УДК 336.76, 519.246

Informed Trading in Cryptocurrency Markets

Grigorii Kuzmin[1], **Alexei Boulatov**[2]

[1] National Research University Higher School of Economics,
11, Pokrovsky Bulvar, Moscow, 109028, Russian Federation.
E-mail: gikuzmin@hse.ru

[2] National Research University Higher School of Economics,
11, Pokrovsky Bulvar, Moscow, 109028, Russian Federation.
E-mail: aboulatov@hse.ru

This paper empirically estimates information asymmetry in cryptocurrency markets using the Probability of Informed Trading (PIN) and Adjusted PIN metrics. These markets, characterized by a high proportion of algorithmic trading and large volumes of high-frequency data, present a promising environment for analyzing informed trading behavior. We introduce a modified estimation procedure for Adjusted PIN, addressing floating-point errors and issues with local extrema, thereby improving its accuracy compared to the traditional naive approaches commonly used in the literature. Additionally, we propose an alternative trade aggregation method at higher frequencies than the conventional daily aggregation to enhance the efficiency of both PIN and Adjusted PIN models. Through analysis of both simulated and real data, we demonstrate that aggregating total buy and sell trades on a daily basis results in less meaningful estimates due to noisy input data, making it difficult to capture informed trader activity. The true optimal trade aggregation frequency is still to be further investigated, as increasing the frequency introduces heterogeneity in order imbalances, and the specific frequencies at which informed traders operate are still unknown. Finally, several empirical studies are conducted to evaluate the behavior of the metrics, revealing that illiquid cryptocurrencies exhibit relatively higher estimated probabilities of informed trading. This finding aligns with similar results observed in equity markets.

***Key words*:** PIN; Adjusted PIN; probability of informed trading; cryptocurrency markets; information asymmetry.

JEL Classification: G12, G14, D53, C13.

Grigorii Kuzmin – Phd Student, Doctoral School of Economics HSE.
Alexei Boulatov – Tenured Professor.

The article was received: 14.05.2024/The article is accepted for publication: 07.11.2024.

DOI: 10.17323/1813-8691-2024-28-4-615-646

For citation: Kuzmin G., Boulatov A. Informed Trading in Cryptocurrency Markets. *HSE Economic Journal.* 2024; 28(4): 615–646.

1. Introduction

Cryptocurrency markets gave additional impetus to the development of trading as well as market making in high frequency environments. Alongside, cryptocurrencies have emerged as a favored asset class for algorithmic trading and various quantitative strategies. These sophisticated trading algorithms often integrate exclusive information and market forecasts beyond the reach of ordinary investors, and if distinguished in the order flow, can reveal the information asymmetry. This in turn makes cryptocurrency markets topical to reconsider a specific class of models, identifying informed trading. A popular approach is PIN measure, proposed by Easley, Kiefer, O'Hara, and Paperman (1996), which uses Glosten and Milgrom (1985) setting to estimate the probability of informed trading. There was a debate whether PIN is a measure of liquidity rather than of informed trading activity as well as its critics that it mathematically fails to explain empirical positive correlation between buy and sell trades. To account for PIN model's limitations, its nested version was introduced Adjusted PIN (APIN) [Duarte, Young, 2007], which is more computationally extensive, but promised to be more accurate. Both models proved to be widely used in equity research and corporate finance as some proxies for insider trading before public events such as mergers and acquisitions and others. The emergence of cryptocurrency trading has revitalized the relevance of these metrics, expanding their scope of interpretation. However, up to date, the literature on their application in this context remains relatively sparse.

The ability to accurately identify informed trading is quite a topical issue which could be potentially applied for trading purposes and market regulation. The latter, combined with the fact, that cryptocurrency markets are still regulated to much lower extent than stock markets, makes such analysis even more crucial. Informed trading includes both insider trading (unlawful acquisition of privileged information by insiders), and informed trading, based on information asymmetry arising from the capacity to access new information ahead of the majority of other market participants. In other words, the informed trading involves one's ability to outpace the majority of the market in finding information. This is crucial for cryptocurrency markets with a wide range of algorithmic funds, trading high volumes, which appear to have much higher computational power and are superior and faster in analyzing numerous market variables than average traders.

Thus, this research extends the application of the informed trading metrics and introduces some novel modifications to their estimation, testing them on both real and simulated data, which in turn leads to the following objectives and results. Firstly, we compare the empirical evidence of informed trading metrics on the cryptocurrency market to the existing ones of the stock market. We find that as in case of stocks less liquid cryptocurrencies tend to have higher probability of informed trading. Secondly, we introduce a modified approach to estimation of APIN model, which substantially improves its theoretical accuracy. Finally, we propose a new technique of trades aggregation for estimation of daily PIN and APIN and try to identify the most efficient frequency to combine trades. Using simulated data we determine that for frequencies higher than 15 minutes the improvement in accuracy is not proportional to increasing computational diffi-

culty. Empirical series of both PIN and APIN prove that using higher frequency aggregation that one day provides more meaningful output, however, the behavior of the metrics at higher frequencies can vary due to potentially arising heterogeneity in the order flow and inability to determine at which exact frequency sophisticated algorithms are more likely to trade. We sidestep discussions regarding the PIN's theoretical drawbacks and the ongoing theoretical discourse on whether solely analyzing order imbalance suffices is enough to detect informed trading, as well as the usage of alternative methodologies.

The remainder is organised in the following way: Section 2 is devoted to literature review, Section 3 describes the models used, Section 4 discusses the estimation novelty of APIN model, Section 5 describes trades bucketing and optimal aggregation frequency to enhance efficiency of metrics, Section 6 presents data overview, while Sections 7 and 8 provide main results and conclusion.

2. Literature Review

Constructing a good metric of informed trading has become a topical question in financial microstructure literature. Hasbrouk (2007) claims it to be one of the integral aims to identify information asymmetry in financial markets. Information asymmetry occurs when information is not delivered to all investors identically in terms of time and costs. This allows for some agents, who are relatively more informed, to benefit from market operations, also affecting other investors. Informed trading has long been pervasive in equity markets, often observed surrounding numerous M&A and takeover transactions [Brennan, Huh, Subrahmanyam, 2017]. Furthermore, its presence is intricately linked to the concept of price stability.

This problem becomes even more topical in the context of the cryptocurrency markets. In traditional equity markets there is a range of tools, smoothing the presence of private information such as regulatory frameworks, under which companies should disclose and update information publicly, to different trading rules, that guarantee punishment for unethical operations. As for crypto markets, they still lack regulation and in the absence of disclosure systems ordinary investors are left with very limited sources of information, which enhances information asymmetry. The latter is further expanded since cryptocurrency systems, blockchain technology and other related issues are still too complicated and not quite transparent to typical users. European Central Bank (2012) argues on the complexity of cryptocurrency frameworks, which combined with availability and accessibility of the software, might lead to high risks. Agents, who do not understand how such systems work, will still download the applications and will conduct operations.

This paper extends the application of most widely used structural models to the cryptocurrency markets with some new modifications, enhancing their efficiency. Current literature counts a limited number of research papers, evaluating information asymmetry and even less via structural models. Feng et al. (2017) use their own version of volume imbalance indicator to identify informed trading, associated with bitcoin (BTC). They estimate the metric around some important public announcements and find evidence of informed trading activity prior to events, associated with both negative and positive news. Felez-Vinas et al. (2022) finds informed trading before 10–25% of cryptocurrency exchange listings, while Westland (2021) identifies trade informativeness as a princial driving force of liquidity in the BTC markets. Regarding structural approaches, Park and Chai (2020) applied PIN metric on several cryptocurrency tickers and concluded that cryptocurrency markets exhibit similar levels of information asymmetry as traditionnal equity markets.

There exist several structural approaches to measuring information asymmetry. Easley et al. (1996) proposed one of the most popular metrics – Probability of Informed Trading (PIN) which captures the posterior probability of informed trading. PIN model is based on the Glosten and Milgrom (1985) framework and uses the observed order imbalance to identify information asymmetry. By order imbalance in this case we understand the difference between buyer and seller initiated trades. Easley et al. (1996) observe that less liquid stocks have larger bid-ask spreads than more liquid ones, which among other reasons, can be also explained by private information, leading to higher risks and wider spreads. This is confirmed by the empirical results of PIN model which showed higher values, on average, for less frequently traded stocks. In this paper we conduct similar analysis, but for cryptocurrencies.

However, original version of PIN was found to have some disadvantages, both technical and empirical. As for technical ones, it performed poorly on the data with large trading volumes, leading to floating point error (FPE). The likelihood function in the model contains factorials (trades are assumed to have Poisson distribution) which cannot be computed for some large numbers. Another problem is related to computer optimisation procedure in general. Modern optimization algorithms suffer from the problem of obtaining local maximum instead of the global one, thus, applying original PIN model without any further modification could lead to biased results. These two inefficiencies were tackled by Lin and Ke (2011) and Yan and Zhang (2012) that introduced modified likelihood and initial parameters algorithm, based on the method of moments conditions, implied by the PIN model. The usage of these two sub-models in PIN estimation substantially increases the computational complexity, but, on the other hand, significantly improves the accuracy. Considering empirical disadvantages, Collin-Dufresne and Fos (2012) and Aktas et al. (2007) show that PIN provides contradictory low values when the detected presence of insider trading was high. This creates an issue similar to joint hypothesis problem since we cannot distinguish between the model itself fails or rather order imbalance itself is not sufficient to reveal informed trading.

Still, the most important flaw of the PIN model was revealed by Duarte and Young (2009). They showed that empirically buyer and seller initiated trades have positive correlation in the stock market, which PIN fails to capture, since this correlation can only be negative theoretically. Thus, they introduced a modified version Adjusted PIN (APIN) which allows for positive correlation. By comparing the results of these two models, they introduce the hypothesis that original PIN model might be a liquidity measure rather than a metric of informed trading. APIN, being the nested model of PIN, inherits both FPE and local maximum problems. We propose the solution in the fashion of Lin and Ke (2011) and Yan and Zhang (2012) for PIN. There is a similar attempt to tackle these issues, outlined in the preprint paper of Ersan and Grachem (2023), who introduced the same modification to the likelihood function, but different version of initial parameters algorithm.

Among other models, not considered in this paper, the most prominent were VPIN [Easley, López de Prado, O'Hara, 2012] and OWR [Odders-White, Ready, 2008]. The first was proved to converge to PIN and, thus, was considered as its approximation due to simple estimation procedure that does not involve estimation of intermediate parameters. As for OWR, in contrast to PIN and APIN, it is based on Kyle (1985) and takes as additional inputs intraday and overnight returns besides order imbalance (only input for PIN and APIN). However, it fails to estimate directly the proportion of informed trading, forecasting only the value the probability of the private signal in the market. This, on the one hand, limits its comparison to PIN and APIN, while still enables to

check the hypothesis whether order imbalance alone is enough to identify informed trading. Unfortunately, OWR is hardly applicable on the cryptocurrency markets since crypto exchanges accepts trades 24 hours and overnight returns, implied by the model, lose their tractability.

3. Theoretical Review (PIN and APIN)

3.1. PIN

This section describes the underlyings of the PIN model applied in further sections. We moved away from the traditional set-up (EKOP) [Easley et al., 1996] to its updated version (EHO) [Easley et al., 2002].

