Trades, Quotes, and Information Shares[1]

Björn Hagströmer and Albert J. Menkveld

March 27, 2023

[1]Björn Hagströmer, Stockholm Business School, Stockholm University, S-106 91 Stockholm, Sweden, tel +46 8 163030, bjh@sbs.su.se. Albert J. Menkveld, Vrije Universiteit Amsterdam, SBE, De Boelelaan 1105, 1081 HV, Amsterdam, Netherlands, tel +31 20 598 6130, albertjmenkveld@gmail.com and Tinbergen Institute. Hagströmer is a visiting research fellow at Swedish House of Finance. He gratefully acknowledges Jan Wallander Foundation and Tom Hedelius Foundation for research funding. Menkveld gratefully acknowledges NBIM for an NFI grant and NWO for a Vici grant. This article benefited greatly from discussions with Fatemeh Aramian, Frank de Jong (discussant), Terrence Hendershott, Joel Hasbrouck (discussant), Anqi Liu, Lars Nordén, and Ryan Riordan (discussant), Lucas Saru as well as seminar participants at Columbia Business School, the SAFE Market Microstructure Conference, Stockholm Business School, Vrije Universiteit Amsterdam, and Vienna University of Economics and Business.

Trades, Quotes, and Information Shares

Abstract

Information arrives at securities markets through price quotes and trades. Informed traders impose adverse-selection costs on quote suppliers. This creates incentives for the latter to identify relatively uninformed groups and trade with them off-exchange. The marketplace turns hybrid, at the cost of thinner, highly informed (toxic) volume at the center. This pattern has largely eluded econometricians, because the standard approach to measuring information shares is biased against finding it. We show why this is the case, and design a bias-free approach. The novel approach shows that, indeed, the conjectured pattern is strongly present in the data.

Online appendix with additional results is available at https://bit.ly/3HoD8oD.

1 Introduction

Market structure seems to be perennially changing. It is subject to two opposite forces. Participation externalities pulls it towards centralization (Pagano, 1989; Chowdhry and Nanda, 1991), but information asymmetry across traders acts as a centrifugal force. The reason is that uninformed traders subsidize informed ones in the single centralized market. The uninformed might therefore be better off trading among themselves in a peripheral market, even if that is at the cost of reduced participation externalities (Easley, Kiefer, and O'Hara, 1996; Battalio and Holden, 2001). Various manifestations of this centrifugal force are dark pools (DPs), single-dealer platforms (SDPs), periodic auctions (PAs), and payment for order flow (PFOF).

Regulators worry about the social cost of an overly fragmented market structure. In the US, the Securities and Exchange Commission (SEC) seems mostly concerned about misuse of private customer information in peripheral markets, commonly referred to as alternative trading systems (SEC, 1998, 2015). Regulators in the European Union (EU) worry about *"unduly harm[ed] price formation"* if too much volume is executed off-exchange (EU, 2014, Article 5).

To determine whether the concerns have merit requires measuring the informativeness of price quotes, on-exchange trades, and off-exchange trades.[1] This calls for a holistic view on how information enters prices. It requires identifying information in time series of quotes and trades. The *de facto* standard approach is the information share framework by Hasbrouck (1995, henceforth H95), which involves estimation of a vector-error correction model (VECM). This approach has two strong advantages. First, a VECM can be estimated with ordinary least squares (OLS). This is particularly appealing when samples become large due to sub-second time resolutions. Second, the approach does not require distributional assumptions other than that price differentials are stationary and ergodic. All

[1]Note that price quotes in this case are on-exchange price quotes, since off-exchange quotes are mostly non-existent (DPs) or private (SDPs).

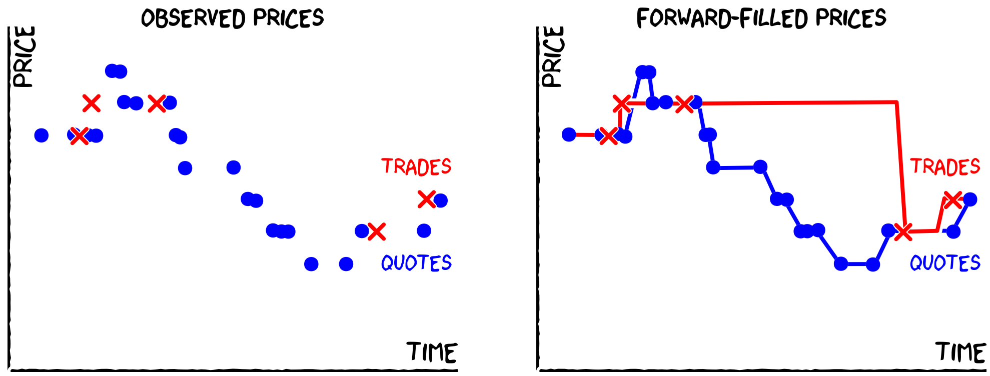


Figure 1: Forward-Filled Prices

results are essentially identified off of the auto-covariance function.

However, whereas H95 is economically reasonable for a multivariate system of price quotes, we believe that it is *not* when one adds trade prices. First, as Hasbrouck (2003) observes, trades happen at either a bid price or at an ask price, which, therefore, adds noise to trade prices, disadvantaging them relative to quote prices.

Second, we believe that a far more damaging disadvantage of trade prices in H95 is the practice of forward-filling (e.g., Hasbrouck (2003, 2021) and Chakravarty, Gulen, and Mayhew (2004)). Estimating a VECM with forward-filled trade prices essentially assumes that such prices somehow remain valid after a trade, until the next trade occurs. Forward-filling price quotes, on the other hand, makes economic sense, because quotes are valid until revised or cancelled. Figure 1 illustrates the practice. Note how the forward-filling can lead to large and persistent price differentials between price quotes and trades, until a next trade arrives. This artificial wedge is then seemingly closed in favor of price quotes, since trade prices seem to move towards quotes, artificially advantaging price quotes relative to trades.[2,3]

[2]More specifically, note how, in the long period between the third and fourth trade, the price seems to move towards the quote series, and away from the (stale) trade price series. This in contrast to other periods, where quotes tend to move *towards* trade prices, which is a common immediate response to reflect the information in trades.

[3]The only note about this problem in previous literature, to our knowledge, is by Hasbrouck (1995,

The contribution of our paper is essentially three-fold. First, we document that, indeed, H95 information shares are biased when both quote and trade prices are included. We do so by running simulations where we know the true information share. Specifically, forward-filling causes information shares of trade prices to be underestimated. We refer to this as a "stale-price bias."

However, if trades occur sporadically, then their information share is *over*estimated. Or, more precisely, if the number of trades in the sample tends to one, then the information share tends to 50%.[4] We believe that this effect is purely mechanical and we refer to it as a "small-sample bias."

Taken together, these biases tend to push the information share estimates of exchange trades downwards and those of off-exchange trades upwards, because there are typically many exchange trades and very few off-exchange trades. Forward-filling, therefore, biases against finding the conjectured pattern of uninformed trades being pulled off of the exchange. Our analysis of real-world data shows that both biases are economically significant.

Second, we develop a generalized, bias-free, H95 framework that recognizes the different nature of price quotes and trades. In a nutshell, the alternative approach avoids forward-filling by replacing trade prices in the VECM with signed volumes. The latter is the sum of all buyer-initiated trade sizes in a period, minus the sum of all seller-initiated trade sizes. In periods without trades, the signed volume is zero. There are thus no missing observations and there is, therefore, no need for forward-filling. And, in addition to avoid-

p. 1187–1188): *"An analysis of an individual stock[,] trading in multiple markets, if based on last sale prices, would labor under problems of autocorrelation induced by infrequent trading. In particular, a market that happened to have relatively infrequent trades would tend to have last-sale prices that were most obsolete, and therefore least informative."* Hasbrouck used this as motivation for using quotes instead of trades in his analysis. An obvious disadvantage of this approach is that one loses the ability to identify information in trades. And, one necessarily excludes dark markets, because they lack price quotes.

[4]The intuition is simply that, for a single trade, subsequent price quotes either tend to revert to this price in the remainder of the sample, or they do not. The information share in these two cases is either 100% or 0%, respectively, thus averaging to 50%.

ing bias, the alternative approach has the advantage that trade *size* enters the system, which recognizes that large trades are potentially more informative than small ones.

Our alternative approach is inspired by Hasbrouck (1991), who uses signed volume to study information shares in a *single*-market setting. Such a setting can be analyzed with a simpler vector auto-regression (VAR) model, which is nested in a VECM. One could, therefore, say that our multi-market approach effectively marries Hasbrouck (1991) with H95. The outcome is a measure of each market's contribution to price discovery, both in terms of its price quotes and its trades. We refer to this information share as *generalized information share (GIS)*. Our approach generalizes H95 in the sense that it yields unbiased estimates of information shares for price both quotes and trades.

Third, we apply the new approach to stocks listed at the London Stock Exchange (LSE). The analysis includes price quotes and trades in the main exchange (i.e., LSE), in other exchanges, and in a variety of off-exchange venues. To our knowledge, we are the first to analyze informativeness of different types of off-exchange venues. We find that, indeed, on-exchange trades are a lot more informative than off-exchange trades. In terms of information per pound traded, exchange trades are at least *four times* more informative than off-exchange trades. Further comparisons across off-exchange venue types do not reveal any differences, except that periodic auctions are even less informative than the rest (per pound traded). Our full analysis yields a granular and complete picture of information arrival. Overall, we find that (on-exchange) price quotes contribute at least 46% to price discovery, on-exchange trades contribute at least 12%, and all types of off-exchange venues jointly contribute *at most* 0.34%.[5]

Literature. In her discussion of the new high-frequency market structure, O'Hara (2015, p. 263) observes that *"trades are not the basic unit of market information – the underlying*

[5]The remaining 42% cannot be uniquely assigned due to commonalities across all time series. This is a statistical artefact and affects both H95 and our approach. Information shares are identified up to a lower and an upper bound.

orders are." Empirical models of information arrival should therefore model price quotes and trades jointly. Brogaard, Hendershott, and Riordan (2014) study how the quotes and trades of high-frequency traders facilitate price discovery. They do so by estimating a version of the state-space model (SSM) that is developed in Hendershott and Menkveld (2014). An attractive feature of SSMs, in this context, is that they deal naturally with non-availables by filtering, and thus avoid forward-filling (Durbin and Koopman, 2012). An important drawback is that estimation is non-trivial for large samples. Our model avoids non-availables and can, therefore, be estimated with OLS, which handles large samples well.[6]

Theoretical studies on informed trading provide various insights into how informed traders trade in the presence of exogenous uninformed trading. Chowdhry and Nanda (1991) show that they split orders across markets in proportion to the amount of uninformed trading in these markets. Zhu (2014) shows that competitive informed traders prefer exchanges over DPs, because execution is uncertain in DPs. The reason is that an order in a DP only executes if there is an opposite-sign order in the DP. Ye and Zhu (2020) argue that Zhu's finding hinges critically on the assumption of competition, because a monopolistic informed trader prefers dark venues. They back up this point up by showing that informed trading by activist hedge funds is positively related to dark pool market share. Further supportive evidence is provided by Albuquerque, Song, and Yao (2020) and Brogaard and Pan (2022). These results suggest that informed traders, to the extent that they have access to DPs, will only avoid them if they are under competitive pressure.

