Modelling Transaction Costs when Trades May Be Crowded: A Bayesian Network Using Partially Observable Orders Imbalance

Marie Brière* Charles-Albert Lehalle † Tamara Nefedova‡
Amine Raboun §

December 28, 2019

Abstract

Using a large database of US institutional investors' trades in the equity market, this paper explores the effect of simultaneous executions on trading cost. We design a Bayesian network modelling the inter-dependencies between investors' transaction costs, stock characteristics (bid-ask spread, turnover and volatility), meta-order attributes (side and size of the trade) and market pressure during execution, measured by the net order flow imbalance of investors meta-orders. Unlike standard machine learning algorithms, Bayesian networks are able to account for explicit inter-dependencies between variables. They also prove to be robust to missing values, as they are able to restore their most probable value given the state of the world. Order flow imbalance being only partially observable (on a subset of trades or with a delay), we show how to design a Bayesian network to infer its distribution and how to use this information to estimate transaction costs. Our model provides better predictions than standard (OLS) models. The forecasting error is smaller and decreases with the investors' order size, as large orders are more informative on the aggregate order flow imbalance (R^2 increases out-of-sample from -0.17% to 2.39% for the smallest to the largest decile of order size). Finally, we show that the accuracy of transaction costs forecasts depends heavily on stock volatility, with a coefficient of 0.78.

Keywords: Trading Costs, Liquidity, Crowding, Bayesian Networks.

*Université Paris-Dauphine, PSL Research University, LEDa, 75016 Paris, France. Amundi Asset Management, 75015 Paris, France. marie.briere@dauphine.psl.eu

†Imperial College London, England. Capital Fund Management, 75007 Paris, France. charles-albert.lehalle@cfm.com

‡Université Paris-Dauphine, PSL Research University, CNRS, UMR [7088], DRM, 75016 Paris, France. tamara.nefedova@dauphine.psl.eu

§Université Paris-Dauphine, PSL Research University, LEDa, 75016 Paris, France. Euronext Paris, 92054 Courbevoie, France. amine.raboun@dauphine.psl.eu

1 Introduction

Transaction costs became of primary importance after the financial crisis. On the one hand, investment banks turned to more standardized products, switching from a high margin, inventory driven business to a low margin, flow business, where transactions costs have to be minimized. On the other hand, the asset management industry concentrated (Haldane et al. (2014)). A common practice has been to organize the execution of large orders around one well structured dealing desk. In 2007, the first Markets in Financial Instruments European directive (MiFID) introduced the concept of "best execution" as a new requirement for market participants. The European best practices, including among others execution reviews, transaction costs analysis, and adequate split of large orders, have spread overseas in this globalized industry.

In this paper we use a unique dataset of institutional investors trades: the ANcerno database, containing a large sample of asset managers meta-orders on the US markets (Angel et al. (2015), Pagano (2008), Briere et al. (2019)). While most other databases contain the meta-orders of only one asset manager, ANcerno records roughly 10% of total institutional investors activity and 8% of total daily traded volume. Because of this specificity, it is possible to estimate the "imbalance of meta-orders", i.e. the aggregated net order flow traded by investors, each day on each stocks. This variable plays a role of primarily importance in the transaction costs (Capponi and Cont (2019), Bucci et al. (2018)). Transaction costs tend to be large when you trade in the same direction as your peers, while you can even have a price improvement (i.e. obtain an average price that is lower than your decision price) if you are almost alone in front of the majority of agents trading that day. Stated differently, you pay to consume liquidity when you are part of the crowd, executing in the same direction as the market, and you are rewarded to provide liquidity to the crowd, when you are executing in the opposite direction of the market.

The specificity of this "imbalance" variable is that it cannot be observed by market participants in real time. Brokers and market makers can have a broad view of the imbalance of their clients' flows and can provide this information to the rest of market participants with a delay, while asset management dealing desks do only observe their own instructions. Therefore, the imbalance is a "*latent variable*" in the sense of Bayesian modelling. It is linked to some observable explanatory variables and it conditions the transaction costs at the same time. For instance: conditionally to the fact that the investor trades a buy meta-order (rather than a sell one), the imbalance is more likely to be large and positive. This dependence can be inferred using the Bayes' rule.

In this paper, we show how to use a specific model belonging to the large toolbox provided by

machine learning: the Bayesian network, adapted to this kind of conditioning, to predict transaction costs, taking into account market information and trade characteristics. This class of models has been created in the golden age of machine learning (Jordan (1998)); it is also known as graphical models, and has been recently used to model analysts predictions (Bew et al. (2018)). Such models have two very interesting characteristics. First, they are able to handle missing data. Second, they can infer the distribution of latent variables given the knowledge of other ones. In our case, a model fitted on ANcerno data can be used to forecast transaction costs when the imbalance is no longer observable. In practice, our model could be fitted on data provided ex-post by brokers[1]. Afterwards, given other explanatory variables and the observed transaction costs, a Bayesian network can infer the expected distribution of the imbalance on a given day. This is a natural feature of the Bayes' rule: once the joint distribution of a set of variable is known, it is possible to obtain the expected value of any subset of other variables given the observations.

The goal of this paper is to show how Bayesian networks can be used to model the relation between transaction costs and stock characteristics (bid-ask spread, average turnover and volatility), meta-order attributes (side and size of the trade) and market pressure (net order flow imbalance). This last variable will be considered as latent because it is only partially observable by investors (typically with a delay, or in real time but only on the investors' own trades). In practice, a possible way to implement our approach would probably be to implement a learning transfer: first learn the graphical model on ANcerno or a similar database provided by brokers, then switch to a database in which the imbalance can not be observed.

We find that institutional investors daily order flow imbalance is a good predictor of transaction costs. Interestingly, because investors' trading tends to be crowded in one direction, and given the fund manager's knowledge of its own meta-order, he can infer the aggregate order flow of the market that day, to better forecast his trading costs. Stated differently, a fund manager could update his beliefs on order flow imbalance distribution of the day, after observing his own trading decision (side and size of his order). We find that his estimation is more accurate when his executed meta-order is large. Besides, we disclose evidence that a sell order is more informative on imbalance distribution than a buy order, probably because a crowded selling context is more informative about specific market conditions than a crowded buying context. We note that when an asset manager takes the decision to sell a stock with high participation rate, he could expect a "rushing towards the exit door" behaviour from his peers and assign a high probability for strong negative imbalance. Our finding confirms that the

[1]Brokers, exchanges and custodians are selling the delayed information on the flows they saw the previous day or week. This Bayesian modelling approach is perfectly suited to this kind of partial information.

dominating variable for implementation shortfall forecast is indeed the order flow imbalance and not the order size. Moreover, the accuracy of transaction costs and market impact estimates are generally very low (Bacry et al. (2015)). Practitioners have long suspected that the difficulty of estimating orders transaction costs is due to the variance of price innovations that is hardly predictable. Thanks to our Bayesian framework, we prove that this is actually true. The Bayesian network explicitly models the dependencies between the variance of the residuals and the rest of network nodes. We find that the dominant variable is, indeed, the price volatility with coefficient 0.78, while other nodes contribution to the variance is insignificant. This allows an investor to assess how confident he could be on each prediction given his meta-order and stock characteristics. Finally, we show that using partially observable order imbalance has value. The Bayesian network provides a better prediction of transaction costs after capturing the conditional dependencies between the nodes and the order flow imbalance, than when this information is not used at all (R^2 increase out-of-sample from 0.38% to 0.50%). Besides, the estimates get more accurate when the order size is large (R^2 is 2.39% for the tenth decile of order size compared to -0.17% for the first decile). These results can explain the recent concentration of institutional investors executions on a few dealing desks. By executing the orders of a large and representative set of institutional investors, these dealing desks would have a better grasp of the aggregate order flow imbalance of the day. This information of paramount importance could then be either used for predicting the transaction cost more accurately, or to design a better optimized execution scheme taking the aggregated market pressure into account.

The structure of this paper is as follows: Section 2 reviews the existing literature on transaction costs modelling and Bayesian networks. Section 3 presents the data. Section 4 provides empirical evidence of the influence of investors trade size and orders imbalance on transaction costs. Section 5 describes the Bayesian network method and its application to transaction costs modelling. Section 6 concludes.

2 Related literature

This paper takes place at the crossing of two fields: the transaction costs and market impact literature on the one hand, and Bayesian modelling on the other hand.

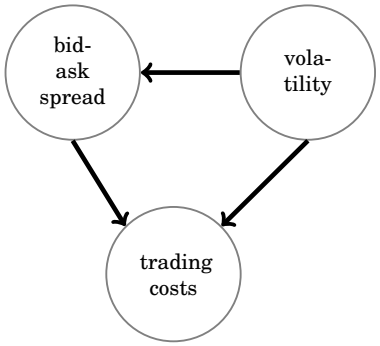
Transaction costs and market impact. Market impact attracted the attention of academics following two papers: economists have been initiated to this crucial concept by Kyle's theoretical paper (1985), while researchers in quantitative finance have been largely influenced by Almgren and Chriss

(2001) empirical results. Kyle (1985) has shown how a market maker should strategically ask informed traders (i.e. asset managers) for a cost to compensate for the difficulty to assess the adverse selection she is exposed to in a noisy environment. This is typically what we observe empirically. Asset managers have to pay for liquidity demand while they can be rewarded for liquidity provision. Other market participants react to the aggregate offer or demand. This aggregate is exactly what we define as *the imbalance of meta-orders for a given day*. Kyle's essential result is that given a linear market-maker pricing rule and within a Gaussian framework, the transaction costs paid by the aggregation of investors are linear in the size of the aggregated meta-order. *Kyle's lambda*, measuring the sensitivity of price impact with respect to volume flow, is a traditional measure of liquidity. This theoretical framework has been sophisticated recently, extending Kyle's game theoretical framework to continuous time, non Gaussian behaviours, and allowing risk aversion in market makers' strategy (Cetin and Rogers (2007)). It is now understood that the informed trader optimal strategy is to try to hide its meta-order in the noise, while the market maker has to slowly digest orders flow to try to extract the information it contains and ask for the corresponding price. However, the resulting market impact is not necessarily linear. Empirical studies that followed showed that in practice market impact is more square root than linear in the size of the order (Collins and Fabozzi (1991), Bouchard et al. (2011) or Robert et al. (2012)).

Almgren and Chriss (2001) seminal paper showed how to split an order optimally to minimize execution cost, making the assumption of concave transient market impact. Bouchard et al. (2011) derived an optimal control scheme to mitigate this cost for large meta-orders. This literature is of primary importance since it answers the regulatory requirements around "proof of best execution" and provides a baseline framework to asset managers and investment banks to improve their best practices and metrics for execution. With the popularity of factor investing, the specific question of the implementation costs of investment strategies following an index or a systematic active strategy has been raised by regulators and market participants. Frazzini et al. (2012), Novy-Marx and Velikov (2015), or Briere et al. (2019) are attempting to answer the question of potential maximum capacity of a trading strategy, by modelling transaction costs for large order sizes and estimating the break-even capacity of factor-driven investment strategies.

Bayesian networks. Machine learning is an extension of statistical learning, born with the seminal paper of Vapnik and Chervonenkis (1971). Following the universal approximation theorem for non linear Perceptrons (a specific class of neural networks) with at least one hidden layer (Hornik et al. (1989)), statisticians and mathematicians started investigating approximation schemes based on

Figure 1. A simple graphical model for trading costs modelling



the minimization of a possibly non-convex loss function, generally using a stochastic gradient descent (Amari (1993)) to reach the global minimum while having good chances to escape from the local minima. Successes in Bayesian statistics, focused on coupling a prior and a posterior distribution via the concept of conjugate (Vila et al. (2000)), opened the door to a mix of neural networks and Bayesian statistics, based on maximum likelihood estimations. Bayesian networks were born (see the seminal paper by Pearl (1986)). Bayesian networks are convenient tools for modelling large multivariate probability models and for making inference. A Bayesian network combines observable explanatory variables with hidden latent variables in an intuitive, graphical representation.

In terms of applications, Bayesian networks have first been used for medical diagnosis, since they have been perceived as a natural extension of *expert systems*. Expert systems emerged with the first wave of artificial intelligence tools: Deterministic decision trees. Adding some probabilistic properties to these trees and reshaping them into graphs is another way to see the emergence of Bayesian networks. These models have also been used with success in troubleshooting of computed components, from printers (Skaanning et al. (2000)) to computer networks (Lauritzen (2003)). They played an important role in the automation of problem solving for computers related questions. Recently, they have been applied in finance. Bew et al. (2018) use Bayesian networks to combine analysts' recommendations to improve asset management decisions.