Model outline. There are two types of traders: informed traders who obtain private information and use it for speculative trading and noise traders, trading for liquidity or other exogenous purposes. Moreover, there is a market maker, setting bids and ask quotes according to buy and sell orders flow and estimating the probability of receiving orders from informed traders. At the beginning of each day there is an independent private information event which occurs with probability **α**. This event can be bad (negative signal) with probability δ and good (positive signal) with probability $1-\delta$. Defining the arrival rates of uninformed buyers and sellers and informed traders are distributed by Poisson process as ε_B, ε_S and μ respectively, on the day with positive signal the total *buy order flow* is $\mu+\varepsilon_B$ and the total *sell order flow* is ε_S. On the day with negative signal everything is vice versa (see the Fig. 1).

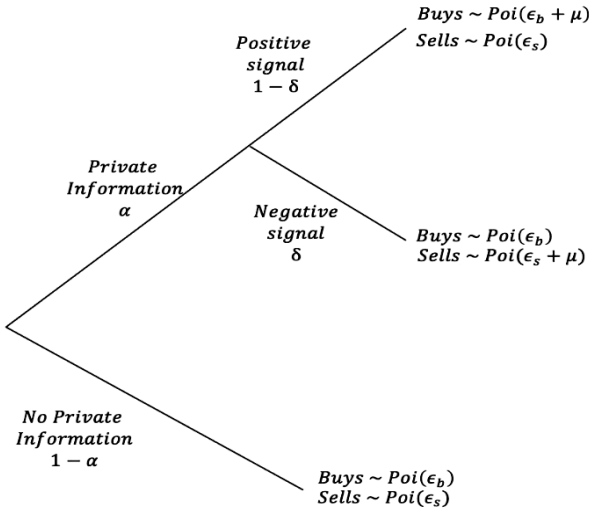


Fig. 1. Trading process tree[1]

[1] This diagram represents the trading mechanics, described on the previous page, where $\alpha, \delta, \varepsilon_B, \varepsilon_S$ and μ stay for probability of private information event, probability of negative signal, rate of uninformed buy and sell operations rates and informed trade arrival respectively.

Using homogenous Poisson processes, the following Likelihood function is derived:

$$L(\Theta|B,S) = (1-\alpha)\cdot e^{-\varepsilon_b}\cdot\frac{\varepsilon_b^B}{B!}\cdot e^{-\varepsilon_s}\cdot\frac{\varepsilon_s^S}{S!} + \\ + \alpha\cdot\delta\cdot e^{-\varepsilon_b}\cdot\frac{\varepsilon_b^B}{B!}\cdot e^{-(\mu+\varepsilon_s)}\cdot\frac{(\mu+\varepsilon_s)^S}{S!} + \\ + \alpha\cdot(1-\delta)\cdot e^{-(\mu+\varepsilon_b)}\cdot\frac{(\mu+\varepsilon_b)^B}{B!}\cdot e^{-\varepsilon_s}\cdot\frac{\varepsilon_s^S}{S!}, \quad (1)$$

where $\Theta = (\alpha,\delta,\mu,\varepsilon_b,\varepsilon_s)$ is a vector of parameters and $(1-\alpha)$, $\alpha\cdot\delta$, $\alpha\cdot(1-\delta)$ are no news, bad news, good news trading days respectively, while B and S are total Buy and Sell operations per day.

Formulating the maximization problem for t trading days is similar to the product of daily Likelihoods functions. Taking *log* (monotonic transformation) makes this equivalent to the sum of daily Log Likelihood functions:

$$V = \prod L(\Theta|M) = \sum \log L(\Theta|M), \quad (2)$$

where $M = ((B_1,S_1),\ldots,(B_n,S_n))$ is vector of Buy and Sell orders.

Finally, having solved the maximization problem, we obtain the optimal parameters values which are used for PIN calculation. PIN (*probability of informed trading*) is calculated as ratio of expected informed arrival rate to expected total arrival rate:

$$PIN = \frac{\alpha\cdot\mu}{\alpha\cdot\mu+\varepsilon_b+\varepsilon_s}. \quad (3)$$

Formula 3 takes into account both insider and noise trading and beliefs of liquidity provider. For instance, if there are only informed trades, based on private information, $(\varepsilon_s = \varepsilon_b = 0)$ then *PIN* = 1 and there is a wide bid-ask spread. Considering the case without private signal or no insider trading $(\mu = 0$ or $\alpha = 0)$ the *PIN* = 0 is obtained and there is no spread. Here we can observe the main distinction in the approach in the EHO model presented above and the original EKOP. In the EKOP model there is no differentiation between uninformed buyers and sellers, they are assumed to act at the same rate $\varepsilon_s = \varepsilon_b = \varepsilon$. However, in considered EHO specification, liquidity buyers and seller participate with unique rates ε_b and ε_s, respectively.

3.1.1. Factorization techniques and initial parameters

PIN model estimation procedure has two technical inefficiencies that affect its output: *floating point error (FPE)* and *false boundary solutions*. The first one appears due to factorials in the likelihood functions which cannot be computed for some large numbers, while the latter is

related to the computer estimation problem, where the optimization algorithm stops at local maximum.

These two problems were solved by modified likelihood and initial parameters algorithm, introduced by Lin and Ke (2011) and Yan and Zhang (2012), respectively. We will return to these submodels in further sections when consider APIN, which being the nested model, suffers from the same issues. However, these two algorithms were not modified, so this paper contributes by provided new versions of initial parameters and likelihood, but for APIN, in Section 4.

3.2. Why should not we stop at PIN?

The main caveat of PIN is that it theoretically does not allow *positive* correlation between Buy and Sell trades In terms of the model, the mathematical expression is always negative, while the empirical correlation appears to be positive (see Table 1). Thus, there is a need for an extension, solving this puzzle – Adjusted PIN.

Table 1.

Correlations between buyer and seller initiated trades (year 2022)

Ticker	Mean	Median	Max	Min
XBT	0.907	0.929	0.996	0.47
XRP	0.825	0.874	0.999	0.107
DOGE	0.692	0.752	0.999	–0.724
SOL	0.701	0.738	0.998	0.019
ADA	0.564	0.634	0.993	–0.297
LINK	0.647	0.67	0.999	–0.067
LTC	0.636	0.735	0.999	–0.164
AXS	0.499	0.573	0.988	–0.340
AVAX	0.526	0.561	0.996	–0.166
BCH	0.525	0.592	0.991	–0.471

3.3. APIN

This model, introduced by Duarte and Young (2007) is an extension of PIN (EHO) model. As in the original model, there are two types of traders (insiders and noise), and with probability α there is a private information event, which can be either positive or negative with underlying probabilities δ and $(1 - \delta)$, respectively. However, we allow informed traders to be heterogeneous and perform Buy/Sell trades at different rates: μ_b and μ_s. Moreover, we introduce an event of symmetric order flow, leading to additional Sell (Δ_s) and Buy (Δ_b) orders at the same time. This modification enables the model to match the empirically observed positive correlation between Buy and Sell orders, which the traditional PIN model fails to do.

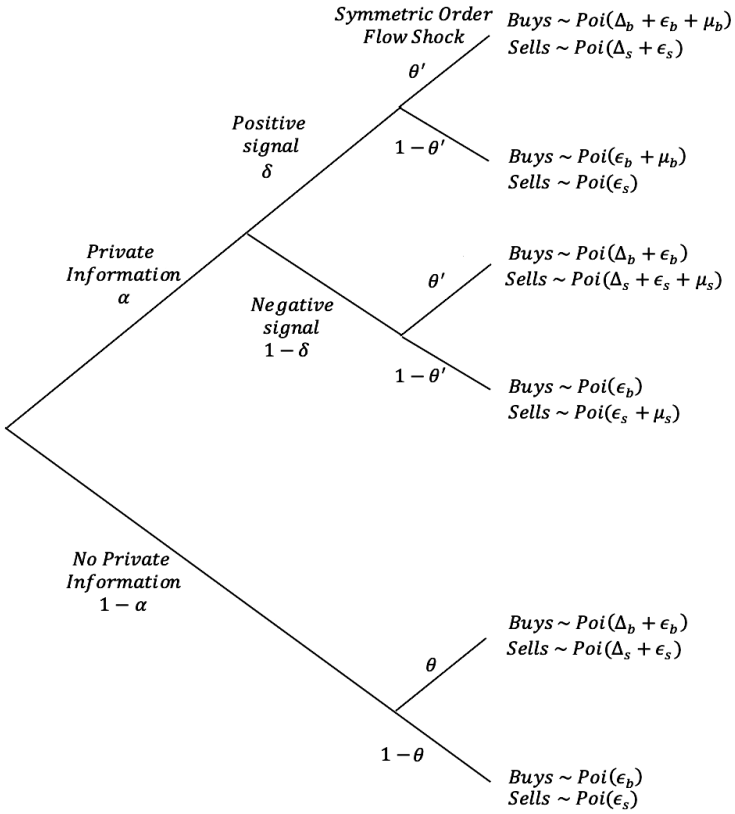


Fig. 2. Adjusted PIN model Prob. Tree[2]

Using the Poisson distribution assumption, we arrive at the following Likelihood function:

$$L\left(\Theta\middle|B_i,S_i\right)=\ln\Bigg[\left(1-\alpha\right)\left(1-\theta\right)\exp\left(-\varepsilon_b-\varepsilon_s\right)\frac{\varepsilon_b^{B_i}\varepsilon_s^{S_i}}{B_i!S_i!}+$$

$$+\left(1-\alpha\right)\theta\exp\left(-\varepsilon_b-\varepsilon_s-\Delta_b-\Delta_s\right)\frac{\left(\varepsilon_b+\Delta_b\right)^{B_i}\left(\varepsilon_s+\Delta_s\right)^{S_i}}{B_i!S_i!}+ \tag{4}$$

$$+\alpha\left(1-\theta'\right)\left(1-\delta\right)\exp\left(-\varepsilon_b-\mu_s-\varepsilon_s\right)\frac{\varepsilon_b^{B_i}\left(\mu_s+\varepsilon_s\right)^{S_i}}{B_i!S_i!}+$$

[2] Adjusted trading process tree. This diagram represents the trading mechanics, where $\alpha,\delta,\theta,\theta'$, $\mu_b,\mu_s,\varepsilon_b,\varepsilon_s,\Delta_b$ and Δ_s stay for probabilities of private information event, of positive signal and of symmetric order flow in case of private event and its absence, rate of informed buy and sell operations, rates of noisy buy and sell operations and symmetric buy and sell rates, respectively.

$$+\alpha\theta'(1-\delta)\exp(-\varepsilon_b-\varepsilon_s-\mu_s-\Delta_b-\Delta_s)\frac{(\varepsilon_b+\Delta_b)^{B_i}(\mu_s+\varepsilon_s+\Delta_s)^{S_i}}{B_i!S_i!}+$$

$$+\alpha(1-\theta')\delta\exp(-\mu_b-\varepsilon_b-\varepsilon_s)\frac{(\mu_b+\varepsilon_b)^{B_i}\varepsilon_s^{S_i}}{B_i!S_i!}+$$

$$+\alpha\theta'\delta\exp(-\mu_b-\varepsilon_b-\varepsilon_s-\Delta_b-\Delta_s)\frac{(\mu_b+\varepsilon_b+\Delta_b)^{B_i}(\varepsilon_s+\Delta_s)^{S_i}}{B_i!S_i!}\Bigg],$$

where $\Theta=(\alpha,\delta,\theta,\theta',\mu_b,\mu_s,\varepsilon_b,\varepsilon_s,\Delta_b,\Delta_s)$ are probability of news, probability of good news, probability of symmetric buy and sell trades, given there is NO private signal, probability of symmetric buy and sell trades, given there is private signal, insider's buy and sell trading rates, noise traders' buy and sell trading rates, additional buy and sell trading rates in case of symmetric trading event, respectively, while B and S are total Buy and Sell operations per day.

