The Probability of Informed Trading: A measure for informed trading on Bitcoin exchanges?

Thesis for the
Bachelor Seminar in Empirical Economics
of Prof. Dr. Martin Biewen and Prof. Dr. Joachim Grammig

January 9, 2020

Marie-Sophie Lappe
Eduard-Spranger-Straße 57/5
72076 Tübingen
Student ID: 4065274
International Economics, B.Sc.

Contents

List of Figures

List of Tables

1 Introduction

With 300,000 verified transactions per day (blockchain.com, 2020) Bitcoin is currently the most heavily traded cryptoasset. As cryptocurrencies are on the radar of various institutions and even the European Central Bank currently investigating the influence of cryptoassets such as Bitcoin on the financial market (Pan, 2019), the attention surely has not yet quite died down after the big ”Bitcoin-Hype” in 2017. The unique microstructure of the Bitcoin market possibly has implications for existing models (e.g. Kyle (1985); Admati and Pfleiderer (1988)) or even demands the adjustment of many classical models to fit the heavily traded and highly volatile nature of the digital asset.

This thesis in particular concerns itself with a measure of information asymmetry, namely the Probability of Informed Trading (PIN), first introduced by Easley, Kiefer, O’Hara, and Paperman (1996). I attempt to use the measure in a multivariate setting to compare three Bitcoin exchanges following the methodology of Grammig, Schiereck, and Theissen (2001).

I intend to estimate the PIN for a highly discussed virtual asset, namely Bitcoin. Furthermore, research on Bitcoin and its market structure is referred to in the next section. What makes Bitcoin intriguing to many is its relation to information. As Bitcoin is neither a currency as it was intended to be nor a regular asset, I aim to reveal its unique properties by relating it to a traditional measure.

The PIN has been a widely used measure in a variety of contexts. These include, but are not limited to, Easley, Hvidkjaer, and O’Hara (2002) relating the PIN to cross-sectional asset returns, or Kang (2010) investigating the impact of seasonality of the PIN-return relationship. Vega (2006) uses the PIN measure to estimate the influence of private and public information on the post-announcement drift of stock prices. Easley, De Prado, and O’Hara (2011) use an extension of the PIN to measure order flow toxicity during the flash crash in May, 2010.

The measure has also been subject to many alterations and extensions. While it is a simple measure with respect to necessary data, it is also prone to biases and issues especially when one works with highly frequently traded assets. Many of these issues revolve around computing accuracy and efficiency. Easley, Hvidkjaer, and O’Hara (2010), as well as Lin and Ke (2011) deal with floating point exception. There are various methods to compute initial values to circumvent the issue of boundary solutions and local maxima (Yan and Zhang, 2010; Gan, Wei, and Johnstone, 2015). I will elaborate on these issues in Section 5.

Further studies include alternations and extensions of the PIN model itself. Duarte and Young (2009) develop a model to account for positively correlated buys and sells. Chang, Chang, and Wang (2014) construct a dynamic intra-day measure of the PIN. Easley, López de Prado, and O’Hara (2012) use a volume-synchronized PIN measure to estimate order flow toxicity. I intend to investigate how these alterations and extensions might help implementing the PIN as an information asymmetry measure for a heavily traded asset with various unique characteristics.

The structure of this thesis is as follows: In section 2 the market structure of Bitcoin is examined, including possible implications for the PIN measure. The model itself is introduced in section 3, followed by section 4 concerning itself with the data and descriptive statistics. In section 5 the estimation methodology is set forth and an applied extension is disclosed. The subsequent section concerns itself with further possible extensions suggested by existing literature. Lastly, the findings are concluded.

I find highly unstable estimation results for the extension on three markets. Various reasons for this are explored while I try to provide possible solutions. I believe to have found an extension that accounts for issues encountered in the estimation. The examination of univariate and bivariate results show low PIN measures on the markets respectively, while when applying restrictions on the model, significantly higher values for the PIN are found.