Empirical studies that use H95 to analyze off-exchange information shares yield the following results. For Australian stocks, Comerton-Forde and Putniņš (2015) find that long-term price impact of DP trades is about 10% lower than that of exchange trades. Hatheway, Kwan, and Zheng (2017) report a 8% lower bound for dark-market informa-

[6]In later work, Brogaard, Hendershott, and Riordan (2019) use VECM to study quote informativeness across exchanges (Table XI). They stop short of adding trades to this analysis. Our approach makes this possible.

tion share for Nasdaq stocks. Chakrabarty, Cox, and Upson (2022) report that US stocks subject to the tick size pilot have economically significant DP information shares (ranging from 11% to 15%), albeit less than their volume shares (27% to 40%). Taken together, these studies suggest that dark trades seem only marginally less informative than exchange trades. The biases we highlight suggest that this wedge is low because it might be underestimated. Hasbrouck (2021), in contrast, finds a wedge that is substantially larger with the informational contribution of dark trades being negligible compared to exchange trades. He, however, studies stocks that have relatively large dark volume shares, more than 25%. His dark trades are, therefore, unlikely to be affected by the positive small-sample bias.[7] The contribution of our study over these studies is that we offer unbiased estimates of the information shares of various types of trades and quotes. We, therefore, implicitly respond to the call of O'Hara (2015) to include quotes and trades in information-share analysis.

Outline. The rest of the paper is organized as follows. Section 2 reviews market structure for equities in the United Kingdom (UK). Section 3 develops the generalized information share framework. Section 4 documents how information shares become biased when trade prices are forward-filled, first in simulations, then in real-world data. Section 5 uses the new approach to analyze price discovery in UK equity markets. Section 6 concludes.

2 UK Equity Markets

We study the informativeness of quotes and trades for a 2021 sample of LSE stocks. The sample includes a variety of venue types, which will all be discussed in this section.

Following the Markets in Financial Instruments Directive (MiFID, which came into force in 2007) and its follow-up directive (MiFID II, 2018), European equity trading is

[7] Other studies avoid VECM estimation at the cost of offering a narrower view. They generally find that dark trades are less informative than exchange trades (Grammig and Theissen, 2012; Aramian and Nordén, 2021; Ernst, Sokobin, and Spatt, 2021).

fragmented across competing "lit" exchanges and several types of off-exchange "dark" venues, including DPs, PAs, and SDPs. This *hybrid* market structure is similar to US equity market structure, and increasingly similar to the market structure of other asset classes (Johnson, 2010).[8]

The reason for picking stocks listed at the LSE is that all the relevant trading venues are located in close proximity to the listing venue. For US stocks or other European stocks, the primary listing venue and the competing trading platforms tend to be located in different cities, which implies that timestamps are not easily comparable due to geographic latency. The close proximity of all relevant venues for LSE stocks implies that, in this case, geographic latency is an order of magnitude smaller than our sampling frequency (which is one millisecond).[9]

The LSE operates a limit order book market during its trading hours from 8:00 AM (GMT) until 4:30 PM. The market opens and closes with a call auction. At noon, trading is interrupted for two minutes to facilitate another call auction. All these auctions are excluded from our analysis because our focus is on regular trading.

During LSE trading hours, there are four other types of venues that trade LSE stocks:

- **Multilateral trading facilities (MTFs).** The on-exchange trading activity in the UK is split between the listing exchange and the MTFs. The latter do not have their own listings, but operate limit order book markets for securities listed at the LSE. The on-exchange markets have in common that they are pre-trade transparent, disseminating price quotes and depths in real time.

- **Dark pools (DPs).** DPs are multilateral trading venues that operate without pre-trade transparency and execute trades at a price determined in the lit markets. Under

[8]For a more detailed survey of market fragmentation and its effects on equity markets in Europe, see Hagströmer (2022).

[9]For example, the geographic latency between the LSE and it main competitor, BATS, was around 0.32 milliseconds in 2015 (Aquilina, Foley, O'Neill, and Ruf, 2016).

MiFID II, which is incorporated in UK legislation, DPs can operate without pre-trade transparency under the "reference price waiver," the "negotiated price waiver," and the "large-in-scale waiver." Under the reference price waiver, the DP trade price is typically set to the bid-ask spread midpoint at the listing venue.[10]

- **Periodic auctions (PAs).** PA orders can be submitted at any time during the continuous trading session. When there is matching PA interest, a short call auction procedure, lasting around a tenth of a second, is triggered. This can happen as often as several times a second (Neumeier, Gozluklu, Hoffmann, O'Neill, and Suntheim, 2021). Continuous trading is not interrupted during these auctions.

  Although PAs, technically, are operated on-exchange—by the LSE and CBOE for UK stocks—we categorize them as off-exchange. PAs are transparent in the sense that indicative execution prices and volumes are transmitted to the market in real time during the auction period. Accordingly, no pre-trade transparency waiver is required. However, pre-trade transparency is limited, because unmatched PA orders are not disclosed. Furthermore, PAs are similar to DPs in that orders can be priced with reference to the midpoint at the time of the auction. Two recent papers on European equities trading, by Johann, Putniņš, Sagade, and Westheide (2019) and Neumeier et al. (2021), find that PAs are close substitutes to DPs.

- **Single-dealer platforms (SDPs).** SDPs, known in Europe as "systematic internalisers," are bilateral trading platforms operated by dealers. The dealer quotes bid and ask prices and thus participates in each trade. The most active SDPs are run by investment banks and high-frequency trading firms. SDPs have to publish their quotes, but some additional rules and limitations apply.[11]

[10]The negotiated price waiver provides an exception from pre-trade transparency for trades that are negotiated bilaterally, but reported to the market in a formalized way. The large-in-scale waiver exempts block trades from pre-trade transparency. The threshold for large-in-scale trades is set relative to the average daily volume.

[11]Specifically, SDP quotes are not required to be made public when the quote volume exceeds *standard*

We analyze trades executed at all of the above venue types. To do so, we require a framework that accommodates both pre-trade transparent venues and dark markets.

3 Approach

This section presents our approach to measuring information shares, which recognizes the different nature of price quotes and trade prices. The key distinction is that while price-quote series are essentially continuous, trade-price series are not. The approach further recognizes that large trades are potentially more informative than small ones.

Most securities markets operate in continuous time. Price series are therefore point processes. When mapped to discrete-time intervals at the wall clock, there will be intervals without "price events." Absence of price events for quotes can reasonably be interpreted as the latest quotes remaining valid. This is due to firm commitments in limit-order book protocols, but even in the polar case of human-intermediated over-the-counter (OTC) markets, there are reputational costs to dealers for not honoring their latest price quotes. It therefore makes economic sense to forward-fill price quotes until a next update. The same logic does not apply to trade prices. We agree with H95 when he observes that, after a trade, the "last-sale price" tends to become "obsolete" (see footnote 3). We show in Section 4 that, indeed, forward-filling trade prices can lead to substantial bias in information share estimates.[12] We avoid forward-filling by using signed volume, as will become clear in the next section.

market size, which is stock-specific and informed by historical volume. However, during LSE trading hours, they *are* required to publish two-way quotes on a continuous basis for a volume of at least 10% of standard market size. We do not have access to such quotes.

[12]There are several econometric alternatives to forward-filling (Lin and Tsai, 2020). For univariate series, for example, one could use backward-filling, interpolation methods, moving average methods, or Kalman smoothing. For multivariate series, one could use neural networks or random forests. None of these methods addresses our economic concern that imputed trade prices are stale and, therefore, do not pertain to the current state of the market.

3.1 Econometric Model

Consider a set of size C that contains venues that disseminate price quotes on a continuous basis, and another set of size E with venues that disseminate trades. A venue that disseminates price quotes typically also reports trades, in which case it is included in both sets. Note that C emphasizes that the series is continuous in nature, whereas E emphasizes the event-type nature of the series. Our application involves trade events, but, more generally, the approach can accommodate any event-type series (e.g., target price announcements by financial analysts or fundamental value claims vented on social media).

Our approach requires that there is at least one venue that disseminates quotes (i.e., $C \geq 1$ and $E \geq 0$). Let p_t^c denote the (column) vector with the C midquotes at time t, where the midquote of a venue is the average of the best bid and ask quote. Let x_t denote the vector with the E signed volumes in the interval $(t-1, t]$, defined as the sum of signed volumes for trades in that interval.[13]

We model price quote changes as

$$\Delta p_t^c = az_{t-1} + b_1 \Delta p_{t-1}^c + \cdots + b_L \Delta p_{t-L}^c + c_1 x_{t-1} + \cdots + c_L x_{t-L} + u_t, \tag{1}$$

where a, b_i, and c_i are coefficient matrices, L is the number of lags, u_t captures unexpected (surprise) price quote changes with variance $\sigma_{i,u}^2$ for venue i, and z_{t-1} is a row vector of price differences across venues, known as "errors" in time-series econometrics:

$$z_{t-1} = [(p_{1,t-1}^c - p_{2,t-1}^c) \quad (p_{1,t-1}^c - p_{3,t-1}^c) \quad \ldots \quad (p_{1,t-1}^c - p_{C,t-1}^c)]. \tag{2}$$

In a setting without trades ($E = 0$), the model in (1) collapses to the VECM proposed by H95. Our model, therefore, generalizes the H95 approach by including trade information.

[13]For example, an interval with one buyer-initiated trade worth £100 and one seller-initiated trade worth £300 would have a signed volume of -£200.

We model signed volume as

$$x_t = dz_{t-1} + e_1 \Delta p^c_{t-1} + \cdots + e_L \Delta p^c_{t-L} + f_1 x_{t-1} + \cdots + f_L x_{t-L} + v_t, \tag{3}$$

where d, e_i, and f_i are coefficient matrices, and v_t captures unexpected (surprise) signed volume with variance $\sigma^2_{j,v}$ for venue j. Note that signed volumes are allowed to respond to the error vector z_{t-1}. One rationale for such response is plain-vanilla arbitrage activity. In this case, signed volume in venue V is expected to load positively on price differentials of V with other venues (i.e., $p^c_{i\neq V,t-1} - p^c_{V,t-1}$), since arbitrageurs buy in "cheap markets" to sell in "expensive markets."