In order to tackle factorials, which cannot be computed for large numbers, Duarte et al. (2007) modifies the Likelihood, using $e^{-\varepsilon_b}\frac{\varepsilon_b^B}{B!}\sim e^{-\varepsilon_b+B\cdot\ln(\varepsilon_b)-\sum_{i=1}^{B}\ln(i)}$ trick.

Given independence of information signals for each particular day, we can reformulate the maximization problem for *t* periods as:

$$V=\prod\{L(\Theta|B,S)\}=\sum\log L(\Theta|B,S). \tag{5}$$

Formula for Adjusted PIN is the ratio of expected insider trading order flow to total order flow (nested PIN formula):

$$Adj\,PIN=\frac{\alpha(\delta\cdot\mu_s+(1-\delta)\cdot\mu_b)}{\alpha((1-\delta)\cdot\mu_b+\delta\cdot\mu_s)+(\Delta_b+\Delta_s)\cdot(\alpha\cdot\theta'+(1-\alpha)\cdot\theta)+\varepsilon_s+\varepsilon_b}. \tag{6}$$

4. Two New Algorithms to Enhance APIN Estimation

Being nested version of PIN model, APIN inherits FPE and local maximum problems, described in the Subsection 3.1.1. However, to our best knowledge, although these problems were solved for PIN, this issue was not widely reconsidered for APIN.

We propose a new algorithm of initial parameters and new modified likelihood for APIN, based on ideas of Yan and Zhang (2012) and Lin and Ke (2011), which demonstrates a significant boost in accuracy on a simulated data sample.

4.1. New factorization technique and initial parameters algorithm for Adjusted PIN

Using the intuition and ideas from Lin and Ke (2011) and Yang and Zhang (2012) models for PIN, in the following two subsections we derive their modified versions for Adjusted PIN framework.

New factorization technique

As in Lin and Ke (2011), we base our derivation on two main ideas:

1. Computer provides more stable estimates for e^{x+y} rather than for $e^x e^y$.
2. We should avoid plugging too large inputs into exp() and too low ones into ln(). For instance, if we want to estimate $\ln\left(e^{x+y}+e^z\right)$ we should better rewrite as:

$$\ln\left[\frac{\left(e^{x+y}+e^z\right)e^k}{e^k}\right]=\ln\left(e^{(x+y)-k}+e^{(z-k)}\right)+k, \tag{7}$$

where $k=\max(x+y,z)$.

This trick guarantees the expression inside logarithm is always greater than one and we do not obtain $ex, x>710$, leading to overflow.

Applying these two principles on the initial Likelihood function, we get the more accurate expression:

$$\begin{aligned} L\left(\Theta\middle|B_i,S_i\right)=\ln\Big[&(1-\alpha)(1-\theta)\exp(-e_{maxi})+(1-\alpha)\theta\exp(e_{1i}-e_{maxi})+\\ &+\alpha(1-\theta')(1-\delta)\exp(e_{2i}-e_{maxi})+\alpha\theta'(1-\delta)\exp(e_{3i}-e_{maxi})+\\ &+\alpha(1-\theta')\delta\exp(e_{4i}-e_{maxi})+\alpha\theta'\delta\exp(e_{5i}-e_{maxi})\Big]-(\varepsilon_b+\varepsilon_s)+\\ &+S_i\ln(\varepsilon_s)+B_i\ln(\varepsilon_b)+e_{maxi}-\ln(B_i!S_i!), \end{aligned} \tag{8}$$

where $e_{1i}=-\Delta_b-\Delta_s+B_i\ln\left(1+\Delta_b/\varepsilon_b\right)+S_i\ln\left(1+\Delta_s/\varepsilon_s\right)$,

$e_{2i}=-\mu_s+S_i\ln\left(1+\mu_s/\varepsilon_s\right)$,

$e_{3i}=-\mu_s-\Delta_b-\Delta_s+B_i\ln\left(1+\Delta_b/\varepsilon_b\right)+S_i\ln\left(1+\left[\mu_s+\Delta_s\right]/\varepsilon_s\right)$,

$e_{4i}=-\mu_b+B_i\ln\left(1+\mu_b/\varepsilon_b\right)$,

$e_{5i}=-\mu_b-\Delta_b-\Delta_s+B_i\ln\left(1+\left[\mu_b+\Delta_b\right]/\varepsilon_b\right)+S_i\ln\left(1+\Delta_s/\varepsilon_s\right)$,

$e_{maxi}=\max\left(e_{1i},e_{2i},e_{3i},e_{4i},e_{5i}\right)$.

New initial parameters algorithm

We will use a more parsimonious specification with 10 parameters to estimate (setting $\theta=\theta'$):

$$\Theta=\left(\alpha,\delta,\theta,\mu_s,\mu_b,\varepsilon_s,\varepsilon_b,\Delta_s,\Delta_b\right).$$

We use 1-st and 2-nd moment conditions:

$$E(B)=\varepsilon_b+\theta\Delta_b+\alpha\delta\mu_b, \tag{9}$$

$$E(S) = \varepsilon_s + \theta\Delta_s + \alpha(1-\delta)\mu_s, \tag{10}$$

$$E(B^2) = \varepsilon_b^2 + \alpha\delta\mu_b^2 + \theta\left(\Delta_b^2 + 2\varepsilon_b\Delta_b\right) + 2\alpha\delta\mu_b(\varepsilon_b + \theta\Delta_b), \tag{11}$$

$$E(S^2) = \varepsilon_s^2 + \alpha(1-\delta)\mu_s^2 + \theta\left(\Delta_s^2 + 2\varepsilon_s\Delta_s\right) + 2\alpha(1-\delta)\mu_s(\varepsilon_s + \theta\Delta_s). \tag{12}$$

As $E(B)$ is always greater than ε_b, so we set the latter to be proportion of the sample analogue $\varepsilon_b = \gamma\overline{B}$, where $\gamma = \{0.1, 0.3, 0.5, 0.7, 0.9\}$. In the same fashion we set, $\varepsilon_s = \gamma\overline{S}$, where $\gamma' = \{0.1, 0.3, 0.5, 0.7, 0.9\}$.

As for probabilities of signal, positive signal and symmetric order flow shock we assign the following set of potential values to them:

$$\alpha = \{0.1, 0.3, 0.5, 0.7, 0.9\},$$
$$\delta = \{0.1, 0.3, 0.5, 0.7, 0.9\},$$
$$\theta = \{0.1, 0.3, 0.5, 0.7, 0.9\}.$$

As a result, we have 5^5 = 3125 initial values to estimate, which makes the algorithm extremely computationally intensive. Thus, in further analysis and testing we set $\gamma = \gamma'$ which decreases the number of points to 625. Moreover, we will also exclude some of them due to elimination of negative roots.

4.2. Monte Carlo simulation

In order to test the increase in accuracy we simulate the order flow, based on theoretical parameters:

Parameter	Description	Value
α	prob. of signal	$U[0,1]$
δ	prob. positive signal	$U[0,1]$
θ	prob. symmetric order flow	$U[0,1]$
I	total trade intensity	2500
p_μ	proportion of informed traders as proportion of I	$U[0,1]$
p_{μ_b}	informed who buy	$U[0,1]$
p	proportion of noise traders	$U[0,1]$
p_{ε_b}	noise who buy	$U[0,1]$
$p\Delta_b$	additional buy trades under symmetric order flow	$U[0,1]$

Thus, we get the following theoretical rates:

$$\mu_b = p_\mu \cdot p_{\mu_b} \cdot I, \tag{13}$$

$$\mu_s = p_\mu \cdot \left(1 - p_{\mu_b}\right) \cdot I, \tag{14}$$

$$\varepsilon_b = \left(1 - p_\mu\right) \cdot p_\varepsilon \cdot p_{\varepsilon_b} \cdot I, \tag{15}$$

$$\varepsilon_s = \left(1 - p_\mu\right) \cdot p_\varepsilon \cdot \left(1 - p_{\varepsilon_b}\right) \cdot I, \tag{16}$$

$$\Delta_b = \left(1 - p_\mu\right) \cdot \left(1 - p_\varepsilon\right) \cdot p_{\Delta_b} \cdot I, \tag{17}$$

$$\Delta_s = \left(1 - p_\mu\right) \cdot \left(1 - p_\varepsilon\right) \cdot \left(1 - p_{\Delta_b}\right) \cdot I. \tag{18}$$

As we know the theoretical rates, we can estimate the implied true value of APIN:

$$Adj\ PIN = \frac{\alpha \cdot \left(\delta \cdot \mu_s + \left(1-\delta\right) \cdot \mu_b\right)}{\alpha \cdot \left(\left(1-\delta\right) \cdot \mu_b + \delta \cdot \mu_s\right) + \left(\Delta_b + \Delta_s\right) \cdot \left(\alpha \cdot \theta' + \left(1-\alpha\right) \cdot \theta\right) + \varepsilon_s + \varepsilon_b}. \tag{19}$$

Finally, using Poisson distribution, we generate the Buy and Sell trades:

$$Buy \sim Poisson\left(\varepsilon_b, \mu_b, \Delta_b \middle| \Theta\right),$$

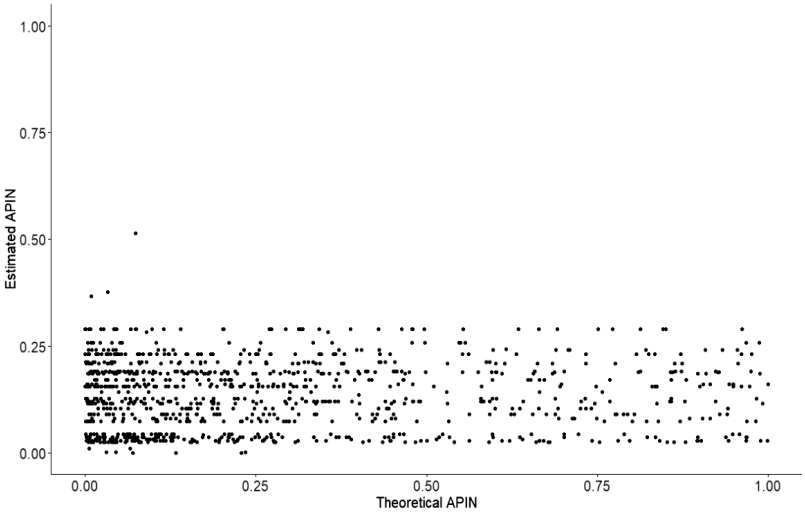
$$Sell \sim Poisson\left(\varepsilon_s, \mu_s, \Delta_s \middle| \Theta\right).$$

If there is no signal (α = 0) then there are only noise traders in the market, $Buy \sim Poisson\left(\varepsilon_b\right)$ and $Sell \sim Poisson\left(\varepsilon_s\right)$ or if there is a positive signal with event of symmetric order flow then $Buy \sim Poisson\left(\varepsilon_b + \mu_b + \Delta_b\right)$, and $Sell \sim Poisson\left(\varepsilon_s + \mu_s + \Delta_s\right)$. Finally, we utilize the generated order flow to obtain an estimate of theoretical APIN.