2 Market structure

Since the introduction of Bitcoin in 2009, its market went through various changes and developments (for more information see Böhme, Christin, Edelman, and Moore (2015)). Bitcoin was initially introduced as "a purely peer-to-peer version of electronic cash [which] would allow online payments to be sent directly from one party to another without going through a financial institution" (Nakamoto et al., 2008), whereas of today, it seems to be classified as a virtual asset rather than a currency (see Cheah and Fry (2015); Yermack (2015); Baur, Hong, and Lee (2018) and more). The unpredictable volatility of prices, caused by the price being at the mercy of market sentiments, is one reasoning behind this notion (Cheah and Fry, 2015). Said market sentiments are also a possible explanation for unsettling trading intensity. Dwyer (2015) finds that the average monthly volatility of Bitcoin prices is higher than for gold or fiat currencies. Unlike many stock exchanges Bitcoin exchanges do not offer volatility interruptions (Dimpfl, 2017), displaying the absence of regulatory instances, which is a core characteristic of the market. It can be argued that regulatory action does not have much impact on the informational market structure of Bitcoin. The idea of "Insider Trading" is simply not existent in Bitcoin markets, as it is not backed by any institution or company that could illegally leak inside information.

Bitcoin has another special feature which might be of interest for this analysis: its missing anonymity. While it seems fairly anonymous at first glance, Bitcoin is constructed in such a way that theoretically every transaction can be traced back to its initiator. Due to privacy concerns of many users, institutions (so-called 'mixers') intend to mask the traders identity by basically pooling multiple sets of transactions (for further information see Böhme et al. (2015)). However, according to Androulaki, Karame, Roeschlin, Scherer, and Capkun (2013), privacy protection can be breached such that still almost half of the users profiles can be traced back. As informed traders tend to avoid transparent market structures in order to profit from their information advantage, the PIN is expected to be fairly low on Bitcoin exchanges.

Another interesting thought arises with regard to the actual information an insider could carry. Since Bitcoin is not linked to any company or backed by an institution, it differs from other assets in terms of background information. It has been found to not follow any price movements of storages of values, like fiat currencies or gold (Yermack, 2015). Based on the intuition of Bitcoin not holding any informational value, there should be no difference between informed and uninformed market participants, as there simply is no information affecting the price. The underlying value it carries seems to be determined by the users' perception of its value (Dimpfl, 2017). Further research on this topic even suggests that the fundamental value of Bitcoin is zero and its price is thus highly reliant on investors sentiments (Cheah and Fry, 2015). This supports the intuition that the PIN estimates for the exchanges should be fairly low, possibly even close to zero. One might argue that carrying information on market sentiments in general, which seem to be a determinant of its value, could be a source for informed trading. This, however, is unlikely, as predicting market sentiments on a volatile asset like Bitcoin revolves around speculating rather than having accurate information.

This thesis intends to compare three exchanges in particular, which seem to be fairly similar in their structure: Bitfinex, Kraken and Poloniex. Trading is available 24 hours on a daily basis, hence a trading day equals a normal calendar day. Trades are executed via an open limit-order book with all bids and asks being displayed with their respective volume in real time. Information on overall price and volume movements is also publicly available on websites like `coindesk.com` or `coinbase.com`. The exchanges' user interfaces also provide different graphs displaying statistics revolving around price and volume development. This often serves as a basis for market participants' predictions on price developments.

While Bitfinex and Kraken offer various order types, Poloniex seems to be more simplistic in its trading process. Bitfinex e.g. offers hidden orders (bitfinex.com), which facilitates concealing one's intentions. Hidden orders might contaminate the estimation of the PIN measure, as the PIN is based on a model where traders are assumed to trade using only market orders, while in reality informed traders may often use limit or hidden orders (Aktas, De Bodt, Declerck, and Van Oppens, 2007). Kraken launched its own dark pool for Token and Cryptocurrency trading (kraken.com, 2015). Dark pools might attract seemingly informed traders, causing them to self-select themselves out of the open limit-order book market. This further supports the notion of the PIN being low.

The three platforms make use of a similar fee structure where fees decrease in monthly traded volume and differing charges with respect to trader role. Maker or liquidity provider fees are lower compared to Taker fees as seen in Table 1. The fee structure encourages traders to trade higher volumes or trade on a more frequent basis as the fees decrease in monthly traded volume. The fee difference between the trader roles favours liquidity traders who do not value immediacy as highly and act on the ”Maker” side of the market. Liquidity traders classify as uninformed traders, hence, they are favoured to make up for possible losses arising when trading with an informed individual. In turn, informed traders do not trade for liquidity reasons and are charged higher fees. They are believed to prefer trading immediately upon information arrival. Obviously, this only occurs when they assume to trade without much of a price impact. If they do fear their order would cause the price to shift, informed traders would use options like hidden orders. The fee structure thus might cause informed traders to be inclined to partially trade for liquidity reasons to reduce their fees, if that is to maximize their returns in total. This could possibly cause the overall number of liquidity trades to rise, keeping the PIN measure low.

