathcal{P}^{D2D} Q_l(\chi_l, \mathcal{P}^{D2D})$ paid for the $Q_l(\chi_l, \mathcal{P}^{D2D})$ units of the asset in the D2D market.

We follow the standard route used in most of the market microstructure literature and confine our attention to linear equilibria, characterized in the following definition.

Definition IA.1 *A linear Nash equilibrium is a collection of the following policies:*

- *price schedules* $p_l(q) \;=\; \alpha_l \;+\; b_l\, q$ *in the D2C market segment,* $l = 1, \cdots, M$ *;*
- *customer demand*

  $$q(\theta_c) = (q_l(\theta_c))_{l=1}^{M}, \;\; q_l(\theta_c) \;=\; \delta_l + \eta_l \theta_c\, ; \;\; \textit{and}$$

- *dealer demand schedules* $Q_l(\mathcal{P}^{D2D}) \;=\; Q_l^{(0)} + Q_l^{(1)}\mathcal{P}^{D2D}$ *in the inter-dealer market*

such that

- *dealer demand schedules form a robust Nash equilibrium in the D2D market;*[39]
- *customers' demand maximizes customers' utility* (IA.6)*; and*

[38] The assumption of homogeneous customer compositions across dealers is made for two reasons: First, it simplifies the analysis and, second, we cannot observe customer composition empirically. In the real world, heterogeneity in customer composition might be responsible for a large fraction of variation in the effective inventory cost. We view the inventory cost Γ as a theoretical shortcut for the unobservable sources of heterogeneity. Allowing for dealers' heterogeneity is crucial for matching their highly heterogeneous empirically observed behavior.

[39] As in Rostek and Weretka (2015, p. 2955), the robust Nash equilibrium is in one in which large traders' demands are optimal even after adding full-support uncertain additive noise to their residual demand.

- *dealers choose α_l and b_l to maximize expected utility (IA.7) given customers' demand functions $q(\theta_c)$ and provided the equilibrium allocation from the second stage of the game.*

Given the definition of equilibrium, we follow the standard backward induction procedure: First, we solve for the unique, robust linear Nash equilibrium of the second stage of the game (the D2D market). Second, we use this equilibrium to calculate dealers' utilities (IA.7). Third, we solve for customers' optimal demand given the linear dealer price schedules. Fourth, we use customers' demand schedules as well as the dealers' utilities from the second trading round to solve for the equilibrium in the liquidity provision game in the D2C market.

IA.3.2 Linking Dealer Characteristics to D2C Prices and Spreads in the Extended Model

This section shows how unobservable dealer characteristics can be linked to prices and spreads in the D2C market. The equilibrium relationship between prices and liquidity in the D2C market and dealers' inventories and risk aversions is generally complex. However, it is possible to derive analytical approximations when dealer heterogeneity is small. Such an approximation allows us to capture the first-order effects of heterogeneity on equilibrium quantities while preserving analytical tractability. Under such approximation, we show that: (i) more risk-averse dealers quote wider spreads, that is, b_l^{-1} is negatively related to Γ_l (see Proposition IA.6); and (ii) dealers with higher prices end up holding higher post-D2C inventories, that is, χ_l is positively related to α_l (see Lemma IA.8). Then, $Y_{mismatch} \stackrel{s}{=} \mathcal{A}_{mismatch}$.

Both (i) and (ii) are intuitive. More risk-averse dealers are less efficient at holding inventory and require higher compensation, resulting in wider spreads. Dealers posting the highest prices would attract a disproportionate share of sell volume from customers (since customers would choose to sell to dealers offering the highest price) and would end up with the highest inventories post-D2C.

We thus formulate the analog to Proposition IA.7, where unobservable dealer characteristics are substituted by observable prices and spreads in the D2C market.

Proposition IA.11 *Suppose that $\|\Gamma - \Gamma^*\|^2 + \|x\|^2 + \bar{\theta}^2$ is sufficiently small and that n is sufficiently large. Then $Y_{mismatch} \stackrel{s}{=} \mathcal{A}_{mismatch}$. Suppose further that $\{\Gamma_l\}_l$ are fixed. Consider three possibilities for the joint distribution of mid-prices and bid-ask spreads in the D2C market: (a) $\mathcal{A}_{mismatch} > 0$, (b) $\mathcal{A}_{mismatch} > 0 = 0$, and (c) $\mathcal{A}_{mismatch} < 0$. Suppose that in all three cases, $E[\alpha_l]$ is the same. Then, $\mathcal{P}^{D2D}_{(a)} < \mathcal{P}^{D2D}_{(b)} < \mathcal{P}^{D2D}_{(c)}$.*

As in Proposition IA.10, Proposition IA.11 states that $Y_{mismatch} \stackrel{s}{=} \mathcal{A}_{mismatch}$ when heterogeneity among dealers is small and when customer order flow is significant. In Proposition IA.11, the last requirement is captured by the condition that n must be large enough.

We also derive the following proposition from the fact that bid-ask spreads in the D2C market are positively related to dealers' risk aversion.

Proposition IA.12 *An increase in the dispersion of D2C bid-ask spreads b_l (defined as a mean-preserving spread) is associated with an increase in the liquidity of the D2D market.*

IA.4 Equilibrium in the Extended Model

IA.4.1 Customers' Problem

In this section, we study customers' optimization problems. Given the M different price schedules offered by the dealers, a customer optimally decides how much to trade with each dealer. Substituting dealers' price schedules into customers' utility (IA.6), we get that customers' optimization problem takes the form of

$$\max_{\{q_{c,l}\}_{l=1}^{M}} \left\{ -\sum_{l=1}^{M} q_{c,l} p_l(q_{c,l}) \; + \; \left(\sum_{l=1}^{M} q_{c,l} + \theta_c \right) d - 0.5\gamma \left(\sum_{l=1}^{M} q_{c,l} + \theta_c \right)^2 \right\} ,$$

taking dealers' schedules

$$p_l(q_{c,l}) = \alpha_l + \lambda_l^{-1} q_{c,l}$$

as given. (Thus, we denote $\lambda_l = b_l^{-1}$). Define $a_l = \lambda_l(d - \alpha_l)$ to be the bid-ask spread-normalized deviation from the fundamental value d of the mid-price quoted by dealer l. It plays the role of the risk premium in our model: the higher the premium, the lower the mid-price relative to the fundamental value d. In the sequel, we will therefore refer to this quantity as the "risk premium". Let also $a_{-l} = \sum_{\ell \neq l} a_\ell$ be the total risk premium of dealers $\ell \neq l$, and let $\lambda_{-l} = \sum_{\ell \neq l} \lambda_l$ be the measure of total liquidity provided by these dealers.

Writing down the first-order conditions, we arrive at the following result.

Lemma IA.1 *The optimal demand of customer j is given by $q_l = \{q_{c,l}\}_{l=1}^{M}$ with*

$$q_{c,l} \;=\; \delta_l(a_l, \lambda_l, \lambda_{-l}, a_{-l}) \;+\; \eta_l(a_l, \lambda_l, \lambda_{-l}, a_{-l})\, \theta_c \tag{IA.8}$$

with

$$\delta_l \;=\; \frac{0.5a_l + 0.25\gamma(a_l\lambda_{-l} - a_{-l}\lambda_l)}{1 + 0.5\gamma(\lambda_l + \lambda_{-l})}$$
$$\eta_l \;=\; -\frac{0.5\gamma\lambda_l}{1 + 0.5\gamma(\lambda_l + \lambda_{-l})} \;,\; l = 1, \cdots, M \,.$$

The intuition behind the optimal order-splitting strategy of Lemma IA.1 is as follows: Ideally, the customer would like to buy from the dealer with the lowest mid-quote (equivalently, the highest risk premium a_l) and the highest offered liquidity. Thus, his average demand addressed to dealer l is increasing in a_l and is decreasing in a_{-l} (the attractiveness of trading with other dealers). The customer's demand curve is naturally downward sloping in his inventory (that is, $\eta_l < 0$ for all l), and the size of the slope is proportional to the liquidity

λ_l offered by dealer i, as well as to the customer's cost of holding inventory, γ. The stylized linear-quadratic setting of the customer problem makes the analysis tractable and explicit while capturing the key realistic features of real-world demand functions in the D2C segment of FX markets: Quoting higher (lower) mid-prices relative to other dealers attracts sell (buy) volume, whereas dealers quoting wider spreads get less customer volume.

IA.4.2 Dealers' Optimal Price Schedules and Equilibrium in the D2C Market

At the time $t = 1$, dealer l selects the optimal price schedule $p_l(q)$ that he quotes to all customers, taking as given other dealers' price schedules (a_{-l}, λ_{-l}), as well as customers' optimal response (Lemma IA.1). Substituting (16) into (IA.7), we get that dealers' objective it to maximize

$$\begin{aligned} U_l(a_l, \lambda_l; a_{-l}, \lambda_{-l}) \;&=\; E\Bigg[\Pi_l(Q_l) \;+\; \chi_l d - 0.5\Gamma_l \chi_l^2 \\ &+\; (0.5\Gamma_l + \beta_l)(\Gamma_l + \beta_l)^{-2}\left(\mathcal{B}^{-1}\sum_{\ell=1}^{M}(\Gamma_\ell + \beta_\ell)^{-1}\Gamma_\ell\chi_\ell - \Gamma_l\chi_l\right)^2\Bigg] \end{aligned} \tag{IA.9}$$

over a_l, λ_l subject to (IA.4) and (IA.8). Thus, the dealer's objective function has three components:

- total revenues from trading with customers, as given by

$$\Pi_l(Q_l) \;=\; \sum_{c=1}^{n} p_l\big(q_{c,l}(a_l, \lambda_l; a_{-l}, \lambda_{-l})\big)\, q_{c,l}(a_l, \lambda_l; a_{-l}, \lambda_{-l})$$

- the expected utility $\chi_l d - 0.5\Gamma_l\chi_l^2$ from holding the *post-D2C trading round* inventory

$$\chi_l(a_l, \lambda_l; a_{-l}, \lambda_{-l}, x_l) \;=\; x_l \;-\; \sum_{c=1}^{n} q_{c,l}(a_l, \lambda_l; a_{-l}, \lambda_{-l})\,,$$

  where x_l is the dealer's initial inventory.

- the utility surplus from trade in the inter-dealer market

$$(0.5\Gamma_l + \beta_l)(\Gamma_l + \beta_l)^{-2}\left(\mathcal{B}^{-1}\sum_{\ell=1}^{M}(\Gamma_\ell + \beta_\ell)^{-1}\Gamma_\ell\chi_\ell - \Gamma_l\chi_l\right)^2. \tag{IA.10}$$

  The latter is determined by the deviation of dealer's post-D2C inventory χ_ℓ from the liquidity-weighted[40] average of post-D2C inventories of other dealers, as given by

$$\mathcal{B}^{-1}\sum_{\ell=1}^{M}(\Gamma_\ell+\beta_\ell)^{-1}\Gamma_\ell\chi_\ell\,. \tag{IA.11}$$

In particular, dealer l has incentives to select a quote policy in the D2C market that pushes his inventory χ_l and the other dealers' inventory (IA.11) apart.

Importantly, dealers' choice of the price schedule characteristics a_l, λ_l influence all three components, including *the gains from trades made in the D2D market.* The latter effect is particularly subtle and itself consists of two sub-components: the dealer's impact on equilibrium price and the impact of the dealer's liquidity provision in the D2C market on the inventories of all other dealers. Indeed, the more liquidity the dealer provides in the D2C market, the fewer clients will trade with other dealers, directly influencing other dealers' inventories $\chi_\ell,\ \ell\neq l$, and, hence, also influencing the surplus from trade in (IA.10) through (IA.11).

Let

$$\Psi \;=\; \left(\frac{\Gamma_1}{\mathcal{B}(\Gamma_1+\beta_1)},\cdots,\frac{\Gamma_M}{\mathcal{B}(\Gamma_M+\beta_M)}\right)^T$$

be the vector of weights that define the "aggregate risk" in the D2D market, and let also $\Psi_l \;\equiv\; \Psi-\Gamma_l\mathbf{1}_{\ell=l}$. Denote also $\boldsymbol{\delta}\equiv(\delta_l)_{l=1}^M,\ \boldsymbol{\eta}\equiv(\eta_l)_{l=1}^M$ to be the vectors of coefficients of customers' demand (see (IA.8)), and let $x=(x_\ell)_{\ell=1}^M$ be the vector of dealer initial inventories. Recall that we use $\bar\theta$ and σ_θ^2 to denote expected customer endowment and the variance of customer endowments, respectively. Evaluating the expectation in (IA.9), we get the following expression for dealers' indirect utility.

Lemma IA.2 *We have*

$$\begin{aligned}U_l(\alpha_l,\lambda_l;\alpha_{-l},\lambda_{-l}) \;&=\; x_l d-n\lambda_l^{-1}a_l(\delta_l+\eta_l\bar\theta)+n\lambda_l^{-1}\Big[(\delta_l+\eta_l\bar\theta)^2+\eta_l^2\sigma_\theta^2\Big]\\ &-0.5\Gamma_l\Big[(x_l-n\delta_l-n\bar\theta\eta_l)^2+n\sigma_\theta^2\eta_l^2\Big]\\ &+\;(0.5\Gamma_l+\beta_l)(\Gamma_l+\beta_l)^{-2}\Big[\Big(\Psi_l\cdot(x\;-\;n\boldsymbol{\delta}-n\bar\theta\boldsymbol{\eta})\Big)^2+n\sigma_\theta^2(\Psi_l\cdot\boldsymbol{\eta})^2\Big]\end{aligned}$$

and a Nash equilibrium in linear price schedules is a collection of $(\alpha_l,\lambda_l)_{l=1}^M$ such that, for all ℓ,

$$(\alpha_l,\lambda_l) \;=\; \arg\max_{\alpha_l,\lambda_l}U_l(\alpha_l,\lambda_l;\alpha_{-l},\lambda_{-l})\,. \tag{IA.12}$$

The first term, $x_l d$, is just the expected payoff from the initial inventory of the dealer. The second term, $n\lambda_l^{-1}a_l(\delta_l+\eta_l\bar\theta)$, is the amount the dealer is losing (relative to the asset payoff

[40]The weights are inversely related to price impacts β_ℓ in the D2D market.

d) from the trades in the D2C market: By quoting the price at a premium of a_l, he is selling the asset "too cheap" and his losses are proportional to the number n of customers. The third term, $n\lambda_l^{-1}\Big[(\delta_l+\eta_l\bar{\theta})^2+\eta_l^2\sigma_\theta^2\Big]$ represents the rents the dealer extracts from the customer using the increasing price schedule. These rents are proportional to the slope λ_l^{-1} : The higher the slope, the more the dealer charges for liquidity provision. The fourth term is just the expected quadratic inventory cost,

$$0.5\Gamma_l\Big[(x_l-n\delta_l-n\bar{\theta}\eta_l)^2+n\sigma_\theta^2\eta_l^2\Big] \;=\; 0.5\Gamma_l\chi_l^2\,.$$

Finally, the last term is the expected utility gain from trading in the D2D market. Importantly, contrary to the quadratic inventory cost term, the other quadratic terms are convex. This is a novel aspect of our model: Dealer market power in both segments introduces incentives for risk-taking whereby amplifying volatility of demand shocks also amplifies the rents the dealer is able to extract in the D2C and the D2D market.

In the Appendix, we write down the system of first conditions for (IA.12). In general, this system cannot be solved explicitly. However, it is possible to derive analytical approximations to its solution when dealer heterogeneity is small. Such an approximation allows us to capture the first-order effects of heterogeneity on equilibrium quantities while preserving analytical tractability.

IA.4.3 The Joint D2C-D2D Equilibrium

Everywhere in the sequel, we assume that dealers' heterogeneity is small. This technical assumption will allow us to derive approximate closed-form expressions for equilibrium objects. We start with the case when dealers do not hold any inventory and risk aversions are homogeneous, $\Gamma_l=\Gamma$. Then, the following is true.

Proposition IA.13 *If $x_l=0$ (zero initial dealer inventories) and $\Gamma_l=\Gamma$ for all $l=1,\cdots,M$, and $\bar{\theta}=0$ (zero expected customer endowment), then there exists a unique symmetric equilibrium, and the inverse of the price function slope is given by*

$$\lambda^*(\Gamma) \;=\; \frac{(M-2)\gamma-2\Gamma+\sqrt{((M-2)\gamma-2\Gamma)^2+8\gamma\Gamma(M-1)}}{2\gamma\Gamma(M-1)}\,,$$

while $a_l=0$ for all l. $\lambda^*(\Gamma)$ is decreasing in both Γ and γ and is increasing in M.

The result of Proposition IA.13 is very intuitive: Liquidity in the D2C market, as captured by the inverse slope λ^*, is determined by two forces: the willingness (and the ability) of dealers to take on risk (that is, their cost of holding inventory, Γ) and the dealers' market power, determined by their number, M. When Γ increases, or when clients are more aggressively trying to get rid of their inventory (that is, γ is large), dealers optimally widen the spread. At the same time, an increase in M creates competitive pressure on equilibrium spreads, driving D2C market liquidity up. In particular, in the competitive limit, as $M\to\infty$, we have $\lambda^* \to 1/\Gamma$, consistent with the standard competitive CAPM, whereby price sensitivity to inventory shocks equals the reciprocal of the risk aversion.

Having solved for the equilibrium in the limiting case $x_l = 0$ and $\Gamma_l = \Gamma$, we can now use Taylor approximation to compute the equilibrium for the case when initial (prior to D2C trading) dealer inventories (x_l), mean customer endowments $\bar{\theta}$, and heterogeneity in Γ_l are all sufficiently small. Let $\bar{\alpha} = M^{-1}\sum_l \alpha_l$ be the average mid quote, and recall the definition of the observable price-based liquidity mismatch, $\mathcal{A}_{mismatch}$, in the D2C market. The following Proposition characterizes the equilibrium link between prices in the two market segments

Proposition IA.14 *The price $P = P^{D2D}$ in the D2D market is linked to prices in the D2C market via*

$$\begin{aligned}\mathcal{P}^{D2D} = & \left[1 + \frac{2\gamma\lambda^*}{(2+M\gamma\lambda^*)[2+(M-1)\gamma\lambda^*]}\right](\bar{\alpha} - \mathcal{A}_{mismatch}) \\ & - (\lambda^*)^2\left[\Gamma^*(M\varphi^A_{0,\lambda} + \varphi^a_1) + \frac{M\varphi^\lambda(\varphi^a_0 - 0.5n)}{M^2 - 2M + 2}\right]\mathcal{A}_{mismatch} \\ & - \frac{2\gamma\lambda^*}{(2+M\gamma\lambda^*)[2+(M-1)\gamma\lambda^*]}d - \Gamma^*\varphi^\theta_0\bar{\theta} - \gamma\frac{1}{2+\gamma\Lambda}\lambda^*\Gamma^*\Theta \\ & + O(\|\Gamma - \Gamma^*\|^2 + \|x\|^2 + \bar{\theta}^2)\end{aligned}$$

Proposition IA.14 provides an explicit expression linking the price in the D2D market to:

- mean and *distribution of quotes* in the D2C market, as captured by $\bar{\alpha}$ and $\alpha_{mismatch}$;
- expected client endowment, $\bar{\theta}$, as well as the aggregate client endowment shock, Θ;[41]
- the fundamental, d.[42]

The most important consequence of Proposition IA.14 is that, in the presence of heterogeneity, the joint cross-sectional distribution of quotes and bid-ask spreads is directly linked to the price level in the D2D market. Namely, in addition to the average mid-quote $\bar{\alpha}$, the price also depends on $\mathcal{A}_{mismatch}$ which is effectively a spread between low bid-ask spread (large λ_l) mid-quotes, and high bid-ask spread (low λ_l) quotes. As we explain in Section IA.1.1, the term originates from (i) strategic interactions between dealers in the D2D market; (ii) dealers' bid shading in the D2C market; and (iii) non-exclusive competition and customers' optimal order splitting across dealers in the D2C market.