3.2 Generalized Information Share

Since the full econometric model defined by (1) and (3) is linear, it can be estimated with ordinary least squares (OLS). The estimated coefficients allow us to define the generalized information share, nesting the H95 information share for price quotes. Since all price quote series are for the same security, the price system is cointegrated, with cointegration rank $C - 1$. This implies that all price quote series can be viewed as the sum of a common (univariate) random walk and a stationary term (H95, p. 1179). Let the time t innovation in this random walk be denoted by w_t. Following in the footsteps of H95, this innovation can be expressed as a linear function of the innovations in the full econometric model.

Before defining GIS formally, let us illustrate the construct in a simple example. Consider the case of two markets that both feature quotes and trades (i.e., $C = E = 2$). In this case, w_t can be written as:

$$w_t = \begin{bmatrix} g_1 & g_2 \end{bmatrix} \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix} + \begin{bmatrix} h_1 & h_2 \end{bmatrix} \begin{bmatrix} v_{1,t} \\ v_{2,t} \end{bmatrix}. \tag{4}$$

The coefficients g_t and h_t can be obtained from the estimated econometric model by computing the long-run impact of the model innovations u_t and v_t on price quotes.[14]

The variance of the efficient price innovation w_t can be written as:

$$\sigma_w^2 = \begin{bmatrix} g_1 & g_2 & h_1 & h_2 \end{bmatrix} \begin{bmatrix} \Sigma_{uu} & \Sigma_{uv} \\ \Sigma_{vu} & \Sigma_{vv} \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \\ h_1 \\ h_2 \end{bmatrix}, \tag{5}$$

where Σ_{ij} is a 2×2 covariance matrix of vectors $i, j \in \{u, v\}$. If all elements of the innovation vectors u_t and v_t are mutually uncorrelated, then (5) can be written as:

$$\sigma_w^2 = g_1^2\sigma_{u_1}^2 + g_2^2\sigma_{u_2}^2 + h_1^2\sigma_{v_1}^2 + h_2^2\sigma_{v_2}^2. \tag{6}$$

In this example, it is therefore straightforward to decompose total efficient price variance across quotes and trades in the two markets.

In general, when the residual covariance matrix is diagonal, the generalized information shares of price quote i and signed volume j are uniquely identified and given by

$$GIS_i^c = \frac{g_i^2\sigma_{i,u}^2}{\sigma_w^2} \tag{7}$$

and

$$GIS_j^e = \frac{h_j^2\sigma_{j,v}^2}{\sigma_w^2}, \tag{8}$$

respectively, where $i = 1, \ldots, C$ and $j = 0, \ldots, E$.

If, however, the covariance matrix is not diagonal, then lower and upper bounds can be obtained by computing GIS for all possible sequences of venues in the system (see H95, p. 1183–1184). In sum, the GIS framework we propose nests the IS framework developed

[14]Technically, one way to obtain a long-run impact is by summing across the thetas for all lags of the vector moving average (VMA) representation of the model (Hasbrouck, 1995, Section II).

in H95, and extends it to include trade information through signed volume. Absent trade information, GIS reduces to IS.

A couple of points deserve some discussion. First, the most attractive feature of GIS is that the accompanying econometric model can be estimated with OLS. We emphasize this point since OLS continues to show robust performance when datasets get large. Estimation does not break down, nor does it require prohibitively large computational resources. This feature is particularly important when the aim is to attribute information to quotes and trades from a variety of venues. Only at extremely high time resolutions will one be able to measure how one variable responds to another. Such resolutions tend to make datasets large. For example, an eight hour trading period contains 28.8 million milliseconds.

Second, in the case of a single price quote series (i.e., $C = 1$), cointegration becomes a mute issue and the GIS approach collapses to a VAR model (i.e., the error vector z_t becomes empty and therefore drops out). This model is similar to Hasbrouck (1991), but with one important difference. Since a trade can trigger a price quote change, Hasbrouck (1991) allows signed volume to contemporaneously impact price quote changes, but not vice versa. We prefer to not assume directionality *ex ante*, since price changes within an interval might also trigger trade. The cost of allowing bi-directionality is a potentially larger distance between lower and upper bounds of GIS. We prefer to be conservative in our results.

4 Information Share Biases

This section illustrates how forward-filling trade prices in an H95 framework causes information share estimates to become biased. First, we simulate from a data-generating process (DGP) where we can observe the true information share. We show that forward-filling trade price results in two types of biases in information share estimates. We further show that these biases disappear when using our generalized approach. Second, we redo the analysis based on our LSE sample to show that these biases in real-world markets are economically sizeable and can therefore not be ignored.

4.1 Simulated Data

To analyze how forward-filling trade prices affects information share estimates, we will vary the probability of trade while keeping the true information share constant. A benefit of this approach is that we can include the "continuous case" of a trade probability that is equal to 100%. This special case does not require forward-filing and, therefore, information share estimates should be unbiased.

4.1.1 Data-Generating Process

It is useful to design the DGP as consisting of two parts: a latent efficient-price process and a set of observed variables (i.e., a state-space structure). This clarifies the dynamics to the economist as well as creates a versatile structure that can be extended in a variety of ways, as will become clear in the remainder of the subsection.

Latent efficient price process. Let m_t be the unobserved efficient price at time t ("m" for martingale). Let the martingale innovations in each period consist of two components:

$$m_t = m_{t-1} + \lambda x_t + y_t, \tag{9}$$

where x_t is signed volume, λ captures its price impact, and y_t captures the information in quote updates. Each period features trade with probability π and, if there is trade, x_t is drawn from standard normal distribution:

$$x_t = \begin{cases} X_t \sim N(0,1) & \text{with probability } \pi \qquad \text{(trade),} \\ 0 & \text{with probability } (1-\pi) \qquad \text{(no trade),} \end{cases} \tag{10}$$

Orthogonal to this signed-volume component of efficient-price changes is the information component in price (mid)quote updates:

$$y_t \sim N(0, \pi). \tag{11}$$

The variance for quote updates is picked to ensure that when we vary the trade probability π, the true information share remains constant. We thus avoid a moving target when we vary the size of the friction from none at all (i.e., $\pi = 100\%$) to higher degrees of the friction (i.e., lower π). These true information shares become:

$$TrueIS^c = \frac{\pi}{(1+\lambda^2)\pi} = \frac{1}{1+\lambda^2}, \qquad \text{(price quotes)} \tag{12}$$

$$TrueIS^e = \frac{\lambda^2\pi}{(1+\lambda^2)\pi} = \frac{\lambda^2}{1+\lambda^2}. \qquad \text{(trades)} \tag{13}$$

Note that these information shares depend solely on the price impact λ, which allows us to study how any pattern varies with the relative informativeness of trades.

Observed price quote and signed volume process. The process described so far is a latent one as it pertains to the efficient price. The *observed* prices associated with this latent process are price quotes and transactions prices:

$$p_t^c = m_{t-1} + y_t, \qquad \text{(price quotes)} \tag{14}$$

$$p_t^e = \begin{cases} m_{t-1} + \lambda x_t & \text{if there is trade,} \\ \varnothing & \text{if there is no trade.} \end{cases} \qquad \text{(trades)} \tag{15}$$

The DGP we propose may be enriched in a variety of ways. One could, for example, allow for transitory price impact, use signed trades instead of signed volume (thus effectively ignoring trade size), or impose a discrete price grid. Appendix A shows that the patterns that we identify in our baseline simulations below also show up in such perturbed DGPs.

Empirical design. We simulate from the baseline DGP as follows. We generate 100 random samples, where each sample is one million observations long. Without loss of generality, we start each sample with the efficient price being equal to zero: $m_0 = 0$. Then, for each subsequent period t, we draw $w_t = (y_t, x_t)$ and compute the observed variables: the price quote p_t^c, the transaction price p_t^e in case there is a trade event, and signed volume x_t (which is zero if there is no trade event).

To study how forward-filling transaction prices affects information share estimates, we vary both the probability of trade π and the relative trade informativeness λ. We consider five different levels of π: 100%, 10%, 1%, 0.1%, and 0.01%. We further pick three levels for λ, $\sqrt{1/99}$, 1/3 and 1/2, so that true trade informativeness becomes 1%, 10%, and 20%, respectively. Note that, by design, true IS only depends on λ, not on π. We analyze all possible pairs of λ and π values, and re-use the w_t draws to generate $5 \times 3 = 15$ versions of our 100 samples. Re-using draws ensures apples for apples comparisons across perturbed DGPs, since any variation in parameter estimates cannot be due to variation in random samples.

4.1.2 Empirical Results: Information Share Estimates

In the remainder of the section, we illustrate the results of IS and GIS estimates based on simulated data. Appendix D contains a detailed discussion of the estimation procedure.

Information shares. Figure 2 shows that the information share of trades becomes biased when trade price series need to be forward-filled. Consider the case where the true information share is 10% ($\lambda = 1/3$, blue lines). When there is trade in every period (i.e., $\pi = 100\%$), no forward-filling is needed and, as expected, the IS estimate is not significantly different from the true value. More precisely, the true value falls within the 95%

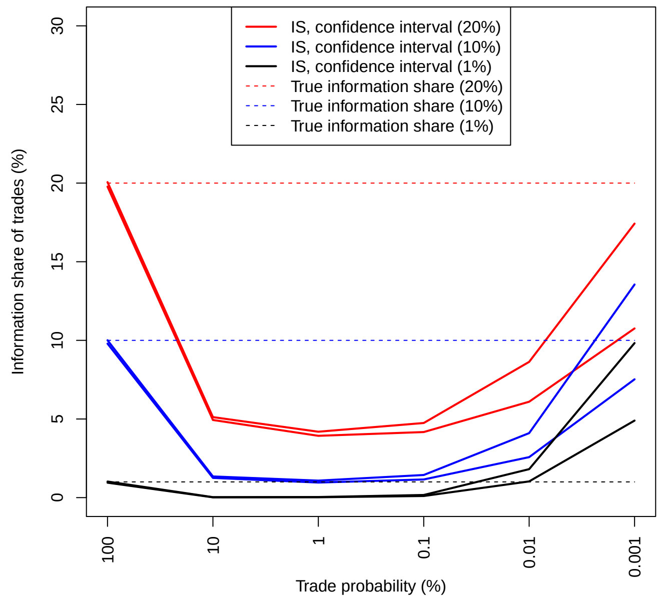


Figure 2: Information Share Estimates by Probability of Trade. This graph plots the 95% confidendence interval of mean H95 information share (IS) estimate of trades, as a function of the probability of trade π (solid line). It is based on 100 sample draws from a data-generating process for which the true IS is known (dotted line). Each sample is 1 million periods long. The $\pi = 100\%$ case represents the special case where forward-filling of trades is not needed, and therefore serves as a no-friction benchmark.

confidence interval of the estimator mean. This confidence interval is based on 100 sample draws, each of which is 1 million periods long.[15]

If, however, forward-filling is needed, then IS estimates become biased. In the figure, this is seen in that the lower and upper bounds are both below true values. The initial decline with a progressively lower probability of trade is intuitive, since forward-filling trade prices creates price staleness. We refer to this as the stale-price bias.