Table 1: Fee structure

Fees in percent	Maker	Taker
Bitfinex	0.000 - 0.100	0.055 - 0.200
Kraken	0.000 - 0.160	0.100 - 0.260
Poloniex	0.000 - 0.150	0.100 - 0.250

The table depicts fees with decrease in monthly traded volume and differ for those who provide (”Maker”) and those who demand (”Taker”) liquidity.

Source: `https://www.bitfinex.com/fees` [2019-12-27]

`https://www.kraken.com/en-us/features/fee-schedule` [2019-12-27]

`https://poloniex.com/fees/` [2019-12-27]

3 Model

The PIN model used in this thesis was first introduced by Easley et al. (1996, 2002) and is based on Easley and O’Hara (1987). It can be classified as a sequential trade model following the principle of Glosten and Milgrom (1985) with a market maker at its center. Furthermore the extension by Grammig et al. (2001) is applied to compare the different trading platforms and test the PIN simultaneously. This section is intended to briefly describe the model and its extension. It will be put to the data in the next section. While Grammig et al. (2001) extend the model to two markets (floor and screen trading on the German stock market), the model applied in this thesis intends to extend their work in such a way that it is fitting for estimating the PIN on three Bitcoin exchanges. The different trading platforms are indicated by an index $n \epsilon [b, k, p]$ respectively.

A period over $i \epsilon [1, I]$ trading days is examined, within which time is indexed by $t \epsilon [1, T]$. Trading is believed to be continuous during trading days. Both informed and uninformed individuals trade a single risky asset by issuing market orders with a competitive and risk neutral market maker who posts bid and ask prices according to her beliefs about the fundamental value of the asset. The market maker is believed to not face any inventory holding costs. Before trading begins each day, nature determines whether an information event occurs. Said events are independently distributed across trading days and occur with probability α. The nature of the event itself can be "good" or "bad" with probability $(1 - \delta)$ and δ, respectively. On a positive (negative) event day the value of the asset is $\overline{V}_i > V_i^*$ ($\underline{V}_i > V_i^*$). If no information event occurs the stock holds the value V_i^*, which naturally lies between the high and low signal values. This process occurs anew everyday with information being fully incorporated into the market makers' beliefs.

Informed traders know whether an information event occurred, as well as the nature of said event and trade accordingly. They buy (sell) when the signal is high (low). Informed traders arrive at the market following a Poisson process with rate μ_n. They only trade when an information event occurs. Uninformed traders' arrivals are determined by an independent Poisson process with rate ε_n. Such traders are believed to trade for liquidity reasons and receive no signals whatsoever concerning information events, hence trades always occur following the arrival rate. Therefore, arrival rates are higher on days where an information event takes place $(\varepsilon_n + \mu_n)$, whereas the arrival rate is simply ε_n, if no informational event took place. Following Grammig et al. (2001), arrival rates are given and allowed to differ across markets $(n \epsilon [b, k, p])$. This is motivated by differing trading activity, which I will further elucidate in the next section.

Overall, trades arrive depending on the information event that occurred that day. On a no event day, only uninformed buyers and sellers arrive according to their rate ε_n.[1] On a good (bad) event day, the sell arrival rate is ε_n $(\varepsilon_n + \mu_n)$, while the buy arrival rate is $\varepsilon_n + \mu_n$ (ε_n), as informed traders react to the respective events. The market maker knows these order arrival rates as well as the probabilities α and δ, however, she does not observe whether or which event actually occurs. The market maker is assumed to update her beliefs using the Bayes' rule and the information she obtains

[1] Many papers allow buy and sell arrival rates for uninformed traders to differ, I however abstract from this distinction due to simplicity.

from each trade. For example, after a trade initiated by a seller arrives, she would revise the probability for a negative information event upwards. At the beginning of the trading day, i.e. $t = 0$, her beliefs concerning the event probabilities, namely no news (N), bad news (B) and good news (G) are the unconditional probabilities: $P_n(t = 0) = (1 - \alpha, \alpha\delta, \alpha(1 - \delta))$. She updates her beliefs conditional on the trade history, such that the probabilities prior to the trade at time t on market n can be written as $P_n(t) = (P_{n,N}(t), P_{n,B}(t), P_{n,G}(t))$. Hence, the expected value of the asset, conditional on the prior trade history, can be written as

$$E[V_i \mid t] = P_{n,N}(t)V_i^* + P_{n,B}(t)\underline{V}_i + P_{n,G}(t)\overline{V}_i. \tag{1}$$