[41] Thus, $E[\Theta] = n\bar{\theta}$.

[42] Note that, although the coefficient on the fundamental value, d, is negative, the price $\mathcal{P}^{D2D}$ depends positively on d (at it should) due to a positive link between α_l and d.

IA.5 Derivations and Proofs for Internet Appendix

IA.5.1 Proof of Proposition IA.7

Proof of Proposition IA.7. Starting with Equation IA.1, we have

$$\mathcal{P}^{D2D} \;=\; d - \mathcal{B}^{-1}ME[(\Gamma_l+\beta_l)^{-1}\Gamma_l]E[\chi_l] \;-\; \mathcal{B}^{-1}M\,\mathrm{Cov}((\Gamma_l+\beta_l)^{-1}\Gamma_l,\chi_l)\,.$$

The first two terms in the R.H.S. of the equation below remain the same if the three cases (a), (b), and (c). Focusing on the last term, we have

$$-\mathcal{B}^{-1}M\,\mathrm{Cov}((\Gamma_l+\beta_l)^{-1}\Gamma_l,\chi_l) \overset{s}{=} \;\mathrm{Cov}(\chi_l,\Gamma_l)\,.$$

The result then follows. Q.E.D.

IA.5.2 Proof of Proposition IA.8

We will need the following technical lemmas.

Lemma IA.3 *The function* $(xy+2+\sqrt{(xy)^2+4})^{-1}$ *is jointly convex* $\mathbb{R}_+$.

Proof. Consider the function

$$f(x) \;=\; \left(x+2+\sqrt{x^2+4}\right)^{-1} \qquad\Longrightarrow\qquad f''(x) \;=\; \frac{\left(\sqrt{x^2+4}-3\right)x^2+4\left(\sqrt{x^2+4}-2\right)}{x^3\left(x^2+4\right)^{3/2}}\,.$$

We show that f'' is positive for $x>0$. Clearly, its denominator is positive for $x>0$. Consider its numerator

$$h(x) \;\equiv\; \left(\sqrt{x^2+4}-3\right)x^2+4\left(\sqrt{x^2+4}-2\right) \qquad\Longrightarrow\qquad h'(x) \;=\; 3x\left(\sqrt{x^2+4}-2\right)\,.$$

Thus, h is increasing for $x\geq 0$. Moreover, $h(0)=0$. Thus f is convex for $x>0$. It follows that the function

$$G(x,y)\equiv\left(xy+2+\sqrt{(xy)^2+4}\right)^{-1}$$

is jointly convex in (x,y).

Q.E.D.

Lemma IA.4 *The function* $\mathcal{B}(\Gamma)$ *is jointly convex in the vector* $\Gamma=(\Gamma_l)_{l=1}^{M}$.

Proof. Let

$$f(\Gamma_l,B)=\left(\Gamma_lB+2+\sqrt{((\Gamma_lB)^2+4)^{-1}}\right)^{-1}$$

Let $B_1(\Gamma^1)$ and $B_2(\Gamma^2)$ be defined implicitly by

$$\sum_l f(\Gamma_l^1, B_1) = 1/2,$$

$$\sum_l f(\Gamma_l^2, B_2) = 1/2.$$

We need to show that $B(\lambda\Gamma_l^1 + (1-\lambda)\Gamma_l^2) > \lambda B_1 + (1-\lambda)B_2$. We have

$$\sum_l f(\lambda\Gamma_l^1 + (1-\lambda)\Gamma_l^2, \lambda B_1 + (1-\lambda)B_2) > \lambda \sum_l f(\Gamma_l^1, B_1) + (1-\lambda)\sum_l f(\Gamma_l^2, B_2) = 1/2.$$

Since $f(\Gamma_l, B)$ decreases in B, the B that solves

$$\sum_l f(\lambda\Gamma_l^1 + (1-\lambda)\Gamma_l^2, \lambda B_1 + (1-\lambda)B_2) = 1/2$$

is greater than $\lambda B_1 + (1-\lambda)B_2$.

Q.E.D.

Proof of Proposition IA.8. We have $BA = 2/\mathcal{B}$. Whew aggregate liquidity $\mathcal{B}$ is given by

$$\sum_l \frac{1}{\Gamma_l + \beta_l} \;=\; \mathcal{B}\,,$$

where the equality was established in the proof of Proposition 4. By definition, a mean preserving spread is such that we move from (Γ_l) to $(\tilde{\Gamma}_l \;=\; \Gamma_l + \varepsilon_l)$ where (ε_l) has $E[\varepsilon_l] = 0$ and is mean-independent of (Γ_l). We know from Lemma IA.4 that $\mathcal{B}(\Gamma_l)$ is convex in (Γ_l). Hence,

$$\mathcal{B}(\Gamma_l) \;<\; \mathcal{B}(\tilde{\Gamma}_l)\,.$$

Furthermore, the function

$$\beta_l^{-1} \;=\; \Gamma_l\mathcal{B} - 2 + \sqrt{(\Gamma_l\mathcal{B})^2 + 4}$$

is clearly increasing and convex in $\mathcal{B}$. Thus, β_l^{-1} is convex in Γ_{-l}, and Jensen's inequality implies

$$\beta_l^{-1}(\Gamma_{-l}) \;\leq\; E[\beta_l^{-1}(\tilde{\Gamma}_{-l})]$$

that is

$$\beta_l \;\geq\; E[\beta_l^{-1}(\tilde{\Gamma}_{-l})]^{-1}\,.$$

Q.E.D.

IA.5.3 Proof of Proposition IA.9

Proof of Proposition IA.9. An increase in dispersion in b_l corresponds to an increase in dispersion in Γ_l, which by Proposition IA.8 is associated with smaller BA in D2D market.

Q.E.D.

IA.5.4 Proof of Proposition IA.10

Proof of Proposition IA.10. We number traders according to their inventories, i.e. x_l increases in l. Then α_l decreases in l. Then $\chi_l - x_l$ decreases in l. (Inventories increase for the dealer with highest midprice and decreae for the dealer with lowest midprice. They remain unchanged for other dealers.)

Then we get

$$\begin{aligned} cov(\chi_l - x_l, \Gamma_l) &\overset{s}{=} cov(\alpha_l, \Gamma_l) \\ &\overset{s}{=} cov(\alpha_l, b_l). \end{aligned}$$

Moreover, for q large enough we get $cov(\chi_l - x_l, \Gamma_l) \approx cov(\chi_l, \Gamma_l)$.

Q.E.D.

IA.5.5 Proof of Proposition IA.11

Proof of Proposition IA.11. Suppose that $\mathcal{A}_{mismatch} \overset{s}{=} Y_{mismatch}$.

Starting with Equation IA.1, we have

$$\mathcal{P}^{D2D} \; = \; d - \mathcal{B}^{-1} M E[(\Gamma_l + \beta_l)^{-1}\Gamma_l] E[\chi_l] \; - \; \mathcal{B}^{-1} M \,\mathrm{Cov}((\Gamma_l + \beta_l)^{-1}\Gamma_l, \chi_l)\,.$$

The first two terms in the R.H.S. of the equation below remain the same if the three cases (a), (b), and (c). Focusing on the last term, we have

$$-\mathcal{B}^{-1} M \,\mathrm{Cov}((\Gamma_l + \beta_l)^{-1}\Gamma_l, \chi_l) \overset{s}{=} \; - \mathrm{Cov}(\chi_l, \Gamma_l) = -Y_{mismatch}\,.$$

Thus,

$$-\mathcal{B}^{-1} M \,\mathrm{Cov}((\Gamma_l + \beta_l)^{-1}\Gamma_l, \chi_l) \;\; = \;\; -\mathcal{A}_{mismatch}\,.$$

Thus, the $\mathcal{P}^{D2D}$ price has the reverse ordering of the $\mathcal{A}_{mismatch}$ under the conditions of the proposition. To complete the proof, we need to show that

$$\mathcal{A}_{mismatch} \overset{s}{=} Y_{mismatch}\,.$$

We prove this result in the next three lemmas. Q.E.D.

Lemma IA.5 *Suppose that*

$$\lambda_l \;=\; a_\lambda \;+\; b_\lambda \Gamma_l \;+\; O(\|\Gamma - \Gamma^*\|^2 + \|x\|^2 + \bar\theta^2)$$

and

$$\chi_l \;=\; a_\chi \;+\; b_\chi \alpha_l \;+\; O(\|\Gamma - \Gamma^*\|^2 + \|x\|^2 + \bar\theta^2)$$

with $b_\lambda \;<\; 0 \;<\; b_\chi$. *Then,*

$$\mathcal{A}_{mismatch} \stackrel{s}{=} Y_{mismatch}\,.$$

Proof.

$$\begin{aligned} Y_{mismatch} =\;& E\left[(\Gamma_l - \Gamma^*)(\chi_l - \chi^*)\right] \\ \stackrel{s}{=}\;& -E\left[(\lambda_l - \lambda^*)(\alpha_l - \alpha^*)\right] \;\stackrel{s}{=}\; E\left[(\lambda_l^{-1} - (\lambda^{-1})^*)(\alpha_l - \alpha^*)\right] \;=\; E\left[(b_l - b^*)(\alpha_l - \alpha^*)\right] \\ =\;& \mathcal{A}_{mismatch}\,. \end{aligned}$$

Q.E.D.

Lemma IA.6 *Suppose that x_l are sufficiently small and Γ_l are sufficiently close to Γ^* and that $\bar\theta$ is sufficiently small. Then, there exists a unique equilibrium with*[43]

$$\lambda_l \;=\; \lambda^*(\Gamma^*) \;+\; \Phi^\Gamma(\Gamma^*)(\Gamma_l - \Gamma^*) \;+\; O(\|\Gamma - \Gamma^*\|^2 + \|x\|^2 + \bar\theta^2)\,,$$

where $\Phi^\Gamma(x)$ is negative and increasing in M for $x > 0$ and $M > 1$. Furthermore,

$$\begin{aligned} a_l =\;& \Phi_0^x\, x_l \;+\; \Phi_0^X\, \bar X \;+\; \Phi_0^\Gamma\, X_{mismatch} \;+\; \Phi_0^\theta\, \bar\theta \;+\; (\Gamma_l - \Gamma^*)[\Phi_1^x\, x_l + \Phi_1^X\, \bar X + \Phi_1^\theta\, \bar\theta] \\ &+\; O((\|\Gamma - \Gamma^*\|^2 + \|x\|^2 + \bar\theta^2)^{3/2})\,, \end{aligned}$$

where for some coefficients $\Phi_0^x, \Phi_0^X, \Phi_0^\Gamma, \Phi_0^\theta, \Phi_1^x, \Phi_1^X, \Phi_1^\Gamma, \Phi_1^\theta$ with $\Phi_0^x \;>\; 0$.

Proof of Lemma IA.6. Recall that

$$\boldsymbol\delta = \frac{[2+\gamma\Lambda]\mathbf{a} \;-\; \gamma A \boldsymbol\lambda}{4 + 2\gamma\Lambda}; \qquad \boldsymbol\eta = -\frac{\gamma\boldsymbol\lambda}{2+\gamma\Lambda};$$

$$\partial_{a_l}\delta_j = \frac{\mathbb{1}_l(j)[2+\gamma\Lambda] \;-\; \gamma\lambda_l}{4+2\gamma\Lambda}; \qquad \partial_{a_l}\boldsymbol\delta = \frac{1}{4+2\gamma\Lambda}\Big[(2+\gamma\Lambda)\boldsymbol{e}_l - \gamma\boldsymbol\lambda\Big]\,,$$

where $\boldsymbol{e}_l$ is the ith coordinate vector. Consider the first FOC:

$$\begin{aligned} 0 =\;& n(\partial_{a_l}\delta_l)\Big[\Gamma_l x_l - \lambda_l^{-1} a_l\Big] + n\lambda_l^{-1}\Big[(2 - n\Gamma_l\lambda_l)(\partial_{a_l}\delta_l) - 1\Big](\delta_l + \eta_l\bar\theta) \\ &-\; n(\Gamma_l + 2\beta_l)(\Gamma_l+\beta_l)^{-2}\Big(\Psi_l\cdot\partial_{a_l}\boldsymbol\delta\Big)\Big(\Psi_l\cdot(x \;-\; n\boldsymbol\delta - n\bar\theta\boldsymbol\eta)\Big)\,. \end{aligned}$$

[43] As usual, we use $\|x\| = \left(\sum_l x_l^2\right)^{1/2}$ to denote the Euclidean norm.

We cancel n and multiply by λ_l and get

$$\begin{aligned}
0 = \ & (\partial_{a_l}\delta_l)\Big[\lambda_l\Gamma_l x_l - a_l\Big] + \Big[(2-n\Gamma_l\lambda_l)(\partial_{a_l}\delta_l) - 1\Big](\delta_l + \eta_l\bar{\theta}) \\
& - \ \lambda_l \frac{\Gamma_l + 2\beta_l}{(\Gamma_l+\beta_l)^2}\Big(\Psi_l \cdot \partial_{a_l}\boldsymbol{\delta}\Big)\Big(\Psi_l \cdot (x \ - \ n\boldsymbol{\delta} - n\bar{\theta}\boldsymbol{\eta})\Big) \\
= \ & (\partial_{a_l}\delta_l)\Big[\lambda_l\Gamma_l x_l - a_l\Big] + \Big[(2-n\Gamma_l\lambda_l)(\partial_{a_l}\delta_l) - 1\Big]\frac{[2+\gamma\Lambda]a_l \ - \ \gamma(A+2\bar{\theta})\lambda_l}{4+2\gamma\Lambda} \\
& - \ \lambda_l \frac{\Gamma_l + 2\beta_l}{(\Gamma_l+\beta_l)^2}\Big(\Psi_l \cdot \partial_{a_l}\boldsymbol{\delta}\Big)\left(\Psi_l \cdot x \ - \ n\Psi_l \cdot \frac{[2+\gamma\Lambda]\mathbf{a} \ - \ \gamma(A+2\bar{\theta})\boldsymbol{\lambda}}{4+2\gamma\Lambda}\right) \\
= \ & (\partial_{a_l}\delta_l)\Big[\lambda_l\Gamma_l x_l - a_l\Big] + \Big[(2-n\Gamma_l\lambda_l)(\partial_{a_l}\delta_l) - 1\Big]\frac{[2+\gamma\Lambda]a_l \ - \ \gamma(A+2\bar{\theta})\lambda_l}{4+2\gamma\Lambda} \\
& - \ \lambda_l \frac{\Gamma_l + 2\beta_l}{(\Gamma_l+\beta_l)^2}\Big(\Psi_l \cdot \partial_{a_l}\boldsymbol{\delta}\Big)\left(\tilde{\chi} \ - \ \frac{n}{2}\tilde{A} \ + \ n\frac{\gamma(A+2\bar{\theta})}{4+2\gamma\Lambda}\tilde{\Lambda} \ - \ \Gamma_l x_l \ + \ \frac{n\Gamma_l}{2}a_l \ - \ \frac{\gamma(A+2\bar{\theta})n\Gamma_l}{4+2\gamma\Lambda}\lambda_l\right) \\
= \ & -\frac{1}{2}\left[n\Gamma_l\lambda_l(\partial_{a_l}\delta_l) \ + \ 1 \ + \ n\lambda_l\Gamma_l\frac{\Gamma_l + 2\beta_l}{(\Gamma_l+\beta_l)^2}\Big(\Psi_l \cdot \partial_{a_l}\boldsymbol{\delta}\Big)\right]a_l \\
+ \ & \lambda_l\Gamma_l\left[(\partial_{a_l}\delta_l) \ + \ \frac{\Gamma_l + 2\beta_l}{(\Gamma_l+\beta_l)^2}\Big(\Psi_l \cdot \partial_{a_l}\boldsymbol{\delta}\Big)\right]x_l \ - \ \lambda_l\frac{\Gamma_l + 2\beta_l}{(\Gamma_l+\beta_l)^2}\Big(\Psi_l \cdot \partial_{a_l}\boldsymbol{\delta}\Big)\tilde{\chi} \\
& + \ \frac{n}{2}\lambda_l\frac{\Gamma_l + 2\beta_l}{(\Gamma_l+\beta_l)^2}\Big(\Psi_l \cdot \partial_{a_l}\boldsymbol{\delta}\Big)\tilde{A} \\
& - \ \lambda_l\left[n\frac{\Gamma_l + 2\beta_l}{(\Gamma_l+\beta_l)^2}\Big(\Psi_l \cdot \partial_{a_l}\boldsymbol{\delta}\Big)\Big(\tilde{\Lambda} \ - \ \Gamma_l\lambda_l\Big) \ + \ (2-n\Gamma_l\lambda_l)(\partial_{a_l}\delta_l) \ - \ 1\right]\frac{\gamma(A+2\bar{\theta})}{4+2\gamma\Lambda} \\
= \ & -\frac{1}{2}\left[1 \ + \ nc_{0x_l}\right]a_l \ + \ c_{0x_l}x_l \ - \ \lambda_l\frac{\Gamma_l + 2\beta_l}{(\Gamma_l+\beta_l)^2}\Big(\Psi_l \cdot \partial_{a_l}\boldsymbol{\delta}\Big)\Big(\tilde{\chi} \ - \ \frac{n}{2}\tilde{A}\Big) \\
& - \ \frac{\gamma\lambda_l}{4+2\gamma\Lambda}\left[n\frac{\Gamma_l + 2\beta_l}{(\Gamma_l+\beta_l)^2}\Big(\Psi_l \cdot \partial_{a_l}\boldsymbol{\delta}\Big)\tilde{\Lambda} \ - \ nc_{0x_l} \ + \ 2(\partial_{a_l}\delta_l) \ - \ 1\right](A+2\bar{\theta}),
\end{aligned}$$

where

$$\begin{aligned}
c_{0x_l} = \ & \lambda_l\Gamma_l\left[(\partial_{a_l}\delta_l) \ + \ \frac{\Gamma_l + 2\beta_l}{(\Gamma_l+\beta_l)^2}\Big(\Psi_l \cdot \partial_{a_l}\boldsymbol{\delta}\Big)\right] \\
= \ & \lambda_l\Gamma_l\left[\frac{[2+\gamma\Lambda] \ - \ \gamma\lambda_l}{4+2\gamma\Lambda} \ + \ \frac{\Gamma_l + 2\beta_l}{(\Gamma_l+\beta_l)^2}\left(\frac{\Gamma_l}{2}\left(\frac{1}{\mathcal{B}(\Gamma_l+\beta_l)} - 1\right) \ - \ \frac{\gamma(\tilde{\Lambda} - \Gamma_l\lambda_l)}{4+2\gamma\Lambda}\right)\right].
\end{aligned}$$