These models can very naturally capture the joint distribution of different variables, specified via a graphical model where nodes represent variables and arrows model the probabilistic dependencies. The very simple example of Figure 1 specifies that the stock bid-ask spread and its volatility both influence trading costs, while at the same time, the stock volatility has an influence on the bid-ask spread (Laruelle and Lehalle (2018)). The translation in a probabilistic language of this graph is the following. The *trading costs* TC, follows a law $\mathscr{L}$ which parameters Θ_{TC} are functions of the *bid-ask spread* ψ and of the *volatility* σ: $\mathrm{TC} \sim \mathscr{L}(\Theta_{TC}(\psi, \sigma))$. The parameters of the law of the *bid-ask spread*

are seen as a random variable, itself a function of the volatility: $\psi \sim \mathscr{L}(\Theta_\psi(\sigma))$.

More details on the mechanisms of Bayesian networks are given in Section 5. At this stage, it is enough to say that *latent variables* can be added to the graph. An intermediate variable that is not always observable, but acting as a probabilistic intermediary (i.e. a conditioning variable) between observed variables, is enough to structure a Bayesian model. In the simple example of Figure 1, we can observe or not the bid-ask spread. When it is not observed, the Bayesian network will use its law $\mathscr{L}(\Theta_\psi(\sigma))$ to infer its most probable value, conditionally to the observed volatility. To do that, the model uses Bayes' conditional probability chain rule. In our analysis, we always observe the bid-ask spread, but the net order flow imbalance of institutional investors meta-orders is usually not known. This paper proposes a Bayesian network to model and forecast transaction costs with a graphical model where the imbalance of institutional meta-orders is a latent variable.

To sum up, our paper makes use of Bayesian networks to model the expected transaction costs of institutional investors as a function of the characteristics of the meta-order (essentially its size and direction), the market environment (stock volatility, bid-ask spread and order flow imbalance). We contribute to the current literature on trading costs estimation by proposing a methodology to account for latent variables, in our case, order flow imbalance. This variable can only be partially observed (with a delay or on a subset of all trades), but is essential to structure the model. Our model has numerous potential applications and could be used to forecast trading costs, estimate the capacity of a strategy or decide on the optimal trading execution.

3 ANcerno database

We obtain institutional trading data for the period from January 1^{st} 2010 to September 30^{th} 2011 from ANcerno Ltd. ANcerno, formerly Abel Noser Corporation, is one of the leading consulting companies in providing Transaction Cost Analysis (TCA) in the US. It provides equity trading costs analysis for more than 500 global institutional investors, including pension funds, insurance companies and asset managers. This database was largely used by academics to investigate institutional investors trading behaviour (see for example Anand et al. (2011), Puckett and Yan (2011) and Eisele et al. (2017)). ANcerno clients send their equity trades in order to monitor their execution quality. ANcerno systematically reports all equity trades it receives. Therefore, costs estimated on ANcerno are representative of what is effectively paid by institutional investors. Besides, previous research have shown that ANcerno is free from any survivorship or backfill bias (see Puckett and Yan (2011)), constitute approximately 8% of the total CRSP daily dollar volume (Anand et al. (2013)), and 10% of total

institutional activity (Puckett and Yan (2011)).

Hence, in our study we use trade-level data from ANcerno on the historical composition of S&P 500 index. For each execution, ANcerno reports information on the CUSIP and ticker of the stock, the execution time at minute precision, the execution date, execution price, side (i.e., buy or sell), number of shares traded, commissions paid, whether the trade is part of a larger order, and a number of trade-level benchmarks to evaluate the quality of the execution. In our sample, we have execution data of 285 institutions (i.e., ANcerno clients). They could be either an individual mutual fund, a group of funds, or a fund manager subscribing to Abel Noser analytical service. Each institution has one or several accounts. In our sample, we successfully track the activity of almost 44 thousands of accounts, responsible of 3.9 trillion dollars of transactions, and using the service of 680 different brokerage firms. Compared to market volume reported in CRSP, ANcerno accounts for an average of 4.5% over the whole period. The traded amount reported in ANcerno is over a trillion dollars every year and is, therefore, large enough to be relevant. We complement ANcerno database with daily bid-ask spread obtained from Reuters Tick History (RTH).

Consistent with machine learning best practices, we split our sample to a training set accounting for 70% of the meta-orders and a testing set accounting for the remaining 30%. The training-set is chosen randomly from meta-orders in our sample such as the number of buy orders and sell orders are equal. This procedure is very important for our study in order to estimate a non biased net order flow imbalance. In the case of unbalance number of buy and sell orders in the training-set, the prior distribution of order flow imbalance will be artificially skewed toward positive values if the number of buy orders is higher or toward negative values otherwise. The training set is used to compute the results of section 4 and 5, while the testing set is used for the out-of-sample predictions in section 6.

4 Transaction cost modelling

We measure trading costs with the traditional measure of implementation shortfall (Perold (1988)). This is the difference between a theoretical or benchmark price and the actual traded price effectively paid for the execution, in percent of the benchmark price. In our study, we define the reference price as the last visible price before the start of the execution (arrival price). The implementation shortfall measures the total amount of slippage a strategy might experience from its theoretical returns. In essence, our cost estimate measures how much of the theoretical returns of a strategy can actually be achieved in practice.

For a parent-ticket m of size $Q_k(m)$ split into N_{trades} child tickets [2] of size $v_{k,m}(i)$ executed at date d in the direction $s_k(m)$, the implementation shortfall is calculated as follows:

$$\text{IS}_k(m,d) = \frac{s_k(m)}{P_k(0)} \left(\sum_{i=1}^{N_{\text{trades}}} \frac{v_{k,m}(i)}{Q_k(m)} \times P_k(i) - P_k^{\text{ref}} \right) \tag{1}$$

where $P_k^{\text{ref}} = P_k(0)$ is the reference price (in our case, the arrival price as provided by ANcerno). In this section, we investigate the effect of order size and order flow imbalance on the implementation shortfall of investors transactions.

4.1 Order size

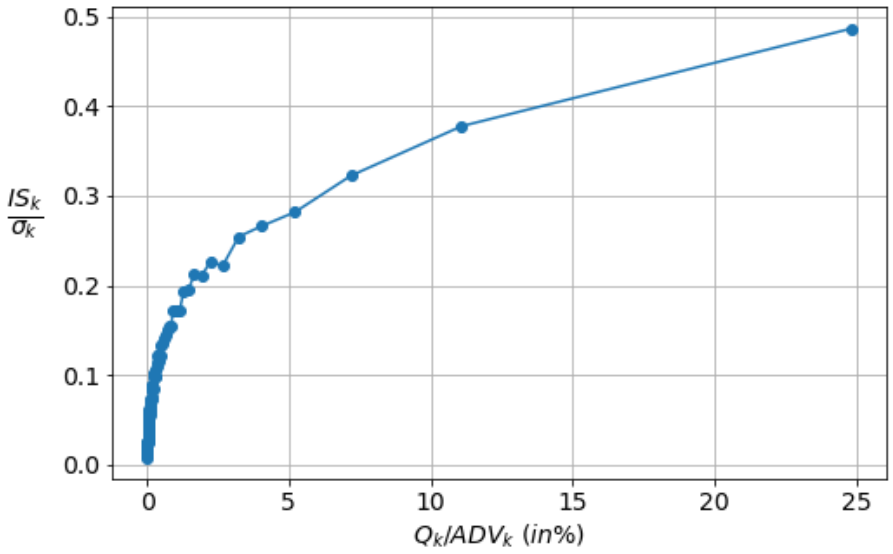
Kyle (1985) introduced the concept that trades by a market participant may have an impact on the market price. Market impact is a direct consequence of the order size effect. A large meta-order may move the price in an unfavourable direction for the trader, resulting in a higher implementation shortfall. The execution cost is then increasing with order size. A series of empirical studies followed Kyle's theoretical work to confirm the existence of order size effect (Torre and Ferrari (1999), Moro et al. (2009), Gomes and Waelbroeck (2015), Bacry et al. (2015), Briere et al. (2019)). To illustrate this effect, we regroup ANcerno tickets in 100 bins based on participation rate Q/ADV and plot in Figure 2 the average implementation shortfall scaled by price volatility of the tickets in each bin. The scaling with stock's price volatility makes estimates of implementation shortfall comparable through time and across the universe of stocks. Otherwise, we can not compute the average on each bin as the effect of participation rate is not the same for large bid-ask spread stocks than those with small bid-ask spreads. ANcerno tickets show a concave relation between the implementation shortfall and order size relative to daily traded volume. We observe a sharp increase of the costs from 0 to 0.2 points of price volatility when order size increases from 0.01% to 2% of the average daily volume. The slope decays afterward. For instance a ticket with 14% participation rate, costs on average 0.4 points of volatility. A power law function captures well the dependence of orders trading costs to order size.

4.2 Order flow imbalance

While most of trading cost models emphasize on the historical dependence of market impact on stock liquidity and order size, It is only recently that order flow imbalance has been recognized as a significant factor in explaining the magnitude of orders transaction costs. Using ANcerno database,

[2] Orders in ANcerno (parent tickets) are split within the execution period into smaller orders (child tickets). For each child ticket, ANcerno reports the executed volume, the price and time of execution.

Figure 2. Order size effect on trading costs
Institutional trading data are obtained from ANcerno Ltd on the period ranging from January 1, 2010 to September 30, 2011. We split our sample into 100 bins based on meta-order participation rate $Q_k(m)/\text{ADV}_k(d)$ and plots the average implementation shortfall scaled by stock's volatility IS_k/σ_k for each bin (blue dots)



Capponi and Cont (2019), compared the explanatory power of order size to the effect of a proxy of market pressure "Order Flow Imbalance" on transaction costs and came to the conclusion that investors should focus on modelling the aggregate dynamics of market pressure during execution period, rather than focusing on optimizing market impact at a trade-by-trade level. Moreover, market pressure is contributed by all market participants present at the trading session. But the traders who are responsible of executing institutional investors orders contribute the most to this pressure and should be specifically taken into account in price movement forecast and transaction costs modelling. These market participants have the same profile as the informed/insider trader introduced by Kyle in 1985. By the end of the trading session, the private information, that was once detained by the insider, spread to the market and get incorporated into the price level. Bucci et al. (2018) argue that price market impact is function of the aggregate net volume, that for shared indiscriminately between all market participants. Consequently, a small sized order would cost nearly the same implementation shortfall as a much larger order if executed in the same direction during the same time frame.

We introduce the Net Order Flow Imbalance, to investigate the impact of institutional investors synchronous trading on the implementation shortfall. For a met-order m executed at date d, the net investors order flow imbalance is defined as the ratio of net volume executed by the other investors at

day d over their total traded volume:

$$\mathrm{Imb}_k(m,d) = \frac{\sum_{m' \neq m} Q_k(m',d).s_k(m',d)}{\sum_{m' \neq m} Q_k(m',d)} \tag{2}$$

Where k designs the stock, $s_k(m',d)$ is the side of the meta-order m' (i.e. 1 for buy orders and -1 for sell orders) and $Q_k(m',d)$ its size.