The market maker sets her quotes according to the expected value of the asset conditional on the trade history at time t *and* on the trade direction of the arriving order. In particular, assuming the market maker observes an incoming sell, the zero profit expected bid price

$$b_n(t) = \frac{P_{n,N}(t)\varepsilon_n V_i^* + P_{n,B}(t)(\varepsilon_n + \mu_n)\underline{V}_i + P_{n,G}(t)\varepsilon_n\overline{V}_i}{\varepsilon_n + P_{n,B}(t)\mu_n}, \tag{2}$$

is derived using the posterior probabilities of the news events as weights for the respective possible realizations of the fundamental value on trading day i. Similarly, the ask price in case of an incoming buy

$$a_n(t) = \frac{P_{n,N}(t)\varepsilon_n V_i^* + P_{n,B}(t)\varepsilon_n\underline{V}_i + P_{n,G}(t)(\varepsilon_n + \mu_n)\overline{V}_i}{\varepsilon_n + P_{n,G}(t)\mu_n}, \tag{3}$$

can be computed. These prices reflect the expected value of the asset conditional on the history prior to t and the observed trade.

From equation (2) and (3) it can be inferred that the spread

$$a_n(t) - b_n(t) = PI_{n,buy}(t)(\overline{V}_i - E[V_i \mid t]) + PI_{n,sell}(t)(E[V_i \mid t] - \underline{V}_i) \tag{4}$$

also implicitly depends on the arrival rates of informed and uninformed traders. $PI_{n,buy}(t)$ and $PI_{n,sell}(t)$ denote the conditional probabilities of informed trading and are defined as

$$PI_{n,buy} = \frac{\mu_n P_{n,G}(t)}{\varepsilon_n + \mu_n P_{n,G}(t)} \tag{5}$$

and

$$PI_{n,sell} = \frac{\mu_n P_{n,B}(t)}{\varepsilon_n + \mu_n P_{n,B}(t)}. \tag{6}$$

Taking a look at boundary solutions, where $\mu_n = 0$ or $\varepsilon_n = 0$, various implications can be made. In the case of no informed traders, the trade would carry no information and the spread would be zero, as bid and ask price would be equal to the prior expected value of the asset. If instead there are no uninformed traders active, $a_n(t) = \overline{V}_i$ and

$b_n(t) = \underline{V}_i$, hence, informed traders would also not participate since they will not be able to profit from their informational advantage. The market would then shut down.

The unconditional probability of an informed trade can be calculated via the probabilities of an informed trade conditioned on the trade directions (see equations (5) and (6)) which then results in

$$PI_n(t) = \frac{\mu_n(P_{n,G}(t) + P_{n,B}(t))}{2\varepsilon_n + \mu_n(P_{n,G}(t) + P_{n,B}(t))}. \tag{7}$$

The unconditional probability for informed trading at the opening can therefore be computed as

$$PI_n(0) = \frac{\alpha\mu_n}{2\varepsilon_n + \alpha\mu_n}, \tag{8}$$

so that only model-specific parameters remain. All parameters $\theta_n = (\alpha, \delta, \varepsilon_n, \mu_n)$ can thus be estimated via Maximum Likelihood as proposed by Easley et al. (1996). In the next section insights on the data itself will be provided, followed by elucidation on the estimation methodology.

4 Data

In this thesis three Bitcoin exchanges are compared, namely Bitfinex, Kraken and Poloniex. Data from said Bitcoin trading platforms over a period of 75 trading days is examined. The datasets include information on trade indicators, volume, prices and timestamps of each transaction from August 23rd 2018 to November 5th 2018.[2] The timestamps follow universal coordinated time (UTC) and the trading pair is BTC-USD. October 14th is excluded due to overly extreme trading activity on that day, leaving us with 74 trading days.

Literature suggests the number of trading days to be around 60, as this usually is in line with business cycle being measured in quarter years (Recktenwald, 2018). Problems might arise when estimating over a substantially longer period of time, since the data might then be serially correlated (Easley et al., 1996). Trading days are additionally separated into 3 hour intervals and thereby 8 PIN measures per market are computed in order to identify possible daily patterns in a later part of the next section. Boehmer, Grammig, and Theissen (2007) suggest that the PIN is often biased due to a misclassification of trades. Since the data includes the trade direction, buys and sells are believed to be accurately labelled, ruling out this issue. To rule out biases due to split trades, an estimation with a modified dataset is also conducted, where those trades which occurred at the same time with the same price are aggregated. The results do not differ with respect to their consistency, hence the unmodified full sample dataset is used.