It follows that

$$\begin{aligned}
a_l = \ & \frac{2c_{0x_l}}{1 \ + \ nc_{0x_l}}x_l \ - \ \frac{2\lambda_l}{1 \ + \ nc_{0x_l}}\frac{\Gamma_l + 2\beta_l}{(\Gamma_l+\beta_l)^2}\Big(\Psi_l \cdot \partial_{a_l}\boldsymbol{\delta}\Big)\Big(\tilde{\chi} \ - \ \frac{n}{2}\tilde{A}\Big) \\
& - \ \frac{2}{1 \ + \ nc_{0x_l}}\frac{\gamma\lambda_l}{4+2\gamma\Lambda}\left[n\frac{\Gamma_l + 2\beta_l}{(\Gamma_l+\beta_l)^2}\Big(\Psi_l \cdot \partial_{a_l}\boldsymbol{\delta}\Big)\tilde{\Lambda} \ - \ nc_{0x_l} \ + \ 2(\partial_{a_l}\delta_l) \ - \ 1\right](A+2\bar{\theta}),
\end{aligned}$$

We now write the Taylor series expansion of this FOC. We shall use the following lemma:

Lemma IA.7 *When the dispersion of $\Gamma_{i,t}$ is sufficiently small, we have*

$$\begin{aligned}
\frac{\Gamma_{i,t}}{\mathcal{B}_t(\Gamma_{i,t}+\beta_{i,t})} &\approx \frac{\Gamma_t^*}{M} + \frac{1}{M^2-2M+2}(\Gamma_{i,t}-\Gamma_t^*) \\
\frac{1}{\Gamma_{i,t}+\beta_{i,t}} &\approx \frac{M-2}{\Gamma_t^*(M-1)} - \frac{(M-2)^2}{(M^2-2M+2)(\Gamma_t^*)^2}(\Gamma_{i,t}-\Gamma_t^*) \\
\frac{\beta_{i,t}}{\Gamma_{i,t}+\beta_{i,t}} &\approx \frac{1}{M-1} - \frac{M(M-2)}{(M^2-2M+2)(M-1)\Gamma_t^*}(\Gamma_{i,t}-\Gamma_t^*)\,.
\end{aligned}$$

It thus follows that

$$\begin{aligned}
\frac{\Gamma_{i,t}}{\Gamma_{i,t}+\beta_{i,t}} &\approx \frac{M-2}{M-1} + \frac{M(M-2)}{(M^2-2M+2)(M-1)\Gamma_t^*}(\Gamma_{i,t}-\Gamma_t^*) \\
\frac{\Gamma_{i,t}+2\beta_{i,t}}{(\Gamma_{i,t}+\beta_{i,t})^2} &= \frac{\Gamma_{i,t}}{(\Gamma_{i,t}+\beta_{i,t})^2} + \frac{2\beta_{i,t}}{(\Gamma_{i,t}+\beta_{i,t})^2} \\
&\approx \frac{M-2}{M-1}\left[1 + \frac{M}{(M^2-2M+2)\Gamma_t^*}(\Gamma_{i,t}-\Gamma_t^*)\right]\frac{M-2}{\Gamma_t^*(M-1)}\left[1 - \frac{(M-1)(M-2)}{(M^2-2M+2)\Gamma^*}(\Gamma_{i,t}-\Gamma_t^*)\right] \\
&+ \frac{2}{M-1}\left[1 - \frac{M(M-2)}{(M^2-2M+2)\Gamma_t^*}(\Gamma_{i,t}-\Gamma_t^*)\right]\frac{M-2}{\Gamma_t^*(M-1)}\left[1 - \frac{(M-1)(M-2)}{(M^2-2M+2)\Gamma_t^*}(\Gamma_{i,t}-\Gamma_t^*)\right] \\
&= \frac{(M-2)^2}{\Gamma_t^*(M-1)^2}\left[1 + \frac{M}{(M^2-2M+2)\Gamma_t^*}(\Gamma_{i,t}-\Gamma_t^*) - \frac{(M-1)(M-2)}{(M^2-2M+2)\Gamma^*}(\Gamma_{i,t}-\Gamma_t^*)\right] \\
&+ \frac{2(M-2)}{\Gamma_t^*(M-1)^2}\left[1 - \frac{M(M-2)}{(M^2-2M+2)\Gamma_t^*}(\Gamma_{i,t}-\Gamma_t^*) - \frac{(M-1)(M-2)}{(M^2-2M+2)\Gamma_t^*}(\Gamma_{i,t}-\Gamma_t^*)\right] \\
&= \frac{M(M-2)}{\Gamma_t^*(M-1)^2}\left[1 - \frac{M(M-2)}{(M^2-2M+2)\Gamma_t^*}(\Gamma_{i,t}-\Gamma_t^*)\right].
\end{aligned}$$

Moreover,

$$\lambda_l\Gamma_l \approx \lambda^*\Gamma^* + [\Phi^\Gamma(\Gamma^*)\Gamma^* + \lambda^*](\Gamma_l - \Gamma^*) \qquad \text{and} \qquad \tilde{\Lambda} \approx \lambda^*\Gamma^*$$

and

$$\begin{aligned}
\tilde{A} &= \sum_{l=1}^{M} \frac{\Gamma_l}{\mathcal{B}(\Gamma_l+\beta_l)} a_j \approx \frac{\Gamma^*}{M}A + \frac{1}{M^2-2M+2}A^\Gamma \\
\tilde{\chi} &= \sum_{l=1}^{M} \frac{\Gamma_l}{\mathcal{B}(\Gamma_l+\beta_l)} x_l \approx \frac{\Gamma^*}{M}X + \frac{1}{M^2-2M+2}X_{mismatch}\,,
\end{aligned}$$

where

$$A^\Gamma = \sum_{l=1}^{M}(\Gamma_l - \Gamma^*)a_j \qquad \text{and} \qquad X_{mismatch} = \sum_{l=1}^{M}(\Gamma_l-\Gamma^*)x_l$$

Using the lemma above, to a first-order approximation, we have

$$
\begin{aligned}
a_l = \; & [k_{0x} + k_{1x}(\Gamma_l - \Gamma^*)]x_l \; - \; [k_{0x} + k_{1x}(\Gamma_l - \Gamma^*)]\left(\tilde{\chi} \; - \; \frac{n}{2}\tilde{A}\right) \; - \; [k_{0x} + k_{1x}(\Gamma_l - \Gamma^*)](A + 2\bar{\theta}) \\
= \; & [k_{0x} + k_{1x}(\Gamma_l - \Gamma^*)]x_l \; - \; [k_{0\theta} + k_{1\theta}(\Gamma_l - \Gamma^*)](A + 2\bar{\theta}) \; - \; [k_{0X} + k_{1X}(\Gamma_l - \Gamma^*)]\frac{\Gamma^*}{M}\left(X \; - \; \frac{n}{2}A\right) \\
- \; & [k_{0X} + k_{1X}(\Gamma_l - \Gamma^*)]\frac{1}{M^2 - 2M + 2}\left(X_{mismatch} \; - \; \frac{n}{2}A^{\Gamma}\right) \\
= \; & [k_{0x} + k_{1x}(\Gamma_l - \Gamma^*)]x_l \; - \; [k_{0\theta} + k_{1\theta}(\Gamma_l - \Gamma^*)](A + 2\bar{\theta}) \; - \; [k_{0X} + k_{1X}(\Gamma_l - \Gamma^*)]\frac{\Gamma^*}{M}\left(X \; - \; \frac{n}{2}A\right) \\
- \; & \frac{k_{0X}}{M^2 - 2M + 2}\left(X_{mismatch} \; - \; \frac{n}{2}A^{\Gamma}\right) \\
= \; & [k_{0x} + k_{1x}(\Gamma_l - \Gamma^*)]x_l \; - \; [k_{0X} + k_{1X}(\Gamma_l - \Gamma^*)]\frac{\Gamma^*}{M}X \; - \; \frac{k_{0X}}{M^2 - 2M + 2}X_{mismatch} \\
- \; & \left[\frac{2Mk_{0\theta} - nk_{0X}\Gamma^*}{2M} + \frac{2Mk_{1\theta} - nk_{1X}\Gamma^*}{2M}(\Gamma_l - \Gamma^*)\right] A \; + \; \frac{n}{2}\frac{k_{0X}}{M^2 - 2M + 2}A^{\Gamma} \\
- \; & [k_{0\theta} + k_{1\theta}(\Gamma_l - \Gamma^*)]2\bar{\theta}\,,
\end{aligned}
$$

Summing over i we get

$$
\begin{aligned}
A^{\Gamma} = \; & k_{0x}X_{mismatch} \\
A = \; & k_{0x}X + k_{1x}X_{mismatch} \; - \; k_{0X}\Gamma^* X \; - \; \frac{Mk_{0X}}{M^2 - 2M + 2}X_{mismatch} \\
& - \; \frac{2Mk_{0\theta} - nk_{0X}\Gamma^*}{2}A \; + \; \frac{n}{2}\frac{Mk_{0X}}{M^2 - 2M + 2}A^{\Gamma} \; - \; 2Mk_{0\theta}\bar{\theta}\,.
\end{aligned}
$$

Thus,

$$
A = \left[1 \; + \; \frac{2Mk_{0\theta} - nk_{0X}\Gamma^*}{2}\right]^{-1}\left(\left[k_{0x} - k_{0X}\Gamma^*\right]X \; + \; \left[k_{1x} - \frac{(2 - nk_{0x})Mk_{0X}}{2(M^2 - 2M + 2)}\right]X_{mismatch} \; - \; 2Mk_{0}\right.
$$

Substituting into the equation for a_l we obtain

$$\begin{aligned}
a_l = \; & [k_{0x} + k_{1x}(\Gamma_l - \Gamma^*)]x_l \; - \; [k_{0X} + k_{1X}(\Gamma_l - \Gamma^*)]\frac{\Gamma^*}{M}X \\
- \; & \left[\frac{2Mk_{0\theta} - nk_{0X}\Gamma^*}{M} + \frac{2Mk_{1\theta} - nk_{1X}\Gamma^*}{M}(\Gamma_l - \Gamma^*)\right] \frac{k_{0x} - k_{0X}\Gamma^*}{2 + 2Mk_{0\theta} - nk_{0X}\Gamma^*}X \\
- \; & \left(1 \; - \; \frac{n}{2}k_{0x}\right) \frac{k_{0X}}{M^2 - 2M + 2}X_{mismatch} \\
- \; & \left[\frac{2Mk_{0\theta} - nk_{0X}\Gamma^*}{M} + \frac{2Mk_{1\theta} - nk_{1X}\Gamma^*}{M}(\Gamma_l - \Gamma^*)\right] \frac{k_{1x} - \frac{(2-nk_{0x})Mk_{0X}}{2(M^2-2M+2)}}{2 + 2Mk_{0\theta} - nk_{0X}\Gamma^*}X_{mismatch} \\
+ \; & \left[\frac{2Mk_{0\theta} - nk_{0X}\Gamma^*}{M} + \frac{2Mk_{1\theta} - nk_{1X}\Gamma^*}{M}(\Gamma_l - \Gamma^*)\right] \frac{2Mk_{0\theta}}{2 + 2Mk_{0\theta} - nk_{0X}\Gamma^*}\bar{\theta} \\
- \; & 2[k_{0\theta} + k_{1\theta}(\Gamma_l - \Gamma^*)]\bar{\theta}\,.
\end{aligned}$$

It follows that

$$\begin{aligned}
a_l = \; & \Phi_0^x\, x_l \; + \; \Phi_0^X\, X \; + \; \Phi_0^\Gamma\, X_{mismatch} \; + \; \Phi_0^\theta\, \bar{\theta} \; + \; (\Gamma_l - \Gamma^*)[\Phi_1^x\, x_l + \Phi_1^X\, X + \Phi_1^\theta\, \bar{\theta}] \\
+ \; & O((\|\Gamma - \Gamma^*\|^2 + \|x\|^2 + \bar{\theta}^2)^{3/2})\,,
\end{aligned}$$

with

$$\begin{aligned}
\Phi_0^x = \; & k_{0x}; \qquad\qquad \Phi_1^x = \; k_{1x} \\
\Phi_0^X = \; & - \; \left[\frac{2k_{0X}\Gamma^* \; + \; k_{0x}\,(2Mk_{0\theta} - nk_{0X}\Gamma^*)}{2 + 2Mk_{0\theta} - nk_{0X}\Gamma^*}\right] \frac{1}{M}; \\
\Phi_1^X = \; & - \; \left[\frac{k_{1X}\Gamma^*(2 + 2Mk_{0\theta}) \; + \; 2Mk_{1\theta}(k_{0x} - k_{0X}\Gamma^*) \; - \; nk_{1X}k_{0x}\Gamma^*}{2 + 2Mk_{0\theta} - nk_{0X}\Gamma^*}\right] \frac{1}{M} \\
\Phi_0^\Gamma = \; & - \; \left[\frac{k_{0X}(2 - nk_{0x})}{M^2 - 2M + 2} \; + \; \frac{k_{1x}(2Mk_{0\theta} - nk_{0X}\Gamma^*)}{M}\right] \frac{1}{2 + 2Mk_{0\theta} - nk_{0X}\Gamma^*} \\
\Phi_0^\theta = \; & \frac{-4k_{0\theta}}{2 + 2Mk_{0\theta} - nk_{0X}\Gamma^*} \\
\Phi_1^\theta = \; & - \; \frac{2\,[2k_{1\theta} \; - \; n\,(k_{0X}k_{1\theta} \; - \; k_{0\theta}k_{1X})\,\Gamma^*]}{2 + 2Mk_{0\theta} - nk_{0X}\Gamma^*}\,.
\end{aligned}$$

Q.E.D.

Lemma IA.8 *After the D2C trading round, dealer inventories become*

$$\begin{aligned}
\chi_l \approx \; & -(\varphi_0^a - 0.5n)\lambda^*\alpha_l \; - \; (\varphi_0^A - 0.5n\gamma\eta_0^*\lambda^*)M\lambda^*\bar{\alpha} \; - \; M\lambda^*[(\varphi_0^A - 0.5n\gamma\eta_0^*\lambda^*) + \varphi_{0,\lambda}^A\lambda^*]\hat{\alpha} \\
+ \; & \lambda^*[\varphi_0^a + M\varphi_0^A + n\eta_0^*]d \; + \; \varphi_0^\theta\bar{\theta} \; - \; \gamma\eta_0^*\lambda^*\Theta \\
+ \; & (\lambda_l - \lambda^*)\left[-(\varphi_0^a + \varphi_1^a\lambda^* - 0.5n)\alpha_l + M\lambda^*(0.5n\gamma\eta_0^* - \varphi_1^A)\bar{\alpha}\right] \\
+ \; & (\lambda_l - \lambda^*)[(\varphi_0^a + n\eta_0^* + \lambda^*\varphi_1^a + M\lambda^*\varphi_1^A)d - \gamma\eta_0^*\Theta + \varphi_1^\theta\bar{\theta}]\,,
\end{aligned}$$

where

$$\bar{\alpha} \;=\; \frac{1}{M}\sum_l \alpha_i; \quad \alpha_{mismatch} \;=\; \frac{1}{\Lambda}\sum_l (\lambda_i - \lambda^*)\alpha_i; \quad \text{and} \quad \eta_0^* \;=\; -\frac{1}{2+\gamma\Lambda}\,.$$

Proof of Lemma IA.8.

$$\begin{aligned}
\chi_l = \;& x_l \;-\; \sum_j (\delta_l \;+\; \eta_l \theta_{j,t-}) \;=\; x_l \;-\; 0.5 n a_l \;-\; \eta_l(\Theta_t + 0.5nA) \\
=\;& \varphi_0^a a_l \;+\; \varphi_0^A A \;+\; \varphi_{0,\lambda}^A A^\lambda \;+\; \varphi_0^\theta \bar{\theta} \;+\; (\lambda_l - \lambda^*)[\varphi_1^a a_l + \varphi_1^A A + \varphi_{1,\lambda}^A A^\lambda + \varphi_1^\theta \bar{\theta}] \\
&-\; 0.5 n a_i \;-\; \eta_i(\Theta + 0.5nA) \\
=\;& (\varphi_0^a - 0.5n) a_l \;+\; (\varphi_0^A - 0.5n\eta_i) A \;+\; \varphi_{0,\lambda}^A A^\lambda \;+\; \varphi_0^\theta \bar{\theta} \;-\; \eta_i \Theta \;+\; (\lambda_l - \lambda^*)[\varphi_1^a a_l + \varphi_1^A A + \varphi_1^\theta \bar{\theta}]\,.
\end{aligned}$$

For simplicity, we assume that the total customer inventory shocks $\Theta_t = \sum_l \theta_{j,t}$ are i.i.d. over time. We rerwite both a_l and A in terms of mid-prices:

$$\begin{aligned}
a_l =\;& \lambda_l(d - \alpha_l) \\
A =\;& \sum_l a_j \;=\; d\sum_l \lambda_l \;-\; \sum_l \lambda_l \alpha_j \;=\; d\Lambda \;-\; \sum_l \lambda_l \alpha_j \\
=\;& M\lambda^*\left[d \;-\; (\alpha_{mismatch} + \bar{\alpha})\right] \\
A^\lambda =\;& \sum_l (\lambda_l - \lambda^*) a_j \;=\; d\sum_l (\lambda_l - \lambda^*)\lambda_l \;-\; \sum_l (\lambda_l - \lambda^*)\lambda_l \alpha_j \;\approx\; -\,\lambda^* \sum_l (\lambda_l - \lambda^*)\alpha_j \\
=\;& -\,M(\lambda^*)^2 \hat{\alpha}\,.
\end{aligned}$$