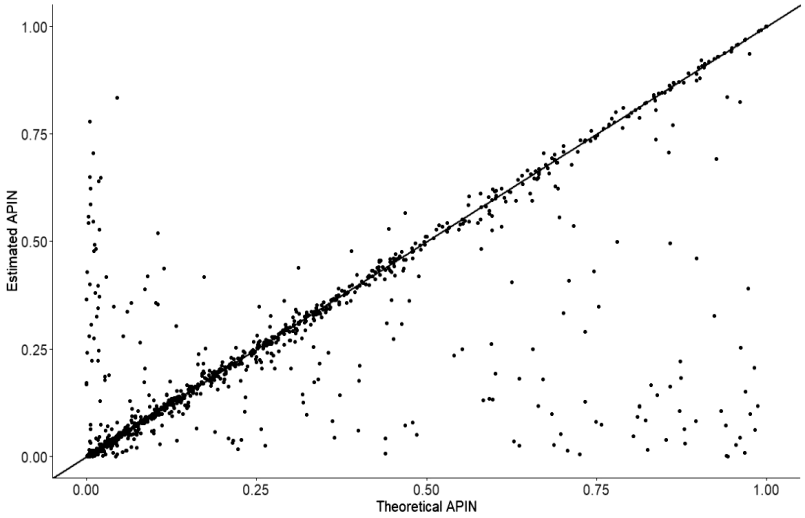
We use this Monte Carlo setting to show that our two estimation procedures for APIN substantially improves the accuracy of model's predictions. To show this we compare two specifications:

1. "Naive APIN" (Used by Duarte and Young):
   - Duarte Likelihood function;
   - Ten random initial points for optimization'
2. Modified APIN (Our version):
   - New likelihood (Section 4.1);
   - New initial parameters algorithm (Section 4.1).

We find that APIN estimates with our proposed algorithms (Modified APIN) are much more precise than the original version (Naive APIN), derived and applied empirically by Duarte and Young.

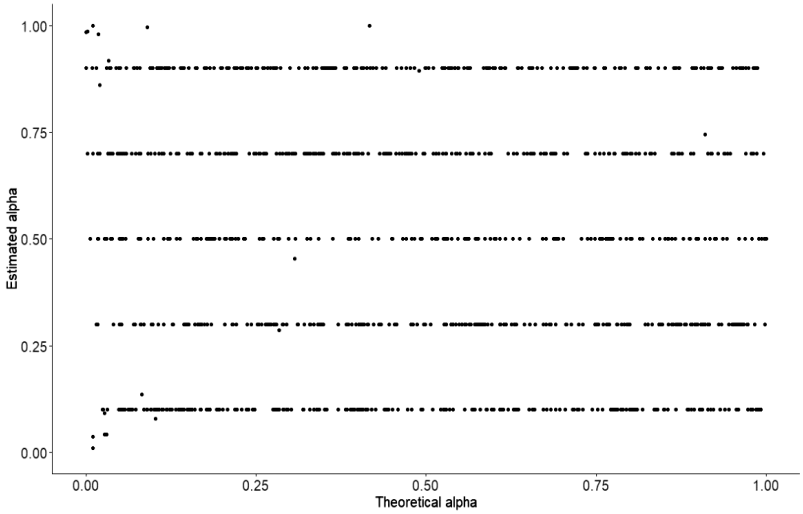


(a) APIN estimate vs theoretical value (Naive APIN)

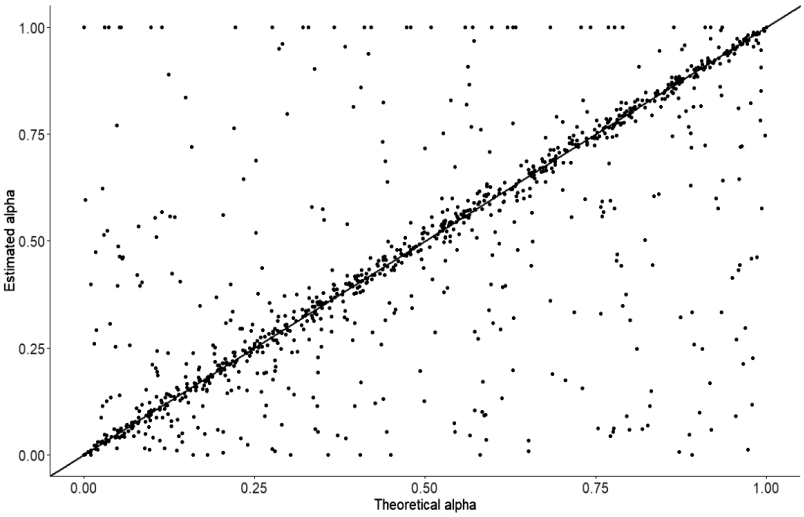


(b) APIN estimate vs theoretical value (Modified APIN)

Fig. 3. Naive APIN vs Modified APIN

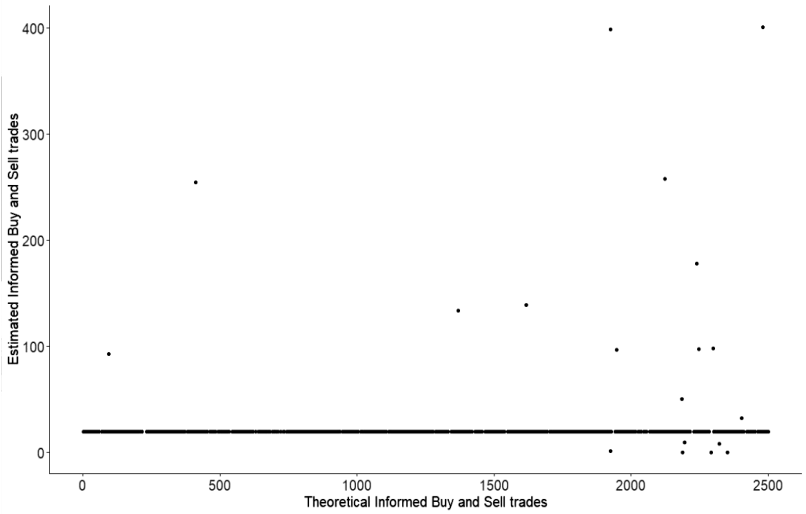


(a) α estimate vs theoretical value (Naive APIN)

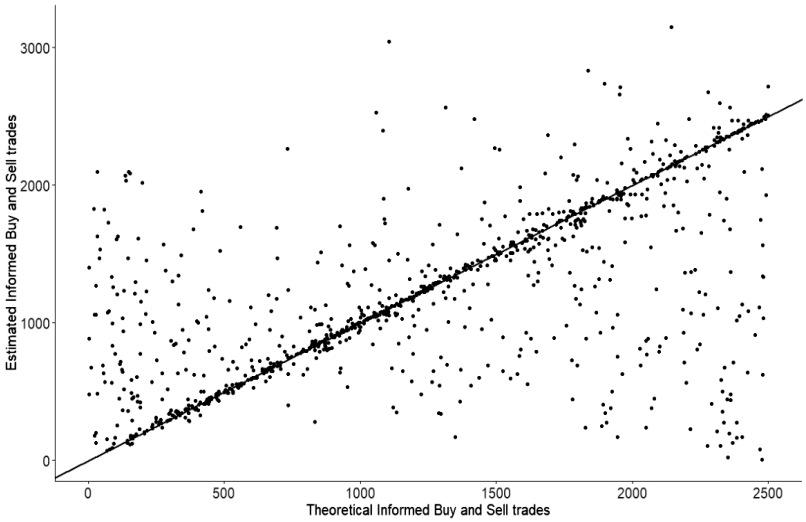


(b) α estimate vs theoretical value (Modified APIN)

Fig. 4. Naive APIN vs Modified APIN (Cont'd)



(a) $\mu_s + \mu_b$ estimate vs theoretical value (Naive APIN



(b) $\mu_s + \mu_b$ estimate vs theoretical value (Modified APIN)

Fig. 5. Naive APIN vs Modified APIN (Cont'd)

From the figures above we can infer a large difference in the results. Naive APIN fails and cannot predict the true value of the metric. However, it is not due to the model being bad itself, but because of estimating inefficiencies that were solved by our approach. Modified APIN has a very high precision and we observe 45 degree line between estimates and theoretical values of metric and underlying parameters almost in all graphs.

5. Identifying Optimal Aggregation Frequency

Previous empirical studies on stock markets typically estimated daily PIN and APIN, using order imbalance, calculated as a difference between buyer and seller initiated trades, which are in turn aggregated for the whole trading day. However, we find this approach not effective, since it uses only vector of two data points for the maximum likelihood optimization which is likely to provide false solutions. Some other papers, such as Cepoi et al. (2023), estimate PIN for some given period, that is taking as set of daily aggregated trades and use it to estimate the value of parameters for several days, weeks or even months, assuming parameters are identically distributed for each day.

Thus, using the idea of period estimation, in this paper we try to mitigate the inefficiencies of previous researches, estimating daily PIN and APIN, by aggregating trades at higher frequencies rather than day, i.e. 2 hours, hour, 30 minutes, and other. Figure 5 illustrates this approach of bucketing trades within the day, which if summed will provide the aggregate daily order imbalance. This approach allows to significantly increase the number of data points used, however, at the cost of assumption that all these trades come from the same distribution, which is still more natural in our setting of daily metric estimation rather than weekly or monthly alternatives.

This assumption is additionally supported by two other arguments. Informed trading implies not only some illegal activity before public events but also cyclical activity of sophisticated algorithms that outperform the market due to increased ability of finding and analysing information. Thus, the assumption that with the same probability there is a signal every hour within one day is not as unrealistic. Another supporting idea is the notion of strategic trading. To avoid large effect on the price, informed traders might trade several times instead of executing the whole order immediately, thus, considering several order imbalances, by aggregating trades frequencies higher than by day, is expected to provide more relevant outcomes.

In order to assess the validity and improvement in accuracy, we apply two Monte Carlo simulations similar to the one described in the previous subsection to generate theoretical order flows for PIN and APIN set-ups and evaluate performance of the models with different frequencies of trades aggregation as inputs. In Section 4 we have shown that our introduced modification for APIN estimation procedure provides more efficient results theoretically, so we use this modified APIN specification in the current and all further sections.