[2] The data is obtained from Coindesk `https://www.coindesk.com/price/bitcoin`.

Table 2 presents some descriptive statistics for the sampling period. The average price is computed from the transaction prices normalized for one Bitcoin. Bitfinex and Kraken report similar figures for price and traded volume (per trade) during the considered sample period, while Poloniex differs. A lower average price as well as a lower average volume per trade is reported. A core difference between the former two and Poloniex are the order types. Poloniex offers only market, limit and stop orders, while Bitfinex and Kraken additionally offer post orders (limit order is either posted or cancelled) and other order types. These might cause the discrepancy in the traded volume as posting higher volume orders on the two markets are not necessarily associated with a higher loss risk, since the order can be customized to avoid potentially high losses.

Table 2: Descriptive Statistics

	Bitfinex	Kraken	Poloniex
average price in USD	6257.53	6222.85	6059.83
average traded volume in BTC	0.4193	0.3870	0.1064
average volume per day in BTC	15941.33	2899.643	1218.765
average number of trades per day	35941	7108	10494

The table depicts the descriptive statistics for the sample period of 74 trading days. The average price is calculated from transaction prices and normalized to the equivalent of one Bitcoin.

Average volume per day and average number of trades show that out of the three exchanges Bitcoin is most heavily traded on Bitfinex. According to Easley et al. (1996), more frequently traded stocks yield lower PINs. At a first glance this finding does not seem to be related to the estimation as only one asset is considered, however, theory from traditional microstructure research can be invoked. Traditional microstructure implies that depth can be linked to informed trading (see Kyle (1985); Admati and Pfleiderer (1988)). An active market creates a favourable environment for informed traders as they can hide their intentions behind uninformed trading. This would obviously suggest that informed trading should occur on Bitfinex to a higher extent. As the PIN measures a relation of informed trading to overall trading,

$$\text{PIN} = \frac{\text{Expected number of information-based transactions}}{\text{Expected total number of transactions}},$$

it can be argued that the measure would in fact even be lower on markets with higher trading activity, as the overall arrival rate for uninformed traders should be fairly high, keeping the measure itself low. This is in line with results presented in the next section. It is important to keep in mind that the PIN does not measure the absolute informed trading activity, but more so the *risk* of encountering an informed trader from a market maker position. This relative measure is thus expected to be fairly equal on the similarly structured markets.

5 Estimation

Following Easley et al. (1996) and Grammig et al. (2001), the parameters are estimated using Maximum Likelihood Estimation. It should be noted again that using daily number of sells and buys as underlying data is sufficient for the estimation of the PIN. While this makes the PIN a simple measure of information asymmetry in terms of required data, its simplicity also bears problems which will be additionally elaborated on in this and the following section.

In order to simultaneously test the PIN on different Bitcoin trading platforms the model extension proposed by Grammig et al. (2001) is used. It is plausible to assume that the probabilities concerning information events (α and δ) are equal across the trading platforms, since the same asset is considered. In this paper wants to examine whether the PIN differs across Bitcoin exchanges, hence it makes sense to allow for different arrival rates of informed and uninformed traders (ε_n and μ_n). This is also in line with the actual trading activities on the three exchanges as these seem to substantially differ. Accordingly, the base model of this thesis (Model 1 in the following) restricts α and δ to be equal across markets and allows ε_n and μ_n to differ. To test the validity of this model and its specifications, models with different restrictions are constructed. Model 2 allows for all parameters to differ across the trading platforms. Model 3 instead restricts all parameters to be equal across markets. Likelihood Ratio is conducted to test these models against each other and to draw conclusions in the following subsections.

Making use of the assumption that trader arrival rates follow independent Poisson, multiple likelihood functions for each type of trading day i can be set up, specifically no news (N)

$$\begin{aligned} l_{N,i} = {} & e^{-\varepsilon_b T}\frac{(\varepsilon_b T)^{B_{b,i}}}{B_{b,i}!} e^{-\varepsilon_b T}\frac{(\varepsilon_b T)^{S_{b,i}}}{S_{b,i}!} \\ & e^{-\varepsilon_k T}\frac{(\varepsilon_k T)^{B_{k,i}}}{B_{k,i}!} e^{-\varepsilon_k T}\frac{(\varepsilon_k T)^{S_{k,i}}}{S_{k,i}!} \\ & e^{-\varepsilon_p T}\frac{(\varepsilon_p T)^{B_{p,i}}}{B_{p,i}!} e^{-\varepsilon_p T}\frac{(\varepsilon_p T)^{S_{p,i}}}{S_{p,i}!}, \end{aligned} \tag{9}$$

where only uninformed traders arrive at the market.