Thus,

$$\begin{aligned}
\tilde{\chi}_l =\ & (\varphi_0^a - 0.5n)\lambda_l(d-\alpha_l) \ + \ (\varphi_0^A - 0.5n\eta_i)M\lambda^*\,[d \ - \ (\alpha_{mismatch} + \bar{\alpha})] \ - \ \varphi_{0,\lambda}^A M(\lambda^*)^2\hat{\alpha} \\
& + \ \varphi_0^\theta\bar{\theta} \ - \ \eta_i\Theta \ + \ (\lambda_l - \lambda^*)[\varphi_1^a\lambda_l(d-\alpha_l) + \varphi_1^A M\lambda^*\,[d \ - \ (\alpha_{mismatch} + \bar{\alpha})] + \varphi_1^\theta\bar{\theta}] \\
=\ & -(\varphi_0^a - 0.5n)\lambda_l\alpha_l \ - \ (\varphi_0^A - 0.5n\gamma\lambda_l\eta_0^*)M\lambda^*\bar{\alpha} \ - \ M\lambda^*[(\varphi_0^A - 0.5n\gamma\lambda_l\eta_0^*) + \varphi_{0,\lambda}^A\lambda^*]\hat{\alpha} \\
& + \ [(\varphi_0^a - 0.5n)\lambda_l \ + \ (\varphi_0^A - 0.5n\gamma\lambda_l\eta_0^*)M\lambda^*]d \ + \ \varphi_0^\theta\bar{\theta} \ - \ \gamma\lambda_l\eta_0^*\Theta \\
& + \ (\lambda_l - \lambda^*)[-\varphi_1^a\lambda_l\alpha_l - \varphi_1^A M\lambda^*(\alpha_{mismatch} + \bar{\alpha}) + (\varphi_1^a\lambda_l + \varphi_1^A M\lambda^*)d + \varphi_1^\theta\bar{\theta}] \\
\approx\ & -(\varphi_0^a - 0.5n)\lambda^*\alpha_l \ - \ (\varphi_0^A - 0.5n\gamma\eta_0^*\lambda^*)M\lambda^*\bar{\alpha} \ - \ M\lambda^*[(\varphi_0^A - 0.5n\gamma\eta_0^*\lambda^*) + \varphi_{0,\lambda}^A\lambda^*]\hat{\alpha} \\
& + \ [(\varphi_0^a - 0.5n)\lambda^* \ + \ (\varphi_0^A - 0.5n\gamma\eta_0^*\lambda^*)M\lambda^*]d \ + \ \varphi_0^\theta\bar{\theta} \ - \ \gamma\eta_0^*\lambda^*\Theta \\
& -(\varphi_0^a - 0.5n)(\lambda_l - \lambda^*)\alpha_l \ + \ 0.5n\gamma\eta_0^*(\lambda_l - \lambda^*)M\lambda^*\bar{\alpha} \\
& + \ [(\varphi_0^a - 0.5n)(\lambda_l - \lambda^*) \ - \ 0.5n\gamma\eta_0^*(\lambda_l - \lambda^*)M\lambda^*]d \ - \ \gamma\eta_0^*(\lambda_l - \lambda^*)\Theta \\
& + \ (\lambda_l - \lambda^*)[-\varphi_1^a\lambda^*\alpha_l - \varphi_1^A M\lambda^*\bar{\alpha} + \lambda^*(\varphi_1^a + M\varphi_1^A)d + \varphi_1^\theta\bar{\theta}] \\
\approx\ & -(\varphi_0^a - 0.5n)\lambda^*\alpha_l \ - \ (\varphi_0^A - 0.5n\gamma\eta_0^*\lambda^*)M\lambda^*\bar{\alpha} \ - \ M\lambda^*[(\varphi_0^A - 0.5n\gamma\eta_0^*\lambda^*) + \varphi_{0,\lambda}^A\lambda^*]\hat{\alpha} \\
& + \ [(\varphi_0^a - 0.5n)\lambda^* \ + \ (\varphi_0^A - 0.5n\gamma\eta_0^*\lambda^*)M\lambda^*]d \ + \ \varphi_0^\theta\bar{\theta} \ - \ \gamma\eta_0^*\lambda^*\Theta \\
& + (\lambda_l - \lambda^*)\,[-(\varphi_0^a - 0.5n)\alpha_l \ + \ 0.5nM\gamma\eta_0^*\lambda^*\bar{\alpha}] \\
& + \ (\lambda_l - \lambda^*)[(\varphi_0^a - 0.5n) \ - \ 0.5nM\gamma\eta_0^*\lambda^*]d \ - \ (\lambda_l - \lambda^*)\gamma\eta_0^*\Theta \\
& + \ (\lambda_l - \lambda^*)[-\varphi_1^a\lambda^*\alpha_l - \varphi_1^A M\lambda^*\bar{\alpha} + \lambda^*(\varphi_1^a + M\varphi_1^A)d + \varphi_1^\theta\bar{\theta}]\,,
\end{aligned}$$

where we used the definition

$$\eta_i \ = \ -\frac{\gamma\lambda_l}{2+\gamma\Lambda} \ = \ \gamma\eta_0^*\lambda_l \qquad \Longrightarrow \qquad 1 \ + \ M\gamma\eta_0^*\lambda^* \ = \ - \ 2\eta_0^*\,.$$

Thus,

$$\begin{aligned}
\tilde{\chi}_l \approx\ & -(\varphi_0^a - 0.5n)\lambda^*\alpha_l \ - \ (\varphi_0^A - 0.5n\gamma\eta_0^*\lambda^*)M\lambda^*\bar{\alpha} \ - \ M\lambda^*[(\varphi_0^A - 0.5n\gamma\eta_0^*\lambda^*) + \varphi_{0,\lambda}^A\lambda^*]\hat{\alpha} \\
& + \ \lambda^*[\varphi_0^a + M\varphi_0^A + n\eta_0^*]d \ + \ \varphi_0^\theta\bar{\theta} \ - \ \gamma\eta_0^*\lambda^*\Theta \\
& + (\lambda_l - \lambda^*)\,[-(\varphi_0^a - 0.5n)\alpha_l + 0.5nM\gamma\eta_0^*\lambda^*\bar{\alpha}] \\
& + \ (\lambda_l - \lambda^*)[(\varphi_0^a + n\eta_0^*)d - \gamma\eta_0^*\Theta] \\
& + \ (\lambda_l - \lambda^*)[-\varphi_1^a\lambda^*\alpha_l - \varphi_1^A M\lambda^*\bar{\alpha} + \lambda^*(\varphi_1^a + M\varphi_1^A)d + \varphi_1^\theta\bar{\theta}]\,,
\end{aligned}$$

Q.E.D.

IA.5.6 Proof of Proposition IA.12

Proof of Proposition IA.12. The proof follows along the same lines as that of Proposition IA.8 so we skip some details.

By Proposition 4, our goal is to show that aggregate liquidity $\mathcal{B}$ is monotone increasing in spread dispersion. Proposition IA.15 shows that, for small heterogeneity in risk aversion, there is a one-to-one mapping between λ_l and Γ_l. Therefore, for small heterogeneity in risk aversion, proving Proposition IA.12 is equivalent to proving that $\mathcal{B}$ is monotone increasing in the dispersion in risk aversions. This follows directly because $\mathcal{B}$ is a convex function of risk aversions (see Lemma IA.4).

Q.E.D.

IA.5.7 Proof of Lemma IA.1

Proof of Lemma IA.1. The first order condition is

$$-\alpha_l - 2\lambda_l^{-1}q_l + d - \gamma(\sum_l q_l + \theta) \;=\; 0$$

where $q = (q_l)$ and this gives

$$q \;=\; (2B + \gamma\mathbf{1})^{-1}((d - \gamma\theta)\mathbf{1} - \alpha)$$

where $B = \mathrm{diag}(\lambda_l^{-1}),\ \alpha = (\alpha_l)$.

Q.E.D.

IA.5.8 Proof of Lemma IA.2

Proof of Lemma IA.2. We have

$$E[\Pi_l(Q_l)] \;=\; E[\sum_l((d-\lambda_l^{-1}a_l)+\lambda_l^{-1}(\delta_l+\eta_l\theta_c))(\delta_l+\eta_l\theta_c)] \;=\; (d-\lambda_l^{-1}a_l)n(\delta_l+\eta_l\bar{\theta})+\lambda_l^{-1}(n(\delta_l+\eta_l\bar{\theta})^2+n\eta_l^2\sigma_\theta^2)\,.$$

Thus, the utility becomes

$$\begin{aligned}
E[U_l(\chi_l;(\chi_l)_{j\neq i})] =&\; (d-\lambda_l^{-1}a_l)n(\delta_l+\eta_l\bar{\theta}) + \lambda_l^{-1}(n(\delta_l+\eta_l\bar{\theta})^2 + n\eta_l^2\sigma_\theta^2) \\
&+ (x_l - n\delta_l - \eta_l n\bar{\theta})d - 0.5\Gamma_l[(x_l - n\delta_l - \eta_l n\bar{\theta})^2 + \eta_l^2 n\sigma_\theta^2] \\
&+\; (0.5\Gamma_l + \beta_l)(\Gamma_l + \beta_l)^{-2}\Big[\Big(\Psi_l\cdot(x \;-\; n\boldsymbol{\delta} - n\bar{\theta}\boldsymbol{\eta})\Big)^2 + n\sigma_\theta^2(\Psi_l\cdot\boldsymbol{\eta})^2\Big] \\
=&\; x_l d - n\lambda_l^{-1}a_l(\delta_l+\eta_l\bar{\theta}) + n\lambda_l^{-1}\Big[(\delta_l+\eta_l\bar{\theta})^2 + \eta_l^2\sigma_\theta^2\Big] \\
&- 0.5\Gamma_l\Big[(x_l - n\delta_l - n\bar{\theta}\eta_l)^2 + n\sigma_\theta^2\eta_l^2\Big] \\
&+\; (0.5\Gamma_l + \beta_l)(\Gamma_l + \beta_l)^{-2}\Big[\Big(\Psi_l\cdot(x \;-\; n\boldsymbol{\delta} - n\bar{\theta}\boldsymbol{\eta})\Big)^2 + n\sigma_\theta^2(\Psi_l\cdot\boldsymbol{\eta})^2\Big]\,.
\end{aligned}$$

where we have defined

$$\Psi \;=\; \left(\frac{\Gamma_1}{\mathcal{B}(\Gamma_1+\beta_1)},\cdots,\frac{\Gamma_M}{\mathcal{B}(\Gamma_M+\beta_M)}\right)^T \qquad \text{and} \qquad \Psi_l \;\equiv\; \Psi - \Gamma_l\mathbf{1}_{l=i}\,.$$

Now, we have

$$\begin{aligned}
\boldsymbol{\delta} \;&=\; \frac{[2+\gamma\Lambda]\mathbf{a} \;-\; \gamma A\boldsymbol{\lambda}}{4+2\gamma\Lambda} \\
\boldsymbol{\eta} \;&=\; -\frac{\gamma\boldsymbol{\lambda}}{2+\gamma\Lambda}\,.
\end{aligned}$$

Q.E.D.

IA.5.9 Proof of Proposition IA.13

Proof of Proposition IA.13. It follows from Lemma IA.1 that

$$
\begin{aligned}
\partial_{a_l}\delta_j &= \frac{\mathbb{1}_l(j)[2+\gamma\Lambda] \; - \; \gamma\lambda_l}{4+2\gamma\Lambda} \\
\partial_{a_l}\eta_j &= 0 \\
\partial_{\lambda_l}\delta_j &= -\frac{2\gamma\Big(2a_j + \gamma[a_j(\Lambda-\lambda_l) - (A-a_j)\lambda_l]\Big)}{(4+2\gamma\Lambda)^2} \; + \; \gamma\frac{a_j - A\mathbb{1}_l(j)}{4+2\gamma\Lambda} \\
\partial_{\lambda_l}\eta_j &= \frac{\gamma^2\lambda_l}{(2+\gamma\Lambda)^2} \; - \; \mathbb{1}_l(j)\frac{\gamma}{2+\gamma\Lambda}\,,
\end{aligned}
$$

where $\mathbb{1}_l(\cdot)$ is the indicator function. We make the following definition:

$$
\begin{aligned}
\partial_{a_l}\boldsymbol{\delta} \equiv \; (\partial_{a_l}\delta_1, \cdots, \partial_{a_l}\delta_n)^T \; &= \; \frac{1}{4+2\gamma\Lambda}\Big[(2+\gamma\Lambda)\boldsymbol{e}_l - \gamma\boldsymbol{\lambda}\Big] \\
\partial_{a_l}\boldsymbol{\eta} = \; 0 & \\
\partial_{\lambda_l}\boldsymbol{\delta} \equiv \; (\partial_{\lambda_l}\delta_1, \cdots, \partial_{\lambda_l}\delta_n)^T \; &= \; -\frac{2\gamma}{(4+2\gamma\Lambda)^2}\Big((2+\gamma\Lambda)\boldsymbol{a} - \gamma A\boldsymbol{\lambda}\Big) \; + \; \frac{\gamma}{4+2\gamma\Lambda}\Big(\boldsymbol{a} - A\boldsymbol{e}_l\Big) \\
\partial_{\lambda_l}\boldsymbol{\eta} \equiv \; (\partial_{\lambda_l}\eta_1, \cdots, \partial_{\lambda_l}\eta_n)^T \; &= \; \frac{\gamma^2}{(2+\gamma\Lambda)^2}\boldsymbol{\lambda} \; - \; \frac{\gamma}{2+\gamma\Lambda}\boldsymbol{e}_l\,,
\end{aligned}
$$

where $\boldsymbol{e}_l$ is the ith coordinate vector. Then, the FOCs are

$$
\begin{aligned}
0 = \; & n(\partial_{a_l}\delta_l)\Big[\Gamma_l x_l - \lambda_l^{-1}a_l\Big] + n\lambda_l^{-1}\Big[(2-n\Gamma_l\lambda_l)(\partial_{a_l}\delta_l) - 1\Big](\delta_l + \eta_l\bar{\theta}) \\
& - \; n(\Gamma_l + 2\beta_l)(\Gamma_l+\beta_l)^{-2}\Big(\Psi_l\cdot\partial_{a_l}\boldsymbol{\delta}\Big)\Big(\Psi_l\cdot(x \; - \; n\boldsymbol{\delta} - n\bar{\theta}\boldsymbol{\eta})\Big) \\
0 = \; & n\lambda_l^{-2}a_l(\delta_l + \eta_l\bar{\theta}) - n\lambda_l^{-2}\Big[(\delta_l+\eta_l\bar{\theta})^2 + \eta_l^2\sigma_\theta^2\Big] + n\lambda_l^{-1}\sigma_\theta^2\Big[2 - \Gamma_l\lambda_l\Big](\partial_{\lambda_l}\eta_l)\eta_l \\
& + \lambda_l^{-1}n\Big[2(\delta_l+\eta_l\bar{\theta}) + \lambda_l\Gamma_l(x_l - n\delta_l - n\bar{\theta}\eta_l) - a_l\Big](\partial_{\lambda_l}\delta_l + \bar{\theta}\partial_{\lambda_l}\eta_l) \\
& - \; n(\Gamma_l+2\beta_l)(\Gamma_l+\beta_l)^{-2}\Big[\Big(\Psi_l\cdot(\partial_{\lambda_l}\boldsymbol{\delta} + \bar{\theta}\partial_{\lambda_l}\boldsymbol{\eta})\Big)\Big(\Psi_l\cdot(x \; - \; n\boldsymbol{\delta} - n\bar{\theta}\boldsymbol{\eta})\Big) - \sigma_\theta^2(\Psi_l\cdot\partial_{\lambda_l}\boldsymbol{\eta})(\Psi_l\cdot\boldsymbol{\eta})\Big]\,.
\end{aligned}
$$

Suppose that

$$
x_l \; = \; 0, \qquad \Gamma_l \; = \; \Gamma \quad \forall l; \qquad \text{and} \qquad \bar{\theta} \; = \; 0\,.
$$

It follows that

$$
\begin{aligned}
\beta_i =&\ \frac{\Gamma}{M-2} \quad \forall\ i; \quad \boldsymbol{\mathcal{B}}(\Gamma_i+\beta_i) = \ M \quad \forall\ i; & \boldsymbol{\mathcal{B}} =&\ \frac{M(M-2)}{\Gamma(M-1)}; & \Psi =&\ \frac{\Gamma}{M}\mathbf{1};\\
\tilde{A} =&\ \frac{\Gamma}{M}A; \qquad \tilde{\Lambda} = \ \frac{\Gamma}{M}\Lambda; & \frac{\Gamma_i+2\beta_i}{(\Gamma_i+\beta_i)^2} =&\ \frac{M(M-2)}{\Gamma(M-1)^2}; & &\\
\delta_l =&\ \frac{2a}{4+2\gamma\Lambda}; \qquad \eta_l = \ -\frac{\gamma\lambda}{2+\gamma\Lambda}; & \Psi_l\cdot\boldsymbol{\delta} =&\ 0; & \Psi_l\cdot\boldsymbol{\eta} =&\ 0;
\end{aligned}
$$

Then, the FOCs become

$$
\begin{aligned}
0 =&\ n(\partial_{a_l}\delta_l)\Big[\Gamma_l x_l - \lambda_l^{-1}a_l\Big] + n\lambda_l^{-1}\Big[(2-n\Gamma_l\lambda_l)(\partial_{a_l}\delta_l)-1\Big](\delta_l+\eta_l\bar{\theta})\\
0 =&\ n\lambda_l^{-2}a_l(\delta_l+\eta_l\bar{\theta}) - n\lambda_l^{-2}\Big[(\delta_l+\eta_l\bar{\theta})^2+\eta_l^2\sigma_\theta^2\Big] + n\lambda_l^{-1}\sigma_\theta^2\Big[2-\Gamma_l\lambda_l\Big](\partial_{\lambda_l}\eta_l)\eta_l\\
&+\lambda_l^{-1}n\Big[2(\delta_l+\eta_l\bar{\theta})+\lambda_l\Gamma_l(x_l-n\delta_l-n\bar{\theta}\eta_l)-a_l\Big](\partial_{\lambda_l}\delta_l+\bar{\theta}\partial_{\lambda_l}\eta_l)\,.
\end{aligned}
$$

We simplify the first of the FOCs and obtain

$$
0 = \ 2\left[\gamma(M-1)(\gamma M+\Gamma n)\lambda^2 + 2(2\gamma M+\Gamma n)\lambda + 4\right]\frac{a}{(4+2\gamma\Lambda)^2}\,.
$$