However, the decline is non-monotonic since for extremely low probabilities of trade, the underestimation seem to become *less* severe. We believe this is due to a small-sample

[15]Note that we do not need to consider lower and upper bounds, since the quote and trade innovations are independent in the DGP. See Hasbrouck (1995, Eqn. 14) for IS and Eqns. (7) and (8) for GIS.

bias. Note that a trade probability of 0.001% corresponds to, on average, 10 trades in a sample with one million periods. After further experimentation, we believe that the relatively high information shares for this case are mechanical in the sense that trade information shares tend to 50% when the expected number of trades in the sample tends to one (see footnote 4). The same pattern is obtained when the true information share is set to 1% or 20% (black and red lines in Figure 2, respectively), which testifies to the robustness of the pattern. In the 1% case, when the trade intensity is low, the small-sample bias even dominates the stale-price bias, leading to overestimated information shares.

The result that information shares might be overestimated in samples with very few trades is not merely a statistical curiosity in artificial data. In present-day hybrid markets, off-exchange markets with very infrequent trades are common. In this paper we include data from DPs, PAs, and SDPs, where trades are often less frequent than what we consider in the simulations. Other settings, such as the upstairs markets, block trading mechanisms, and voice markets, are also susceptible to infrequent trading.

To further demonstrate robustness of the reported bias patterns, we add two sets of analyses. First, Appendix A repeats the above analysis for various perturbations of the baseline DGP: signed trades drive prices instead of signed volume, transitory shocks are added to price and trade innovations, or prices are discrete instead of continuous. The patterns that emerge are very similar to the one in Figure 2.

Second, the biases remain even if one reverts to alternative approaches of analyzing information in a one-security-many-markets setting. Appendix B shows results for the same DGP when applied to component shares (Gonzalo and Granger, 1995; Harris, McInish, and Wood, 2002) and information leadership shares (Yan and Zivot, 2010; Putniņš, 2013).

Finally, other than (H95) IS producing biased estimates, which we consider a severe limitation of the approach, we also note that forward-filled prices slows down the convergence of prices after shocks. Table 1 reports, for both IS and GIS, the number of periods it takes for per-period price changes in the impulse-response function (IRF) to become

		Trade probability π (%)					
		100	10	1	0.1	0.01	0.001
Periods needed for IRF convergence:							
IS	Trade-price shock	320	194	907	6,752	43,383	132,131
	Price-quote shock	217	193	916	6,858	44,463	142,767
GIS	Signed-volume shock	176	187	179	175	177	169
	Price-quote shock	176	187	179	175	177	169

Table 1: Convergence of Impulse Response Functions. This table reports on the speed of convergence of impulse response functions for prices responding to either price shocks or to signed volume shocks. The results are based on simulations where we can vary the size of the friction, from a trade probability π equal to one to ever lower probabilities. The results are shown for IS where trade prices are forward-filled and for GIS that does not need forward-filling. The table shows the average number of periods that are needed for per-period price changes to be smaller than one millionth of the initial shock. The results are based on 100 draws of samples that each contain one million observations.

smaller than one millionth of the initial shock. This number is similar in magnitude for IS and GIS for the case that does not need forward-filling (i.e., the case where trade probability π is equal to 100%). These numbers range between 100 to 400 periods. For IS, this number becomes progressively larger for lower trade probabilities, while there is essentially no change for GIS. For example, for a trade probability of 0.001%, IS needs more than 130,000 periods to converge while GIS still requires only 169 periods. Note that for millisecond samples, this case corresponds to one trade every 100 minutes, and thus a few trades per day. This is not uncommon for off-exchange venues.

Generalized information shares. Figure 3 mirrors Figure 2, but plots the generalized information shares instead of information shares. For all levels of trade probability, and for all true information shares, the 95% confidence interval of the GIS mean contains the true value. GIS therefore produces unbiased estimates. The bounds are tight in the sense that they are at least within 20% of the true value overall, but much tighter for a trade probability that is at least 0.1%. Appendix A shows that these findings hold up when considering the

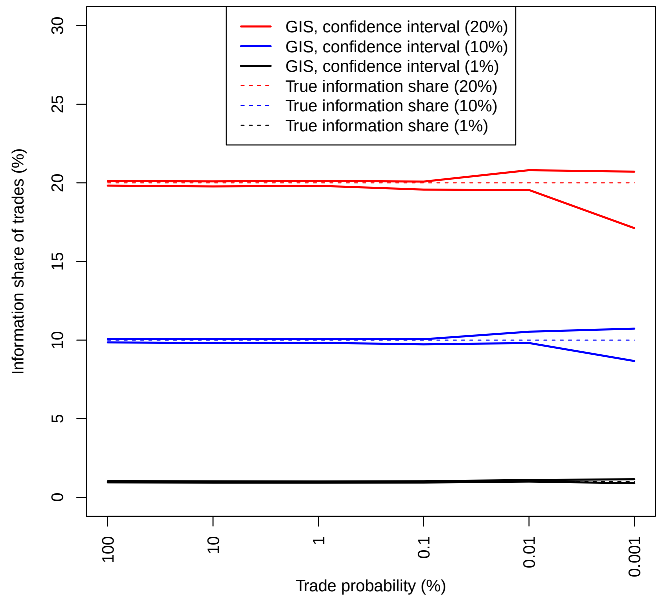


Figure 3: Generalized Information Share Estimate by Probability of Trade. This figure echoes Figure 2, but plots generalized information shares instead of H95 information shares. It plots the 95% confidendence interval of the mean GIS estimate of trades, as a function of the probability of trade π (solid line). It is based on 100 sample draws from a data-generating process for which the true IS is known (dotted line). Each sample is 1 million periods long. The π = 100% case represents the special case where forward-filling of trades is not needed, and therefore serves as a no-friction benchmark.

various perturbations of the DGP discussed in the previous subsection.

4.2 UK Equity Market Data

We now turn to real-world data to estimate information shares and generalized information shares. We will learn if differentials between the two are economically sizeable.

4.2.1 Data and Sample

We select ten large-cap stocks with a primary listing at the LSE. We source price quotes from the LSE as well as the the four largest MTFs trading the these stocks: Aquis Exchange, CBOE BXE, CBOE CXE, and LSEG Turquoise. We source trades from the same venues, as well as from dark pool, periodic auctions, and single-dealer platforms. The sample covers three months of trading: April 1 through June 30, 2021. In all analyses, the sample includes the trading hours during which the LSE operates its continuous limit order book. The data are downloaded from the *Tick History* database, provided by Refinitiv. Details about sample selection, stock properties, filtering procedures, and timestamp considerations are given in Appendix C.

To familiarize the reader with the sample, Table 2 presents various descriptive statistics. Across stocks, the average relative bid-ask spread ranges between 2.9 and 5.5 basis points (bps). As is common for large-cap stocks, the minimum tick size often imposes a lower bound on the spread. The frequency of milliseconds for which this constraint binds ranges between 10% and 41%. Volatility measured as standard deviation of one-minute midquote returns, varies between 4.2 bps and 6.6 bps. Trade frequency ranges between 29.4 and 71.1 trades per minute. We believe that these trade characteristics are representative for trading in large-cap stocks around the world.

Given the forward-filling biases identified above, variation in trade frequency across venue types is of particular interest. Although our large-cap stocks are actively traded, most of the trades materialize in the exchanges: LSEs and MTFs. For all stock-days, these venues record multiple trades per minute, on average. The only exception is the stock RKT, where the MTFs have less than one trade per minute. For the off-exchange venue types – PAs, DPs, and SDPs – such low trade frequencies are common. The most extreme case is the NG stock, where 26% of all trading days have at most one trade per minute in the PAs. These observations coupled with the one-millisecond sample frequency needed to trace out information shares illustrate that forward-filling trade prices will be paramount, in particular for dark venue types.

Ticker	Daily volume (£, millions)	Bid-ask spread (bps)	Tick constr. (%)	Vola-tility (bps)	#Trades per min.	≤ 1 trade / minute LSE (%)	MTFs (%)	PAs (%)	DPs (%)	SDPs (%)
BARC	1145	3.3	10	6.6	61	0	0	8	2	2
GSK	1229	2.9	30	4.6	56	0	0	5	2	0
HSBA	922	3.0	15	4.9	53	0	0	16	8	2
LLOY	750	3.7	11	5.9	46	0	0	3	3	5
LSEG	650	5.0	31	6.0	29	0	0	23	16	8
NG	572	3.1	18	4.2	37	0	0	26	8	3
PRU	661	5.5	41	5.7	30	0	0	20	16	10
RIO	1800	3.4	21	6.4	71	0	0	8	0	0
RKT	737	3.4	27	5.0	32	0	2	21	13	3
VOD	988	3.0	31	5.0	63	0	0	3	0	0

Table 2: Descriptive Statistics. This table presents trade statistics for the sample stocks from April 1 through June 30, 2021. *Bid-ask spread* is the difference between the best bid and ask prices, divided by the midpoint. *Tick constr.* is the fraction of milliseconds where the bid-ask spread equals minimum tick size. *Volatility* is the standard deviation of one-minute midquote returns. *#Trades per min.* is the number of trades per minute. The five rightmost columns report the percentage of trading days that, on average, have less than one trade per minute in the different venue types.

Empirical design. We estimate IS and GIS based on stock-days sampled at a one-millisecond frequency, which is the highest timestamp precision available in our raw data. Each sample then consists of more than 30 million observations. We base our statistical inference on variation across estimates, following the tradition of Bartlett (1950), Fama and MacBeth (1973), and Hasbrouck (2003).

The ultra-high sampling frequency is necessary. Hasbrouck (2021) emphasizes the importance of applying high enough timestamp resolution to identify information shares. In his application to US equities, however, he obtains close to identical dark pool information shares when using sampling frequencies up to 10 microseconds, as he gets with milliseconds. In our application, we seem to have sufficiently high resolution since we obtain tight ranges between the lower and upper bounds of information shares.

Statistical testing is done conservatively. In all our analysis, we account for multiple testing by setting significance threshold levels at the family of tests (Šidák, 1967). For example, panel (a) of Table 3 contains six statistical tests. When determining if one of those test statistics is significant at the 5% level, we apply a p-value threshold of 0.85%. We further use robust standard errors that are clustered by stocks and date.