Furthermore, a bad news (B) day can be depicted as

$$\begin{aligned} l_{B,i} = {} & e^{-\varepsilon_b T}\frac{(\varepsilon_b T)^{B_{b,i}}}{B_{b,i}!} e^{-(\varepsilon_b+\mu_b)T}\frac{((\varepsilon_b+\mu_b)T)^{S_{b,i}}}{S_{b,i}!} \\ & e^{-\varepsilon_k T}\frac{(\varepsilon_k T)^{B_{k,i}}}{B_{k,i}!} e^{-(\varepsilon_k+\mu_k)T}\frac{((\varepsilon_k+\mu_k)T)^{S_{k,i}}}{S_{k,i}!} \\ & e^{-\varepsilon_p T}\frac{(\varepsilon_p T)^{B_{p,i}}}{B_{p,i}!} e^{-(\varepsilon_p+\mu_p)T}\frac{((\varepsilon_p+\mu_p)T)^{S_{p,i}}}{S_{p,i}!}, \end{aligned} \tag{10}$$

while a good news (G) day can be expressed as

$$\begin{aligned} l_{G,i} = {} & e^{-(\varepsilon_b+\mu_b)T}\frac{((\varepsilon_b+\mu_b)T)^{B_{b,i}}}{B_{b,i}!} e^{-\varepsilon_b T}\frac{(\varepsilon_b T)^{S_{b,i}}}{S_{b,i}!} \\ & e^{-(\varepsilon_k+\mu_k)T}\frac{((\varepsilon_k+\mu_k)T)^{B_{k,i}}}{B_{k,i}!} e^{-\varepsilon_k T}\frac{(\varepsilon_k T)^{S_{k,i}}}{S_{k,i}!} \\ & e^{-(\varepsilon_p+\mu_p)T}\frac{((\varepsilon_p+\mu_p)T)^{B_{p,i}}}{B_{p,i}!} e^{-\varepsilon_p T}\frac{(\varepsilon_p T)^{S_{p,i}}}{S_{p,i}!}. \end{aligned} \tag{11}$$

These functions can be combined by weighting them with the respective event probabilities to obtain the conditional maximum likelihood function for day i as

$$l_i = \alpha\delta l_{B,i} + \alpha(1-\delta) l_{G,i} + (1-\alpha) l_{N,i}. \tag{12}$$

Due to the frequency of Bitcoin trades and the large number of daily buys and sells, computational issues arise when estimating the PIN parameters. The factorials and the term including $\varepsilon_n + \mu$ in equation (9) to (11) cannot be computed due to floating-point exception. There are different suggestions in the literature to solve this issue by making use of factorization. Easley et al. (2010) suggest a factorization (EHO-factorization) when testing the relationship between returns and the PIN measure. Lin and Ke (2011) introduce another factorization (LK-factorization) to further increase computing efficiency. While both methods follow the same principle, I find that only using LK factorization yields finite results. Literature also finds that LK is more stable for heavily traded assets (Lin and Ke, 2011; Gan et al., 2015), hence LK factorization will be used. The log-likelihood function for trading day i can be set up accordingly,

$$\begin{aligned} log(l_i) = {} & log[\alpha\delta\, exp(e_{1b,i} + e_{1k,i} + e_{1p,i} - e_{maxb,i} - e_{maxk,i} - e_{maxp,i}) \\ & + \alpha(1-\delta)\, exp(e_{2b,i} + e_{2k,i} + e_{2p,i} - e_{maxb,i} - e_{maxk,i} - e_{maxp,i}) \\ & + (1-\alpha)\, exp(e_{3b,i} + e_{3k,i} + e_{3p,i} - e_{maxb,i} - e_{maxk,i} - e_{maxp,i})] \\ & + B_{b,i} log((\varepsilon_b+\mu_b)T) + B_{k,i} log((\varepsilon_k+\mu_k)T) + B_{p,i} log((\varepsilon_p+\mu_p)T) \\ & + S_{b,i} log((\varepsilon_b+\mu_b)T) + S_{k,i} log((\varepsilon_k+\mu_k)T) + S_{p,i} log((\varepsilon_p+\mu_p)T) \\ & - 2T(\varepsilon_b + \varepsilon_k + \varepsilon_p) + e_{maxb,i} + e_{maxk,i} + e_{maxp,i} \\ & - log(B_{b,i}! B_{k,i}! B_{p,i}! S_{b,i}! S_{k,i}! S_{p,i}!) \end{aligned} \tag{13}$$