Figure 3. Net order flow imbalance effect on trading costs
Institutional trading data are obtained from ANcerno Ltd on the period ranging from January 1, 2010 to September 30, 2011. We split our sample into 100 bins based on net order flow imbalance multiplied by the side of the trade and plots the average implementation shortfall scaled by stock's volatility IS_k/σ_k for each bin (blue dots)

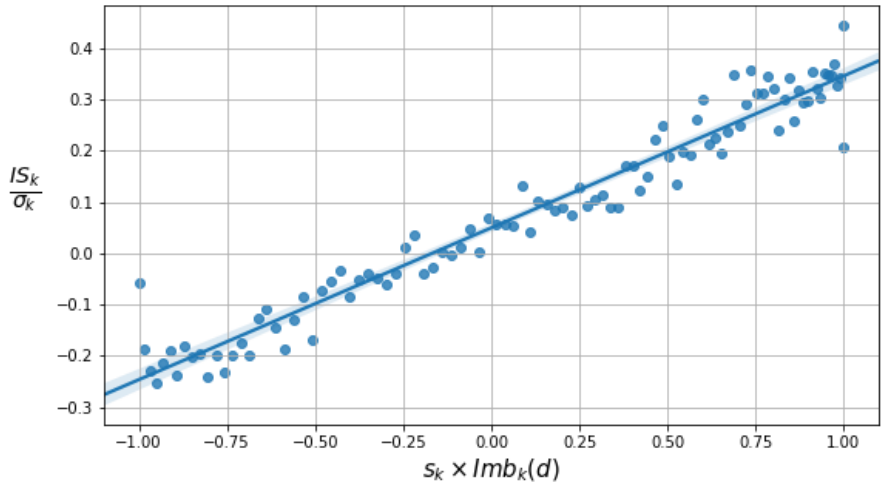
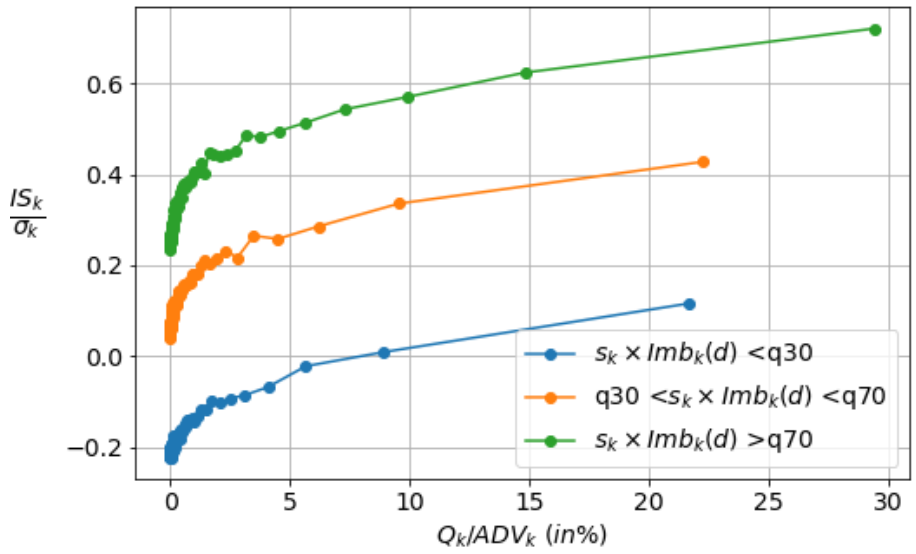


Figure 3 illustrates the dependence of the implementation shortfall to institutional investors trading imbalance. First, we note that the relationship is linear. The stronger the absolute imbalance, the higher the absolute value of price deviation during the execution. But depending on whether the trade is on the same direction as the net order flow imbalance, thus contributes to the existing market pressure, or on the opposite side, and provides liquidity to the market, one could expect either to pay a significant trading cost up to 0.4 points of price volatility when investors are trading synchronously toward the same directions ($\mathrm{Imb}_k(m,d) = 1$) or benefit from a price improvement of 0.3 points of volatility when the trader is almost alone in front of his competitors aggregate flow ($\mathrm{Imb}_k(m,d) = -1$). Also worth noting that the implementation shortfall at zero imbalance is slightly positive. At neutral market pressure, the investors pays a positive transaction cost depending on stock traded and meta-order size.

4.3 Joint effect of order size and order flow imbalance

Figure 4. Joint effect of order size and net order flow imbalance on trading costs
Institutional trading data are obtained from ANcerno Ltd on the period ranging from January 1, 2010 to September 30, 2011. First, We split our sample on 3 buckets w.r.t meta-order signed imbalance ($s_k(m) \cdot \text{Imb}_k(m,d)$) 30% and 70% quantiles. We sort meta-orders within each bucket into 100 bins based on meta-order participation rate ($Q_k(m)/\text{ADV}_k(d)$) and plots the average implementation shortfall scaled by stock's volatility IS_k/σ_k for each of the bins.



The results in subsection 4.1 and 4.2 show that the implementation shortfall depends on both the size of the executed order, and market pressure during execution period. Market pressure being approached by investors net order flow imbalance. To disentangle the two effects, we split our sample on 3 distinct buckets with respect to meta-order signed imbalance ($s_k(m) \cdot \text{Imb}_k(m,d)$) 30% and 70% quantiles. Within each bucket, we sort meta-orders into 100 bins based on meta-order participation rate ($Q_k(m)/\text{ADV}_k(d)$) and compute the average implementation shortfall scaled by stock's volatility for each of the bins. Figure 4 plots the result, where the blue, line shows order size effect for meta-orders executed against high market pressure (signed imbalance is lower than the 30% quantile). Orange line illustrate the effect for meta-orders executed under standard market pressure (signed imbalance between the 30% and 70% quantiles). Whereas the green line shows the result for orders executed in the same direction as the market (signed imbalance larger than the 70% quantile). We observe the impact of meta-order size is persistent in the 3 buckets and the power law remains valid even after conditioning on net order flow imbalance. The linear effect of the signed imbalance is visible in the difference of transaction cost level between the 3 buckets. Which proves that these two explanatory

factors does not cancel one another. We also note that most of meta-orders executed against investors net order flow benefit from a price improvement between the moment the execution starts and the moment it ends. During strongly unbalanced markets, the provider of liquidity is rewarded with a better execution price. However, for larger meta-orders ($Q_k(m)/\mathrm{ADV}_k(d) = 23\%$) the market impact of the trade prevail and the trader pays on average a positive transaction cost. The opposite is also true, when traders seek liquidity on the same direction as the remainder of institutional investors, the trading cost gets more expensive than usual.

To further explore the joint effect of order size and net order flow imbalance on the implementation shortfall, we run the following step-wise multivariate regression. First, we perform the regression of order implementation shortfall on stock bid-ask spread and the square root of the order participation rate scale by stock volatility as described in equation (3). Then, a regression of the implementation shortfall on the bid-ask spread and the signed imbalance, also scaled by stock volatility (equation 4). Finally, we gather the 3 factors on the same regression as in equation (5).

$$\mathrm{IS}_k(m,d) = \alpha\ \psi_k(d) + \beta\ \sigma_k^{\mathrm{GK}}(d)\sqrt{\frac{Q_k(m)}{\mathrm{ADV}_k(d)}} + \varepsilon_k(m,d) \tag{3}$$

$$\mathrm{IS}_k(m,d) = \alpha\ \psi_k(d) + \gamma\ \sigma_k^{\mathrm{GK}}(d)\ s_k(m)\ \mathrm{Imb}_k(m,d) + \varepsilon_k(m,d) \tag{4}$$

$$\mathrm{IS}_k(m,d) = \alpha\ \psi_k(d) + \beta\ \sigma_k^{\mathrm{GK}}(d)\ \sqrt{\frac{Q_k(m)}{\mathrm{ADV}_k(d)}} + \gamma\ \sigma_k^{\mathrm{GK}}(d)\ s_k(m)\ \mathrm{Imb}_k(m,d) + \varepsilon_k(m,d) \tag{5}$$

where $\mathrm{IS}_k(m,d)$ is the implementation shortfall of meta-order m submitted on stock k at day d. $\psi_k(d)$ is the quoted intraday bid-ask spread of stock k averaged on the month. $\sigma_k^{\mathrm{GK}}(d)$ is the Garman and Klass (1980) intraday volatility of stock k estimated on a 12 month rolling window. $Q_k(m)$ and $s_k(m)$ are respectively size and side (Buy/Sell) of the order. $\mathrm{ADV}_k(d)$ is the daily traded volume averaged on a 12 months rolling window, and $Q_k(m)/\mathrm{ADV}_k(d)$ is the participation rate. $\mathrm{Imb}_k(m,d)$ is the net investors order flow imbalance estimate for order m at day d. Finally, α, β and γ are model parameters and $\varepsilon_k(m,d)$ is the respective error term.

The results of these three regressions are presented in Table 1. In the first regression the coefficient of bid-ask spread and order size term $\left(\sigma_k^{\mathrm{GK}}(d)\ \sqrt{Q_k(m)/\mathrm{ADV}_k(d)}\right)$ are respectively 0.4 and 0.95, both statistically significant at the 1% level. Consistently with Briere et al. (2019) we find that for small orders, institutional investors pay only 0.4 times the bid-ask spread. In the second regression, we replace the order size term by the market pressure term. We notice that the coefficient of the bid-ask

Table 1. Transaction cost model
Institutional trading data are obtained from ANcerno Ltd on the period ranging from January 1, 2010 to September 30, 2011 on the S&P 500 historical components. $\psi_k(d)$ is the quoted intraday bid-ask spread of stock k averaged on the month, obtained from RTH database. $\sigma_k^{\mathrm{GK}}(d)$ and $\mathrm{ADV}_k(d)$ are respectively the Garman Klass intraday volatility and the average daily volume of stock k estimated on a 12 month rolling window. $Q_k(m)$ and $s_k(m)$ are respectively size and side (Buy/Sell) of the order. $\mathrm{Imb}_k(m,d)$ is the net order flow imbalance for order m at day d.

Model	Dependent variable: $\mathrm{IS}_k(m,d)$		
$\psi_k(d)$	0.399***	0.708***	0.180***
	(0.032)	(0.028)	(0.032)
$\sigma_k^{\mathrm{GK}}(d)\ \sqrt{Q_k(m)/\mathrm{ADV}_k(d)}$	0.951***		0.712***
	(0.021)		(0.021)
$\sigma_k^{\mathrm{GK}}(d)\ s_k(m)\ \mathrm{Imb}_k(m,d)$		0.234***	0.224***
		(0.002)	(0.002)
Observations	7421548	7421548	7421548
R^2	0.005	0.016	0.017
Adjusted R^2	0.005	0.016	0.017
Residual Std. Error	0.017	0.017	0.017
F Statistic	1892.157	5993.682	4391.073
AIC	-3964520	-3972637	-3973801
Note:			*p $<$ 0.1; **p $<$ 0.05; ***p $<$ 0.01

spread increases (from 0.4 to 0.7) and its confidence interval becomes tighter (lower standard deviation 0.028 vs 0.032). The determination coefficient for the second regression is also much higher (1.6% vs 0.5%). Finally, when we put all explanatory variables together, we find that coefficient of the order imbalance does not change (0.22-0.23) while both order size term and bid-ask spread have much lower parameters (0.18 for bid-ask spread and 0.71 for order size term) compared to the first two models. Besides, the determination coefficient of the second and third regressions are comparable. The net order flow imbalance seems to be much better predictor of expected implementation shortfall than the size of the order. Although, all coefficient are statistically significant a the 1% level.

5 Bayesian network modelling with net order flow imbalance as latent variable

Institutional investors net order flow imbalance is a key factor in the estimation of meta-orders transaction costs. However, this variable is only observable with a delay, for example through brokers or custodians' reports, or on a subset of trades only (the investor's own trades). Thus, it can not be used for production purposes. To remedy this issue, we propose a Bayesian network to incorporate all information we could get before the execution of the meta-order, and update our beliefs on the

probabilistic distribution of the latent variable. We then use the most probable value of the net order flow imbalance to estimate the meta-order transaction cost. One of the interesting features of Bayesian networks is that they can be explored in both directions, thanks to the Bayes' rule. Therefore, we can give an estimate of the latent variables, by probabilistic inference, before and after the variable of interest is observed. In the context of this study, this means that:

- given the characteristics of the meta-order (side and size of the trade) and stock attributes (bid-ask spread, average daily traded volume and price volatility), we can compute a first estimate of the imbalance and forecast the transaction costs that should be paid by the investor.
- Once we get the effectively paid trading cost, we can recover a more accurate estimate of the distribution of investors net order flow of the day, and for example incorporate it in the estimation of order flow imbalance of the following day.

5.1 Bayesian inference

The main difference between the frenquentist approach and the Bayesian approach is that in the latter, the parameters of the models are no longer unknown constants that need to be estimated, but random variables which parameters have to be estimated. The statistician has the possibility to incorporate his prior belief on the probabilistic distribution of the variable and update his belief step by step as soon as new data becomes available. From one step to the other: the former "posterior distribution" is used as a prior for the next step estimate.