The empirical approach outlined here makes estimation feasible, but it remains computationally challenging. To obtain a holistic yet feasible view of the fragmented marketplace, we analyze venue types rather than individual platforms. For example, we aggregate the data from four MTFs to one price quotes series and one trade series.[16]

Appendix D presents more detailed information on how estimation is implemented for all analysis based on real-world data. It discusses, among other things, the precise construction of price and signed volume variables, the specification of the VECM lag structure, the implementation of VECM estimation, and the convergence criteria for IRFs.

4.2.2 Empirical Results: Information Share Estimates

To assess the magnitude of the biases in the real-world data, we consider an application with LSE price quotes, LSE trades, and dark pool trades (the full set of venue types is analyzed in Section 5). Table 3 presents the lower and upper bounds of information shares and generalized information shares estimates.

Panel (a) reports overall averages which show that differentials across IS and GIS are both economically sizeable and statistically significant. The results are in line with the two types of forward-filling biases we have identified. First, the IS for LSE trades is estimated to be between 29.24% and 35.12%. The corresponding interval for GIS is between 35.52% and 43.38%. The differentials for both the lower and the upper bound are statistically significant. In relative terms, the GIS lower bound is one fifth higher than the IS lower bound. For the upper bound, GIS exceeds IS by almost one third. These findings suggest

[16] The full-sample analysis presented below includes two quote series and five trade series. The GIS estimation of that system takes a full week even when relying on parallel processing in state-of-the-art cloud computing services. Because each series depends on lags from all other series, the number of parameters to estimate increases exponentially with the number of quote and trade series.

	LSE quotes				LSE trades				DP trades			
	IS (%)	GIS (%)	Diff. (%)	t-stat.	IS (%)	GIS (%)	Diff. (%)	t-stat.	IS (%)	GIS (%)	Diff. (%)	t-stat.
LBound	63.76	56.46	-7.30	-10.2**	29.24	35.52	6.27	10.7**	0.89	0.03	-0.86	-6.6**
UBound	69.63	64.33	-5.31	-9.1**	35.12	43.38	8.26	11.4**	1.23	0.21	-1.01	-6.5**

(a) Information-share estimates, lower and upper bounds

	LSE quotes				LSE trades				DP trades			
DP trades	IS (%)	GIS (%)	Diff. (%)	t-stat.	IS (%)	GIS (%)	Diff. (%)	t-stat.	IS (%)	GIS (%)	Diff. (%)	t-stat.
Q1 (fewest)	63.54	55.91	-7.63	-9.9**	29.12	36.08	6.97	11.1**	1.09	0.02	-1.07	-5.9**
Q2	63.33	55.97	-7.36	-6.9**	29.38	35.84	6.46	8.7**	0.99	0.03	-0.96	-4.2**
Q3	64.39	56.95	-7.44	-11.0**	28.76	35.12	6.37	9.4**	0.91	0.03	-0.88	-5.2**
Q4 (most)	63.79	57.02	-6.77	-7.9**	29.71	35.02	5.31	6.4**	0.56	0.03	-0.53	-4.4**

(b) Information-share estimates by number of DP trades, lower bounds

	LSE quotes				LSE trades				DP trades			
	IS (%)	GIS (%)	Diff. (%)	t-stat.	IS (%)	GIS (%)	Diff. (%)	t-stat.	IS (%)	GIS (%)	Diff. (%)	t-stat.
BARC	66.71	55.36	-11.35	-7.1**	26.20	35.57	9.37	6.2**	0.88	0.01	-0.86	-7.5**
GSK	63.43	56.68	-6.74	-9.2**	30.71	36.27	5.57	7.5**	0.41	0.02	-0.39	-6.7**
HSBA	66.27	57.67	-8.61	-13.2**	27.10	33.46	6.36	10.3**	0.44	0.01	-0.42	-5.3**
LLOY	64.69	56.03	-8.67	-10.6**	27.69	34.18	6.48	8.3**	1.33	<0.01	-1.33	-5.3**
LSEG	57.60	53.78	-3.82	-4.5**	33.98	38.64	4.66	6.8**	1.54	0.04	-1.50	-4.5**
NG	64.70	56.79	-7.91	-9.3**	27.29	34.07	6.78	9.0**	1.07	0.03	-1.04	-3.7**
PRU	63.98	58.55	-5.43	-6.3**	29.98	34.91	4.93	6.1**	0.93	0.06	-0.87	-4.3**
RIO	64.55	57.98	-6.57	-12.6**	29.40	34.50	5.10	10.5**	0.36	0.03	-0.33	-5.7**
RKT	61.89	53.10	-8.79	-10.7**	30.12	39.33	9.22	11.5**	0.98	0.04	-0.94	-4.8**
VOD	63.79	58.68	-5.11	-8.4**	29.95	34.23	4.28	7.3**	0.93	0.02	-0.91	-5.2**

(c) Information-share estimates by stock, lower bounds

Table 3: Information Share Estimates for LSE Stocks. This table presents information share estimates of quotes and trades for a sample of LSE stocks from April 1 through June 30, 2021. The LSE midquote is used for the price quote series. Trades include LSE trades and dark pool trades. The H95 information shares (IS) estimates are based on forward-filled trade prices, while generalized information shares (GIS) use signed volume. Panel (a) presents overall averages of the information-share lower and upper bounds. Panel (b) sorts stock-days by the number of dark-pool trades, and reports lower bounds by quartile. Panel (c) presents lower bounds by stock. Panel (d) presents the lower bound results by month. All panels test for the difference between IS and GIS, where ** and * denote statistical significance at the 1% and 5% level, respectively. The standard errors are clustered by stock and date, except for panel (c), where they are clustered by date. Estimation details are in Appendix D.

	LSE quotes				LSE trades				DP trades			
	IS (%)	GIS (%)	Diff. (%)	t-stat.	IS (%)	GIS (%)	Diff. (%)	t-stat.	IS (%)	GIS (%)	Diff. (%)	t-stat.
April	64.29	57.31	-6.98	-7.1**	28.38	33.79	5.41	6.1**	0.79	0.02	-0.77	-5.3**
May	64.27	56.44	-7.83	-10.2**	28.67	35.18	6.51	9.6**	0.76	0.02	-0.74	-6.1**
June	62.84	55.71	-7.13	-9.3**	30.52	37.37	6.86	12.0**	1.09	0.04	-1.05	-5.6**

(d) Information-share estimates by month, lower bounds

Table 3: Information Share Estimates for LSE Stocks. Continued from previous page.

that the informational contribution of LSE trades is underestimated by IS, due to the stale-price bias induced by forward-filling.

Second, for dark-pool trades, IS is estimated to be between 0.89% and 1.23%. GIS, on the other hand, is estimated to be between 0.03% and 0.21%. For both the lower and the upper bound, the differentials are statistically significant. In relative terms, the IS lower bound is overestimated by 97%. The IS upper bound is overestimated by 82%. The findings suggest that, as a result of forward filling, IS is overestimated due to a small-sample bias.

The finding of a stale-price bias for LSE trades and a small-sample bias for dark pools is in line with the relative trade frequencies reported in Table 2. Each minute, there are multiple trades at the LSE but often at most one for DPs. To study whether trade frequency is the driver of small-sample bias, we bucket stock-days by the number of DP trades. We redo the analysis and the results in panel (b) show that, indeed, the extent that IS overestimates is largest for stock-day quartile with fewest DP trades, and declines monotonically across quartiles as the number of DP trades increase.

Finally, to verify the robustness of the pattern that IS underestimates the informativeness of exchange trades and overestimates the informativeness of dark pools, we repeat the analysis by stock in panel (c) and by month in panel (d). The results show that the pattern is robust. All differentials between IS and GIS bear the same sign, and are statistically and economically significant.

In sum, these findings could explain why most studies on the informativeness of dark-pool trades find them to be only marginally less informative than exchange trades. As discussed in the introduction, these studies rely on IS, which we show tends to underestimate exchange trade informativeness, and overestimate dark pool trade informativeness.[17]

5 Informativeness of Quotes and Trades by Venue Type

The full sample pools quotes and trades by venue types. It contains two midquote series based on best bid and ask prices separately, LSE, the listing venue, and MTFs. We further have signed-volume series for both of these venue types, but also for DPs, SDPs, and PAs. The sample therefore features seven series: $C = 2$ and $E = 5$. For feasibility, we here analyze generalized information shares only (see footnote 16).

5.1 Generalized Information Share Estimates

Table 4 presents unbiased estimates of the relative informativeness of price quotes and trades in all venue types. The results lead to the following insights. First, price quotes are more informative than trade. The GIS lower and upper bounds for LSE price quotes are 14.13% and 28.92%, respectively. The corresponding interval for MTFs is 28.20% to 61.61%. Since lower bounds can be uniquely assigned, these results show that at least 14.13+28.20=42.33% of the information is incorporated through price quotes. The corresponding number for trades across all venue types is

[17]These studies include Comerton-Forde and Putniņš (2015), Hatheway, Kwan, and Zheng (2017), and Chakrabarty, Cox, and Upson (2022). Comerton-Forde and Putniņš (2015) study a sample of Australian stocks that with many exchange trades, but relatively few dark-pool trades. The median stock-day features only 268 trades, which are either block trades or dark pool trades (they do not further break it down in their reported statistics). Hatheway, Kwan, and Zheng (2017) report that small-cap US stocks have 258 dark-pool trades per day, compared to 941 for mid-caps, and 10,680 for large-caps. They find that small-caps, which have the *lowest* dark-pool market share, have the *highest* dark-pool information share. This aligns with our findings when sorting on dark-pool activity (Table 3, panel (b)). Chakrabarty, Cox, and Upson (2022) do not report the number of trades, but their sample consists of small-caps and mid-caps, where the number of trades is relatively low to begin with.

	LSE quotes	MTF quotes	LSE trades	MTF trades	DP trades	SDP trades	PA trades
GIS lower bound (%)	14.13	28.20	10.17	4.39	0.01	0.05	<0.01
GIS upper bound (%)	28.92	61.61	38.89	29.18	0.22	0.10	<0.01

Table 4: Generalized Information Shares of Quotes and Trades, by Venue Type. This table presents generalized information share estimates by venue type. It presents both the lower and the upper bound of GIS. The estimates are based on a sample of LSE stocks that runs from April 1 through June 30, 2021. The reported statistics are averages across stock-days. Estimation details are in Appendix D.

10.17+4.39+0.01+0.05+0.01=14.63%. This finding is in line with a general trend of price quotes becoming important sources of information revelation (Brogaard, Hendershott, and Riordan, 2019).