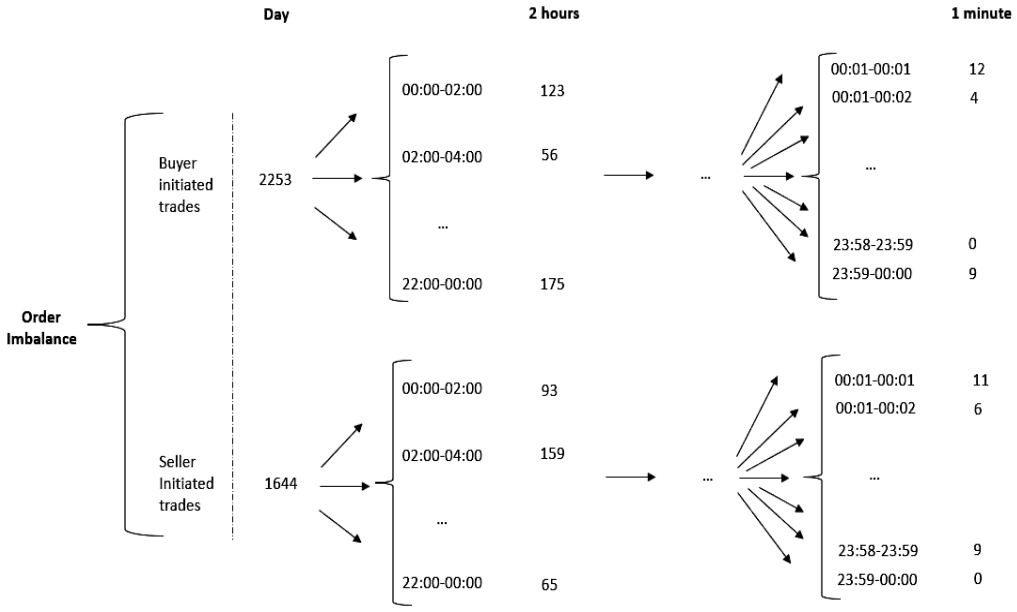


Fig. 6. Aggregation of trades (different levels)

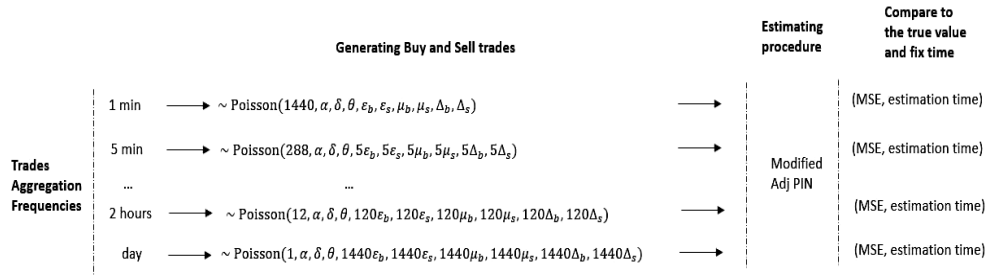


Fig. 7. APIN Monte Carlo simulation scheme for optimal frequency testing

The identification of the potentially optimal aggregation frequency is based on the tradeoff between accuracy and estimation time. By increasing number of data points, we might obtain a more accurate estimate, however, more time to run optimization is needed. Thus, we suppose that after some frequency the improvement in accuracy is so small that it does not offset the increase in the computation time, implying the existence of "optimal" frequency. If estimation *time* increases at higher rate than decrease in the Mean Squared Error (*MSE*) between theoretical value and the predicted one, then the choice of higher frequency to increase data points might be sub-optimal.

Initially we generate 1440 buyer and seller initiated trades that represent the trading frequency of 1 minute within 24 hours. Then we aggregate them on the levels of 5 minutes, 15 minutes up to the daily volume. Since our approach implies that the underlying distribution is

the same, for all frequencies we have the same parameters α, δ, θ, the only thing that changes is the rates of trading which due to additive nature of Poisson distribution increase proportionally with the increase of the respective time interval. For PIN model we generate a separate order flow in the same fashion as for APIN, but with the reduced vector of parameters that correspond to PIN only: $\Theta = (\alpha, \delta, \mu, \varepsilon_b, \varepsilon_s)$.

Based on our simulations, we obtained the following results:

Table 2.

APIN and PIN
(estimation time vs accuracy)

Agg. Freq	APIN MSE	APIN time	PIN MSE	PIN time
Day	0.13	0.00	0.20	0.00
2h	0.09	0.02	0.01	0.01
1h	0.09	0.03	0.01	0.02
30 min	0.08	0.04	0.00	0.04
15 min	0.08	0.08	0.00	0.08
5 min	0.08	0.21	0.00	0.23
1min	0.04	1.00	0.00	1.00

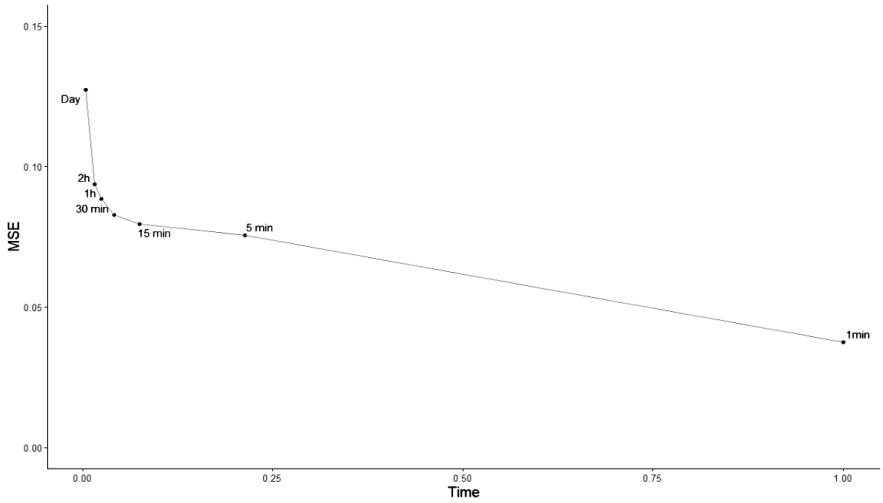


Fig. 8. APIN Optimal Frequency

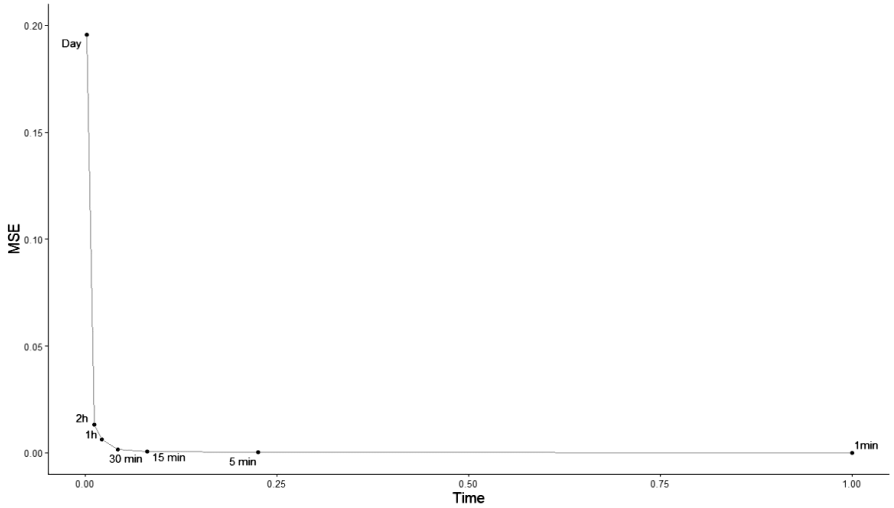


Fig. 9. PIN Optimal Frequency

According to the figures above, for both PIN and APIN, we can detect convergence: at some frequencies the decrease in MSE is relatively much smaller compared to the increase in estimation time (normalised to be between 0 and 1). As for APIN, aggregation of trades at 15 and 5 minutes gives almost the same precision level, while estimation duration increases almost in 3 times. Even sharp decrease in mse at 1 min does not offset higher estimation time. Considering PIN this convergence is even higher and we obtain similar MSE estimates at even lower frequencies.

Thus, based on model's accuracy and estimation complexity trade-off, we find it optimal to evaluate both models, using data aggregation frequencies not higher that 15 minutes. Taking higher frequencies substantially increases the computation time without providing the proportional improvement in precision. Moreover, such conclusion is only valid if we assume the same distribution of information signal and of the parameters within the same trading day. The analysis of optimal frequency could also be further improved by taking into account liquidity issues that could theoretically affect the order flow distribution.

6. Data

This paper considers 10 cryptocurrencies, analyzed over year 2022. The number might seem limited due to high computational intensity, however, these tickers account for the major part of the total market volume. The data was taken from BitMex exchange which is considered one of the largest and focuses on professional investors, by offering leveraged products and derivatives. The latter makes it a good candidate for searching for informed trading activity of HFT funds and sophisticated traders. As already mentioned above, on ordinary computers some model specification may take ages to run, thus, main part of results is obtained using HSE supercomputer Charisma with parallel computing on 48 cores.

In order to contrast some empirical conclusions about stocks markets, we also divide the cryptocurrencies into two groups in terms of liquidity, using Amihud Liquidity measure by Amihud (2002).

Table 3.

Liquid group:
averaged daily statistics summary

Symbol	ADV	Daily volatility	Daily turnover	Daily number of trades
XBTUSD	20,443	0.03	578,098,937	105,580
XRPUSD	16,490,744	0.04	9,766,290	12,230
SOLUSD	154,058	0.06	7,162,003	10,704
DOGEUSD	45,645,322	0.05	5,287,247	5,781
LTCUSD	47,551	0.04	4,272,178	6,760

Table 4.

Illiquid group:
averaged daily statistics summary

Symbol	ADV	Daily volatility	Daily turnover	Daily number of trades
ADAUSD	5,407,374	0.05	3,540,267	5,336
BCHUSD	11,182	0.04	2,194,354	3,748
LINKUSD	205,863	0.05	2,332,850	3,944
AVAXUSD	31,044	0.06	1,166,538	1,322
AXSUSD	12,497	0.06	380,240	593

Note: Average Daily Volume (ADV) is estimated based on units of each cryptocurrency, while Daily Turnover is measured in USD.

7. Main Results

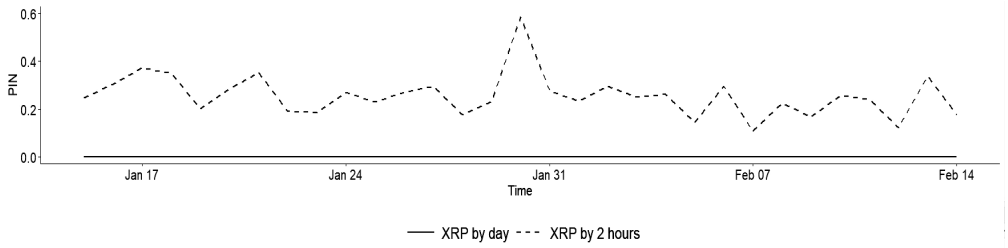
7.1. Different aggregation frequencies

7.1.1. Aggregation of trades by Day vs Higher aggregation frequencies

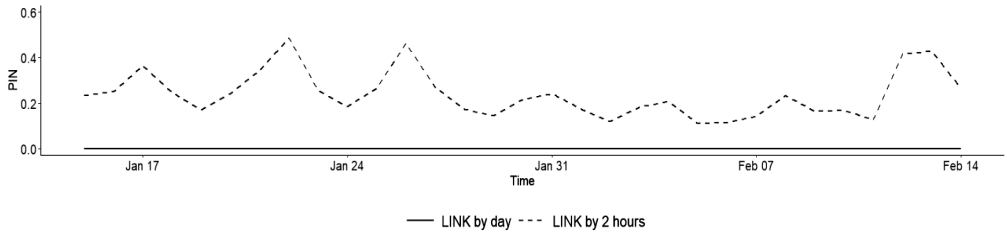
As it was shown in the previous subsections, the lowest type of trades aggregation *by Day* is hardly likely to be efficient and provide meaningful results on large volumes. Empirical data support this idea and it can be inferred from Fig. 10 that PIN takes value of zero for XRP and LINK, which is also true for other cryptocurrencies. This is quite intuitive since daily aggregated trades contain a lot of noise and are unable to capture informed trading activity. On the contrary, unifying trades even every two hours, provides much more information about order imbalance dynamics within the day and results into much smoother non-zero series.