The only solution is

$$
a \ = \ 0\,,
$$

since the quadratic equation in λ admits only negative solutions (if any). Substituting $a=0$ into the second of the FOCs and simplifying yields

$$
0 = \ 1 + (2-\Gamma\lambda)\left[\frac{\gamma\lambda}{2+M\gamma\lambda} \ - \ 1\right]\,.
$$

This equation has a unique positive solution, which is

$$
\lambda \ = \ \frac{\gamma(M-2)-2\Gamma+\sqrt{\gamma^2(M-2)^2+4\Gamma(\Gamma+4\gamma)M}}{2\gamma(M-1)\Gamma}.
$$

We now present the proof of the monotonicity results.

$$\begin{aligned}
\frac{\partial \lambda^*}{\partial \gamma} &= \frac{\sqrt{4\Gamma^2+\gamma^2(M-2)^2+4\gamma\Gamma M} \;-\; (\gamma M+2\Gamma)}{\gamma^2(M-1)\sqrt{4\Gamma^2+\gamma^2(M-2)^2+4\gamma\Gamma M}} \\
&= \frac{4\Gamma^2+\gamma^2(M-2)^2+4\gamma\Gamma M \;-\; (\gamma M+2\Gamma)^2}{\gamma^2(M-1)\sqrt{4\Gamma^2+\gamma^2(M-2)^2+4\gamma\Gamma M}[\sqrt{4\Gamma^2+\gamma^2(M-2)^2+4\gamma\Gamma M} \;+\; (\gamma M+2\Gamma)]} \\
&= \frac{-4\gamma^2(M-1)}{\gamma^2(M-1)\sqrt{4\Gamma^2+\gamma^2(M-2)^2+4\gamma\Gamma M}[\sqrt{4\Gamma^2+\gamma^2(M-2)^2+4\gamma\Gamma M} \;+\; (\gamma M+2\Gamma)]} \\
&< \; 0
\end{aligned}$$

since $M \geq 2$.

$$\frac{\partial \lambda^*}{\partial \Gamma} = \frac{-(M-2)\sqrt{4\Gamma^2+\gamma^2(M-2)^2+4\gamma\Gamma M} - \gamma(M-2)^2 - 2M\Gamma}{2\Gamma^2(M-1)\sqrt{4\Gamma^2+\gamma^2(M-2)^2+4\gamma\Gamma M}} \;<\; 0\,.$$

$$\frac{\partial \lambda^*}{\partial M} = \frac{\gamma^2(M-2) + (\gamma+2\Gamma)\sqrt{4\Gamma^2+\gamma^2(M-2)^2+4\gamma\Gamma M} - 2\Gamma(\gamma+2\Gamma+\gamma M)}{2\gamma\Gamma(M-1)^2\sqrt{4\Gamma^2+\gamma^2(M-2)^2+4\gamma\Gamma M}}\,.$$

To show that this derivative is always positive for $M \geq 2$ it is enough to show that its numerator is positive for $M \geq 2$. We do so by showing that the numerator is increasing and positive for $M \geq 2$. Consider the numerator. Its derivative with respect to M is

$$\gamma\left(\gamma - 2\Gamma + (\gamma+2\Gamma)\frac{2\Gamma+\gamma(M-2)}{\sqrt{4\Gamma^2+\gamma^2(M-2)^2+4\gamma\Gamma M}}\right)\,.$$

We now show that the term in brackets is positive. First,

$$1 \geq \frac{\gamma(M-2)+2\Gamma}{\sqrt{\gamma^2(M-2)^2+4\gamma\Gamma M+4\Gamma^2}} \;\geq\; 1 - \frac{2\gamma}{\gamma M+2\Gamma}\,.$$

Then,

$$\begin{aligned}
\gamma+2\Gamma \geq &\; \frac{(\gamma+2\Gamma)[\gamma(M-2)+2\Gamma]}{\sqrt{\gamma^2(M-2)^2+4\gamma\Gamma M+4\Gamma^2}} \;\geq\; (\gamma+2\Gamma) - \frac{2\gamma(\gamma+2\Gamma)}{\gamma M+2\Gamma} \\
2\gamma \geq &\; \gamma - 2\Gamma + \frac{(\gamma+2\Gamma)[\gamma(M-2)+2\Gamma]}{\sqrt{\gamma^2(M-2)^2+4\gamma\Gamma M+4\Gamma^2}} \;\geq\; 2\gamma\left[1 - \frac{\gamma+2\Gamma}{\gamma M+2\Gamma}\right] \;>\; 0
\end{aligned}$$

for $M > 1$. Thus, to show that λ^* is increasing, it remains to

$$\gamma^2(M-2) + (\gamma+2\Gamma)\sqrt{4\Gamma^2+\gamma^2(M-2)^2+4\gamma\Gamma M} - 2\Gamma(\gamma+2\Gamma+\gamma M)$$

evaluated at $M = 2$ is non-negative. When $M = 2$, the expression above is

$$2\Big[(\gamma + 2\Gamma)\sqrt{\Gamma^2 + 2\gamma\Gamma} \ - \ \Gamma(2\Gamma + 3\gamma)\Big] \ > \ 0$$

for $\gamma, \Gamma > 0$.

Q.E.D.

IA.5.10 Proof of Proposition IA.14

Proposition IA.15 *Suppose that a_l, the dispersion in bid-ask spreads λ_l^{-1}, and $\bar{\theta}$ are sufficiently small, then,*

$$\Gamma_l \ = \ \Gamma^*(\lambda^*) \ + \ \varphi^{\lambda}(\lambda^*)(\lambda_l - \lambda^*) \ + \ O(\|\lambda - \lambda^*\|^2 + \|a\|^2 + \bar{\theta}^2)$$

where

$$\lambda^* \ = \ \frac{1}{M}\sum_{l=1}^{M}\lambda_l \qquad \text{and} \qquad \varphi^{\lambda}(\cdot) \ = \ \frac{1}{\Phi^{\Gamma}(\cdot)} \ < \ 0\,.$$

Furthermore,

$$\begin{aligned} x_l = \ & \varphi_0^a a_l \ + \ \varphi_0^A A \ + \ \varphi_{0,\lambda}^A A_{mismatch} \ + \ \varphi_0^{\theta}\bar{\theta} \ + \ (\lambda_l - \lambda^*)[\varphi_1^a a_l + \varphi_1^A A + \varphi_{1,\lambda}^A A^{\lambda} + \varphi_1^{\theta}\bar{\theta}] \\ & + \ O((\|\Gamma - \Gamma^*\|^2 + \|x\|^2 + \bar{\theta}^2)^{3/2})\,, \end{aligned}$$

where

$$A \ = \ \sum_{l=1}^{M} a_l; \qquad A_{mismatch} \ = \ \sum_{l=1}^{M}(\lambda_l - \lambda^*)a_l,$$

and $\varphi_0^a, \varphi_0^A, \varphi_{0,\lambda}^A, \varphi_0^{\theta}, \varphi_1^a, \varphi_1^A, \varphi_{1,\lambda}^A$, and φ_1^{θ} are rational functions of n, M, γ and λ^*. Moreover,

$$\varphi_0^a(\cdot) \ = \ 1/\Phi_0^x(\cdot) \ > \ 0\,.$$

Proof of Proposition IA.15. The proof is similar to that of Proposition IA.6. The FOC is

$$\begin{aligned} 0 = \ & -\frac{1}{2}\,[1 \ + \ nc_{0x_l}]\,a_l \ + \ c_{0x_l}x_l \ - \ \lambda_l\frac{\Gamma_l + 2\beta_l}{(\Gamma_l + \beta_l)^2}\Big(\Psi_l \cdot \partial_{a_l}\boldsymbol{\delta}\Big)\Big(\tilde{\chi} \ - \ \frac{n}{2}\tilde{A}\Big) \\ & - \ \frac{\gamma\lambda_l}{4 + 2\gamma\Lambda}\left[n\frac{\Gamma_l + 2\beta_l}{(\Gamma_l + \beta_l)^2}\Big(\Psi_l \cdot \partial_{a_l}\boldsymbol{\delta}\Big)\tilde{\Lambda} \ - \ nc_{0x_l} \ + \ 2(\partial_{a_l}\delta_l) \ - \ 1\right](A + 2\bar{\theta})\,. \end{aligned}$$

It follows that

$$
\begin{aligned}
x_l = &\ \frac{1 \ + \ nc_{0x_l}}{2c_{0x_l}} a_l \ + \ \frac{\lambda_l}{c_{0x_l}} \frac{\Gamma_l + 2\beta_l}{(\Gamma_l + \beta_l)^2} \Big(\Psi_l \cdot \partial_{a_l} \boldsymbol{\delta} \Big) \left(\tilde{\chi} \ - \ \frac{n}{2} \tilde{A} \right) \\
&+ \ \frac{1}{c_{0x_l}} \frac{\gamma \lambda_l}{4 + 2\gamma\Lambda} \left[n \frac{\Gamma_l + 2\beta_l}{(\Gamma_l + \beta_l)^2} \Big(\Psi_l \cdot \partial_{a_l} \boldsymbol{\delta} \Big) \tilde{\Lambda} \ - \ nc_{0x_l} \ + \ 2(\partial_{a_l} \delta_l) \ - \ 1 \right] (A + 2\bar{\theta}) \\
\approx &\ [h_{0a} + h_{1a}(\lambda_l - \lambda^*)] a_l \ + \ [h_{0X} + h_{1X}(\lambda_l - \lambda^*)] \left(\tilde{\chi} \ - \ \frac{n}{2} \tilde{A} \right) \ + \ [h_{0\theta} + h_{1\theta}(\lambda_l - \lambda^*)](A + 2\bar{\theta}) \\
= &\ [h_{0a} + h_{1a}(\lambda_l - \lambda^*)] a_l \ + \ \frac{\Gamma^*[h_{0X} + h_{1X}(\lambda_l - \lambda^*)]}{M} \left(X \ - \ \frac{n}{2} A \right) \ + \ [h_{0\theta} + h_{1\theta}(\lambda_l - \lambda^*)](A + 2\bar{\theta}) \\
&+ \ \frac{\varphi^\lambda [h_{0X} + h_{1X}(\lambda_l - \lambda^*)]}{M^2 - 2M + 2} \left(X^\lambda \ - \ \frac{n}{2} A^\lambda \right) ,
\end{aligned}
$$

where

$$
A^\lambda \ = \ \sum_{l=1}^{M} (\lambda_l - \lambda^*) a_j \qquad \text{and} \qquad X^\lambda \ = \ \sum_{l=1}^{M} (\lambda_l - \lambda^*) x_l \, ,
$$

and we used

$$
\begin{aligned}
\tilde{A} = &\ \sum_{l=1}^{M} \frac{\Gamma_l}{\mathcal{B}(\Gamma_l + \beta_l)} a_j \ \approx \ \frac{\Gamma^*}{M} A \ + \ \frac{\varphi^\lambda}{M^2 - 2M + 2} A^\Gamma \\
\tilde{\chi} = &\ \sum_{l=1}^{M} \frac{\Gamma_l}{\mathcal{B}(\Gamma_l + \beta_l)} x_l \ \approx \ \frac{\Gamma^*}{M} X \ + \ \frac{\varphi^\lambda}{M^2 - 2M + 2} X_{mismatch} \, .
\end{aligned}
$$

Summing over i yield

$$
\begin{aligned}
X^\lambda = &\ h_{0a} A^\lambda \\
X = &\ h_{0a} A + h_{1a} A^\lambda \ + \ \Gamma^* h_{0X} \left(X \ - \ \frac{n}{2} A \right) \ + \ M h_{0\theta}(A + 2\bar{\theta}) \ + \ \frac{M h_{0X} \varphi^\lambda}{M^2 - 2M + 2} \left(X^\lambda \ - \ \frac{n}{2} A^\lambda \right) ,
\end{aligned}
$$

implying that

$$
X \ = \ \frac{h_{0a} - \frac{n}{2} \Gamma^* h_{0X} + M h_{0\theta}}{1 - \Gamma^* h_{0X}} A \ + \ \frac{2M h_{0\theta}}{1 - \Gamma^* h_{0X}} \bar{\theta} \ + \ \frac{h_{1a} \ + \ \frac{M h_{0X} \varphi^\lambda}{M^2 - 2M + 2} \left(h_{0a} \ - \ \frac{n}{2} \right)}{1 - \Gamma^* h_{0X}} A^\lambda \, .
$$

Substituting into the equation for x_l we obtain

$$
\begin{aligned}
x_l = \; & [h_{0a} + h_{1a}(\lambda_l - \lambda^*)]a_l \; + \; \frac{\Gamma^*[h_{0X} + h_{1X}(\lambda_l - \lambda^*)]}{M} \frac{h_{0a} - \frac{n}{2}\Gamma^* h_{0X} + M h_{0\theta}}{1 - \Gamma^* h_{0X}} A \\
& - \; \frac{n}{2} \frac{\Gamma^*[h_{0X} + h_{1X}(\lambda_l - \lambda^*)]}{M} A \; + \; [h_{0\theta} + h_{1\theta}(\lambda_l - \lambda^*)]A \; + \; \left(h_{0a} - \frac{n}{2}\right) \frac{\varphi^\lambda [h_{0X} + h_{1X}(\lambda_l - \lambda^*)]}{M^2 - 2M + 2} A^\lambda \\
& + \; \frac{\Gamma^*[h_{0X} + h_{1X}(\lambda_l - \lambda^*)]}{M} \frac{h_{1a} \; + \; \frac{M h_{0X} \varphi^\lambda}{M^2 - 2M + 2}\left(h_{0a} \; - \; \frac{n}{2}\right)}{1 - \Gamma^* h_{0X}} A^\lambda \\
& + \; 2[h_{0\theta} + h_{1\theta}(\lambda_l - \lambda^*)]\bar{\theta} \; + \; \frac{2 h_{0\theta} \Gamma^* [h_{0X} + h_{1X}(\lambda_l - \lambda^*)]}{1 - \Gamma^* h_{0X}} \bar{\theta} \, .
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\varphi_0^a = \; & h_{0a} \\
\varphi_0^A = \; & h_{0\theta} \; + \; \frac{\Gamma^* h_{0X}}{M} \frac{h_{0a} - \frac{n}{2}\Gamma^* h_{0X} + M h_{0\theta}}{1 - \Gamma^* h_{0X}} \; - \; \frac{n}{2} \frac{\Gamma^* h_{0X}}{M} \\
= \; & h_{0\theta} \; + \; \frac{\Gamma^* h_{0X}}{M} \frac{1}{1 - \Gamma^* h_{0X}} \left[h_{0a} \; + \; M h_{0\theta} \; - \; \frac{n}{2} \right] \\
\varphi_{0,\lambda}^A = \; & \left(h_{0a} - \frac{n}{2}\right) \frac{\varphi^\lambda h_{0X}}{M^2 - 2M + 2} \; + \; \frac{\Gamma^* h_{0X}}{M} \frac{h_{1a} + \frac{M h_{0X} \varphi^\lambda}{M^2 - 2M + 2}\left(h_{0a} - \frac{n}{2}\right)}{1 - \Gamma^* h_{0X}} \\
\varphi_0^\theta = \; & 2 h_{0\theta} \left[1 \; + \; \frac{\Gamma^* h_{0X}}{1 - \Gamma^* h_{0X}} \right] \; = \; \frac{2 h_{0\theta}}{1 - \Gamma^* h_{0X}} \\
\varphi_1^a = \; & h_{1a} \\
\varphi_1^A = \; & h_{1\theta} \; + \; \frac{\Gamma^* h_{1X}}{M} \frac{h_{0a} - \frac{n}{2}\Gamma^* h_{0X} + M h_{0\theta}}{1 - \Gamma^* h_{0X}} \; - \; \frac{n}{2} \frac{\Gamma^* h_{1X}}{M} \\
= \; & h_{1\theta} \; + \; \frac{\Gamma^* h_{1X}}{M} \frac{1}{1 - \Gamma^* h_{0X}} \left[h_{0a} \; + \; M h_{0\theta} \; - \; \frac{n}{2} \right] \\
\varphi_{1,\lambda}^A = \; & \left(h_{0a} - \frac{n}{2}\right) \frac{\varphi^\lambda h_{1X}}{M^2 - 2M + 2} \; + \; \frac{\Gamma^* h_{1X}}{M} \frac{h_{1a} + \frac{M h_{0X} \varphi^\lambda}{M^2 - 2M + 2}\left(h_{0a} - \frac{n}{2}\right)}{1 - \Gamma^* h_{0X}} \\
= \; & \frac{1}{1 - \Gamma^* h_{0X}} \left[\left(h_{0a} - \frac{n}{2}\right) \frac{\varphi^\lambda h_{1X}}{M^2 - 2M + 2} \; + \; \frac{\Gamma^* h_{1X} h_{1a}}{M} \right] \\
\varphi_1^\theta = \; & 2 h_{1\theta} \; + \; \frac{2 h_{0\theta} \Gamma^* h_{1X}}{1 - \Gamma^* h_{0X}} \, .
\end{aligned}
$$

Q.E.D.

Proof of Proposition IA.14. We have

$$
P^{D2D} \; = \; d \; - \; \widehat{X} \qquad \text{and} \qquad \widehat{X} \; = \; \Psi \cdot \tilde{\chi} \, .
$$

To first order,

$$
\begin{aligned}
\frac{\Gamma_l}{\mathcal{B}(\Gamma_l+\beta_l)} &\approx \frac{\Gamma^*}{M} + \frac{\varphi^\lambda}{M^2-2M+2}(\lambda_l-\lambda^*) \\
\sum_{l=1}^{M}\frac{\Gamma_l}{\mathcal{B}(\Gamma_l+\beta_l)} &\approx \Gamma^* \\
\sum_{l=1}^{M}\frac{\Gamma_l}{\mathcal{B}(\Gamma_l+\beta_l)}\alpha_l &\approx \Gamma^*\bar{\alpha} + \frac{M\varphi^\lambda\lambda^*}{M^2-2M+2}\alpha_{mismatch} \\
\sum_{l=1}^{M}\frac{\Gamma_l}{\mathcal{B}(\Gamma_l+\beta_l)}(\lambda_l-\lambda^*) &\approx 0 \\
\sum_{l=1}^{M}\frac{\Gamma_l}{\mathcal{B}(\Gamma_l+\beta_l)}\alpha_l(\lambda_l-\lambda^*) &\approx \lambda^*\Gamma^*\alpha_{mismatch}\,.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\Psi\cdot\tilde{\chi} = &-(\varphi_0^a-0.5n)\lambda^*\left[\Gamma^*\bar{\alpha} + \frac{M\varphi^\lambda\lambda^*}{M^2-2M+2}\alpha_{mismatch}\right] \\
&- (\varphi_0^A-0.5n\gamma\eta_0^*\lambda^*)M\lambda^*\Gamma^*\bar{\alpha} - M\lambda^*\Gamma^*[(\varphi_0^A-0.5n\gamma\eta_0^*\lambda^*)+\varphi_{0,\lambda}^A\lambda^*]\hat{\alpha} \\
&+ \lambda^*\Gamma^*[\varphi_0^a+M\varphi_0^A+n\eta_0^*]d + \Gamma^*\varphi_0^\theta\bar{\theta} - \gamma\eta_0^*\lambda^*\Gamma^*\Theta \\
&- (\varphi_0^a+\varphi_1^a\lambda^*-0.5n)\lambda^*\Gamma^*\alpha_{mismatch} \\
= &-\lambda^*\Gamma^*\left[\varphi_0^a+M\varphi_0^A+n\eta_0^*\right](\bar{\alpha}+\alpha_{mismatch}) - (\lambda^*)^2\left[\Gamma^*(M\varphi_{0,\lambda}^A+\varphi_1^a) + \frac{M\varphi^\lambda(\varphi_0^a-0.5n)}{M^2-2M+2}\right]\alpha_{mismatch} \\
&+ \lambda^*\Gamma^*[\varphi_0^a+M\varphi_0^A+n\eta_0^*]d + \Gamma^*\varphi_0^\theta\bar{\theta} - \gamma\eta_0^*\lambda^*\Gamma^*\Theta\,.
\end{aligned}
$$

Q.E.D.