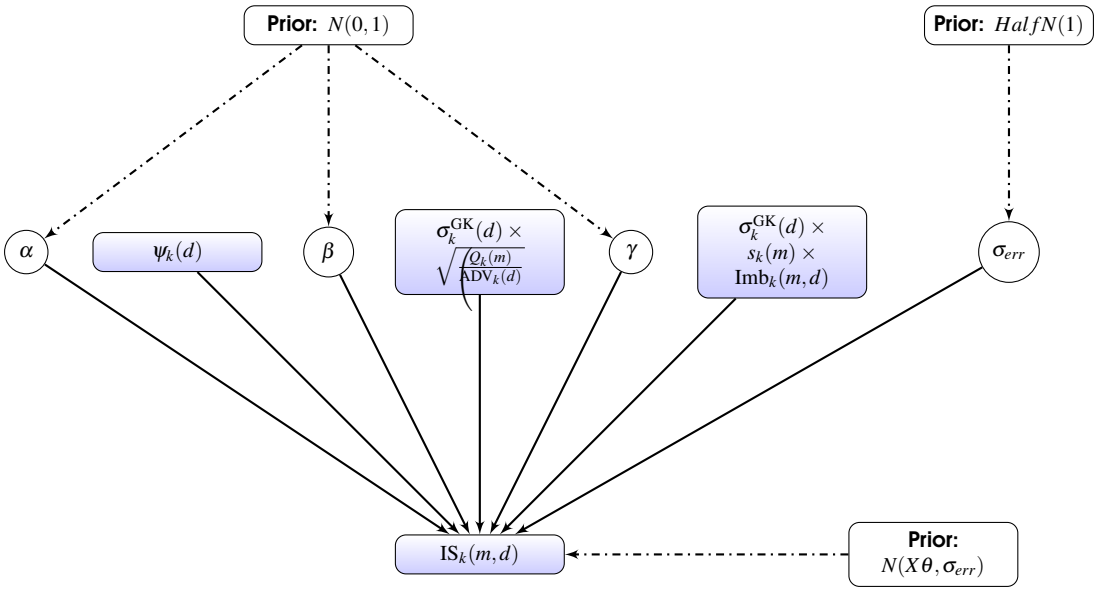
For instance, if y stands for the unknown random variable, x for the observed data, and $\mathbf{P}(y)$ for the prior. The posterior distribution $\mathbf{P}(y|x)$ is obtained as the multiplication of the prior $\mathbf{P}(y)$ with the likelihood $\mathbf{P}(x|y)$ of observing the data, scaled by $\mathbf{P}(x)$. The definition of conditional probabilities applied on this procedure reads:

$$\mathbf{P}(y|x) = \frac{\mathbf{P}(y)\mathbf{P}(x|y)}{\mathbf{P}(x)} \propto \mathbf{P}(y)\mathbf{P}(x|y). \tag{6}$$

Figure 5 shows how to estimate the coefficients of the Bayesian linear regression specified in equation (5). First, we start by incorporating our prior beliefs, if any, on the distribution of the parameters $\theta = (\alpha, \beta, \gamma)^T$. Without any belief, a good choice is to take a non-informative prior like the normal distribution $N(0,1)$. The best initialisation for priors is hence a law close to the empirical repartition function of the considered variable. The variable of interest $\mathrm{IS}_k(m,d)$ follows a normal distribution centered at the estimated value $\hat{y} = X\theta$ and has variance σ^2_{err} of the error term $\varepsilon_k(m,d)$. σ^2_{err} requires a non negative prior distribution, such as the positive part of a Gaussian (i.e. *HalfNormal*) or the positive part of a

Cauchy (i.e. *HalfCauchy*). The Bayesian setup gives direct interpretation of the results: The mean of the posterior distribution is the most probable value of the parameters θ, and the 5% confidence interval is limited by the 2.5% and 97.5% quantiles of the posterior distribution.

Figure 5. Bayesian inference of a linear regression
Blue rectangle represent observed variables. Circles are the parameters that need to be calibrated. Each have a prior distributions detailed in white rectangle.
$X\theta = \alpha\ \psi_k(d) + \beta\ \sigma_k^{\mathrm{GK}}(d)\ \sqrt{Q_k(m)/\mathrm{ADV}_k(d)} + \gamma\ \sigma_k^{\mathrm{GK}}(d) \cdot s_k(m) \cdot \mathrm{Imb}_k(m,d)$



MCMC (Markov Chain Monte Carlo, see Hastings (1970) for one of the first references) methods offers an easy way to sample from the posterior, especially when the posterior does not obey to a well known expression or when we know the expression has a multiplicative term. It is very convenient for the Bayesian approach, where the posterior distribution is proportional to the multiplication of the prior and the likelihood. MCMC algorithms makes computations tractable for parametric models. The intuition behind MCMC is to define a Markov Chain $(x_0, x_1, \ldots)$ on the support of x, such that as when the size n of this chain goes to infinity, the new drawn point x_n is distributed accordingly to the law $\mathbf{P}_x$. The most famous algorithms to generate Markov Chains having this very nice property are the Hasting-Metropolis one, explained in Appendix F, that we use in this study, and the Gibbs sampler. [3]. The marginal distribution of regression coefficients of the calibrated model is shown in the right panel of Table 2, while the result of the OLS regression is in the left panel. As expected from Bayesian models when the sample size is large, we end up with the same results. Beside, when the priors are Gaussian, the maximum a posteriori of the parameters is equivalent to a ridge estimate with a quadratic

[3] We use the PyMC3 python package implementation of Hasting-Metropolis algorithm described in Salvatier et al. (2016) with a large number of iterations $N_{iter} = 10000$

regularization $\left(\mathbb{E}_{\theta|X,Y}\ [\theta] = \arg\max_\theta\ \mathbf{P}(\theta|X,Y) = \arg\min_\theta\ ||Y - X\theta||^2 + \sigma_{err}^2||\theta||^2\right)$. This formula, similar to the one of Ridge regression (see Hoerl and Kennard (1970)), makes the Bayesian regression more robust to outliers than OLS. It is the case for example for the order size term $\sigma_k^{\mathrm{GK}}(d)\sqrt{Q_k(m)/\mathrm{ADV}_k(d)}$ distribution, which explain the minor difference in coefficient estimate (0.71 vs 0.69).

Table 2. OLS regression vs Bayesian regression
Institutional trading data are obtained from ANcerno Ltd on the period ranging from January 1, 2010 to September 30, 2011 on the S&P 500 historical components. $\psi_k(d)$ is the quoted intraday bid-ask spread of stock k averaged on the month, obtained from RTH database. $\sigma_k^{\mathrm{GK}}(d)$ and $\mathrm{ADV}_k(d)$ are respectively the Garman Klass intraday volatility and the average daily volume of stock k estimated on a 12 month rolling window. $Q_k(m)$ and $s_k(m)$ are respectively size and side (Buy/Sell) of the order. $\mathrm{Imb}_k(m,d)$ is the net order flow imbalance for order m at day d.

	OLS Regression				Bayesian Regression			
	coef	std err	Q 2.5%	Q 97.5%	coef	std err	Q 2.5%	Q 97.5%
$\psi_k(d)$	0.18	0.03	0.12	0.24	0.18	0.03	0.12	0.24
$\sigma_k^{\mathrm{GK}}(d)\sqrt{\frac{Q_k(m)}{\mathrm{ADV}_k(d)}}$	0.71	0.02	0.67	0.75	0.69	0.02	0.65	0.73
$\sigma_k^{\mathrm{GK}}(d)s_k(m)\mathrm{Imb}_k(m,d)$	0.22	0.00	0.22	0.23	0.22	0.00	0.22	0.23
RMSE (%)	1.66				1.66			
R^2 (%)	1.77				1.77			

5.2 Bayesian network modelling

Most of the OLS assumptions are violated. As shown in Appendix E, the marginal distribution of trading costs has a peaky shape, with fat tails (excess-kurtosis of 23.46). The assumption of homoscedasticity is also violated. The variance of the error term is hardly constant across orders. Forecasting errors are smaller for small orders (implemented in a few minutes) compared to large ones (split over days) that got exposed for a longer period to market volatility. Finally, it is difficult to assume that the observations are independent of one on-other. Meta-orders on the same stock, whatever the execution day, share some common variance related the stock characteristics. Similarly orders executed at the same trading session on different stocks face the same market conditions, and thus cannot be considered independent of one another.

In addition, Bayesian Networks have the advantage of not relying on Normal error distributions (Zuo and Kita (2012)), as do most other machine learning algorithms. Furthermore, Bayesian networks have the advantage of giving a human-readable description of dependencies between considered variables, whereas other more complex models, such as Neural Networks, suffer from being considered as ”black box” models.

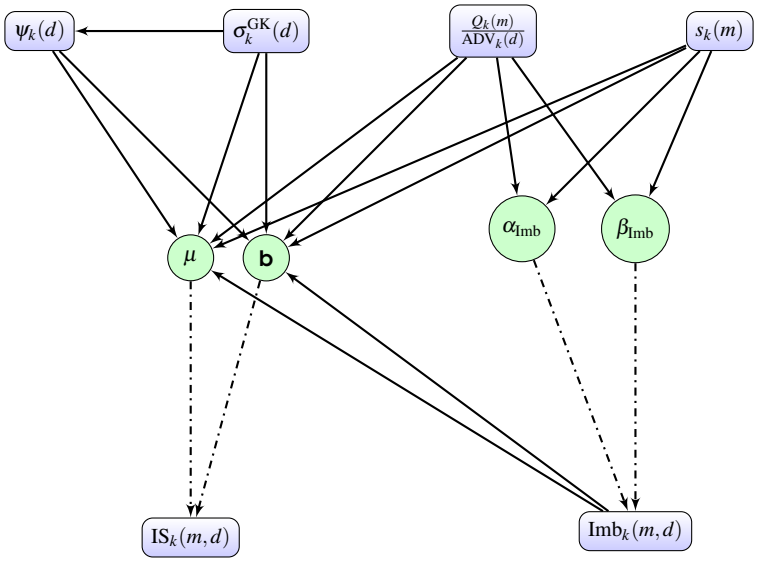
5.2.1 The structure of our Bayesian network

Our goal is to estimate the Implementation Shortfall of an order. We would like the Bayesian network takes into account:

- Attributes of the traded stock, such as average bid-ask spread, price volatility and average turnover;
- Characteristics of the meta-order, mainly order size and side of the trade (Buy/ Sell);
- And the level of crowding during the execution: the net order flow of large institutional firms.

Figure 6 shows the Bayesian network we engineered. We distinguish 3 key dependencies. First, the bid-ask spread depends on the level of stock volatility. Second, the marginal probability distribution of order flow imbalance is function of the meta-order size and side. Finally the implementation shortfall is function of all network nodes. In the following section, we detail the nature of these dependencies and we set the priors for each group of variable separately.

Figure 6. Bayesian network for transaction costs modelling



5.2.2 Bid-ask spread dependencies

The relation between stock volatility and bid-ask spread is well documented. Theoretically, it is justified by Wyart et al. (2008) that, deriving the P&L of traders submitting market orders and those submitting limit orders, an equilibrium price is only achievable if the bid-ask spread is proportional to price volatility (i.e. $\psi_k(d) \propto \sigma_k^{\mathrm{GK}}(d)$). In the same fashion, Dayri and Rosenbaum (2015) study the

optimal tick size, and find that the bid-ask spread that the market would prefer to pay if not constraint by the tick size verifies $\frac{\psi_k(d)}{2} \propto \frac{\sigma_k^{\mathrm{GK}}(d)}{\sqrt{M}}$. The rational is that market makers, setting the best limits of the order book, accept to provide tight bid-ask spreads not only when the volatility (i.e. the risk of a given inventory level) is low, but also when they have more opportunities within the day to unwide their position. The relation between the bid-ask spread and the volatility is confirmed empirically on our data, as illustrated in Figure 11 of Appendix B.

Consistent with the literature of stochastic models for volatility, we set the prior of stocks volatility to a log normal distribution $\sigma_k^{\mathrm{GK}}(d)) \sim \mathit{LogNormal}$. Consequently, the bid-ask spread should follow a log normal distribution too, and the conditional probability of bid-ask spread given price volatility is detailed in equation (7), where $c^{\psi\sigma}, \rho^{\psi\sigma}, \sigma_{\psi,\sigma}$ are model parameters.

$$\psi_k(d)\ \sigma_k^{\mathrm{GK}}(d) \sim \mathit{LogNormal}\left(c^{\psi\sigma} + \rho^{\psi\sigma} log(\sigma_k^{\mathrm{GK}}(d)),\ \ \sigma_{\psi,\sigma}\right) \tag{7}$$

Table 3. Bayesian inference: Bid-Ask spread, volatility dependencies
The table summaries the posterior distribution of model parameters described in equation (7). $\mathbb{E}[X]$, std(X), Q2.5% and Q97.5% are respectively the mean, the standard deviation, the 2.5% and 97.5% quantile of parameters posterior distribution. The results are obtained from Hasting-Metropolis sampler with $N_{iter} = 10000$ iterations (PyMC3 implementation). The institutional investors trading data are obtained from ANcerno Ltd on the period ranging from January 1^{st}, 2010 to September 30^{th}, 2011.