From Tables 5 to 8 we can infer that Illiquid cryptocurrencies, on average, have higher PIN values which is consistent with results for stocks by Easley et al. (1996), however, the average

probabilities of signal (α) and whether it is negative (δ) are approximately the same. The higher PIN value illustrates the fact of higher information asymmetry, compared to liquid cryptos, that results into higher probability of informed trades which investors incorporate into risks and spreads.



(a) Aggregation by Day (Liquid)



(b) Aggregation by 2 hours (Illiiquid)

Fig. 10. PIN (Aggregation of trades by Day vs 2 hours)

Table 5.

(PIN day) Mean values for 2022 (Liquid Group)

Ticker	α	δ	μ	ε_b	ε_s	PIN
XBT	0.77	0.39	1,628,523,555.11	288,366,367.48	284,193,731.31	0.01
XRP	0.95	0.39	1.65	1,375,244.04	1,475,855.48	0.00
DOGE	0.96	0.34	1.00	750,952.08	813,102.24	0.00
SOL	0.94	0.39	4.75	3,272,167.48	3,474,144.79	0.00
LTC	0.95	0.36	0.51	414,825.29	426,370.17	0.00

Table 6.

(PIN 2h) Mean values for 2022 (Liquid Group)

Ticker	α	δ	μ	ε_b	ε_s	PIN
XBT	0.33	0.56	32,224,410.10	20,098,959.30	18,896,418.32	0.18
XRP	0.31	0.65	240,970.23	94,715.38	86,625.47	0.24
DOGE	0.37	0.58	128,192.90	47,079.93	49,645.83	0.26
SOL	0.34	0.56	521,216.93	213,391.47	203,714.50	0.24
LTC	0.37	0.55	55,568.65	28,080.93	26,368.23	0.22

Table 7.

(PIN day) Mean values for 2022 (Illiquid Group)

Ticker	α	δ	μ	ε_b	ε_s	PIN
LINK	0.94	0.39	0.40	362,753.11	366,862.86	0.00
ADA	0.97	0.36	1.21	1,034,714.10	1,069,847.69	0.00
AXS	0.94	0.32	0.22	229,249.37	230,249.46	0.00
AVAX	0.92	0.36	0.07	60,732.26	63,513.44	0.00
BCH	0.95	0.37	0.27	219,845.93	223,297.79	0.00

Table 8.

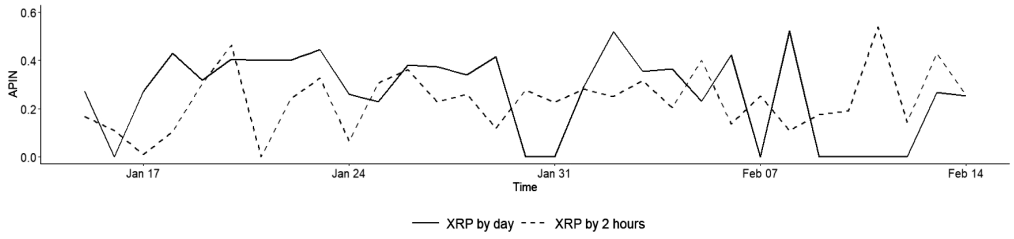
(PIN 2h) Mean values for 2022 (Illiquid Group)

Ticker	α	δ	μ	ε_b	ε_s	PIN
LINK	0.38	0.51	61,992.57	21,418.35	22,997.43	0.26
ADA	0.36	0.55	200,742.19	64,032.58	63,801.74	0.26
AXS	0.33	0.53	76,825.21	11,356.58	10,266.97	0.48
AVAX	0.36	0.54	13,672.18	3,452.49	3,369.22	0.40
BCH	0.39	0.57	30,879.36	14,373.89	13,416.71	0.25

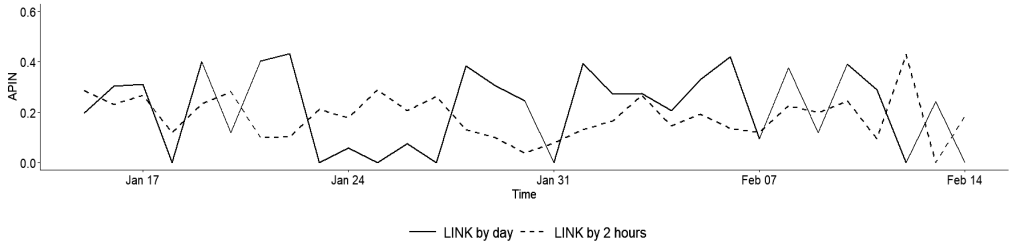
As for APIN, we observe a different behavior. Even at the lowest trade aggregation frequency of one day (Fig. 11), they show non-zero pattern in contrast to PIN. This might be considered as an additional confirmation that PIN model fails due to its theoretical inefficiencies especially in a high trading volume environment. Still, the behavior at daily aggregation is quite volatile, constantly jumping from zero to higher values and although APIN-s are mainly non-zero, we can observe that parameters α, δ, θ tend to have boundary values of 1 (Tables 9 and 11).

Higher aggregation frequency at 2 hours, results into much smoother and ergodic behavior and provides lower values for APIN with parameters being far from boundary solutions. Combined with other facts considered, this provides another argument for confirming that aggregation at low frequencies such as one day might not be efficient.

Interestingly, the phenomenon of higher informed trading for not liquid cryptocurrencies does not exist at daily trades aggregation rather than at higher frequencies such as 2 hours (Tables 9–12). We can attribute this to the fact that daily order imbalance is too noisy and does not provide meaningful insight into informed trading activity, so the APIN tends to be affected by large absolute values of trading volume (volume for liquid cryptos is much higher), leading to a higher metric's value for liquid groups. Trades aggregation at higher frequencies decreases the absolute value of individual order imbalances and provides more information about their dynamics throughout the trading day.



(a) Aggregation by Day (Liquid)



(b) Aggregation by 2 hours (Illiiquid)

Fig. 11. APIN (Aggregation of trades by Day vs 2 hours)

Table 9.

(APIN day) Mean values for 2022 (Liquid group)

Ticker	α	1 – δ	θ	APIN
XBT	1.00	0.50	1.00	0.36
XRP	0.82	0.48	0.81	0.27
DOGE	0.78	0.44	0.78	0.25
SOL	0.82	0.47	0.81	0.27
LTC	0.81	0.45	0.82	0.27

Table 10.

(APIN 2h) Mean values for 2022 (Liquid group)

Ticker	α	1 – δ	θ	APIN
XBT	0.69	0.50	0.44	0.17
XRP	0.69	0.51	0.42	0.19
DOGE	0.70	0.50	0.44	0.21
SOL	0.71	0.47	0.45	0.2
LTC	0.69	0.48	0.45	0.19

Table 11.

(APIN day) Mean values for 2022 (Illiquid group)

Ticker	α	1 – δ	θ	APIN
LINK	0.77	0.49	0.75	0.24
ADA	0.82	0.48	0.78	0.26
AXS	0.82	0.42	0.71	0.29
AVAX	0.79	0.48	0.62	0.25
BCH	0.85	0.44	0.76	0.28

Table 12.

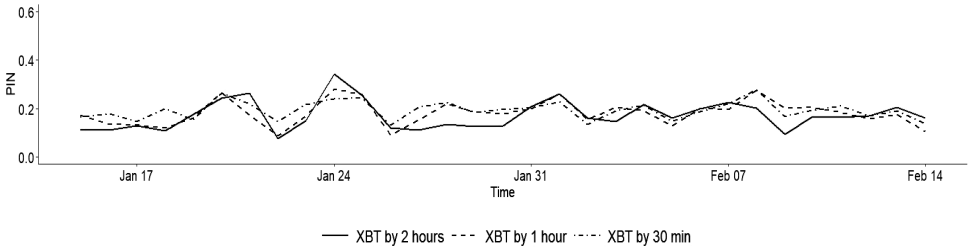
(APIN 2h) Mean values for 2022 (Illiquid group)

Ticker	α	1 – δ	θ	APIN
LINK	0.69	0.51	0.44	0.20
ADA	0.71	0.51	0.47	0.20
AXS	0.67	0.48	0.38	0.26
AVAX	0.61	0.52	0.38	0.25
BCH	0.68	0.51	0.41	0.19

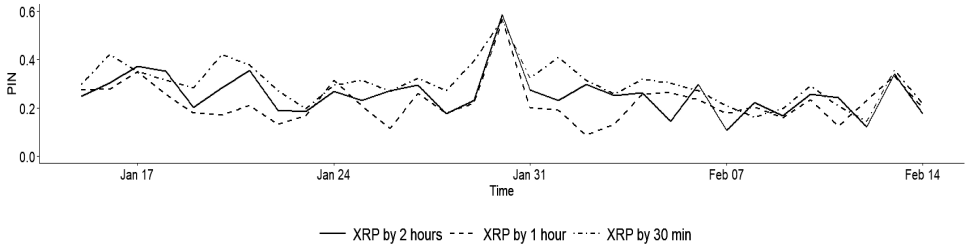
7.2. Theoretically optimal higher aggregation frequency in practice

From Monte Carlo simulation in Section 5, we found that it is not optimal to go higher aggregating frequency than 15 minutes for both PIN and APIN models. According to our theoretical results, we expect to obtain more or less the same outputs for "medium" frequencies (2 hours to 15–30 min), which means we can use lower frequency to decrease estimation time without sacrificing much in accuracy.

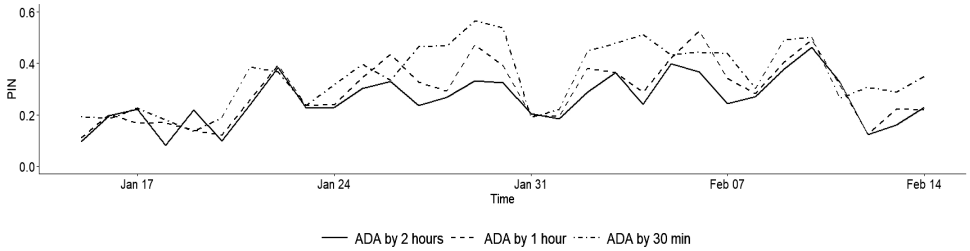
Figure 12 confirms our hypothesis and shows that trades aggregation at 2, 1 and 0.5 hours provide very similar PIN series. Even when the lines do not coincide they appear to have similar peaks and troughs and similar shape. However, it is worth noting 2 hours aggregation, on average, provides lower PIN values, while 30 minutes – higher. This provides empirical difficulty of selecting optimal aggregation frequency since such pattern is either a sign of potential difference in trading volumes, e.g. trading algorithms trade on higher frequency intervals such as every 30 minutes rather than 2 hours or an indication of increasing heterogeneity in order imbalances that the model considers to be the sign of informed trading activity.