IA.6 Fragmented Double Auction Model/Double Auction Benchmark

There are a few existing theoretical papers on two-tiered OTC markets, with different assumptions than our model. Our model tracks real-world FX markets and its predictions are supported by empirical evidence. However, it remains a question whether or not existing models would make the same predictions as ours. We examine this question in this Appendix and show that our model is essential in understanding the dynamics of exchange rates. In particular, we show that a leading two-tiered OTC model, developed by Babus and Parlatore (2022), yields predictions that are not supported by the data from the FX market. This result points to the importance of specific the market structure used in each market, as the model in Babus and Parlatore (2022) was not developed for the specific markets we consider here.

We will not reproduce the model of Babus and Parlatore (2022) here. The key difference between their model and ours is the request for quote mechanism in our model (and the heterogeneity in the Dealer's holding cost/risk aversion). In their model, customers exogenously trade with only one dealer, thus there is no competition in liquidity provision between dealers in the D2C market.

We now re-derive the predictions about price dynamics in a two-tiered OTC market where, instead of the request-for-quote mechanism, customers can only trade with one dealer.

In this appendix, we modify the model of our paper along the lines of Babus and Parlatore (2022) and assume that customers are split into K groups of equal size of L so that $n = KL$. Each group only trades with a single dealer. This drastically simplifies the problem because it eliminates strategic competition in liquidity provision between dealers in the D2C market.

The derivation of the equilibrium follows that of the main model and we will provide fewer details.

Lemma IA.9 *Dealer number l trades with n customers with risk aversions γ. Hence, his post-D2C trade inventory is given by*

$$\chi_l \;=\; (\Gamma_l^{D2C} + \beta_l^{D2C})^{-1}\mathbf{Q}_l^{D2C} + (\Gamma_l^{D2C} + \beta_l^{D2C})^{-1}\beta_l^{D2C}\tilde{x}_l\,.$$

The inter-dealer market does not change, and we can write down the dealer utility as

$$U_l \;=\; -(\chi_l - x_l)\mathcal{P}_l^{D2C} \;+\; \chi_l d - 0.5\Gamma_l\chi_l^2 \;+\; \frac{1}{2}E\Big[\frac{\Gamma_l + 2\beta_l}{(\Gamma_l+\beta_l)^2}\left(\mathcal{B}^{-1}\sum_{\ell=1}^{M}(\Gamma_\ell+\beta_\ell)^{-1}\Gamma_\ell\chi_\ell - \Gamma_l\chi_l\right)^2\Big]$$

and we can rewrite this utility as

$$U_l \;=\; -(\chi_l - x_l)\mathcal{P}_l^{D2C} \;+\; (d + Z_{-l})\chi_l \;-\; 0.5\Gamma_l^{D2C}\chi_l^2 \;+\; const,$$

where

$$\Gamma_l^{D2C} \;=\; \Gamma_l \;-\; \frac{\Gamma_l + 2\beta_l}{(\Gamma_l+\beta_l)^2}\left[\left(\frac{1}{\mathcal{B}(\Gamma_l+\beta_l)} - 1\right)\Gamma_l\right]^2 \;\approx\; \frac{2\Gamma^*}{M} \;+\; \frac{2}{M^2-2M+2}(\Gamma_l - \Gamma^*)\,,$$

and

$$Z_{-l} \;=\; \Delta_l \sum_{\ell\neq l}\frac{\Gamma_\ell}{\mathcal{B}(\Gamma_\ell+\beta_\ell)}E[\chi_\ell]$$

where we have defined

$$\Delta_l = \frac{\Gamma_l + 2\beta_l}{(\Gamma_l+\beta_l)^2}\left(\frac{1}{\mathcal{B}(\Gamma_l+\beta_l)} - 1\right)\Gamma_l \;\approx\; -\frac{M-2}{M-1} \;-\; \frac{M(M-2)}{\Gamma(M-1)(M^2-2M+2)}(\Gamma_l - \Gamma^*)\,,$$

Given this quadratic objective, the optimal demand schedule for the dealer in the D2C market is given by

$$Q_l(\mathcal{P}_l^{D2C}) \;=\; (\Gamma_l^{D2C} + \beta_l^{D2C})^{-1}(d + Z_{-l} - \mathcal{P}_l^{D2C} - \Gamma_l^{D2C} x_l)$$

Lemma IA.10 *We have*

$$\begin{aligned}
\mathcal{B}_l^{D2C} =\; & \frac{4nM\gamma\Gamma_l^{D2C}(n^2-3n+1) + 4n^3(n-2)(\Gamma_l^{D2C})^2 - (2n-1)M^2\gamma^2}{4n(n-1)(\gamma M + 2n\Gamma_l^{D2C})\gamma\Gamma_l^{D2C}} \\
&+\; \frac{(M\gamma + 2n^2\Gamma_l^{D2C})\sqrt{\gamma^2M^2(1-2n)^2 + 4Mn(2n^2-3n+2)\gamma\Gamma_l^{D2C} + 4(n-2)^2n^2(\Gamma_l^{D2C})^2}}{4n(n-1)(\gamma M + 2n\Gamma_l^{D2C})\gamma\Gamma_l^{D2C}} \\
\beta_l^{D2C} =\; & \frac{2\Gamma_l^{D2C}}{\Gamma_l^{D2C}\mathcal{B}_l^{D2C} - 2 + \sqrt{(\Gamma_l^{D2C}\mathcal{B}_l^{D2C})^2 + 4}} \\
\beta_{c,l}^{D2C} =\; & \frac{2\gamma}{\gamma\mathcal{B}_l^{D2C} - 2 + \sqrt{(\gamma\mathcal{B}_l^{D2C})^2 + 4}}\,.
\end{aligned}$$

Now, we obtain the post-D2C round inventory

$$\chi_l \;=\; x_l \;+\; (\Gamma_l^{D2C} + \beta_l^{D2C})^{-1}(d + Z_{-l} - \mathcal{P}_l^{D2C} - \Gamma_l^{D2C} x_l)$$

and the price

$$\mathcal{P}_l^{D2C} \;=\; d + (\mathcal{B}_l^{D2C})^{-1}(\Gamma_l^{D2C} + \beta_l^{D2C})^{-1} Z_{-l} - Q_l^{D2C}$$

where

$$Q_l^{D2C} \;=\; (\mathcal{B}_l^{D2C})^{-1}\left((\Gamma_l^{D2C} + \beta_l^{D2C})^{-1}\Gamma_l^{D2C} x_l + (\gamma + \beta_{c,l}^{D2C})^{-1}\gamma\sum_{i=1}^n \theta_i\right).$$

Note that

$$Z_{-l} \;=\; \Delta_l(\bar{Z} - \mathcal{B}^{-1}(\Gamma_l + \beta_l)^{-1}\Gamma_l E[\chi_l]) \tag{IA.13}$$

where we have defined

$$\bar{Z} \;=\; \sum_\ell \mathcal{B}^{-1}(\Gamma_\ell + \beta_\ell)^{-1}\Gamma_\ell E[\chi_\ell] \tag{IA.14}$$

Taking expectations and substituting, we get

$$E[\chi_l] \;=\; x_l \;+\; (\Gamma_l^{D2C} + \beta_l^{D2C})^{-1}(d + Z_{-l} - E[\mathcal{P}_l^{D2C}] - \Gamma_l^{D2C} x_l) \tag{IA.15}$$

First, we compute the expected price:

$$
\begin{aligned}
E[\mathcal{P}_l^{D2C}] \; &= \; d + (\mathcal{B}_l^{D2C})^{-1}(\Gamma_l^{D2C} + \beta_l^{D2C})^{-1} E[Z_{-l}] - E[Q_l^{D2C}] \\
&= \; d + (\mathcal{B}_l^{D2C})^{-1}(\Gamma_l^{D2C} + \beta_l^{D2C})^{-1} E[Z_{-l}] \\
&- E\left[(\mathcal{B}_l^{D2C})^{-1}\left((\Gamma_l^{D2C} + \beta_l^{D2C})^{-1}\Gamma_l^{D2C} x_l + (\gamma + \beta_{c,l}^{D2C})^{-1}\gamma \sum_{i=1}^{n} \theta_i\right)\right] \\
&= \; d + (\mathcal{B}_l^{D2C})^{-1}(\Gamma_l^{D2C} + \beta_l^{D2C})^{-1} Z_{-l} \\
&- (\mathcal{B}_l^{D2C})^{-1}\left((\Gamma_l^{D2C} + \beta_l^{D2C})^{-1}\Gamma_l^{D2C} x_l \; + \; (\gamma + \beta_{c,l}^{D2C})^{-1}\gamma n E[\theta]\right)
\end{aligned}
\tag{IA.16}
$$

Substituting (IA.16) into (IA.15), we get

$$
\begin{aligned}
E[\chi_l] \; &= \; x_l \; + \; (\Gamma_l^{D2C} + \beta_l^{D2C})^{-1}\Bigg(d + Z_{-l} - \Bigg(d + (\mathcal{B}_l^{D2C})^{-1}(\Gamma_l^{D2C} + \beta_l^{D2C})^{-1} Z_{-l} \\
&- (\mathcal{B}_l^{D2C})^{-1}\left((\Gamma_l^{D2C} + \beta_l^{D2C})^{-1}\Gamma_l^{D2C} x_l \; + \; (\gamma + \beta_{c,l}^{D2C})^{-1}\gamma n E[\theta]\right)\Bigg) - \Gamma_l^{D2C} x_l\Bigg) \\
&= \; C_l^Z Z_{-l} \; + \; C_l^x x_l \; + \; C_l^\theta E[\theta]
\end{aligned}
$$

where

$$
\begin{aligned}
C_l^Z \; &= \; \frac{1}{\Gamma_l^{D2C} + \beta_l^{D2C}}\left[1 - \frac{1}{\mathcal{B}_l^{D2C}(\Gamma_l^{D2C} + \beta_l^{D2C})}\right] \\
C_l^x \; &= \; 1 \; + \; \frac{\Gamma_l^{D2C}}{\Gamma_l^{D2C} + \beta_l^{D2C}}\left[-1 \; + \; \frac{1}{\mathcal{B}_l^{D2C}(\Gamma_l^{D2C} + \beta_l^{D2C})}\right] \\
C_l^\theta \; &= \; \frac{n\gamma}{\mathcal{B}_l^{D2C}(\Gamma_l^{D2C} + \beta_l^{D2C})(\gamma + \beta_{c,l}^{D2C})}\,.
\end{aligned}
$$

Now, we are ready to solve for $E[\chi_l]$. We have from (IA.13) that

$$
E[\chi_l] \; = \; C_l^Z \Delta_l(\bar{Z} - \mathcal{B}^{-1}(\Gamma_l + \beta_l)^{-1}\Gamma_l E[\chi_l]) \; + \; C_l^x x_l \; + \; C_l^\theta E[\theta]
$$

which gives

$$
E[\chi_l] \; = \; (1 + C_l^Z \Delta_l \mathcal{B}^{-1}(\Gamma_l + \beta_l)^{-1}\Gamma_l)^{-1}\left(C_l^Z \Delta_l \bar{Z} \; + \; C_l^x x_l \; + \; C_l^\theta E[\theta]\right)
$$

Using (IA.14), we get

$$
\bar{Z} \; = \; \sum_\ell \mathcal{B}^{-1}(\Gamma_\ell + \beta_\ell)^{-1}\Gamma_\ell(1 + C_\ell^Z \Delta_\ell \mathcal{B}^{-1}(\Gamma_\ell + \beta_\ell)^{-1}\Gamma_\ell)^{-1}\left(C_\ell^Z \Delta_\ell \bar{Z} \; + \; C_\ell^x x_\ell \; + \; C_\ell^\theta E[\theta]\right),
$$

which is equivalent to

$$
\begin{aligned}
\bar{Z} \ &= \ \left(1 - \sum_\ell \mathcal{B}^{-1}(\Gamma_\ell + \beta_\ell)^{-1}\Gamma_\ell(1 + C_\ell^Z\Delta_\ell\mathcal{B}^{-1}(\Gamma_\ell + \beta_\ell)^{-1}\Gamma_\ell)^{-1}C_\ell^Z\Delta_\ell\right)^{-1} \\
&\left(\sum_\ell \mathcal{B}^{-1}(\Gamma_\ell + \beta_\ell)^{-1}\Gamma_\ell(1 + C_\ell^Z\Delta_\ell\mathcal{B}^{-1}(\Gamma_\ell + \beta_\ell)^{-1}\Gamma_\ell)^{-1}\left(C_\ell^x x_\ell \ + \ C_\ell^\theta E[\theta]\right)\right) \\
&= \left(1 - \sum_\ell \left[1 + (C_\ell^Z\Delta_\ell\mathcal{B}^{-1}(\Gamma_\ell + \beta_\ell)^{-1}\Gamma_\ell)^{-1}\right]^{-1}\right)^{-1} \\
&\left(\sum_\ell((\mathcal{B}^{-1}(\Gamma_\ell + \beta_\ell)^{-1}\Gamma_\ell)^{-1} + C_\ell^Z\Delta_\ell)^{-1}\left(C_\ell^x x_\ell \ + \ C_\ell^\theta E[\theta]\right)\right) \\
&= \ \bar{C}_{\bar{Z}}^x \cdot x \ + \ \bar{C}_{\bar{\theta}}^{\bar{Z}} E[\theta]
\end{aligned}
$$

where we have defined

$$
\begin{aligned}
Dem_{\bar{Z}} \ &= \ 1 \ - \ \sum_\ell \frac{1}{1 + \frac{1}{C_\ell^Z\Delta_\ell\frac{\Gamma_\ell}{\mathcal{B}(\Gamma_\ell+\beta_\ell)}}} \\
\bar{C}_{\bar{Z},l}^x &= \ \frac{1}{Dem_{\bar{Z}}}\left(\left(\frac{\Gamma_l}{\mathcal{B}(\Gamma_l + \beta_l)}\right)^{-1} + C_l^Z\Delta_l\right)^{-1} C_l^x \\
\bar{C}_{\bar{Z},l}^{\bar{\theta}} &= \ \frac{1}{Dem_{\bar{Z}}}\left(\left(\frac{\Gamma_l}{\mathcal{B}(\Gamma_l + \beta_l)}\right)^{-1} + C_l^Z\Delta_l\right)^{-1} C_l^\theta \\
\bar{C}_{\bar{Z}}^x \ &= \ (\bar{C}_{\bar{Z},1}^x, \bar{C}_{\bar{Z},2}^x, \cdots, \bar{C}_{\bar{Z},M}^x)^T \\
\bar{C}_{\bar{Z}}^{\bar{\theta}} \ &= \ \sum_\ell \bar{C}_{\bar{Z},\ell}^{\bar{\theta}}\,.
\end{aligned}
$$

We proceed with linking the prices in D2D and D2C markets. We have

$$
\begin{aligned}
\chi_l \ &= \ x_l \ + \ (\Gamma_l^{D2C} + \beta_l^{D2C})^{-1}(d + Z_{-l} - \mathcal{P}_l^{D2C} - \Gamma_l^{D2C}x_l) \\
&= \ (\Gamma_l^{D2C} + \beta_l^{D2C})^{-1}(d + Z_{-l} - \mathcal{P}_l^{D2C}) \ + \ (\Gamma_l^{D2C} + \beta_l^{D2C})^{-1}\Gamma_l^{D2C}x_l
\end{aligned}
$$