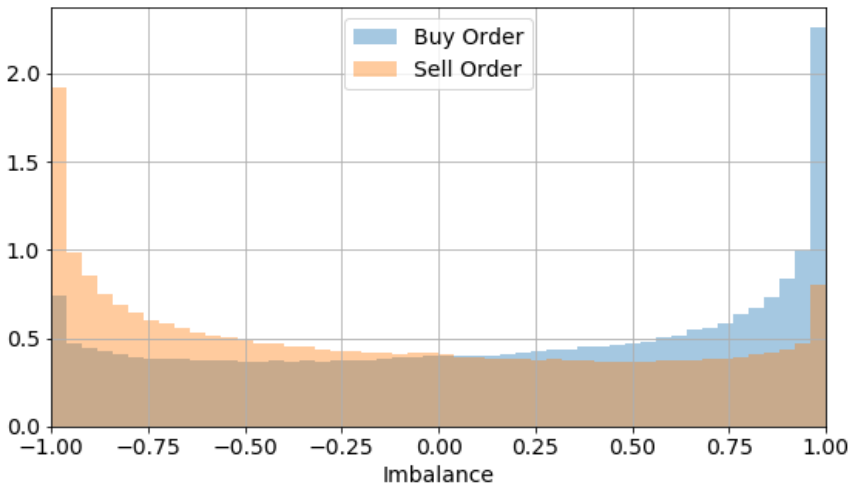
	$\mathbb{E}[X]$	Std[X]	Q2.5%	Q97.5%
$c_{\psi,\sigma}$	-4.137	0.006	-4.150	-4.126
$\rho_{\psi,\sigma}$	0.777	0.001	0.775	0.780
$\sigma_{\psi,\sigma}$	0.402	0.000	0.401	0.402

5.2.3 Net order flow imbalance dependencies

In this section, we quantify the dependence of net order flow imbalance to the remainder variables in the network. Figure 7 shows the marginal distribution of the imbalance depending on the sided of the meta-order. The U-shape of the plot confirms that institutional investors have indeed correlated executions, and tend to execute the same stocks toward the same directions during the same periods, which intensifies the pressure on price movements. This correlation in trade execution can be explained by various factors. Asset managers compete for the same base of customers and can implement similar strategies (Greenwood and Thesmar (2011), Koch et al. (2016)). Thus, they face similar inflows and outflows, depending on liquidity needs and investment opportunities. Moreover, the asset management industry is subject to a series of regulatory constraints that can push funds to buy or sell simultaneously the same kind of assets. We note also, that the U-shape is decomposed in two

skewed distributions depending on the side of the meta-order. So, given the side of the meta-order of an institutional investor, the remainder of large arbitrageurs executions constitute either a positively (for sell) or negatively (for buy) skewed imbalance distribution. Besides, conditional on the level of a meta-order participation rate, Figure 8 shows that the intensity of absolute net order flow imbalance of investors meta-orders gets stronger, a confirmation of institutional investors crowding.

Figure 7. Net order flow imbalance distribution given meta-order side
Institutional trading data are obtained from ANcerno Ltd on the period ranging from January 1, 2010 to September 30, 2011. Given a meta-order m submitted by an institutional investor, the figure plots the distribution of the net order flow imbalance generated by the remainder of investors as defined in equation (2) given the side $s_k(m)$ of meta-order m.



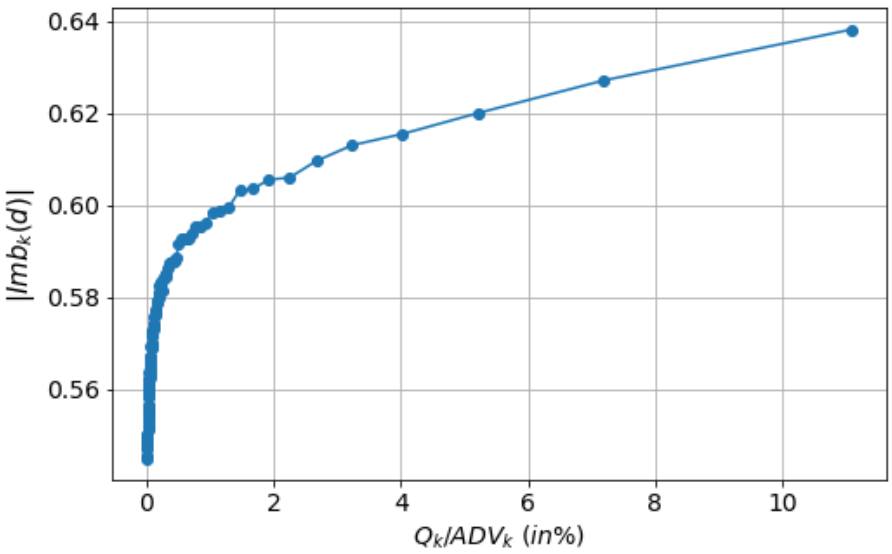
The Side takes two values Buy and Sell. It is modelled with a Bernoulli distribution $s_k(m) \sim Bernoulli(p_{side})$ of parameter $p_{side} = \frac{1}{2}$. The data shows that a Beta function is a good approximation of the U-shape for variables defined between [0,1]. After applying the linear transformation $x \to \frac{x+1}{2}$, $Beta(\alpha, \beta)$ is a plausible distribution of the transformed net order flow imbalance. Moreover, the beta distribution have the particularity to produce different shapes depending on the parameters α and β. It produces a symmetric U-shape when $\alpha = \beta$ and $(\alpha - 1)(\alpha - 2) > 0$, a positive skew when $\alpha < \beta$ and a negative skew when $\alpha > \beta$.

The probability density function of the transformed order flow imbalance PDF_{Beta} is given by:

$$\text{PDF}_{\text{Beta}}(\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\text{B}(\alpha,\beta)} \quad \text{where} \quad \text{B}(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{8}$$

$$\mathbf{P}\left(\text{Imb}_k(m,d)\ s_k(m), \frac{Q_k(m)}{\text{ADV}_k(d)}\right) \Big\{ = \text{B}(\alpha_{\text{Imb}}, \beta_{\text{Imb}}) \tag{9}$$

Figure 8. Net order flow imbalance as a function of participation rate
Institutional trading data are obtained from ANcerno Ltd on the period ranging from January 1, 2010 to September 30, 2011. We sort meta-orders into 100 bins based on meta-order participation rate ($Q_k(m)/\mathrm{ADV}_k(d)$) and plots the average absolute net order flow imbalance for each of the bins



The dependence to order side and order participation rate should be taken into account at the level of the parameters of the Beta function (α, β).

$$\alpha_{\mathrm{Imb}} = c^{\alpha} + \rho_s^{\alpha} \cdot s_k(m) + \rho_p^{\alpha} \cdot s_k(m) \cdot \frac{Q_k(m)}{\mathrm{ADV}_k(d)} \tag{10}$$

$$\beta_{\mathrm{Imb}} = c^{\beta} + \rho_s^{\beta} \cdot s_k(m) + \rho_p^{\beta} \cdot s_k(m) \cdot \frac{Q_k(m)}{\mathrm{ADV}_k(d)} \tag{11}$$

The result of the Bayesian inference of the imbalance dependencies are summarized in Table 4. When $s_k(m) = 0$ and $Q_k(m)/\mathrm{ADV}_k(d) = 0$, the posterior distribution of net order flow imbalance is given by $\mathrm{B}(0.67, 0.68)$ which produces a U-shape. This means that when the asset manager has no signal or information on price movement, he can only assume the synchronization of institutional activity. Thus, the symmetric distribution with higher probability at the extreme values of the imbalance. But once he detains a signal, since the process leading to generate this signal is independent from the execution process, he can use his own meta-order as an observation to update his belief on the distribution of the expected market pressure. Note also that ρ_s^{β} and ρ_p^{β} are very low compared to ρ_s^{α} and ρ_p^{α}. This is not an issue, because what determines the strength of the skew for the Beta function is the difference $(\beta - \alpha)$ (see Appendix C for Beta function properties).

Table 4. Bayesian inference: Net order flow imbalance dependencies
The table summaries the posterior distribution of model parameters described in equations (10) and (11). $\mathbb{E}[X]$, std(X), Q2.5% and Q97.5% are respectively the mean, the standard deviation, the 2.5% and 97.5% quantile of parameters posterior distribution. The results are obtained from Hasting-Metropolis sampler with $N_{iter} = 10000$ iterations (PyMC3 implementation). The institutional investors trading data are obtained from ANcerno Ltd on the period ranging from January 1^{st}, 2010 to September 30^{th}, 2011.

	$\mathbb{E}[X]$	Std[X]	Q2.5%	Q97.5%
c_α	0.666	0.001	0.664	0.667
ρ_s^α	0.101	0.001	0.099	0.102
ρ_p^α	0.884	0.021	0.846	0.928
c^β	0.675	0.001	0.673	0.677
ρ_s^β	0.000	0.000	4.2e-08	0.000
ρ_p^β	0.001	0.001	1.4e-07	0.003

To better interpret the results of the table, we plot the posterior distribution of the net order flow imbalance, given two levels of participation rate (0.1% and 30%) for buy and sell trades. As expected, we observe that the side of the trade skews the distribution positively for a buy order and negatively for a sell order. The information of the meta-order participation rate intensifies the skew and increases the probability of having a full synchronization of investors executions $|\text{Imb}| = 1$. However, the shape of the distribution is not symmetrical between buy orders and sell orders. The skew of imbalance distribution is much stronger for sell orders. This means that when an investor is selling massively with large $\frac{Q}{\text{ADV}}$, he could expect a high selling pressure from the market due to investors synchronous inflows and outflow. Because institutional investors are natural buyers, implementing more long only strategies than short selling ones, a high selling pressure correspond to a "Rushing toward the exit door" situations. While on the opposite scenario, a buying order with high participation rate although informative on the market does not give as much evidence on market participants behaviour.

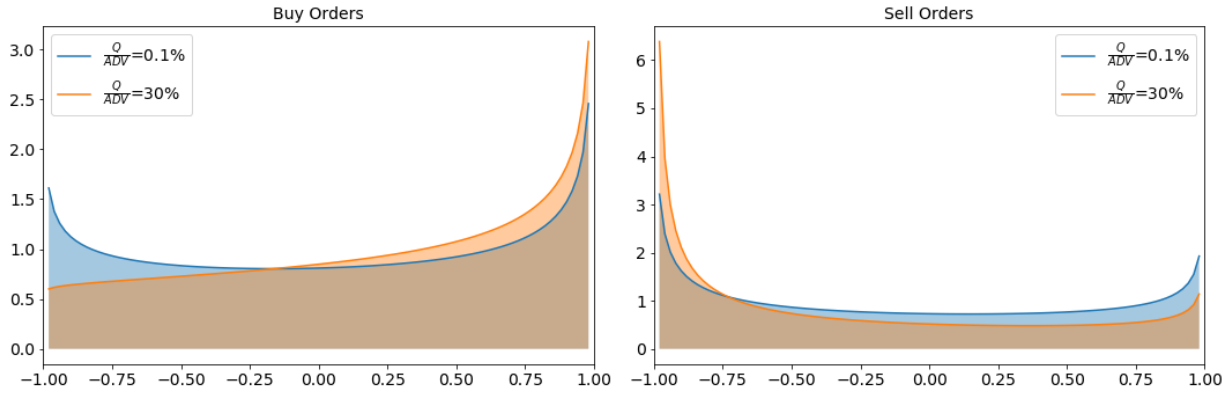
5.2.4 Implementation shortfall dependencies

Similarly, we model the implementation shortfall as a function of all the other nodes of the network. the data shows the historical distribution of implementation shortfall displays fat tails with pronounced non-Gaussian peaky shape. Thus, a double exponential (Laplace) probability density is a good prior of IS distribution. The probability density function (PDF) of Laplace is given by:

$$\text{IS}_k(m,d) \sim \textit{Laplace}(\mu,b), \quad \text{PDF}_{\text{Laplace}}(x,\mu,b) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)\Bigg(\tag{12}$$

The location parameter μ is given by equation (13). As in the linear regression, we condition the magnitude of transaction cost to stock bid-ask spread, the participation rate scaled by the volatility and