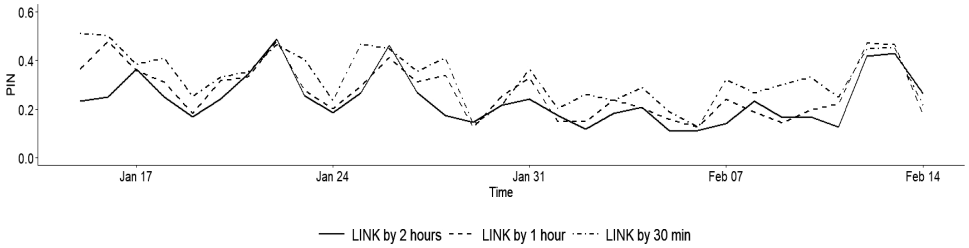
(a) XBT (Liquid Group)



(b) LTC (Liquid Group)



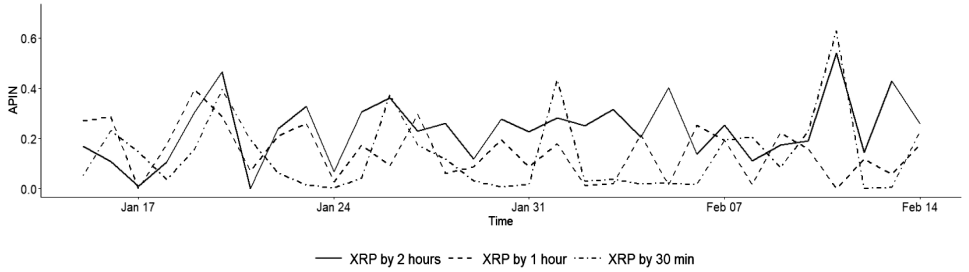
(c) ADA (Illiquid Group)



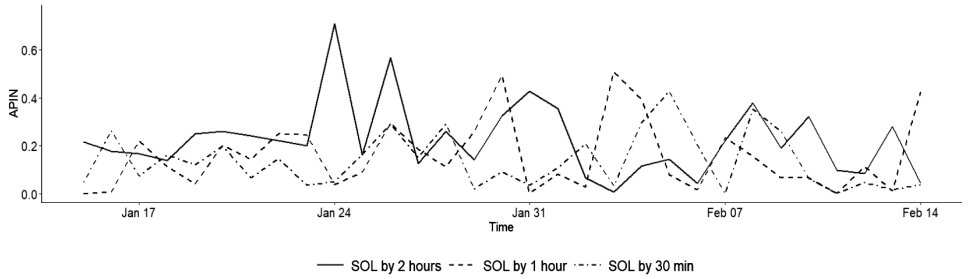
(d) AXS (Illiquid Group)

Fig. 12. PIN (Aggregation of trades by 2 hours vs 1 hour vs 30 minutes)

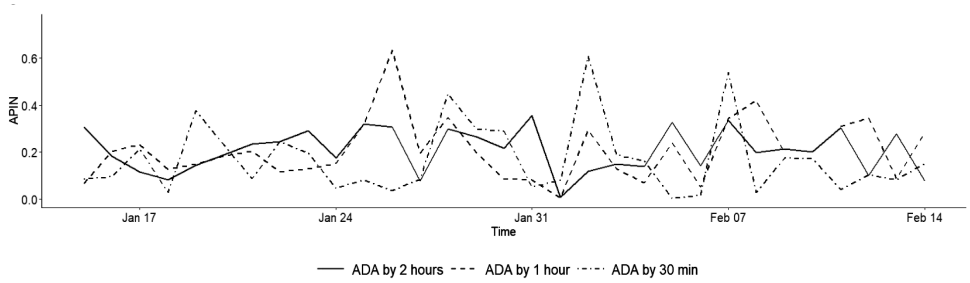
As for APIN, from Fig. 13 we can infer that the notion of medium frequencies of data aggregation being similar still holds, however, to a lower extent. There is some degree of prevailing similarity in behavior, while at different moments of time series diverge from each other. This further proves the frequency puzzle we described before. Various aggregation frequencies might make the resulting order imbalance more or less heterogeneous and exante we do not know at which frequencies informed trades are executed: they might be every hour or every 5 minutes. That it is why, 1 hour trades aggregation could be slightly different from 2 hour and 30 minutes aggregation, since informed activity might happen at the first frequency and not at the latter ones. We will elaborate further on this discussion in further sections.



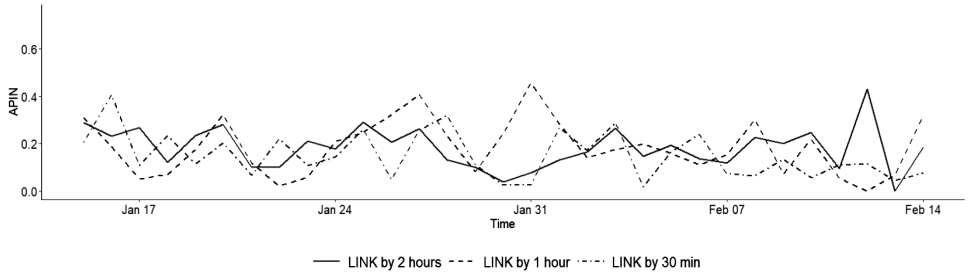
(a) XRP (Liquid Group)



(b) SOL (Liquid Group)



(c) ADA (Illiquid Group)



(d) LINK (Illiquid Group)

Fig. 13. APIN (Aggregation of trades by 2 hours vs 1 hour vs 30 minutes)

7.2.1. Potential implications of frequency aggregation technique

Although PIN and APIN have approximately similar behavior at medium frequencies up to 30 minutes, that does not mean they have a similar shape at higher frequencies such as 1 or 5 minutes. Figure 14 shows that PIN estimates, based on 1 minute (blue dash dotted) and 15 minutes (black dashed) trades aggregation frequencies are different from the ones, based on 30 minutes (green). They are more volatile and have higher amplitude, on average. One may claim that this proves higher frequencies of trades bucketing may be more efficient and accurate, however, higher accuracy on the simulated data does not automatically imply the same on the real one. This difference more probably shows that after some frequencies the model simply breaks and starts to behave weirdly. The potential reason behind this issue is that at higher frequencies an increasing number of considered order imbalances becomes more and more heterogeneous, thus, the assumption of the same underlying theoretical parameters appears to be less valid. Another explanation is that trading activity at different frequencies is also different. This argument is supported by the mean values and standard deviations of different specifications actually divide them into two groups: they are approximately the same for 2 hours, 1 hour and 30 minutes (medium frequencies) and also similar for 15 minutes, 5 minutes, and 1 minute (high frequencies), respectively.

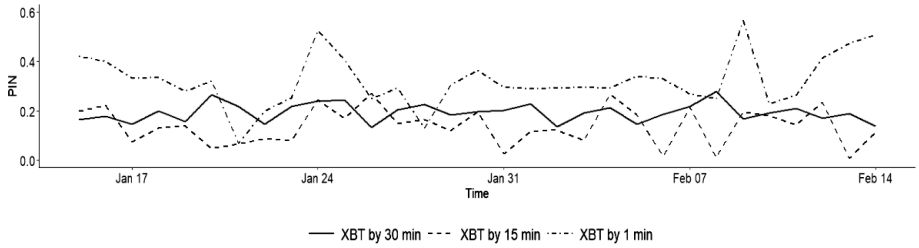


Fig. 14. PIN XBT (30 min Aggr vs 5 and 1 min Aggr)

Table 13.

PIN BTC Averages (Different data)

Ticker	Avg(α)	Avg(δ)	Avg(PIN)	sd(α)	sd(δ)	sd(PIN)
XBT day	0.77	0.39	0.01	0.28	0.35	0.06
XBT 2 hours	0.33	0.56	0.18	0.15	0.4	0.06
XBT 1 hour	0.27	0.55	0.20	0.12	0.35	0.05
XBT 30 min	0.23	0.53	0.21	0.09	0.29	0.05
XBT 15 min	0.12	0.49	0.15	0.09	0.49	0.08
XBT 5 min	0.12	0.5	0.21	0.07	0.49	0.09
XBT 1 min	0.11	0.47	0.33	0.05	0.49	0.10

Similar pattern we observe for APIN, where 15 minutes trades bucketing leads to different shape, compared to 30 minutes and 1 hour, which are quite alike. However, in contrast to PIN, in terms of average values parameter and metric itself there is a convergence. On average, all frequencies from 2 hours to 15 minutes give similar first and second moments values, while 15 minutes results into distinct behavior than medium frequencies that mimic each other. The increased heterogeneity in order imbalance with increased frequency, discussed above, in this model averages, since agents that create heterogeneity come randomly at different time intervals and throughout time offset each other. That is why, if we are interested in the metric as an overall average market indicator there might be no need to consider super high aggregation frequencies. However, the arguments for the model, breaking after some frequency still applies.

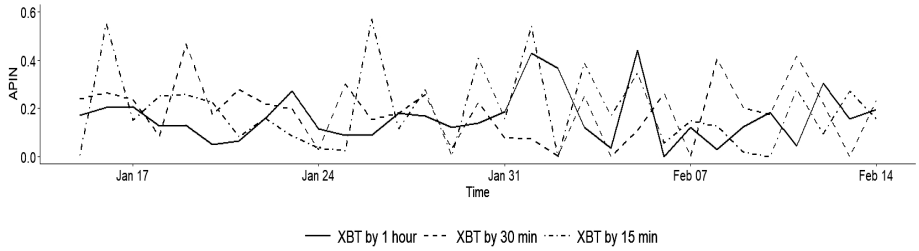


Fig. 15. APIN XBT

Table 14.

APIN BTC Averages (Different data)

Ticker	Avg(α)	Avg(δ)	Avg(PIN)	sd(α)	sd(δ)	sd(θ)	sd(PIN)
XBT day	1.00	0.50	1.00	0	0.5	0.05	0.1
XBT 2 hours	0.69	0.50	0.44	0.19	0.31	0.21	0.08
XBT 1 hour	0.66	0.48	0.45	0.26	0.34	0.26	0.1
XBT 30 min	0.66	0.52	0.41	0.35	0.42	0.32	0.14
XBT 15 min	0.63	0.5	0.46	0.38	0.42	0.37	0.17
XBT 5 min	1.00	0.50	1.00	0	0.5	0.05	0.1
XBT 1 min	0.69	0.50	0.44	0.19	0.31	0.21	0.08

7.3. Crypto Empirics vs Stocks Empirics

Table 15 summarizes empirical results of this paper and compares them to the existing ones in the literature. Primarily, we are interested in comparison of stocks and cryptocurrencies and our results indicate that, on average, cryptocurrencies tend to have higher and more volatile informed trading activity than stocks. In terms of methodology used, the most intriguing comparison is with the Duarte and Young (2018) which adopts APIN in their paper. They obtain much higher values for stocks than our research identifies for cryptos. We consider that difference comes from the methodology and the simplified estimation procedure utilized by Duarte and Young. They use only 10 random starting points for optimization and make almost no changes to the likelihood. In subsection 4.2 we actually compare their approach to our modified one on simulated data and conclude that our new estimation procedure provides more accurate results. Thus, they might get biased estimates, making our outputs less comparable. Moreover, the majority of the papers does not have data about the seller and buyer initiated trades for the stock markets, which is directly available for the cryptocurrencies. To solve this issue they have to apply Lee and Ready (1991) algorithm which makes their results less sound. According to K. Ellis et al. (2000) and E. Theissen (2001), the Lee and Ready algorithm proves to have an accuracy of 72.8% and 81.05%, respectively, in their tests. Such level of precision in turn provides additional noise to the model's input, leading to a less valid output.