with $e_{1n,i} = -\mu_n T - B_{n,i} log(1 + \mu_n T/\varepsilon_n T)$, $e_{2n,i} = -\mu_n T - S_{n,i} log(1 + \mu_n T/\varepsilon_n T)$, $e_{3n,i} = -B_{n,i} log(1+\mu_n T/\varepsilon_n T) - S_{n,i} log(1+\mu_n T/\varepsilon_n T)$ and $e_{maxn,i} = max(e_{1n,i}, e_{2n,i}, e_{3n,i})$. The last term containing the factorials can be dropped as it is a constant with respect to the parameter vector $\theta = (\alpha, \delta, \varepsilon_b, \mu_b, \varepsilon_k, \mu_k, \varepsilon_p, \mu_p)$. The log-likelihood function used for the estimation is then obtained as

$$L(Y|\theta) = \sum_{i=1}^{I} log(l_i), \tag{14}$$

where $Y = \{B_{b,i}, B_{k,i}, B_{p,i}, S_{b,i}, S_{k,i}, S_{p,i}\}_{i=1}^{I}$.
The factorization is mainly based on two principles:

- Computing e^{x+y} or $sign(x)e^{log(|x|)+y}$ is more stable than computing $e^x e^y$ or xe^y.
- Absolute computing errors are larger for functions with a large value first-order derivative.

The first principle is mainly explained by the fact that numbers larger than e^{710} lead to floating-point exception. By using e.g. e^{x+y} or $sign(x)e^{log(|x|)+y}$ instead of $e^x e^y$ or xe^y, large numbers can be circumvented in the exponential (assuming that y is negative). The second principle mainly states that input values for $exp(\cdot)$ and $log(\cdot)$ should not be too positively large and too positively small, respectively. These two principles are ensured by subtracting the $e_{maxn,i}$ in the logarithmic term of the equation and rearranging terms.

The estimation is conducted with the statistical software R. As a guideline and benchmark for initial values used in the numeric maximization, the pinbasic package is used (Recktenwald, 2018), which yields estimates for a univariate estimation following Easley et al. (1996) and Lin and Ke (2011). It can be argued that this method is too approximative and prone to be biased. Research suggests various ways to deal with this issue: Yan and Zhang (2010) deal with boundary solutions which arise due to the estimation of parameters depicting probabilities by using a grid search algorithm for choosing initial values. Hierarchical agglomerative clustering has also been used as a way to deal with boundary solutions and local maxima, also providing initial values (Gan et al., 2015; Ersan and Alıcı, 2016). However, these options will not be included as it would extend the scope of this thesis.

The maxLik package (Henningsen and Toomet, 2011) is used for the Maximum Likelihood estimation. It has various options concerning the maximization method and the computation of the Hessian which is needed for the numeric derivation of standard errors. In this estimation the maximization algorithm yielding the highest Maximum Likelihood value is adopted, which in this case is the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm. Standard errors are estimated by using the Berndt-Hall-

Hall-Hausman (BHHH) algorithm, which seems to be more reliant than obtaining standard errors via the numeric BFGS-Hessian, as the second derivatives do not have to be numerically approximated. The expected second derivatives as the content of the information matrix are approximated by the covariance matrix of the first derivatives vector (see (Greene, 2012, p.562)). It is important to note that the estimation of standard errors is still highly approximative.

Results from the estimation for the base model using the full sample are reported in Table 4 in Appendix A. Starting values were supposed to be taken from the resulting estimates from the pinbasic package. This turned out to be problematic for the full sample, since these initial values did not yield finite results for every trading day function, hence starting values had to be adjusted. The reported results are thus unreliable and an attempt to use different starting values revealed that the function has a very restricted domain of definition. Another issue of infinite standard errors being reported (BFGS) as well as unreportable standard errors computed via the Hessian indicate a misspecification or very noisy data. The BHHH-Hessian denotes rows and columns containing only zeros, causing the matrix to be singular. This is due to the parameter δ not being iterated over. This indicates a misspecification in the model which causes the likelihood function to not depend on the parameter δ. The same applies for the aggregated datasets for which neither pinbasic initial values were applicable, nor were all parameters iterated over when using different starting values.