Figure 9. Inferred net order flow imbalance given the side and the size of the meta-order The figure shows the posterior distribution of net order flow imbalance given the meta-order characteristics. On the left panel we plots the distribution for buy orders with two levels of participation rate, blue line corresponds to small orders $Q_k(m)/\text{ADV}_k(d) = 0.1\%$ and orange line for large orders $Q_k(m)/\text{ADV}_k(d) = 30\%$. The right panel shows the result for sell orders with the same levels of participation rate



investors order flow imbalance signed by meta-order side and scaled by stock volatility. Nevertheless, we don't assume the square root function for meta-order size but a power law highlighted by the exponent γ that we estimate.

$$\mu = a_0 + a_\psi \psi_k(d) + a_{\sigma,\frac{Q}{\text{ADV}}} \exp\left\{\log(\sigma_k^{\text{GK}}(d)) + \gamma \log\left(\frac{Q_k(m)}{\text{ADV}_k(d)}\right)\right\}\left(+ a_{s,\text{Imb}}\ \sigma_k^{\text{GK}}(d)\ s_k(m)\ \text{Imb}_k(m,d)\right. \quad (13)$$

Estimation accuracy is function of market condition, speed and duration of the execution algorithm, the aggressiveness in seeking liquidity. This heteroscedasticity of the implementation shortfall is taken into account by making the standard deviation of Laplace distribution b depend on stock attributes (spread and volatility), meta-order characteristic (participation rate) and market condition (absolute imbalance) as follows:

$$b = \exp(b_{ln}), \quad b_{ln} = b_0 + b_\psi \log(\psi_k(d)) + b_\sigma \log(\sigma_k^{\text{GK}}(d)) + b_{\frac{Q}{\text{ADV}}} \log\left(\frac{Q_k(m)}{\text{ADV}_k(d)}\right)\left(+ b_{\text{Imb}} \log\left(|\text{Imb}_k(m,d)|\right)\right. \quad (14)$$

Table 5 summarizes the first two moments and the 2.5% and 97.5% quantiles of the posterior distribution of the parameters. First, we note that the exponent of the order size term is a bit lower than the square root $\hat{\gamma} = 0.41$, consistent with the previous finding of Bacry et al. (2015) that used a proprietary database of a broker execution in Europe. The literature usually document a power law with exponent between 0.4 and 0.5 (Gomes and Waelbroeck (2015) and Briere et al. (2019)). The parameters relative to the location term are consistent with the ones estimated with OLS regression. As expected, the

Table 5. Bayesian inference: Implementation shortfall dependencies
The table summaries the posterior distribution of model parameters described in equations (13) and (14). $\mathbb{E}[X]$, std(X), Q2.5% and Q97.5% are respectively the mean, the standard deviation, the 2.5% and 97.5% quantile of parameters posterior distribution.The results are obtained from Hasting-Metropolis sampler with $N_{iter} = 10000$ iterations (PyMC3 implementation). The institutional investors trading data are obtained from ANcerno Ltd on the period ranging from January 1^{st}, 2010 to September 30^{th}, 2011.

	$\mathbb{E}[X]$	Std[X]	Q2.5%	Q97.5%
a_0	0.00	0.00	-0.00	0.00
a_ψ	-0.60	0.37	-1.36	0.12
$a_{\sigma,\frac{Q}{\text{ADV}}}$	0.67	0.24	0.26	1.15
γ	0.41	0.11	0.20	0.62
$a_{S,\text{Imb}}$	0.20	0.02	0.17	0.24
b_0	0.06	0.17	-0.27	0.40
b_ψ	0.09	0.03	0.05	0.14
b_σ	0.78	0.03	0.71	0.85
$b_{\frac{Q}{\text{ADV}}}$	0.05	0.01	0.04	0.06
b_{Imb}	0.01	0.01	-0.01	0.03

intercept parameter is null, the coefficient of the order size and order flow imbalance terms are similar to those estimated by the OLS regression. Only the coefficient of bid-ask spread differs significantly. The parameters of the scale of Laplace distribution are small except the stock volatility coefficient. It proves that the main contributor to the heteroscedasticity is not the order size but stock volatility, consistent with the findings of Capponi and Cont (2019) suggesting that conditionally to the level of stock volatility and execution duration, the order size have small impact on transaction costs.

5.3 Forecasting implementation shortfall

We gather the different blocks of variable dependencies to constitute the Bayesian network of Figure 6. The parameters $(\mu, b, \alpha_{imb}, \beta_{imb})$ are estimated via Bayesian inference using Hasting-Metropolis algorithm. Once the network is calibrated on 70% of the meta-orders, we use it to infer the latent variable of net order flow imbalance given meta-order and stock characteristics and estimate orders implementation shortfall both in-sample (on the training set) and out-of-sample (on the testing set, the remaining 30% of the meta-orders not yet seen by the algorithm). Table 6 displays the results for both the linear regression and the Bayesian network predictions in- and out-of-sample. For the linear regression, we compare a model without order flow imbalance (equation (3), column 1) and one with order flow imbalance (equation (5), column 2). In this last model, the realized imbalance is fully observed in real time, which is never achievable in practice, but can serve as a benchmark case. We then show the results of three Bayesian networks: The first network (column 3) has never seen the in-

formation of the imbalance, neither during the training phase nor the prediction phase. In that sense, it is comparable to the first linear regression (OLS when imbalance is not available) in the first column of Table 6. The second network (column 4) was trained with the information of order flow imbalance. Once this information is captured by the conditional probabilities of network edges, the network is exploited without the use of the imbalance. In this regard, order flow imbalance is partially observed. The last network (column 5) has full information on the imbalance, both at the training and testing phases, and is similar to the second OLS regression displayed in column 2. Adding information on imbalance improves the forecasting accuracy, both for the OLS regression and the Bayesian network. In-sample, it increases the R^2 from 0.52% to 1.77% for the two models, and reduces the forecasting error (RMSE from 1.67% to 1.66% and MAPE from 98.74% to 98.48%). For the same set of information (imbalance observable or not), the Bayesian network has the same accuracy than the OLS on the training set. However, the absolute average error is much smaller for the Bayesian network (-0.43 bps vs -0.86 when imbalance is available, -1.41 bps vs -1.43 bps when it is not). In-sample, and when all explanatory variables are observable, the Bayesian network has only a limited advantage over simple linear regressions in terms of prediction accuracy. Out-of-sample, when the imbalance is not available (Panel B, columns 1 and 4), the Bayesian network is also similar to the linear regression (lower average error = 0.08 bps vs 0.16 bps, but similar RMSE= 1.39 % and R^2 =0.38%, and slightly higher MAPE= 99.3% vs 99.0%. But when imbalance is available (Panel B, columns 2 and 5), the Bayesian network has higher forecasting accuracy than the linear regression on all criteria (R^2= 1.20% vs 1.10%, average error= -0.43 bps vs -1.08 bps, RMSE= 1.388 % vs 1.389%, and MAPE= 99.41 % vs 99.57%).

The Bayesian Network is particularly valuable when a subset a variables are only partially observable. In this case, the network captures the conditional dependencies between the nodes, and fills the missing information with the most probable values of the latent variables. In our case, the realized imbalance is not used for the prediction, but the Bayesian network is trained on imbalance to infer its distribution given meta-orders characteristics. This gives a better forecast for the realized transaction cost, both in-sample and out-of-sample (for example higher R^2 =0.56% vs 0.52% in-sample, 0.50% vs 0.38% out-of-sample) than OLS or Bayesian networks that could not rely on this information.

Table 7 provides similar results to those in Table 6, but for ten deciles of orders size, and for Bayesian networks using partial or full information on imbalance. We split the training and testing sets in 10 bins with respect to the training set order size. The first bin contains small orders, lower than 0.01% of daily volume, while the last one contains very large orders, higher than 4.34% of daily volume. We assess the accuracy of the Bayesian network within the three configurations of information availability (order flow imbalance fully, partially or not available). Consistent with intuition, we find

Table 6. Performance of the Bayesian network compared to the standard OLS regression
Institutional trading data are obtained from ANcerno Ltd on the period ranging from January 1, 2010 to September 30, 2011. In-sample predictions are computed on 70% of the data such us the number of buy orders is equal to the number of sell orders. The remaining 30% serves for the out-of-sample prediction. RMSE and MAPE are respectively the Root Mean Squared Error and the Mean Absolute Percentage Error of the estimates

Imbalance	OLS Regression		Bayesian Network		
Available	No	Yes	No	Partial	Yes
Panel A: In-sample Estimation					
$\mathbb{E}[IS]$ (bps)	9.020	9.020	9.020	9.020	9.020
$\mathbb{E}[\hat{IS}]$ (bps)	7.590	8.161	7.606	8.617	8.588
$\mathbb{E}[\hat{IS} - IS]$ (bps)	-1.430	-0.859	-1.414	-0.403	-0.431
RMSE (%)	1.669	1.659	1.669	1.669	1.659
MAPE (%)	98.743	98.476	98.739	98.686	98.476
R^2 (%)	0.517	1.773	0.517	0.558	1.771
Panel B: Out-of-sample Estimation					
$\mathbb{E}[IS]$ (bps)	6.394	6.394	6.394	6.394	6.394
$\mathbb{E}[\hat{IS}]$ (bps)	6.557	5.317	6.482	8.030	5.960
$\mathbb{E}[\hat{IS} - IS]$ (bps)	0.163	-1.076	0.088	1.637	-0.434
RMSE (%)	1.394	1.389	1.394	1.393	1.388
MAPE (%)	99.022	99.570	99.301	99.340	99.410
R^2 (%)	0.377	1.104	0.378	0.502	1.204

that the inferred order flow imbalance distribution is more accurate when the investor holds a larger order. The posterior distribution of order flow imbalance given a small order is a symmetric U-shape function (Figure 9). At best it is slightly skewed, either positively or negatively, depending on the direction of the order. Thus, the larger the investors' trade, the more informative it is on the estimation of order flow imbalance, and as a consequence the more accurate is the forecast of resulting implementation shortfall. We observe that the R^2 increases steadily, whatever the configuration of information availability (partial or full), in-sample and out-of-sample, starting at the seventh bin. For example, the in-sample estimation of transaction costs when imbalance is partially available, goes from an R^2 of 0.18% for the seventh decile to 2.13% for the tenth decile, while smaller deciles of order size have relatively small R^2 (from -0.03% to 0.09% for the first 4 bins). Actually, for small order size, the market impact is very limited and disappears in market noise. Even if the Bayesian network is trained using information on order flow imbalance, it has no advantage when the investor uses its trades attributes, if he trades only small order sizes. Said differently, it is hard to make good prior predictions of the order flow and thus the transaction cost when executing small orders. But we see how information on the investors' own orders becomes more informative on the aggregate net order flow as the investors' own order size get larger. This is in line with the recent concentration of institutional investors executions on few dealing desks. Because the large dealing desk has a more accurate picture on investors'

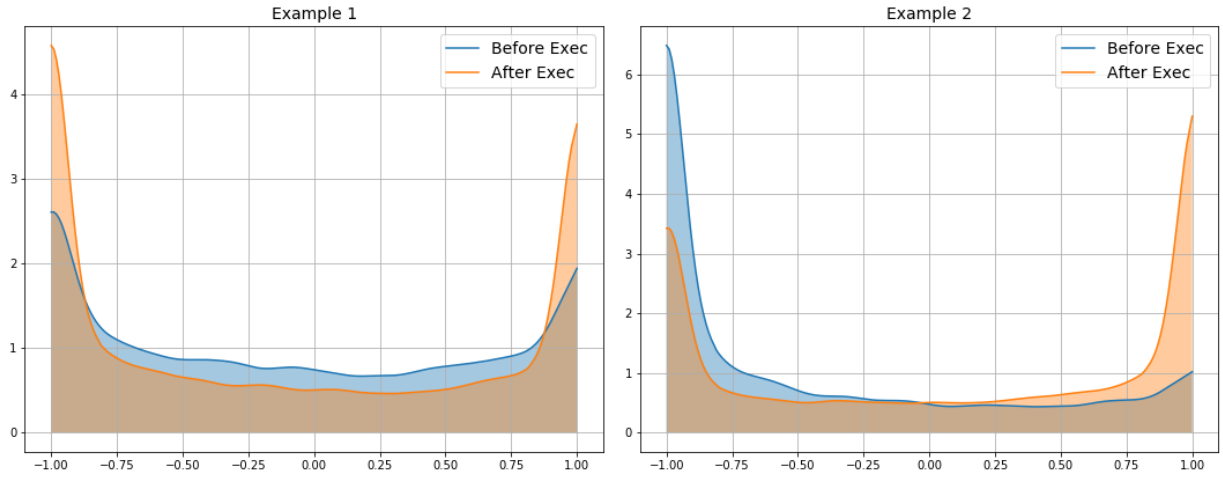
order flow imbalance of the day, it can assess the expected transaction cost more accurately and potentially design a better optimized executing scheme using this information. Note also that the RMSE does not drop, because higher order size bins have few orders with large implementation shortfall that increases the average transaction cost for the bin. This is visible in the difference between the mean and the median realized trading cost (30.64 bps vs 24.79 bps in-sample for the tenth bin and 1.89 bps vs 1.67 bps for the first bin).The MAPE on the other hand, not suffering from this bias, gets smaller as the order size increase.