We need to invert (x_l) from the vector of $\mathcal{P}_l^{D2C}$. We have

$$
\begin{aligned}
Z_{-l} \;&=\; \Delta_l\Bigg(\bar{Z} - \mathcal{B}^{-1}(\Gamma_l+\beta_l)^{-1}\Gamma_l(1+C_l^Z\Delta_l\mathcal{B}^{-1}(\Gamma_l+\beta_l)^{-1}\Gamma_l)^{-1}\left(C_l^Z\Delta_l\bar{Z} \;+\; C_l^x x_l \;+\; C_l^\theta E[\theta]\right)\Bigg)\\
&=\; \Delta_l\Bigg(\bar{Z} - \left(\frac{1}{\mathcal{B}^{-1}(\Gamma_l+\beta_l)^{-1}\Gamma_l} + C_l^Z\Delta_l\right)^{-1}\left(C_l^Z\Delta_l\bar{Z} \;+\; C_l^x x_l \;+\; C_l^\theta E[\theta]\right)\Bigg)\\
&=\; \frac{\mathcal{B}(\Gamma_l+\beta_l)}{\mathcal{B}(\Gamma_l+\beta_l)+\Gamma_l\Delta_l C_l^Z}\Delta_l\bar{Z} \;-\; \frac{\Gamma_l\Delta_l}{\mathcal{B}(\Gamma_l+\beta_l)+\Gamma_l\Delta_l C_l^Z}C_l^x x_l \;-\; \frac{\Gamma_l\Delta_l}{\mathcal{B}(\Gamma_l+\beta_l)+\Gamma_l\Delta_l C_l^Z}C_l^\theta\bar{\theta}\\
&=\; \frac{\mathcal{B}(\Gamma_l+\beta_l)}{\mathcal{B}(\Gamma_l+\beta_l)+\Gamma_l\Delta_l C_l^Z}\Delta_l(\bar{C}_{\bar{Z}}^x\cdot x \;+\; \bar{C}_{\bar{\theta}}^{\bar{Z}}\bar{\theta})\\
&\quad -\; \frac{\Gamma_l\Delta_l}{\mathcal{B}(\Gamma_l+\beta_l)+\Gamma_l\Delta_l C_l^Z}C_l^x x_l \;-\; \frac{\Gamma_l\Delta_l}{\mathcal{B}(\Gamma_l+\beta_l)+\Gamma_l\Delta_l C_l^Z}C_l^\theta\bar{\theta}\\
&=\; \frac{\mathcal{B}(\Gamma_l+\beta_l)}{\mathcal{B}(\Gamma_l+\beta_l)+\Gamma_l\Delta_l C_l^Z}\Delta_l(\bar{C}_{\bar{Z}}^x\cdot x) \;-\; \frac{\Gamma_l\Delta_l}{\mathcal{B}(\Gamma_l+\beta_l)+\Gamma_l\Delta_l C_l^Z}C_l^x x_l\\
&\quad +\left[\frac{\mathcal{B}(\Gamma_l+\beta_l)}{\mathcal{B}(\Gamma_l+\beta_l)+\Gamma_l\Delta_l C_l^Z}\Delta_l\bar{C}_{\bar{\theta}}^{\bar{Z}} \;-\; \frac{\Gamma_l\Delta_l}{\mathcal{B}(\Gamma_l+\beta_l)+\Gamma_l\Delta_l C_l^Z}C_l^\theta\right]\bar{\theta}\\
&=\; C_x^{Z_{-l}}\cdot x \;+\; C_{x_l}^{Z_{-l},l}x_l \;+\; C_{\bar{\theta}}^{Z_{-l},l}\bar{\theta}\,,
\end{aligned}
$$

where we have defined

$$
\begin{aligned}
&C_x^{Z_{-l},l} = \frac{\mathcal{B}(\Gamma_l+\beta_l)}{\mathcal{B}(\Gamma_l+\beta_l)+\Gamma_l\Delta_l C_l^Z}\Delta_l\bar{C}_{\bar{Z},l}^x; \qquad C_x^{Z_{-l}} = \frac{\mathcal{B}(\Gamma_l+\beta_l)}{\mathcal{B}(\Gamma_l+\beta_l)+\Gamma_l\Delta_l C_l^Z}\Delta_l\bar{C}_{\bar{Z}}^x;\\
&C_{x_l}^{Z_{-l},l} = \;-\frac{\Gamma_l\Delta_l}{\mathcal{B}(\Gamma_l+\beta_l)+\Gamma_l\Delta_l C_l^Z}C_l^x; \quad C_{\bar{\theta}}^{Z_{-l},l} = \frac{\Delta_l}{\mathcal{B}(\Gamma_l+\beta_l)+\Gamma_l\Delta_l C_l^Z}\left[\mathcal{B}(\Gamma_l+\beta_l)\bar{C}_{\bar{\theta}}^{\bar{Z}} \;-\; \Gamma_l C_l^\theta\right].
\end{aligned}
$$

Returning to the D2C price for Dealer l, we have

$$
\begin{aligned}
\mathcal{P}_l^{D2C} \;&=\; d + (\mathcal{B}_l^{D2C})^{-1}(\Gamma_l^{D2C} + \beta_l^{D2C})^{-1} Z_{-l} - Q_l^{D2C} \\
&=\; d + (\mathcal{B}_l^{D2C})^{-1}(\Gamma_l^{D2C} + \beta_l^{D2C})^{-1} Z_{-l} - (\mathcal{B}_l^{D2C})^{-1}\left((\Gamma_l^{D2C} + \beta_l^{D2C})^{-1}\Gamma_l^{D2C} x_l + (\gamma + \beta_{c,l}^{D2C})^{-1}\gamma \sum_{i=1}^{n} \theta_i\right) \\
&=\; d + \frac{1}{\mathcal{B}_l^{D2C}(\Gamma_l^{D2C} + \beta_l^{D2C})}(C_x^{Z_{-l}} \cdot x \;+\; C_{x_l}^{Z_{-l},l} x_l \;+\; C_{\bar\theta}^{Z_{-l}}\bar\theta) \\
&\quad - \left(\frac{\Gamma_l^{D2C}}{\mathcal{B}_l^{D2C}(\Gamma_l^{D2C} + \beta_l^{D2C})} x_l \;+\; \frac{\gamma}{\mathcal{B}_l^{D2C}(\gamma + \beta_{c,l}^{D2C})}\sum_{i=1}^{n}\theta_i\right) \\
&=\; d \;+\; \frac{1}{\mathcal{B}_l^{D2C}(\Gamma_l^{D2C} + \beta_l^{D2C})}(C_x^{Z_{-l}} \cdot x) \;+\; \frac{1}{\mathcal{B}_l^{D2C}(\Gamma_l^{D2C} + \beta_l^{D2C})}(C_{x_l}^{Z_{-l}} x_l) \;-\; \frac{\Gamma_l^{D2C}}{\mathcal{B}_l^{D2C}(\Gamma_l^{D2C} + \beta_l^{D2C})} x_l \\
&\quad + \frac{1}{\mathcal{B}_l^{D2C}(\Gamma_l^{D2C} + \beta_l^{D2C})}(C_{\bar\theta}^{Z_{-l}}\bar\theta) \;-\; \frac{\gamma}{\mathcal{B}_l^{D2C}(\gamma + \beta_{c,l}^{D2C})}\sum_{i=1}^{n}\theta_i \\
&=\; d \;+\; \frac{1}{\mathcal{B}_l^{D2C}(\Gamma_l^{D2C} + \beta_l^{D2C})}(C_x^{Z_{-l}} \cdot x) \;+\; \frac{1}{\mathcal{B}_l^{D2C}(\Gamma_l^{D2C} + \beta_l^{D2C})}\left[C_{x_l}^{Z_{-l},l} \;-\; \Gamma_l^{D2C}\right] x_l \\
&\quad + \frac{1}{\mathcal{B}_l^{D2C}(\Gamma_l^{D2C} + \beta_l^{D2C})}(C_{\bar\theta}^{Z_{-l},l}\bar\theta) \;-\; \frac{\gamma}{\mathcal{B}_l^{D2C}(\gamma + \beta_{c,l}^{D2C})}\sum_{i=1}^{n}\theta_i \\
&=\; d \;+\; C_x^{\mathcal{P},l} \cdot x \;+\; C_{x_l}^{\mathcal{P},l} x_l \;+\; C_\Theta^{\mathcal{P},l}\Theta \;+\; C_{\bar\theta}^{\mathcal{P},l}\bar\theta\,,
\end{aligned}
$$

where

$$
\begin{aligned}
C_{x,l}^{\mathcal{P},l} \;&=\; \frac{1}{\mathcal{B}_l^{D2C}(\Gamma_l^{D2C} + \beta_l^{D2C})} C_x^{Z_{-l},l}; \qquad & C_x^{\mathcal{P},l} \;&=\; \frac{1}{\mathcal{B}_l^{D2C}(\Gamma_l^{D2C} + \beta_l^{D2C})} C_x^{Z_{-l}}; \\
C_{x_l}^{\mathcal{P},l} \;&=\; \frac{1}{\mathcal{B}_l^{D2C}(\Gamma_l^{D2C} + \beta_l^{D2C})}\left[C_{x_l}^{Z_{-l},l} \;-\; \Gamma_l^{D2C}\right]; & & \\
C_\Theta^{\mathcal{P},l} \;&=\; -\frac{\gamma}{\mathcal{B}_l^{D2C}(\gamma + \beta_{c,l}^{D2C})}; & C_{\bar\theta}^{\mathcal{P},l} \;&=\; \frac{1}{\mathcal{B}_l^{D2C}(\Gamma_l^{D2C} + \beta_l^{D2C})} C_{\bar\theta}^{Z_{-l},l}\,.
\end{aligned}
$$

It follows that

$$
\mathcal{P}^{D2C} \;=\; d\mathbf{1} \;+\; \mathcal{A}^x x \;+\; \mathcal{V}^\Theta \Theta \;+\; \mathcal{V}^{\bar\theta}\bar\theta\,,
$$

with

$$
\mathcal{A}^x \;=\; \mathcal{A}_0^x + diag(C_{x_l}^{\mathcal{P},l}); \quad \mathcal{A}_0^x \;=\; \sum_{\ell=1}^{M} e_\ell \otimes C_x^{\mathcal{P},\ell}; \quad \mathcal{V}^\Theta \;=\; \left(C_\Theta^{\mathcal{P},\ell}\right)_{\ell=1}^{M}; \quad \mathcal{V}^{\bar\theta} \;=\; \left(C_{\bar\theta}^{\mathcal{P},\ell}\right)_{\ell=1}^{M}.
$$

Inverting, we get

Lemma IA.11 *We have*

$$
x = \; (\mathcal{A}^x)^{-1}\left[\mathcal{P}^{D2C} \;-\; d\mathbf{1} \;-\; \mathcal{V}^\Theta\Theta \;-\; \mathcal{V}^{\bar\theta}\bar\theta\right].
$$

Now we can write the link between D2D and D2C prices:

$$
\begin{aligned}
\mathcal{P}^{D2D} &= d - \mathcal{B}^{-1}\sum_{l=1}^{M}(\Gamma_l + \beta_l)^{-1}\Gamma_l\chi_l \\
&= \; d - \mathcal{B}^{-1}\sum_{l=1}^{M}(\Gamma_l + \beta_l)^{-1}\Gamma_l\left((\Gamma_l^{D2C} + \beta_l^{D2C})^{-1}(d + Z_{-l} - \mathcal{P}_l^{D2C}) \; + \; (\Gamma_l^{D2C} + \beta_l^{D2C})^{-1}\Gamma_l^{D2C}x_l\right) \\
&= \; \left[1 \; - \; \mathcal{B}^{-1}\sum_{l=1}^{M}\frac{\Gamma_l}{\Gamma_l + \beta_l}\frac{1}{\Gamma_l^{D2C} + \beta_l^{D2C}}\right]d \; + \; \mathcal{B}^{-1}\sum_{l=1}^{M}\frac{\Gamma_l}{\Gamma_l + \beta_l}\frac{1}{\Gamma_l^{D2C} + \beta_l^{D2C}}\mathcal{P}_l^{D2C} \\
&\quad - \; \mathcal{B}^{-1}\sum_{l=1}^{M}(\Gamma_l + \beta_l)^{-1}\Gamma_l\frac{1}{\Gamma_l^{D2C} + \beta_l^{D2C}}\left[Z_{-l} \; + \; \Gamma_l^{D2C}x_l\right] \\
&= \; \left[1 \; - \; \mathcal{B}^{-1}\sum_{l=1}^{M}\frac{\Gamma_l}{\Gamma_l + \beta_l}\frac{1}{\Gamma_l^{D2C} + \beta_l^{D2C}}\right]d \; + \; \mathcal{B}^{-1}\sum_{l=1}^{M}\frac{\Gamma_l}{\Gamma_l + \beta_l}\frac{1}{\Gamma_l^{D2C} + \beta_l^{D2C}}\mathcal{P}_l^{D2C} \\
&\quad - \; \mathcal{B}^{-1}\sum_{l=1}^{M}\frac{\Gamma_l}{\Gamma_l + \beta_l}\frac{1}{\Gamma_l^{D2C} + \beta_l^{D2C}}\left[C_x^{Z_{-l}} \cdot x \; + \; C_{x_l}^{Z_{-l},l}x_l \; + \; C_{\bar{\theta}}^{Z_{-l},l}\bar{\theta} \; + \; \Gamma_l^{D2C}x_l\right] \\
&= \; \left[1 \; - \; \mathcal{B}^{-1}\sum_{l=1}^{M}\frac{\Gamma_l}{\Gamma_l + \beta_l}\frac{1}{\Gamma_l^{D2C} + \beta_l^{D2C}}\right]d \; + \; \mathcal{B}^{-1}\sum_{l=1}^{M}\frac{\Gamma_l}{\Gamma_l + \beta_l}\frac{1}{\Gamma_l^{D2C} + \beta_l^{D2C}}\mathcal{P}_l^{D2C} \\
&\quad - \; \mathcal{B}^{-1}\sum_{l=1}^{M}\frac{\Gamma_l}{\Gamma_l + \beta_l}\frac{1}{\Gamma_l^{D2C} + \beta_l^{D2C}}C_x^{Z_{-l}} \cdot x \; - \; \mathcal{B}^{-1}\sum_{l=1}^{M}\frac{\Gamma_l}{\Gamma_l + \beta_l}\frac{1}{\Gamma_l^{D2C} + \beta_l^{D2C}}(C_{x_l}^{Z_{-l},l} \; + \; \Gamma_l^{D2C})x_l \\
&\quad - \; \mathcal{B}^{-1}\sum_{l=1}^{M}\frac{\Gamma_l}{\Gamma_l + \beta_l}\frac{1}{\Gamma_l^{D2C} + \beta_l^{D2C}}C_{\bar{\theta}}^{Z_{-l},l}\bar{\theta} \\
&= \widehat{C}_d^{\mathcal{P},D2D}d \; + \; \widehat{C}_{\mathcal{P}}^{\mathcal{P},D2D}\mathcal{P}^{D2C} \; + \; \widehat{C}_x^{\mathcal{P},D2D}x \; + \; \widehat{C}_{\bar{\theta}}^{\mathcal{P},D2D}\bar{\theta}\,,
\end{aligned}
$$

where we defined

$$
\begin{aligned}
\tilde{C}_x^{\mathcal{P},D2D} &= -\left(\frac{\Gamma_1}{\mathcal{B}(\Gamma_1+\beta_1)}\frac{C_{x_1}^{Z_{-1},1}+\Gamma_1^{D2C}}{\Gamma_1^{D2C}+\beta_1^{D2C}},\cdots,\frac{\Gamma_M}{\mathcal{B}(\Gamma_M+\beta_M)}\frac{C_{x_M}^{Z_{-M},M}+\Gamma_M^{D2C}}{\Gamma_M^{D2C}+\beta_M^{D2C}}\right)^T\\
\widehat{C}_d^{\mathcal{P},D2D} &= 1 \;-\; \mathcal{B}^{-1}\sum_{l=1}^M \frac{\Gamma_l}{\Gamma_l+\beta_l}\frac{1}{\Gamma_l^{D2C}+\beta_l^{D2C}}\\
\widehat{C}_{\mathcal{P}}^{\mathcal{P},D2D} &= \left(\frac{\Gamma_1}{\mathcal{B}(\Gamma_1+\beta_1)}\frac{1}{\Gamma_1^{D2C}+\beta_1^{D2C}},\cdots,\frac{\Gamma_M}{\mathcal{B}(\Gamma_M+\beta_M)}\frac{1}{\Gamma_M^{D2C}+\beta_M^{D2C}}\right)^T\\
\widehat{C}_x^{\mathcal{P},D2D} &= \tilde{C}_x^{\mathcal{P},D2D} \;-\; \mathcal{B}^{-1}\sum_{l=1}^M \frac{\Gamma_l}{\Gamma_l+\beta_l}\frac{1}{\Gamma_l^{D2C}+\beta_l^{D2C}}C_x^{Z_{-l}}\\
\widehat{C}_{\bar{\theta}}^{\mathcal{P},D2D} &= -\mathcal{B}^{-1}\sum_{l=1}^M \frac{\Gamma_l}{\Gamma_l+\beta_l}\frac{1}{\Gamma_l^{D2C}+\beta_l^{D2C}}C_{\bar{\theta}}^{Z_{-l},l}.
\end{aligned}
$$

Substituting

$$
x \;=\; (\mathcal{A}^x)^{-1}\left[\mathcal{P}^{D2C} \;-\; d\mathbf{1} \;-\; \mathcal{V}^{\Theta}\Theta \;-\; \mathcal{V}^{\bar{\theta}}\bar{\theta}\right],
$$

we will get a linear relationship between $\mathcal{P}^{D2D}$ and the vector of $(\mathcal{P}_l^{D2C})_l$:

$$
\begin{aligned}
\mathcal{P}^{D2D} \;&=\; \widehat{C}_d^{\mathcal{P},D2D}d \;+\; \widehat{C}_{\mathcal{P}}^{\mathcal{P},D2D}\mathcal{P}^{D2C} \;+\; \widehat{C}_x^{\mathcal{P},D2D}x \;+\; \widehat{C}_{\bar{\theta}}^{\mathcal{P},D2D}\bar{\theta}\\
&=\; \widehat{C}_d^{\mathcal{P},D2D}d \;+\; \widehat{C}_{\mathcal{P}}^{\mathcal{P},D2D}\mathcal{P}^{D2C} \;+\; \widehat{C}_x^{\mathcal{P},D2D}(\mathcal{A}^x)^{-1}\left[\mathcal{P}^{D2C} \;-\; d\mathbf{1} \;-\; \mathcal{V}^{\Theta}\Theta \;-\; \mathcal{V}^{\bar{\theta}}\bar{\theta}\right] \;+\; \widehat{C}_{\bar{\theta}}^{\mathcal{P},D2D}\bar{\theta}\\
&=\; \left[\widehat{C}_d^{\mathcal{P},D2D} \;-\; (\widehat{C}_x^{\mathcal{P},D2D})^T(\mathcal{A}^x)^{-1}\mathbf{1}\right]d \;+\; \left[\widehat{C}_{\mathcal{P}}^{\mathcal{P},D2D} \;+\; ((\mathcal{A}^x)^{-1})^T\widehat{C}_x^{\mathcal{P},D2D}\right]\mathcal{P}^{D2C}\\
&\quad+\; \left[\widehat{C}_{\bar{\theta}}^{\mathcal{P},D2D} \;-\; (\widehat{C}_x^{\mathcal{P},D2D})^T(\mathcal{A}^x)^{-1}\mathcal{V}^{\bar{\theta}}\right]\bar{\theta} \;-\; (\widehat{C}_x^{\mathcal{P},D2D})^T(\mathcal{A}^x)^{-1}\mathcal{V}^{\Theta}\Theta\\
&=\; C_d^{\mathcal{P},D2D}d \;+\; C_{\mathcal{P}}^{\mathcal{P},D2D}\cdot\mathcal{P}^{D2C} \;+\; C_{\Theta}^{\mathcal{P},D2D}\Theta \;+\; C_{\bar{\theta}}^{\mathcal{P},D2D}\bar{\theta},
\end{aligned}
$$

where we have defined

$$
\begin{aligned}
C_d^{\mathcal{P},D2D} &= \widehat{C}_d^{\mathcal{P},D2D} \;-\; (\widehat{C}_x^{\mathcal{P},D2D})^T(\mathcal{A}^x)^{-1}\mathbf{1}\\
C_{\mathcal{P}}^{\mathcal{P},D2D} &= ((\mathcal{A}^x)^{-1})^{\top}\widehat{C}_x^{\mathcal{P},D2D} \;+\; \widehat{C}_{\mathcal{P}}^{\mathcal{P},D2D}\\
C_{\bar{\theta}}^{\mathcal{P},D2D} &= \widehat{C}_{\bar{\theta}}^{\mathcal{P},D2D} \;-\; (\widehat{C}_x^{\mathcal{P},D2D})^T(\mathcal{A}^x)^{-1}\mathcal{V}^{\bar{\theta}}\\
C_{\Theta}^{\mathcal{P},D2D} &= \;-\,(\widehat{C}_x^{\mathcal{P},D2D})^T(\mathcal{A}^x)^{-1}\mathcal{V}^{\Theta}.
\end{aligned}
$$

Lemma IA.12 *For small heterogeneity in risk aversion, we have*

$$
C_{\mathcal{P}}^{\mathcal{P},D2D} \;=\; \bar{\pi}^*\mathbf{1} \;+\; \pi^{(1)}(\mathbf{\Gamma}-\mathbf{\Gamma}^*) \;+\; O(\|\Gamma-\Gamma^*\|^2),
$$

where $\mathbf{1} = (1, 1, \cdots, 1)^T$, $\mathbf{\Gamma} = (\Gamma_\ell)_{\ell=1}^M$, *and both* $\bar{\pi}^*$ *and* $\pi^{(1)}$ *are constant.*

Proof. It is straightforward to show that, for small heterogeneity in risk aversion, we have

$$C_{\mathcal{P}}^{\mathcal{P},D2D} \;=\; \bar{\Pi}^* \;+\; \Pi^{(1)}(\mathbf{\Gamma} - \mathbf{\Gamma}^*) \;+\; O(\|\Gamma - \Gamma^*\|^2)\,,$$

where $\bar{\Pi}^*$ is a vector and $\Pi^{(1)}$ is a matrix, and both are constant. Equilibrium considerations imply that

$$\begin{aligned} \bar{\Pi}^* &= \bar{\pi}^*\,\mathbf{1} \\ \Pi^{(1)} &= a\,\mathbf{1}\otimes\mathbf{1} \;+\; \pi^{(1)}\,\mathrm{Id}\,. \end{aligned}$$

for scalars $a, \bar{\pi}^*$, and $\pi^{(1)}$. Moreover,

$$\mathbf{1}\otimes\mathbf{1}(\mathbf{\Gamma} - \mathbf{\Gamma}^*) \;=\; \mathbf{0}\,.$$

The result then follows. Q.E.D.