Second, exchange trades are an order of magnitude more informative than off-exchange trades. The relative informativeness of LSE trades is between 10.17% and 38.89%. For MTF trades, these bounds are 4.39% and 29.18%. Exchange trades therefore contribute at least 14.56% of the information. For off-exchange venue types, the lower and upper bounds are all at most 0.22. Collectively, their relative informativeness is between 0.06% and 0.34% (conservative estimates obtained by simple sums of lower and upper bounds). The most that off-exchange trades contribute is thus much less than *the least* that exchange trades contribute: 0.34% $\ll$ 14.56%.[18]

5.2 Trade Informativeness per Pound Traded

Is the finding that exchange trades contribute more information than off-exchange trades simply a volume result? In other words, if one were to assess informativeness on a *per*

[18]We have redone Table 4 based on signed trades instead of signed volume to verify whether the results hinge on trade size variation. The signed number of trades is the number of buyer-initiated trades in an interval, minus the number of seller-initiated trades. For example, if there are three buyer-initiated trades and one seller-initiated trade in the same interval $(t-1, t]$, the signed number of trades is +2.

The results are largely unchanged, with the baseline signed-volume version showing generally higher GIS for trade size, suggesting that there is information in trade size. See Table OA.1 in the Online Appendix.

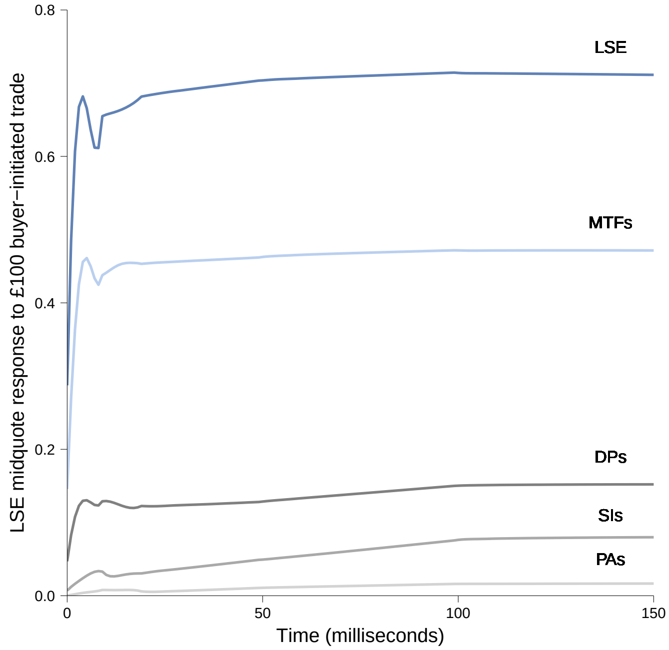


Figure 4: Price Response to Buyer-Initiated Trade. This figure depicts how the LSE midquote responds to a buyer-initiated trade of £100 in the various venue types. The responses are based on a representative stock-day: Rio Tinto (RIO) on April 1, 2021. Technically, the curves depict the cumulative IRF based on a VECM parameter estimates. Estimation details are in Appendix D.

pound basis, would both types of trades be equally informative? And, are there any differentials between dark venue types? We take on these questions in the remainder of this section.

Following the lead of H95, we set the stage by showing the result for a single stock-day before turning to the full sample. Figure 4 plots how the LSE midquote responds to a £100 trade by venue type. Technically, the plot graphs IRFs based on the VECM estimates for Rio Tinto on April 1, 2021 (chosen to be representative for the full sample). The figure suggests that volume differentials cannot be the full explanation, as price impacts differ substantially across venue types. A £100 trade in LSE results in a long-term price impact of about 0.7 basis points. The impact for a similar size trade in MTFs is 0.5 bps. For off-

	LSE	MTFs	DPs	SDPs	PAs
Long-term price impact to £100 trade (bps)	0.50	0.34	0.08	0.06	<0.01
Difference relative to dark pools (bps)	0.42**	0.27**	-	-0.03	-0.07**

Table 5: Long-Term Price Impact of Trades. This table reports the long-term response of LSE midquotes to a £100 trade, by venue type (i.e., h_i in (4)). This response is based on VECM estimates for each stock-day sampled at a one-millisecond frequency. The full sample covers trading in ten LSE stocks from April 1 through June 30, 2021. The table reports differentials relative to DPs, where ** and * denote statistical significance at the 1% and 5% level, respectively. The standard errors are clustered by stock and date. Estimation details are in Appendix D.

exchange trades, these impacts are less than 0.2 bps. DP trades seem to be most informative, followed by SI and PA trades.

Table 5 extends the price-impact analysis of Figure 4 to the full sample. It adds tests on the differential between all venue types and dark pools. LSE and MTF trades are significantly more informative than DP trades. The differentials are significant in size as well. Per pound traded, LSE trades are more than six times more informative than DP trades. MTF trades are at least four times more informative.

Comparing dark pools with other off-exchange venue shows that, quantitatively, dark pools seem to contain the most informative trades. Trades in SDPs and PAs are even less informative than DP trades, although the differential is statistically significant only for the latter. The reason for PAs being significantly less informative than DPs might lie in the batching nature of PA trades. If one submits an order to PAs, then it triggers a 100-millisecond auction period if there is a match on the other side (see Section 2). Others could join the trade, which might dilute the value of trading on information. In other words, if one aims to trade on time-sensitive private information, and others have a correlated signal but are slightly slower, then the latter could join a trade in PAs at the same price in the 100-millisecond interval. This is not possible in any of the other venue types.

6 Conclusion

In this paper we analyze how information is getting revealed in today's hybrid securities markets, where exchanges compete with various types of off-exchange venues. To get unbiased estimates of information shares in such a setting, we develop an approach that recognizes that trades are inherently different economic events than price-quote updates. Our approach avoids any bias that might result from forward-filling transaction prices, which is what is required in the widely used *de facto* standard approach: Hasbrouck (1995).

In an application to UK equities, we show that exchange trades are an order of magnitude more informative than off-exchange trades. All off-exchange venue types feature relatively uninformative trades, providing strong evidence for an assumed centrifugal force acting on uninformed trades. This finding strengthens worries about overly fragmented markets where exchange trading might become toxic to the point of no-trade theorem types of dry-ups (Milgrom and Stokey, 1982). This might explain why modern markets are seemingly prone to flash crashes. It further disincentives investors from costly information acquisition, since it becomes harder to trade on it.

We believe that our evidence on the informativeness of various venue types is relevant for the public debate on market fragmentation. Hybrid markets, where transparent limit order book mechanisms co-exist with opaque venues, are emerging in virtually all asset classes. Our approach is designed for such settings. Other potential applications include the informativeness of spot versus derivative markets and the relative informativeness of various index products, such as exchange-traded funds and index futures.

Appendix

A Various Perturbations of the Baseline DGP

This appendix redoes the analysis of Section 4.1.2, which illustrates how H95 information share estimates become biased due to forward-filling. It further shows that our generalized information share estimates remain unbiased. Here, we redo the analysis based on simulations for perturbed versions of the baseline DGP. For ease of comparison, we report IS results in Figure A.1, which mirrors Figure 2 in the main text. We do the same for GIS in Figure A.2, which corresponds to Figure 3. The four panels in both figures present the following cases:

- Panel (a) repeats the results of the baseline case for ease of comparison.
- Panel (b) replaces signed volume by signed trade. In other words, the trade flow variable becomes binary instead of Gaussian. If there is a trade, then signed trade is one for a buy trade and minus one for a sell trade. Both outcomes are equally likely.
- Panel (c) adds transitory shocks to the baseline specification. That is, part of the innovation in each time period t, both in the price quotes and in the trades, is reversed in the subsequent period.

  Formally, the setup in (9), (14), and (15) can be generalized to

$$
\begin{aligned}
m_t &= m_{t-1} + \lambda\eta^e x_t + \eta^c y_t, && (16)\\
p_t^c &= m_{t-1} + y_t - (1-\eta^e) w_{t-1}, && (17)\\
p_t^e &= \begin{cases} m_{t-1} + \lambda(x_t - (1-\eta^e)x_{t-1}) & \text{if there is trade,}\\ \varnothing & \text{if there is no trade.}\end{cases} && (18)
\end{aligned}
$$

  where $0 \leq \eta^c < 1$ indicates the fraction of innovations in the price quotes that is informative, and $0 \leq \eta^e < 1$ is the corresponding parameter for signed volume. The generalized framework collapses to the baseline case when $\eta^c = \eta^e = 1$. For the simulation depicted in Panel (c), we set $\eta^c = \eta^e = 0.5$. As long as $\eta^c = \eta^e$, the theoretical information shares of trades and quotes remain as specified in (12) and (13).

- Panel (d) considers the case where prices are discrete instead of continuous. The tick size is set to a tenth of one standard deviation of price quote returns. For simplicity, we draw continuous prices and round them to the nearest price on a discrete grid. The rounding introduces correlation between discrete price changes. Information shares are then no longer uniquely identified, but only up to a lower and upper bound (see footnote 5). To capture this in panel (d), we report the lower bound of the 95% confidence interval of the IS lower bound mean, and the upper bound of the 95% confidence interval of the IS upper bound mean.

Figure A.1 and Figure A.2 show that the findings for all perturbed DGPs are similar to those of the baseline DGP. The patterns we document therefore seem to be robust.

B Bias in Alternatives for H95 Information Share

Although the H95 information share analysis is used mostly to assess informativeness in a fragmented setting, there are alternative approaches. In this section, we assess whether these alternatives produce unbiased estimates. We consider two well-known alternatives. First, Gonzalo and Granger (1995) propose a "component share" analysis, which has since been used by, for example, Harris, McInish, and Wood (2002). Second, Yan and Zivot (2010) develop an "information leadership" measure, which Putniņš (2013) puts on a scale between zero and one and refers to as "information leadership share." Both approaches require forward-filling of transaction prices and we, therefore, conjecture that they are subject to the same biases as IS.

Panel (a) of Figure A.3 illustrates how the component share of trade seems to be subject to the two types of forward-filling bias: stale-price bias and a small-sample bias. We compute the component share of trades based on simulations from the baseline DGP outlined in Section 4.1.1. The case where trade probability π is equal to 100% is the no-friction case, and the component share estimate of trades is therefore unbiased. If one increases the friction by reducing π, the estimates should hover around this level to be unbiased. The plot, however, shows that they do not. They first decline, consistent with a stale-price bias, and then increase, in line with a small-sample bias. In other words, the plot exhibits the same pattern as the equivalent plot for IS in Figure 2.

Panel (b) of Figure A.3 does the same analysis as panel (a), but this time for the information leadership share of trades. The findings are very similar although the stale-price bias seems to be smaller and the small-sample bias seems to be larger. Importantly, however, the pattern is the same.