Table 7. Performance of the Bayesian network given order size
Institutional trading data are obtained from ANcerno Ltd on the period ranging from January 1, 2010 to September 30, 2011. In sample predictions are computed on 70% of the data such us the number of buy orders is equal to the number of sell orders. The remaining 30% serves for the out-of sample prediction. The sample are split in 10 bins w.r.t training set orders size. Q50 is the 50% quantile of implementation shortfall realized distribution. RMSE and MAPE are respectively the Root Mean Squared Error and the Mean Absolute Percentage Error of the estimates

Bins	1	2	3	4	5	6	7	8	9	10
$\mathbb{E}[\frac{Q}{ADV}]$ (%)	0.01	0.02	0.03	0.04	0.06	0.09	0.15	0.28	0.64	4.34
Panel A: In-sample Bayesian Estimation										
					Effective Trading Costs					
$\mathbb{E}$[IS] (bps)	1.89	2.87	2.61	4.85	6.88	7.82	7.01	9.25	16.50	30.64
Q50 (bps)	1.67	2.67	2.34	2.89	5.09	6.11	5.99	8.06	11.54	24.79
					Imbalance Partially Available					
$\mathbb{E}[\hat{\text{IS}}]$ (bps)	4.26	4.52	4.82	5.25	5.78	6.51	7.58	9.32	12.59	25.62
RMSE (%)	1.41	1.44	1.51	1.52	1.56	1.61	1.68	1.74	1.92	2.17
MAPE (%)	100.19	99.78	100.11	99.69	98.92	98.59	98.46	97.86	96.91	93.60
R^2 (%)	-0.03	0.04	-0.00	0.09	0.17	0.24	0.18	0.27	0.77	2.13
					Imbalance Available					
$\mathbb{E}[\hat{\text{IS}}]$ (bps)	3.72	4.09	4.36	4.86	5.54	6.36	7.63	9.54	13.07	26.80
RMSE (%)	1.39	1.43	1.50	1.51	1.55	1.60	1.67	1.73	1.90	2.17
MAPE (%)	98.76	98.72	98.76	99.02	98.66	98.70	98.59	98.54	98.06	96.89
R^2 (%)	1.50	1.69	1.43	1.37	1.55	1.35	1.33	1.35	2.09	2.89
Panel B: Out-of-sample Bayesian Estimation										
					Effective Trading Costs					
$\mathbb{E}$[IS] (bps)	-1.44	-0.14	0.75	2.70	3.80	4.83	8.09	8.81	12.71	28.09
Q50 (bps)	0.00	0.00	0.00	1.65	1.28	3.62	5.49	6.73	8.01	19.09
					Imbalance Partially Available					
$\mathbb{E}[\hat{\text{IS}}]$ (bps)	4.53	4.77	5.05	5.39	5.87	6.50	7.46	8.98	11.85	22.73
RMSE (%)	1.22	1.24	1.27	1.37	1.32	1.35	1.40	1.45	1.61	1.71
MAPE (%)	101.28	101.25	101.23	100.99	101.33	98.96	98.23	97.59	97.22	93.65
R^2 (%)	-0.17	-0.09	-0.05	0.06	0.13	0.17	0.39	0.41	0.72	2.39
					Imbalance Available					
$\mathbb{E}[\hat{\text{IS}}]$ (bps)	2.19	2.28	2.69	2.98	3.67	4.33	5.57	6.97	9.91	22.13
RMSE (%)	1.21	1.23	1.26	1.36	1.32	1.35	1.39	1.45	1.60	1.72
MAPE (%)	99.47	99.69	99.51	99.78	99.73	99.37	99.62	99.38	99.23	97.95
R^2 (%)	1.18	1.25	1.15	0.85	0.92	1.18	0.81	0.96	1.34	2.08

5.4 Inference of investors order flow imbalance given post-trade cost and market conditions

Figure 10. Bayesian Inference of Net Order Flow Imbalance
The figure plots in two panels the posterior distribution of net order flow imbalance given two example of market conditions and order characteristics. Each time, the blue curve plots the inferred distribution when only meta-orders attributes are considered and the orange line is the updated distribution once the resulting transaction cost is observed.



Investors net order imbalance is a latent variable, thus not observable by the asset manager before executing his trade. His best prediction of market pressure is the inferred imbalance, after observing his own trading decision. However, his decision although usually in line with investors' trading because of the crowd effect, can depart from what is actually traded by his peers. One of the interests of our Bayesian network model is that it can be used to recover the aggregate order flow imbalance prevailing during the investor's order execution, knowing his transaction costs. After receiving his *Transaction Cost Analysis* report, the investor could update his belief on investors imbalance during his execution using the calibrated Bayesian Network. We explore two cases as an example. First case: the investor sells a stock $s_k(m)$= -1; with a small participation rate $Q_k(m)/\text{ADV}_k(d)$ = 0.01%. His order is not very informative on market pressure since his trade is small, so his best estimate using the Bayesian network is a U-shape slightly skewed towards negative values of mean -0.10 (blue distribution of the left panel of Figure 10). Unexpectedly the resulting trading cost is huge $\text{IS}_k(m,d)$ = 3.02% because the imbalance is very large and negative $\text{Imb}_k(m,d)$= -0.94. The investor could update his belief on the true distribution prevailing during his execution. The posterior distribution after incorporating the realized trading cost gives higher probability to values at -1 (Orange line of the left panel). The average posterior imbalance distribution is -0.17 (Table 8). Second case: the investor takes the decision

to sell massively a stock, $Q_k(m)/\text{ADV}_k(d)$ = 31.8%. This is usually happening during market panic where other investors sell massively as well. Therefore, his prior distribution is highly skewed to the left ($\mathbb{E}[\text{Imb}] = -0.40$, blue distribution of the right panel, Figure 10). While the investor expects a high transaction cost, he got lucky to be against the aggregate order flow ($\text{Imb}_k(m,d)$= 0.94) and benefited from a price improvement of 2.6%. The posterior imbalance distribution after incorporating this information is displayed on the right panel of Figure 10 (orange line) with a 0.05 average.

Table 8. Bayesian inference of net order flow imbalance
The table summaries the first 2 moment and the 2.5% and 97.5% quantiles of net order flow imbalance inferred distribution for two scenarios before and after order implementation shortfall become available

	$\mathbb{E}[\hat{\text{Imb}}]$	Std[$\hat{\text{Imb}}$]	Q2.5%	Q97.5%
		Inferred Imbalance: Example 1		
Before exec	-0.097	0.333	-1.0	0.953
After exec	-0.175	0.381	-1.0	0.999
		Inferred Imbalance: Example 2		
Before exec	-0.400	0.328	-1.0	0.909
After exec	0.050	0.386	-1.0	1.000

6 Conclusion

In this paper, we use a Bayesian network to model transaction costs on US equity markets using ANcerno data, a large database of asset managers' instructions. Our main motivation is to make use of a variable of paramount importance for transaction costs, the *Net Order Flow Imbalance*. This variable is not observed by all market participants. Brokers and market makers have access to the imbalance of their clients' flows, while dealing desks of asset managers do only observe their own instructions. Nevertheless, brokers, custodians and even exchanges started recently reselling aggregate information on their clients flows with a delay. Bayesian networks open new perspectives to model transaction costs using *latent variables*, i.e. variables that are not always known when the model has to be used, but can be partially observed during the learning process. They enable to design a model linking observed and latent variables, based on conditional probabilities. The partially observable data can then be used to train the model.

Bayesian networks are able to estimate not only expected values, but the whole probability distribution of a given variable. They are thus able to estimate the variance of the residuals of their estimation. Because of the heteroskedasticity of the error term, market impact models and transaction costs estimates have traditionally a very small R^2. A common belief among practitionners is that

the effect of small mechanical price pressure is disappearing in the ”market noise” (i.e. innovation on prices). We confirm this intuition in our model, by allowing the accuracy of trading costs forecasts to depend on market conditions and the investors' order characteristics. We find that the main variable explaining the variance of the residuals is the stock volatility, with a coefficient of 0.78.

Last but not least, we show several advantages of Bayesian networks for transaction costs forecasting. First, even when the latent variables (in our case, the imbalance of institutional orders at the start of the day) cannot be observed, the estimation relies on its pre-captured relationships with other observable variables (like the size and side of the investor's order to be executed). This allows the model to provide a better prediction than standard (for example OLS) models. Second, we show that the estimates get more accurate with the size of the meta-order the investor has to execute, because the larger the meta-order, the better the estimation of the order flow imbalance. This gives an informational advantage for large dealing desks in charge of executing the orders of numerous or large investors as they have a better picture of the aggregate imbalance. This finding is consistent with the current evolution of market practice. Small asset managers increasingly use the services of large dealing desks to benefit from this information, leading to the recent concentration of institutional investors orders on a few dealing desks. Finally, these models can use Bayesian inference to deduce the expected distribution of the latent variable. We show how it is feasible to ask the Bayesian network the expected distribution of large orders of other investors, either at the start or at the end of the day, once the resulting trading costs are observed.

Bayesian networks are very promising models to account for partial information. They could prove particularly valuable for “alternative datasets”, like airlines activity, web traffic, or financial flows, than often provide very detailed information on a small subset of transactions. They are difficult to use in standard models, that do not accept missing values. Bayesian networks structurally model relationship between missing and known variables. They could naturally fill this gap.

References

Almgren, R. and Chriss, N. (2001). Optimal execution of portfolio transactions. *Journal of Risk*, 3:5–40.

Amari, S.-i. (1993). Backpropagation and stochastic gradient descent method. *Neurocomputing*, 5(4-5):185–196.

Anand, A., Irvine, P., Puckett, A., and Venkataraman, K. (2011). Performance of institutional trading desks: An analysis of persistence in trading costs. *The Review of Financial Studies*, 25(2):557–598.

Anand, A., Irvine, P., Puckett, A., and Venkataraman, K. (2013). Institutional trading and stock resiliency: Evidence from the 2007–2009 financial crisis. *Journal of financial Economics*, 108(3):773–797.

Angel, J. J., Harris, L. E., and Spatt, C. S. (2015). Equity trading in the 21st century: An update. *The Quarterly Journal of Finance*, 5(01):1550002.

Bacry, E., Iuga, A., Lasnier, M., and Lehalle, C.-A. (2015). Market impacts and the life cycle of investors orders. *Market Microstructure and Liquidity*, 1(02):1550009.

Bew, D., Harvey, C. R., Ledford, A., Radnor, S., and Sinclair, A. (2018). Modelling analysts' recommendations via bayesian machine learning. *Available at SSRN*.

Bouchard, B., Dang, N.-M., and Lehalle, C.-A. (2011). Optimal control of trading algorithms: a general impulse control approach. *SIAM Journal on financial mathematics*, 2(1):404–438.

Bouchaud, J-P, T. B., Lemperiere, Y., Deremble, C., De Lataillade, J., and Kockelkoren, J. a. (2011). Anomalous price impact and the critical nature of liquidity in financial markets. *Physical Review X*, 1(2):021006.

Briere, M., Lehalle, C.-A., Nefedova, T., and Raboun, A. (2019). Stock market liquidity and the trading costs of asset pricing anomalies. *Available at SSRN 3380239*.

Bucci, F., Mastromatteo, I., Eisler, Z., Lillo, F., Bouchaud, J.-P., and Lehalle, C.-A. (2018). Co-impact: Crowding effects in institutional trading activity. *arXiv preprint arXiv:1804.09565*.

Capponi, F. and Cont, R. (2019). Trade duration, volatility and market impact. *Available at SSRN 3351736 (April 17, 2019)*.