Table 15.

Cryptocurrencies vs Stocks

Asset	№ tickers	Exchange	Year	Paper	Metric	Avg Value	Std dev
Crypto	10	BitMex	2022	Authors	PIN (day)	0.0011	0.02
Crypto	10	BitMex	2022	Authors	PIN (2h)	0.2917	0.14
Crypto	10	BitMex	2022	Authors	PIN (1h)	0.3190	0.15
Crypto	10	BitMex	2022	Authors	PIN (30 min)	0.3758	0.15
Crypto	10	BitMex	2022	Authors	APIN (day)	0.2740	0.16
Crypto	10	BitMex	2022	Authors	APIN (2h)	0.2081	0.13

Continuation

Asset	№ tickers	Exchange	Year	Paper	Metric	Avg Value	Std dev
Crypto	10	BitMex	2022	Authors	APIN (1h)	0.2018	0.15
Crypto	10	BitMex	2022	Authors	APIN (30 min)	0.2017	0.16
Stocks	243	NYSE	2012	Duarte, Hu and Young (2018)	PIN (day)	0.382	0.135
Stocks	243	NYSE	2012	Duarte, Hu and Young (2018)	APIN (day)	0.455	0.092
Stocks	10	NYSE	Nov. 1990 – Jan. 1991	Dey and Radhakrishna (2015)	PIN (day)	0.209	0.048
Stocks	15	SIBE	2009	Abad and Yague (2012)	PIN (day)	0.172	0.029
Stocks	12	BSE	Feb. 2020 – Oct. 2020	Cepoi et al (2023)	PIN (week)	0.227	NA

From subsection 7.1 we inferred that illiquid cryptocurrencies tend to have higher metrics' values than more liquid ones. This is consistent with stocks markets – results of Easley et al. (1996) about US stocks and of Abad and Yague (2012) about Spanish equities.

8. Conclusion

This paper presents the new methodology to estimation of APIN and makes an empirical study of PIN and APIN behavior in the cryptocurrency markets. We show that our method of initial parameters and modified likelihood may substantially increase the accuracy of the model, compared to its simplified procedure currently applied in the literature. Moreover, new estimation procedure for both PIN and APIN was introduced by aggregating buyer and seller trades at higher frequencies than day to estimate metrics' daily values. It was shown that theoretically and empirically it is optimal to use medium frequencies for trades bucketing not higher than 15–30 minutes as the models may become more volatile due to more heterogeneous order imbalances.

Availability of high frequency trading data as well as large proportion of sophisticated traders and algorithmic funds make cryptocurrency markets a promising alternative for informed trading analysis. We find that as in the stocks markets both metrics indicate that less liquid cryptocurrencies tend to have higher probability of informed trading. However, cryptocurrencies as an asset class have higher PIN values than stocks in different regions.

We also conduct several event studies and conclude that at medium aggregation frequencies PIN and APIN provide some meaningful results, however, there is no single consistent pattern, making it ambiguous whether models really work and can detect the presence of informed trading precisely.

This work could be modified by considering several alternative issues that would make the analysis even more profound. Firstly, it is worth considering more cryptocurrencies and wider time horizon. Secondly, this work does not cover the topical debate whether PIN model measures

informed trading or it is rather a liquidity measure, that would make its comparison to APIN more solid. Thirdly, it is work extending optimal frequency analysis by incorporating the liquidity consideration which might affect the order flow distribution. Fourthly, one could look at metrics' behaviour not on the daily basis, but within trading day, that is apply our aggregation technique to conduct trading bucketing within, for instance, one hour to predict hourly PIN and APIN. Finally, it is worth analyzing in terms of cryptocurrency markets whether order imbalance alone is enough to detect informed trading or we should consider alternative models that take into consideration more input variables. One of the potential approaches is OWR model [Odders-White, Ready, 2008], based on Kyle (1985) framework. However, it is hardly applicable since it implies the usage of overnight returns the notion of which does not exist in the cryptocurrency markets as they accept trades 24 hours.

* *
*

References

Abad D., Yagüe J. (2012) From PIN to VPIN: An Introduction to Order Flow Toxicity. *The Spanish Review of Financial Economics*, 10, pp. 74–83.

Aktas N., de Bodt E., Declerck F., Van Oppens H. (2007) The PIN Anomaly around M & A Announcements. *Journal of Financial Markets*, 10, pp. 169–191.

Amihud Y. (2002) Illiquidity and Stock Returns: Cross Section and Time-Series Effects. *Journal of Financial Markets,* 5, pp. 31–56.

Andersen T.G., Bondarenko O. (2014) VPIN and the Flash Crash. *Journal of Financial Markets*, 17, pp. 1–46.

Brennan M., Huh SW, Subrahmanyam A. (2016) Asymmetric Effects of Informed Trading on the Cost of Equity Capital. *Management Science*, 62, pp. 2460–2480.

Cepoi C.O., Dragoťa V., Trifan R. et al. (2023) Probability of Informed Trading during the COVID-19 Pandemic: The Case of the Romanian Stock Market. *Financial Innovation*, 9, 34.

Collin-Dufresne P., Fos V. (2012) Do Prices Reveal the Presence of Informed Trading? *Journal of Finance Forthcoming*.

Dey M.K., Radhakrishna B. (2015) Informed Trading, Institutional Trading, and Spread. *Journal of Economics and Finance, 39*, pp. 288–307.

Duarte J., Young L. (2009) Why Is PIN Priced? *Journal of Financial Economics*, 91, pp. 119–138.

Easley D., Hvidkjaer S., O'Hara M. (2010) Factoring Information into Returns. *Journal of Financial and Quantitative Analysis*.

Easley D., Kiefer N., O'Hara M., Paperman J. (1996) Liquidity, Information, and Infrequently Traded Stocks. *Journal of Finance*, 51, pp. 1405–1436.

Easley D., López de Prado M., O'Hara M. (2012) Flow Toxicity and Liquidity in a High-Frequency World. *Review of Financial Studies*, 25, pp. 1457–1493.

Easley D., O'Hara M. (1987) Price, Trade Size, and Information in Securities Markets. *Journal of Financial Economics*, 19, pp. 69–90.

Easley D., Engle R., O'Hara M., Wu L. (2002) Time-Varying Arrival Rates of Informed and Uninformed Trades, Working Paper. *Journal of Financial Econometrics*, 6.

Ellis K., Michaely R., O'Hara M. (2000) The Accuracy of Trade Classification Rules: Evidence from NASDAQ. *Journal of Financial and Quantitative Analysis*, 35, 4, 529551.

Ersan O., Grachem M. (2023) *A Methodological Approach to the Computational Problems in the Estimation of Adjusted PIN Mode*. Pre-print paper.

European Central Bank (2012) *Virtual Currency Schemes*.

Felez-Vinas E., Johnson L., Putnins T.J. (2022) *Insider Trading in Cryptocurrency Markets*. Available at SSRN: https://ssrn.com/abstract=4184367 or http://dx.doi.org/10.2139/ssrn.4184367

Feng W., Wang Y., Zhang Z. (2017) Informed Trading in the Bitcoin Market. *Finance Research Letters*, 26, pp. 63–70.

Gan Q., Wei W.C., Johnstone D. (2015) A Faster Estimation Method for the Probability of Informed Trading Using Hierarchical Agglomerative Clustering. *Quantitative Finance*, 15, 11, pp. 1805–1821.

Glosten L., Milgrom P. (1985) Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders. *Journal of Financial Economics*, 13, pp. 71–100.

Hasbrouck J. (2007) Empirical Market Microstructure: The Institutions, Economics, and Econometrics of Securities Trading. *OUP Catalogue*, Oxford University Press.

Hellwig M. (1980) On the Aggregation of Information in Competitive Markets. *Journal of Economic Theory*, 22, pp. 477–498.

Lee Charles M.C, Ready M.J. (1991) Inferring Trade Direction from Intraday Data. *The Journal of Finance*, 46, pp. 733–746.

Kyle A.S. (1985) Continuous Auctions and Insider Trading. *Econometrica*, 53, pp. 1315–1335.

Lin H.-W.W., Ke W.-C. (2011) A Computing Bias in Estimating the Probability of Informed Trading. *Journal of Financial Markets*, 14, 4, pp. 625–640.

Odders-White E.R., Ready M.J. (2008) The Probability and Magnitude of Information Events. *Journal of Financial Economics*, 87, pp. 227–248.

Park M., Chai S. (2020) *The Effect of Information Asymmetry on Investment Behavior in Cryptocurrency Market*. 10.24251/HICSS.2020.494.

Theissen E. (2001) A Test of the Accuracy of the Lee/Ready Trade Classification Algorithm. *Journal of International Financial Markets*, Institutions and Money, Elsevier, 11, 2, pp. 147–165.

Westland J.C. (2021) Trade Informativeness and Liquidity in Bitcoin Markets. *PLOS ONE*, 16, 8.

Yan Y., Zhang S. (2012) An Improved Estimation Method and Empirical Properties of the Probability of Informed Trading. *Journal of Banking & Finance,* 36, pp. 454–467.

Информированная торговля на рынках криптовалют

Кузьмин Григорий Иванович[1], **Булатов Алексей Эрикович**[2]

[1] Аспирант, Аспирантская школа по экономике,
Национальный исследовательский университет «Высшая школа экономики»,
11, Покровский б-р, Москва, 109028, Россия.
E-mail: gikuzmin@hse.ru

[2] Ординарный профессор,
Национальный исследовательский университет «Высшая школа экономики»,
11, Покровский б-р, Москва, 109028, Россия.
E-mail: aboulatov@hse.ru

В данной работе эмпирически оценивается асимметрия информации на рынках криптовалют с использованием метрик вероятности информированного трейдинга (PIN) и Adjusted PIN. Эти рынки, характеризующиеся высоким уровнем алгоритмической торговли и большими объемами данных высокой частоты, представляют собой перспективную среду для анализа информированной торговой активности. Мы вводим модифицированную процедуру оценки для Adjusted PIN, устраняя ошибки с плавающей запятой и проблемы с локальными экстремумами, что повышает точность модели по сравнению с традиционными подходами, широко применяемыми в литературе. Кроме того, мы предлагаем альтернативный метод агрегации сделок на более высоких частотах, чем традиционное ежедневное агрегирование, с целью повышения эффективности как моделей PIN, так и Adjusted PIN. На основе анализа как симулированных, так и реальных данных мы показываем, что агрегирование общего числа покупок и продаж за день приводит к менее значимым оценкам из-за шумности входных данных, что затрудняет выявление активности информированных трейдеров. Истинную оптимальную частоту агрегирования сделок еще предстоит исследовать, так как с увеличением частоты возрастает гетерогенность дисбалансов ордеров, а конкретные частоты, на которых действуют информированные трейдеры, пока не установлены. Наконец, в ходе ряда эмпирических исследований мы оцениваем поведение метрик, выявляя, что неликвидные криптовалюты демонстрируют относительно более высокие вероятности информированного трейдинга. Этот вывод соответствует аналогичным результатам, полученным на рынках акций.

***Ключевые слова*:** PIN; Adjusted PIN; вероятность информированной торговли; рынки криптовалют; асимметрия информации.

JEL Classification: G12, G14, D53, C13.