C Real-World Sample

Sample selection. We analyze information revelation for the ten largest LSE stocks without dual listings on non-UK exchanges in the same timezone, because adding non-UK data would add significant timestamp synchronization issues. The sample runs from April 1 through June 30, 2021. Table A.1 lists the stocks, their tickers, and their 2021 market capitalizations (which is taken from the *Eikon* database maintained by Refinitiv). We have cross-checked the aggregate trading volume obtained from our Refinitiv sample with the weekly volume statistics that are publicly available from the Fidessa Fragulator.

Timestamps. Tick History provides two timestamps for each trade and quote, one provided by the exchange and one assigned by Refinitiv when receiving the data. We use

Firm name	Ticker	Market capitalization (in billiion £)
Barclays	BARC	44.1
GlaxoSmithKline	GSK	89.3
HSBC Holdings	HSBA	119.1
Lloyds Banking Group	LLOY	41.6
London Stock Exchange Group	LSEG	53.2
National Grid	NG	42.3
Prudential	PRU	55.5
Rio Tinto	RIO	126.6
Reckitt Benckiser Group	RKT	63.9
Vodafone Group	VOD	51.3

Table A.1: Sample Stocks. This table lists all sample stocks, their tickers, and their 2021 market capitalizations.

exchange timestamps throughout to avoid any noise due to reporting latencies. Exchange timestamps are reported in millisecond granularity.

Data filtering. To avoid any contamination of regular trading by the LSE opening, closing and midday auction, we exclude the first and last minute of continuous trading and the period of the midday auction. We implement two additional sets of filters.

1. We exclude quote updates that imply a negative spread or a spread that is larger than 10%. Negative spreads, referred to as "crossed markets," can occur for MTFs when we aggregate quotes across venues. The negative- and large-spread filters turn out to be relatively mild, because they remove less than 20 basis points of the sample.

2. We exclude dark trades that are not transmitted to the market in real time. Timestamps of such trades are unreliable and can therefore not be used in an information-share analysis for which timestamps accuracy is a *sine qua non*. We therefore exclude trades that are flagged as non-price forming trades, trades not contributing to the price discovery process, trades with non-immediate publication, and manual trades. These filters are restrictive, since they exclude all OTC trades and almost half of the SDP

trades.

Figure A.4 illustrates the effect of our data filters. Panel (a) plots the average volume per stock, decomposed by venue type. 59% of total volume is due to dark trades. Panel (b) redoes the plot, but this time based on the sample after applying our data filters. The volume share of dark trades is reduced to 29% of total volume, mostly due to the removal of OTC trades. Importantly, we like to emphasize that this is only due to unreliable timestamps, not because the methodology cannot handle any of these types of trades. Quite the contrary, including them in a generalized information share analysis is trivial once timestamps become reliable.

D Details on VECM Estimation

D.1 Simulated Data

The following comments pertain to VECM estimation based on simulated data (Section 4.1).

1. Denote the forward-filled trade price $\Delta \tilde{p}_t^e$. The H95 return vector is then given by $\Delta p_t = [\Delta p_t^c, \Delta \tilde{p}_t^e]$, and the error vector becomes $z_{t-1} = p_{t-1}^c - \tilde{p}_{t-1}^e$.
2. We estimate a VECM with 100 lags.
3. For the IRFs that are required to obtain information shares, we iterate forward until the per-period absolute price change is smaller than 10^{-6} in both markets.

D.2 Real-World Data

The following comments pertain to VECM estimation based on the real-world sample of LSE-listed stocks (Section 4.2 and 5).

1. For LSE price quotes, we pick the best bid and ask at the end of each period. We then compute the average of these two prices to obtain the LSE midquote.

2. For MTF price quotes, at the end of each period, we pick the highest bid and lowest ask across all MTFs. We then compute the average of these two prices to obtain the MTF midquote.

3. For the LSE trade prices needed to estimate H95 information shares, if there are multiple trades in a period (i.e., a millisecond), we pick the last trade to obtain the trade price. For DP trade prices, we pick the price of the last trade across all DPs in each period.

4. Signed volume is the sum of all buyer-initiated trade sizes in a period, minus the sum of all seller-initiated trade sizes. We sign trades at all venues relative to LSE midquotes, using the algorithm laid out by Lee and Ready (1991). We match trades to the midquote holding in the end of the previous millisecond, following Holden and Jacobsen (2014). Notably, dark trades are frequently midquote executions. To sign such trades, we rely on the second step of the Lee-Ready algorithm, where a midquote trade is considered buyer-initiated (seller-initiated) if the closest previous trade with a different price has a lower (higher) price.

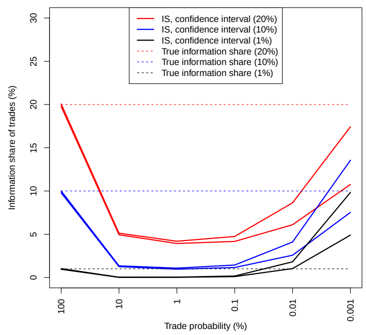
5. All prices and volumes in the data are entered as natural logarithms. For prices, this implies that returns are expressed in relative terms, making them comparable across stocks. For signed volume, this implies that the natural logarithm is taken over the absolute value of signed volume, and then the sign is re-applied so that the result has the same sign as signed volume itself.

6. We estimate a VECM with 100 lags. With all the trading venues being geographically close to each other (see footnote 9), we consider 100 milliseconds sufficient for short-term responses between venues. However, including 100 lags for both quotes and

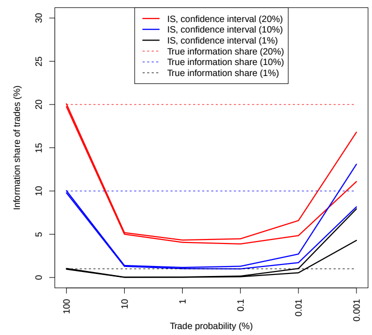
signed volumes in VECMs of up to seven markets would imply a model with around $7{\times}7{\times}100 = 4,900$ parameters. This is computationally infeasible, in particular given the high number of observations. Following Hasbrouck (1995, 2003), we reduce the number of coefficients by applying polynomial distributed lags. Similar to Hasbrouck (2003), we use second-degree polynomials for lags 1 to 10 and 11 to 20, and zero-degree polynomials for lags 21 to 50 and 51 to 100.

7. The VECM is estimated with OLS.

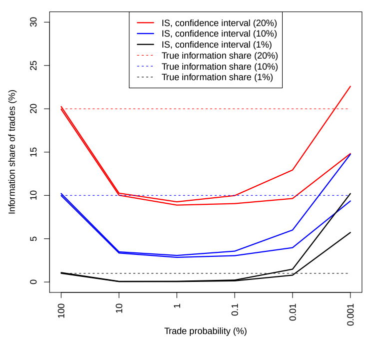
8. We need IRFs to converge in order to obtain the long-term price impact of shocks (i.e., the parameters g and h in Eqn. (4)). We consider IRFs to have converged if the per-period absolute change to a one unit shock is smaller than 10^{-6}.



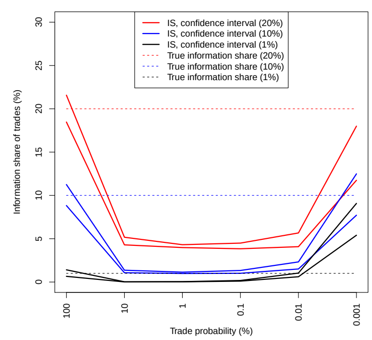
(a) Baseline specification



(b) Baseline with signed trades

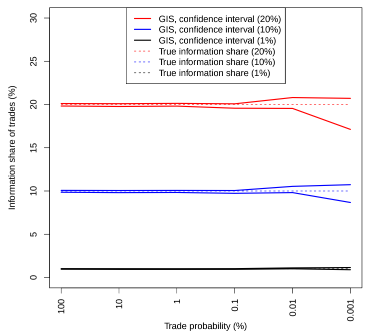


(c) Baseline with transitory shocks

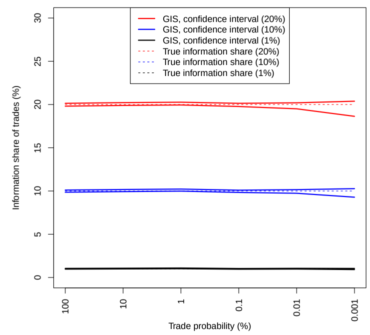


(d) Baseline with discrete prices

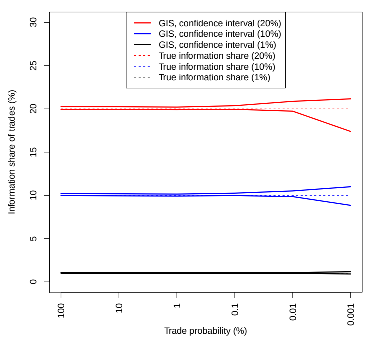
Figure A.1: Information Shares for Perturbed DGPs. This figure presents information share estimates for based on simulating from the baseline DGP, and perturbed versions of it. Panel (a) graphs baseline estimates and therefore mirrors Figure 2. Panels (b) through (d) perturb the baseline DGP by using signed trades instead of signed volume, by adding transitory shocks to both price quotes and transactions prices, and by using a discrete price grid instead of continuous prices. For panel (d), the definition of the confidence interval bounds is based on lower and upper IS bounds, see the appendix text.



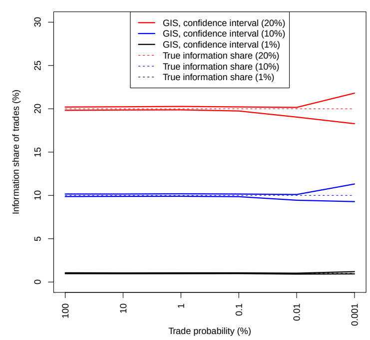
(a) Baseline specification



(b) Baseline with signed trades

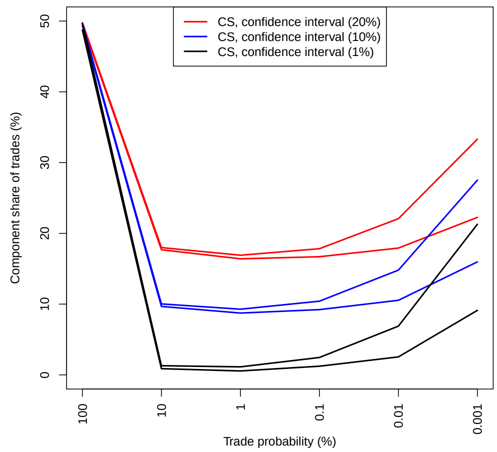


(c) Baseline with transitory shocks

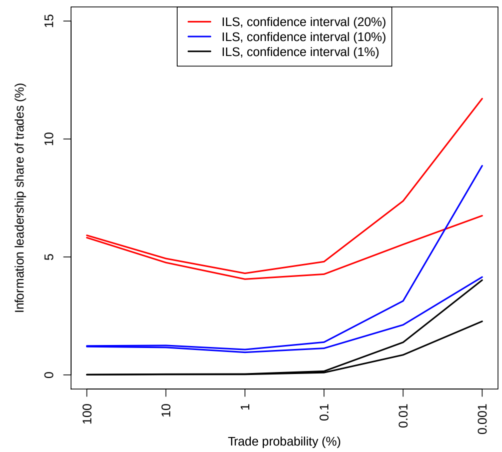


(d) Baseline with discrete prices

Figure A.2: Generalized Information Shares for Perturbed DGPs. This figure mirrors Figure A.1, but plots generalized information shares instead of information shares. Panel (a) graphs baseline estimates and therefore mirrors Figure 3. Panels (b) through (d) perturb the baseline DGP by using signed trades instead of signed volume, by adding transitory shocks to both price quotes and transactions prices, and by using a discrete price grid instead of continuous prices. For panel (d), the definition of the confidence interval bounds is based on lower and upper GIS bounds, see the appendix text.