Cetin, U. and Rogers, L. (2007). Modeling liquidity effects in discrete time. *Mathematical Finance*, 17(1):15–29.

Collins, B. M. and Fabozzi, F. J. (1991). A methodology for measuring transaction costs. *Financial Analysts Journal*, 47(2):27–36.

Dayri, K. and Rosenbaum, M. (2015). Large tick assets: implicit spread and optimal tick size. *Market Microstructure and Liquidity*, 1(01):1550003.

Eisele, A., Nefedova, T., Parise, G., and Peijnenburg, K. (2017). Trading out of sight: An analysis of cross-trading in mutual fund families.

Frazzini, A., Israel, R., and Moskowitz, T. J. (2012). Trading costs of asset pricing anomalies. *Fama-Miller working paper*, pages 14–05.

Garman, M. B. and Klass, M. J. (1980). On the estimation of security price volatilities from historical data. *Journal of business*, pages 67–78.

Gomes, C. and Waelbroeck, H. (2015). Is market impact a measure of the information value of trades? market response to liquidity vs. informed metaorders. *Quantitative Finance*, 15(5):773–793.

Greenwood, R. and Thesmar, D. (2011). Stock price fragility. *Journal of Financial Economics*, 102(3):471–490.

Haldane, A. G. et al. (2014). The age of asset management? *speech at the London Business School*, 4(4).

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12(1):69–82.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.

Jordan, M. I. (1998). *Learning in graphical models*, volume 89. Springer Science & Business Media.

Koch, A., Ruenzi, S., and Starks, L. (2016). Commonality in liquidity: a demand-side explanation. *The Review of Financial Studies*, 29(8):1943–1974.

Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica: Journal of the Econometric Society*, pages 1315–1335.

Laruelle, S. and Lehalle, C.-a. (2018). *Market microstructure in practice*. World Scientific.

Lauritzen, S. L. (2003). Some modern applications of graphical models. *Oxford Statistical Science Series*, pages 13–32.

Moro, E., Vicente, J., Moyano, L. G., Gerig, A., Farmer, J. D., Vaglica, G., Lillo, F., and Mantegna, R. N. (2009). Market impact and trading profile of large trading orders in stock markets.

Novy-Marx, R. and Velikov, M. (2015). A taxonomy of anomalies and their trading costs. *The Review of Financial Studies*, 29(1):104–147.

Pagano, M. S. (2008). Which factors influence trading costs in global equity markets? *The Journal of Trading*, 4(1):7–15.

Pearl, J. (1986). Fusion, propagation, and structuring in belief networks. *Artificial intelligence*, 29(3):241–288.

Perold, A. F. (1988). The implementation shortfall: Paper versus reality. *Journal of Portfolio Management*, 14(3):4.

Puckett, A. and Yan, X. (2011). The interim trading skills of institutional investors. *The Journal of Finance*, 66(2):601–633.

Robert, E., Robert, F., and Jeffrey, R. (2012). Measuring and modeling execution cost and risk. *The Journal of Portfolio Management*, 38(2):14–28.

Salvatier, J., Fonnesbeck, C., et al. (2016). Pymc3: Python probabilistic programming framework. *Astrophysics Source Code Library*.

Skaanning, C., Jensen, F. V., and Kjærulff, U. (2000). Printer troubleshooting using bayesian networks. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 367–380. Springer.

Torre, N. and Ferrari, M. (1999). The market impact model tm. barra research insights. 1999 barra.

Vapnik, V. N. and Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pages 11–30. Springer.

Vila, J.-P., Wagner, V., and Neveu, P. (2000). Bayesian nonlinear model selection and neural networks: a conjugate prior approach. *IEEE Transactions on neural networks*, 11(2):265–278.

Wyart, M., Bouchaud, J.-P., Kockelkoren, J., Potters, M., and Vettorazzo, M. (2008). Relation between bid–ask spread, impact and volatility in order-driven markets. *Quantitative Finance*, 8(1):41–57.

Zuo, Y. and Kita, E. (2012). Stock price forecast using bayesian network. *Expert Systems with Applications*, 39(8):6729–6737.

Appendix A Garman Klass volatility definition

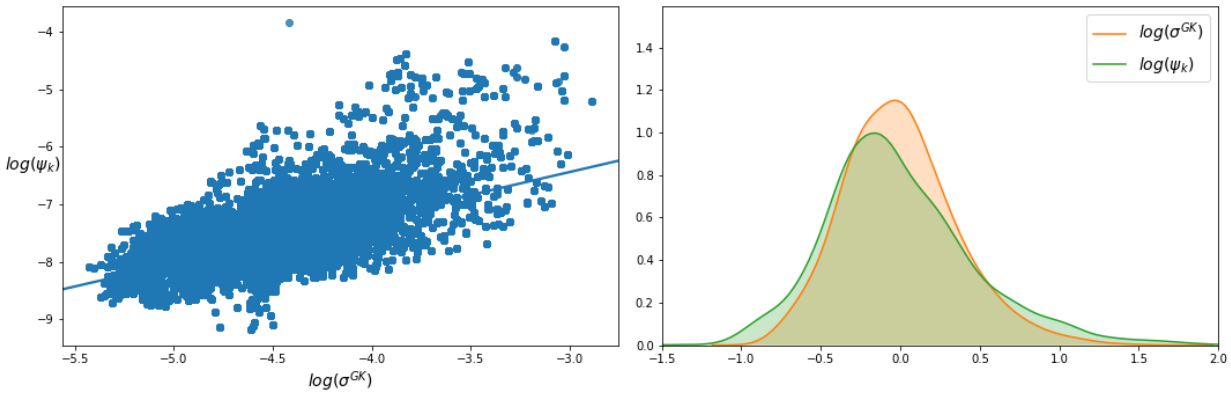
Garman-Klass estimate of the volatility uses the open, high, low and close prices of the day. This estimate is robust and very close in practice to more sophisticated ones. The formula is given by:

$$\sigma_k^{\mathrm{GK}}(d) = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(\frac{1}{2}\log\left(\frac{H_{d-t}^k}{L_{d-t}^k}\right)\right)^2 - (2\log(2)-1)\log\left(\frac{C_{d-t}^k}{O_{d-t}^k}\right)^2} \tag{15}$$

where the indexation k refers to the stock. d to the calculation day. N is the length of the rolling window in day. In our case 252 trading days. O_t^k, H_t^k, L_t^k, C_t^k are respectively the open, high, low, close prices of stock k at day t

Appendix B Bid-Ask spread and volatility distribution dependencies

Figure 11. Bid-Ask spread and volatility distribution dependencies



The left panel of figure 11 shows the scatter plot of the log bid-ask spread and log volatility of S&P 500 components of 2010 and 2011. It proves that the variables are related. The right panel displays the centered distributions $(X - \overline{X})$ of log bid-ask spread and log volatility.

Appendix C Beta distribution properties

The probability density function of the Beta distribution PDF_{Beta} is given by:

$$\forall x \in [0,1] \quad \mathrm{PDF}_{\mathrm{Beta}}(\alpha,\beta,x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)} \quad \textit{where} \quad \mathrm{B}(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{16}$$

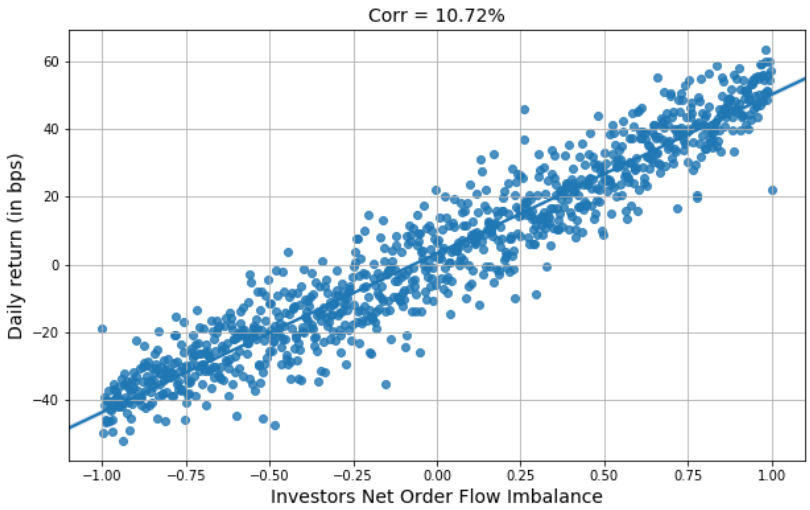
The first 3 moments of the distribution are as following:

$$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta} \qquad \text{Var}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \qquad \text{Skew}[X] = \frac{2(\beta-\alpha)\sqrt{\alpha+\beta+1}}{(\alpha+\beta+2)\sqrt{\alpha\beta}}$$

Note that the skew of the distribution is proportional to $(\beta-\alpha)$. So when $\alpha << \beta$ the probability density function is significantly skewed toward values at 0 and in the opposite case $\alpha << \beta$ the probability density function is skewed toward values at 1. The particular case where $\alpha = \beta$ the distribution is symmetric around the mean $\mathbb{E}[X] = \frac{1}{2}$. $(\text{PDF}_{\text{Beta}}(\alpha,\beta,\ \frac{1}{2}+x) = \text{PDF}_{\text{Beta}}(\alpha,\beta,\ \frac{1}{2}-x))$ and the skew is null. if in addition the condition $(\alpha-1)(\alpha-2) > 0$ is fulfilled the distribution has a U-shape. Otherwise the Beta distribution produces a concave function.

Appendix D Net order flow imbalance properties

Figure 12. Net order flow imbalance, daily returns correlation



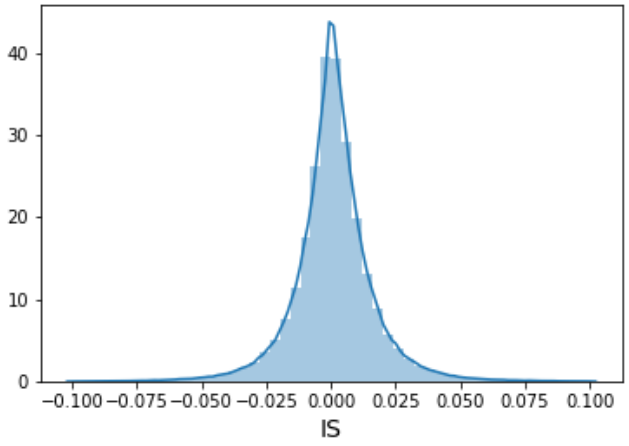
Net order flow imbalance has a strong predictive power of daily returns. The cross sectional average correlation for S&P 500 index components on our period of study is significantly positive up to 10.72% (Figure 12). Furthermore, investors trading imbalance prevail through time. Table 9 shows that the daily imbalance auto-correlation decays slowly from 12.03% for the first lag to 7.44% after 5 days. Since order flow imbalance is only available with a delay, the long memory of the imbalance is appreciated.

Table 9. Net Order Flow Imbalance auto-correlation

	Imb_{t-1}	Imb_{t-2}	Imb_{t-3}	Imb_{t-4}	Imb_{t-5}
Imb_t	12.03	9.11	8.37	7.69	7.44

Appendix E Implementation shortfall distribution

Figure 13. Implementation shortfall marginal distribution



The implementation shortfall estimated on ANcerno meta-orders on S&P 500 components of 2010 and 2011, displays a non-normal distribution centred at 0, with standard deviation equal to 0.64, a positive skew of 0.34 and highly significant excess kurtosis of 23.46. These moments are more comparable to a double exponential distribution.

Appendix F Hasting-Metropolis algorithm

Hasting-Metropolisis one of pioneer Markov Chain Monte Carlo algorithm developed in early 90s to sample from an unknown distribution. Given a function f proportional to the desired probability distribution $P(x)$ (a.k.a the target distribution) and a proposal distribution $q() = q(.|x)$ easy to simulate, the algorithm construct a series of variable $(x_1, x_2, ..., x_n)$ such as given x_n

1. Generate $y_n \sim q(y|x_n)$,
2. Generate $u \sim \mathscr{U}[0,1]$ a uniform distribution
3. Compute the acceptance rate $\alpha = min\left\{\frac{f(y_n)q(x_n|y_n)}{f(x_n)q(y_n|x_n)}, 1\right\}$(
4. Accept the new candidate y_n with probability α if $u \leq \alpha$ Otherwise reject.

$$X_n = \begin{cases} y_n, & \text{if } u \leq \alpha \\ x_n, & \text{otherwise} \end{cases} \tag{17}$$