The natural proxy for the (half) bid-ask spread in the D2C market is the price impact that customers have when trading with a particular dealer, given by $\beta_{c,l}^{D2C}$. By direct calculation,

$$\beta_{c,l}^{D2C} \;=\; \beta_{c,*}^{D2C} \;+\; \beta_c^{(1)}(\Gamma_l - \Gamma^*) \;+\; O(\|\Gamma - \Gamma^*\|^2)\,.$$

The sign of $\beta_c^{(1)}$ will play an important role later on. The proposition below follows directly the results in Malamud and Rostek (2017):

Proposition IA.16 *We always have* $\beta_c^{(1)} > 0$.

Thus, we can write

$$\Gamma_l - \Gamma^* \;\approx\; (\beta_{c,l}^{D2C} - \beta_{c,*}^{D2C})/\beta_c^{(1)}\,.$$

This formula allows us to express the latent risk aversion through the observable bid-ask spreads. Thus, we rewrite

$$\begin{aligned} \mathcal{P}^{D2D} &= C_d^{\mathcal{P},D2D} d \;+\; C_{\mathcal{P}}^{\mathcal{P},D2D}\cdot\mathcal{P}^{D2C} \;+\; C_\Theta^{\mathcal{P},D2D}\Theta \;+\; C_{\bar{\theta}}^{\mathcal{P},D2D}\bar{\theta} \\ &\approx C_d^{\mathcal{P},D2D} d \;+\; (\bar{\pi}^*\mathbf{1} \;+\; \pi^{(1)}(\mathbf{\Gamma} - \mathbf{\Gamma}^*))\cdot\mathcal{P}^{D2C} \;+\; C_\Theta^{\mathcal{P},D2D}\Theta \;+\; C_{\bar{\theta}}^{\mathcal{P},D2D}\bar{\theta} \\ &\approx C_d^{\mathcal{P},D2D} d \;+\; \left(\bar{\pi}^*\mathbf{1} \;+\; \frac{1}{\beta_c^{(1)}}\pi^{(1)}((\beta_{c,l}^{D2C})_{l=1}^M - \beta_{c,*}^{D2C})\right)\cdot\mathcal{P}^{D2C} \;+\; C_\Theta^{\mathcal{P},D2D}\Theta \;+\; C_{\bar{\theta}}^{\mathcal{P},D2D}\bar{\theta}\,. \end{aligned}$$

We have

$$\beta_{c,l}^{D2C} \;=\; \lambda_l^{-1}$$

and hence, up to second-order terms in the size of heterogeneity,

$$((\beta_{c,l}^{D2C})_{l=1}^{M} - \beta_{c,*}^{D2C}) \cdot \mathcal{P}^{D2C} = ((\beta_{c,l}^{D2C})_{l=1}^{M} - \beta_{c,*}^{D2C}) \cdot \alpha \; M\,\mathrm{Cov}(\lambda_l^{-1}, \alpha)$$
$$\approx\; -M\lambda_*^{-2}\,\mathrm{Cov}(\lambda, \alpha) \;=\; -M\lambda_*^{-1}\alpha_{mismatch}\,.$$

Hence,

$$\mathcal{P}^{D2D} \approx\; C_d^{\mathcal{P},D2D} d \;+\; \bar{\pi}^* \bar{\alpha} \;-\; M\lambda_*^{-1}(\pi^{(1)}/\beta_c^{(1)})\alpha_{mismatch} \;+\; C_\Theta^{\mathcal{P},D2D}\Theta \;+\; C_{\bar{\theta}}^{\mathcal{P},D2D}\bar{\theta}\,.$$

We know that $\beta_c^{(1)}$ is positive, and hence we just need to figure out the sign of $\pi^{(1)}$. While we cannot solve for the sign of $\pi^{(1)}$ analytically, we use extensive numerical analysis to show that:

Numerical Claim:

$$\pi^{(1)} > 0$$

when n is sufficiently large.

We now determine $\pi^{(1)}$. Recall from Lemma IA.12 that $\pi^{(1)}$ is defined via the Taylor series expansion of $C_{\mathcal{P}}^{\mathcal{P},D2D}$ and that

$$C_{\mathcal{P}}^{\mathcal{P},D2D} = \; ((\mathcal{A}^x)^{-1})^\top \widehat{C}_x^{\mathcal{P},D2D} \;+\; \widehat{C}_{\mathcal{P}}^{\mathcal{P},D2D}\,,$$

with

$$\mathcal{A}^x = \; \mathcal{A}_0^x \;+\; diag(C_{x_l}^{\mathcal{P},l}); \qquad \mathcal{A}_0^x \;=\; \sum_{\ell=1}^{M} e_\ell \otimes C_x^{\mathcal{P},\ell}\,.$$

We examine the components of matrix $\mathcal{A}^x$, starting with the terms $C_x^{\mathcal{P},l}$.

$$C_x^{\mathcal{P},l} = \; \frac{1}{\mathcal{B}_l^{D2C}(\Gamma_l^{D2C} + \beta_l^{D2C})} C_x^{Z_{-l}} \;=\; \frac{1}{\mathcal{B}_l^{D2C}(\Gamma_l^{D2C} + \beta_l^{D2C})} \frac{\mathcal{B}(\Gamma_l + \beta_l)}{\mathcal{B}(\Gamma_l + \beta_l) + \Gamma_l \Delta_l C_l^Z} \Delta_l \bar{C}_{\bar{Z}}^x\,.$$

It follows that

$$\mathcal{A}_0^x = \; \sum_{\ell=1}^{M} e_\ell \otimes C_x^{\mathcal{P},\ell} \;=\; v \otimes \bar{C}_{\bar{Z}}^x$$

where

$$v \;=\; \left(\frac{1}{\mathcal{B}_\ell^{D2C}(\Gamma_\ell^{D2C} + \beta_\ell^{D2C})} \frac{\mathcal{B}(\Gamma_\ell + \beta_\ell)}{\mathcal{B}(\Gamma_\ell + \beta_\ell) + \Gamma_\ell \Delta_\ell C_\ell^Z} \Delta_\ell \right)_{\ell=1}^{M}$$

and recall that

$$
\begin{aligned}
Dem_{\bar{Z}} \;&=\; 1 \;-\; \sum_{\ell} \frac{1}{1+\frac{1}{C_\ell^Z \Delta_\ell \frac{\Gamma_\ell}{\mathcal{B}(\Gamma_\ell+\beta_\ell)}}} \\
\bar{C}^x_{\bar{Z},l} &= \;\frac{1}{Dem_{\bar{Z}}}\left(\left(\frac{\Gamma_l}{\mathcal{B}(\Gamma_l+\beta_l)}\right)^{-1} + C_l^Z \Delta_l\right)^{-1} C_l^x \\
\bar{C}^x_{\bar{Z}} \;&=\; (\bar{C}^x_{\bar{Z},1}, \bar{C}^x_{\bar{Z},2}, \cdots, \bar{C}^x_{\bar{Z},M})^T \\
C_l^x \;&=\; 1 \;+\; \frac{\Gamma_l^{D2C}}{\Gamma_l^{D2C}+\beta_l^{D2C}}\left[-1 \;+\; \frac{1}{\mathcal{B}_l^{D2C}(\Gamma_l^{D2C}+\beta_l^{D2C})}\right]
\end{aligned}
$$

In addition,

$$
\begin{aligned}
\Delta_l \;&=\; \frac{\Gamma_l+2\beta_l}{(\Gamma_l+\beta_l)^2}\left(\frac{1}{\mathcal{B}(\Gamma_l+\beta_l)}-1\right)\Gamma_l \\
C_l^Z \;&=\; \frac{1}{\Gamma_l^{D2C}+\beta_l^{D2C}}\left[1-\frac{1}{\mathcal{B}_l^{D2C}(\Gamma_l^{D2C}+\beta_l^{D2C})}\right] .
\end{aligned}
$$

We use all these expressions in the numerical work regarding $\mathcal{A}_0^x$. Now, we turn our attention to the term $diag(C_{x_l}^{\mathcal{P},l})$:

$$
\begin{aligned}
C_{x_l}^{\mathcal{P},l} &= \;\frac{1}{\mathcal{B}_l^{D2C}(\Gamma_l^{D2C}+\beta_l^{D2C})}\left[C_{x_l}^{Z_{-l},l} \;-\; \Gamma_l^{D2C}\right] \\
&= \;-\frac{1}{\mathcal{B}_l^{D2C}(\Gamma_l^{D2C}+\beta_l^{D2C})}\left[\frac{\Gamma_l\Delta_l}{\mathcal{B}(\Gamma_l+\beta_l)+\Gamma_l\Delta_l C_l^Z}C_l^x \;+\; \Gamma_l^{D2C}\right] .
\end{aligned}
$$

We use these expressions to evaluate $\mathcal{A}^x$ and thus

$$
C_{\mathcal{P}}^{\mathcal{P},D2D} = \;((\mathcal{A}^x)^{-1})^\top \widehat{C}_x^{\mathcal{P},D2D} \;+\; \widehat{C}_{\mathcal{P}}^{\mathcal{P},D2D} ,
$$

where,

$$\widehat{C}_{\mathcal{P}}^{\mathcal{P},D2D} = \left(\frac{\Gamma_\ell}{\mathcal{B}(\Gamma_\ell+\beta_\ell)}\frac{1}{\Gamma_\ell^{D2C}+\beta_\ell^{D2C}}\right)_{\ell=1}^{M}$$

$$\widehat{C}_{x}^{\mathcal{P},D2D} = \tilde{C}_{x}^{\mathcal{P},D2D} - \mathcal{B}^{-1}\sum_{l=1}^{M}\frac{\Gamma_l}{\Gamma_l+\beta_l}\frac{1}{\Gamma_l^{D2C}+\beta_l^{D2C}}C_x^{Z_{-l}}$$

$$\tilde{C}_{x}^{\mathcal{P},D2D} = -\left(\frac{\Gamma_\ell}{\mathcal{B}(\Gamma_\ell+\beta_\ell)}\frac{C_{x_\ell}^{Z_{-\ell},\ell}+\Gamma_\ell^{D2C}}{\Gamma_\ell^{D2C}+\beta_\ell^{D2C}}\right)_{\ell=1}^{M}$$

$$C_x^{Z_{-l}} = \frac{\mathcal{B}(\Gamma_l+\beta_l)}{\mathcal{B}(\Gamma_l+\beta_l)+\Gamma_l\Delta_l C_l^Z}\Delta_l\bar{C}_{\bar{Z}}^{x}$$

$$\widehat{C}_{x}^{\mathcal{P},D2D} = \tilde{C}_{x}^{\mathcal{P},D2D} - \left[\sum_{l=1}^{M}\frac{\Gamma_l}{\mathcal{B}(\Gamma_l+\beta_l)}\frac{1}{\Gamma_l^{D2C}+\beta_l^{D2C}}\frac{1}{1+\frac{\Gamma_l}{\mathcal{B}(\Gamma_l+\beta_l)}\Delta_l C_l^Z}\Delta_l\right]\bar{C}_{\bar{Z}}^{x}\,.$$

We are now in the position the evaluate $\pi^{(1)}$ numerically. The exogenous parameters are (i) the number of customers n, (ii) the number of dealers M, (iii) the customers' coefficient of risk-aversion γ, and (iv) the dealers' average coefficient of risk-aversion Γ^*. Following Vayanos and Woolley (2013), we choose

$$\Gamma^* = 30\,.$$

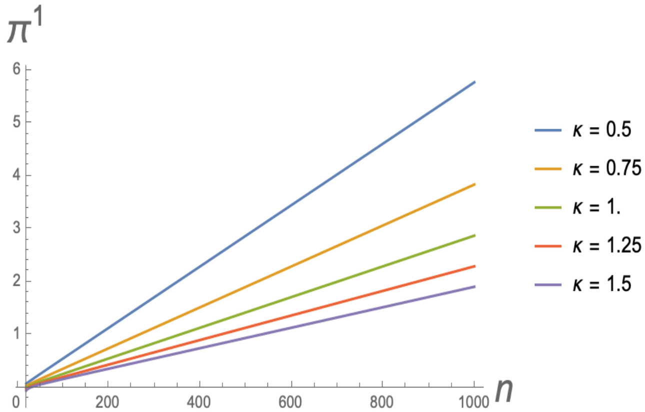
We define

$$\kappa = \frac{\gamma}{\Gamma^*}\,,$$

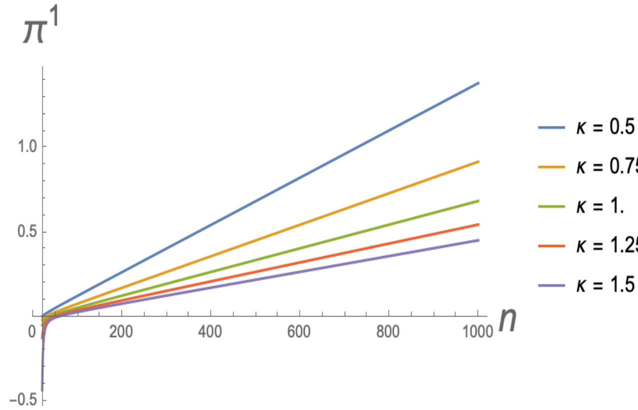
and use the values

$$\kappa \in \{0.5, 0.75, 1, 1.25, 1.5\}\,.$$

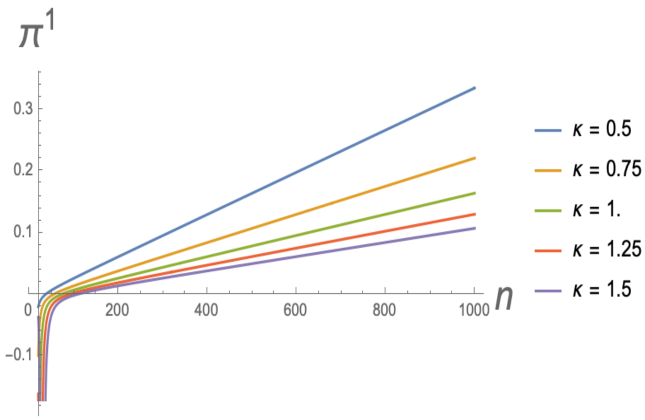
Figure IA.1 plots $\pi^{(1)}$ as a function of n, for $M \in \{5, 10, 20, 40\}$. The figure shows that $\pi^{(1)}$ is negative for n sufficiently large. In our data, the number of dealers M fluctuates between 8 and 12.



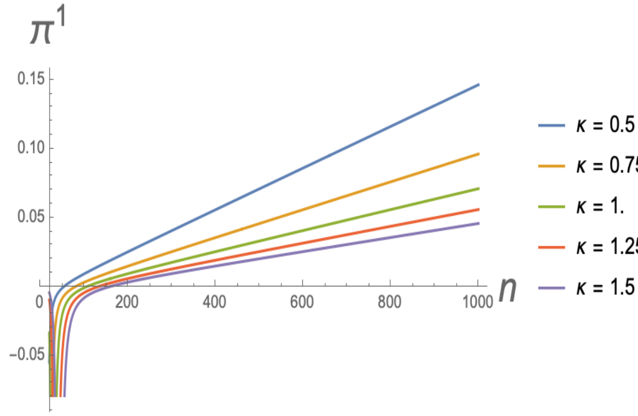
(a) $M = 5$.



(b) $M = 10$.



(c) $M = 20$.



(d) $M \;=\; 30$.

Figure IA.1: Liquidity Mismatch in a Fragmented Double Auction Model The relation between the D2D price and $\alpha_{mismatch}$ in a market such as the one in Babus and Parlatore (2022) is given by

$$\mathcal{P}^{D2D} \;\approx\; C_d^{\mathcal{P},D2D} d + \bar{\pi}^* \bar{\alpha} - \frac{M\lambda_*^{-1}}{\beta_c^{(1)}} \pi^{(1)}) \alpha_{mismatch} + C_\Theta^{\mathcal{P},D2D} \Theta + C_{\bar{\theta}}^{\mathcal{P},D2D} \bar{\theta},$$

where λ_* and $\beta_c^{(1)}$ are positive. We plot $\pi^{(1)}$ as a function of n. The other parameters are

$$\Gamma^* \;=\; 30 \qquad \text{and} \qquad \kappa = \frac{\gamma}{\Gamma^*}.$$