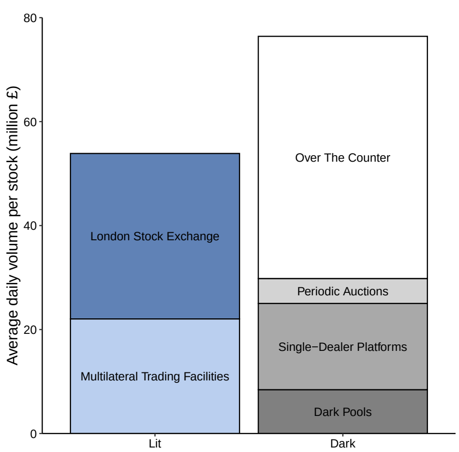


(a) Component share

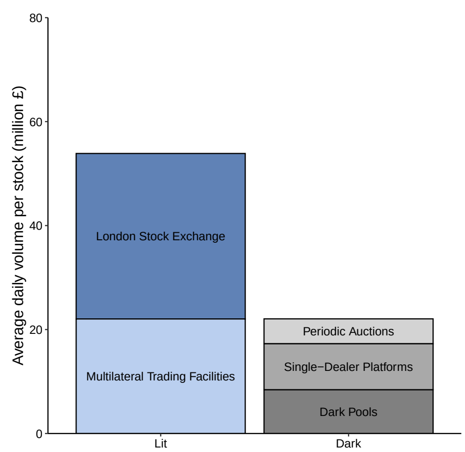


(b) Information leadership share

Figure A.3: Bias in Component Shares and Information Leadership Shares. This figure presents component shares and information leadership shares estimated on simulated data. The plot format follows the same structure as in Figure 2, but presenting the alternative methods instead of the information shares.



(a) All trades



(b) Sample trades

Figure A.4: Trading Volume by Venue Type. This figure plots the average volume per stock, decomposed by venue type. Panel (a) is based on the raw data sample. Panel (b) is based on cleaned sample that enters all of our analysis. The data filters are discussed in detail in Appendix C.

References

Albuquerque, Rui, Shiyun Song, and Chen Yao. 2020. "The price effects of liquidity shocks: A study of the SEC's tick size experiment." *Journal of Financial Economics* 138 (3):700–724.

Aquilina, Matteo, Sean Foley, Peter O'Neill, and Thomas Ruf. 2016. "Asymmetries in dark pool reference prices." Occasional paper, FCA.

Aramian, Fatemeh and Lars L Nordén. 2021. "Costs and Benefits of Trading with Stock Dealers: The Case of Systematic Internalizers." Working paper, Stockholm University.

Bartlett, Maurice S. 1950. "Periodogram Analysis and Continuous Spectra." *Biometrika* 37 (1–2):1–16.

Battalio, Robert and Craig W Holden. 2001. "A simple model of payment for order flow, internalization, and total trading cost." *Journal of Financial Markets* 4 (1):33–71.

Brogaard, Jonathan, Terrence Hendershott, and Ryan Riordan. 2014. "High Frequency Trading and Price Discovery." *Review of Financial Studies* 27:2267–2306.

———. 2019. "Price discovery without trading: Evidence from limit orders." *Journal of Finance* 74 (4):1621–1658.

Brogaard, Jonathan and Jing Pan. 2022. "Dark pool trading and information acquisition." *Review of Financial Studies* 35 (5):2625–2666.

Chakrabarty, Bidisha, Justin Cox, and James E Upson. 2022. "Tick size pilot program and price discovery in US stock markets." *Journal of Financial Markets* 59:100658.

Chakravarty, S., H. Gulen, and S. Mayhew. 2004. "Informed Trading in Stock and Option Markets." *Journal of Finance* 59(3):1235–1258.

Chowdhry, Bhagwan and Vikram Nanda. 1991. "Multimarket Trading and Market Liquidity." *Review of Financial Studies* 4:483–511.

Comerton-Forde, Carole and Tālis Putniņš. 2015. "Dark Trading and Price Discovery." *Journal of Financial Economics* 118:70—92.

Durbin, Jim and Siem Jan Koopman. 2012. *Time Series Analysis by State Space Models*. Oxford: Oxford University Press.

Easley, David, Nicholas M. Kiefer, and Maureen O'Hara. 1996. "Cream-Skimming or Profit-Sharing? The Curious Role of Purchased Order Flow." *Journal of Finance* 51:811–833.

Ernst, Thomas, Jonathan Sokobin, and Chester Spatt. 2021. "The value of off-exchange data." *Working paper* .

EU. 2014. "Markets in Financial Instruments Directive (MiFID 2)." Tech. rep., European Union, Directive 2014/65/EU.

Fama, Eugene F. and James D. MacBeth. 1973. "Risk, Return, and Equilibrium: Empirical Tests." *Journal of Political Economy* 81:607–636.

Gonzalo, J. and C.W.J. Granger. 1995. "Estimation of Common Long-Memory Components in Cointegrated Systems." *Journal of Business and Economic Statistics* 13(1):27–36.

Grammig, Joachim and Erik Theissen. 2012. "Is Best Really Better? Internalization of Orders in an Open Limit Order Book." *Schmalenbach Business Review* 64:82–100.

Hagströmer, Björn. 2022. "Market fragmentation in Europe." Working paper, Stockholm University.

Harris, F.H.deB., T.H. McInish, and R.A. Wood. 2002. "Security Price Adjustment across Exchanges: an Investigation of Common Factor Components for Dow stocks." *Journal of Financial Markets* 5(3):277–308.

Hasbrouck, Joel. 1991. "The Summary Informativeness of Stock Trades: An Econometric Analysis." *Review of Financial Studies* 4:571–595.

———. 1995. "One Security, Many Markets: Determining the Contributions to Price Discovery." *Journal of Finance* 50:1175–1199.

———. 2003. "Intraday Price Formation in U.S. Equity Index Markets." *Journal of Finance* 58(6):2375–2399.

———. 2021. "Price discovery in high resolution." *Journal of Financial Econometrics* 19 (3):395–430.

Hatheway, Frank, Amy Kwan, and Hui Zheng. 2017. "An empirical analysis of market segmentation on US equity markets." *Journal of Financial and Quantitative Analysis* 52 (6):2399–2427.

Hendershott, Terrence and Albert J. Menkveld. 2014. "Price Pressures." *Journal of Financial Economics* 114 (3):405–423.

Holden, Craig W and Stacey Jacobsen. 2014. "Liquidity measurement problems in fast, competitive markets: Expensive and cheap solutions." *Journal of Finance* 69 (4):1747–1785.

Johann, Thomas, Tālis J Putniņš, Satchit Sagade, and Christian Westheide. 2019. "Quasi-dark trading: The effects of banning dark pools in a world of many alternatives." Working paper, SAFE.

Johnson, Barry. 2010. *Algorithmic Trading & DMA*. London: 4Myeloma Press.

Lee, C. and M. Ready. 1991. "Inferring Trade Direction from Intraday Data." *Journal of Finance* 46:733–746.

Lin, Wei-Chao and Chih-Fong Tsai. 2020. "Missing value imputation: a review and analysis of the literature (2006–2017)." *Artificial Intelligence Review* 53 (2):1487–1509.

Milgrom, Paul and Nancy Stokey. 1982. "Information, Trade and Common Knowledge." *Journal of Economic Theory* 26:17–27.

Neumeier, Christian, Arie Gozluklu, Peter Hoffmann, Peter O'Neill, and Felix Suntheim. 2021. "Banning Dark Pools: Venue Selection and Investor Trading Costs." *Financial Conduct Authority's Occasional Papers* .

O'Hara, Maureen. 2015. "High Frequency Market Microstructure." *Journal of Financial Economics* 116:257–270.

Pagano, Marco. 1989. "Trading Volume and Asset Liquidity." *Quarterly Journal of Economics* 104:255–274.

Putniņš, Tālis J. 2013. "What do price discovery metrics really measure?" *Journal of Empirical Finance* 23:68–83.

SEC. 1998. "Regulation Alternative Trading Systems." Manuscript, Securities and Exchange Commission.

———. 2015. "Regulation of NMS Stock Alternative Trading Systems." Manuscript, Securities and Exchange Commission.

Šidák, Zbyněk. 1967. "Rectangular Confidence Regions for the Means of Multivariate Normal Distributions." *Journal of the American Statistical Association* 62 (318):626–633.

Yan, Bingcheng and Eric Zivot. 2010. "A structural analysis of price discovery measures." *Journal of Financial Markets* 13 (1):1–19.

Ye, Mao and Wei Zhu. 2020. "Strategic Informed Trading and Dark Pools." Working paper, University of Illinois at Urbana-Champaign.

Zhu, Haoxiang. 2014. "Do Dark Pools Harm Price Discovery?" *Review of Financial Studies* 27:747–789.

Online appendix to

Trades, Quotes, and Information Shares

The online appendix contains the following additional material:

- Table 4: GIS estimates based on signed number of trades instead of signed volumes.

	LSE quotes	MTF quotes	LSE trades	MTF trades	DP trades	SDP trades	PA trades
Based on signed volume							
GIS lower bound	14.13	28.20	10.17	4.39	0.01	0.05	<0.01
GIS upper bound	28.92	61.61	38.89	29.18	0.22	0.10	<0.01
Based on signed number of trades							
GIS lower bound	16.13	30.22	8.53	3.95	0.03	0.05	<0.01
GIS upper bound	30.82	60.80	34.83	27.40	0.45	0.09	<0.01

Table OA.1: Signed Volume vs. Signed Number of Trades This table presents the lower and upper bound of generalized information share estimates, by venue type. It repeats baseline GIS estimates (Table 4), but adds GIS estimates that use signed number of trades as opposed to signed volume. If all trades were of equal size, the GIS estimates would be the same.