The interval datasets using the pinbasic initial values did not yield results either, ruling out the idea that the absolute number of buys and sells are too high for this estimation to work. Using the naturally lower absolute values from the interval datasets does not cause the estimation to yield finite results. This does however not rule out the possible problem of the variation within one dataset or across the three datasets being too high.

Note again that the way of choosing initial values is highly simplistic and these issues could possibly be resolved by using more sophisticated methods like grid search (Yan and Zhang, 2010) or HAC (Gan et al., 2015) to avoid boundary solutions and local maxima.

5.1 Misspecification and computational issues

To rule out a misspecification due to the combination of the extension on various markets and the LK factorization, a bivariate model is set up to then attempt to replicate the results of Grammig et al. (2001) using their dataset.[3] The estimates yield similar results. The difference can be explained by the usage of the LK factorization. Standard

[3] The data for the stock Bayer was kindly provided by Professor Grammig and Professor Dimpfl (University of Tübingen).

errors also differ, most likely due to the different maximization algorithm.[4] This gives us reason to believe that the model for the bivariate case using LK factorization is specified in a correct manner.

The question where the unstable results originate from thus still remains. Another possibility are the discrepancies within the data. As stated earlier, Bitfinex is by far the most active market, which is also reflected in the dataset. Therefore, the parameters for the exchanges are estimated first in the bivariate environment. Testing Kraken and Poloniex yields somewhat stable results, e.g. iterations over all parameters and evaluable function at pinbasic starting values for all model variations for the full sample dataset. The results can be found in Table 8 to 10 for the different models later used in the Likelihood Ratio tests.

When testing Bitfinex against one of the other markets, it is noticeable that initial values cannot be implemented from the pinbasic single estimation results, as the domain of definition seems to be quite restricted. This aligns with the earlier problem when simultaneously testing all three markets. As absolute values do not seem to be the issue (see earlier subsection), the relative number of buys and sells could be an issue with Bitfinex having too high numbers in comparison to Kraken or Poloniex or the high variance within the number of trades themselves.

Therefore Kraken and Poloniex are tested with a fictional dataset resulting from dividing the Bitfinex dataset by 10. This dataset then has roughly the same absolute dimension as Kraken and Poloniex. The pinbasic initial values do not yield finite values for all trading days either, as the function cannot be evaluated at said initial values. This possibly indicates that either the high variance of the number of trades in the Bitcoin datasets themselves or the extension on three markets cause the function to not yield finite values in its initial gradient. These infinite values tend to apply to trading days with extreme (high or low) trading activity. In theory, this flaw could be eradicated by taking out multiple outliers. As this method would however erase the distinct feature of the Bitcoin market being highly volatile and contaminate our results, it is refrained from doing so in this setting.

The issue is likely to lie in the extension on the three markets as the estimation equation yields a very restricted domain of definition, which obviously biases the results. Gan, Wei, and Johnstone (2017) argue that the mixture of Poisson distributions causes said restrictions. The density mass of the overlapping distributions depends on the arrival rates of traders. Substantially differing arrival rates for differently categorized trading days cause the overlapping mass to be quite small. For similar arrival rates the overlapping masses are bigger, however, extreme values are then not accounted for. This is a possible explanation for infinite values resulting in the likelihood equation, as days

[4] Grammig et al. (2001) used BFGS for the entire estimation using GAUSS CML library.

with extreme trading activity are necessary to identify the nature of the information event that day. They will however not be accounted for in the overlapping density mass. Since the PIN measure relies on these extreme days to identify information events, this poses a deficiency causing the issues mentioned in this section. The estimation hence is also prone to boundary solutions and local maxima. As already mentioned, it might be worth looking into the grid search algorithm introduced by Yan and Zhang (2010) or the HAC approach by Gan et al. (2015) to work around this issue in particular. While it does not solve the concern on the mixture of Poisson distributions itself, it provides a mean to circumvent issues resulting from the mixture.

5.2 Interpretation of results

The overall restriction of the domain of definition in the parallel trading environment on three markets causes the likelihood function to not yield finite values for the starting values obtained by using the pinbasic package. Single and bivariate estimation results are thus analysed to gain a general overview on the PIN for Bitcoin markets. Possible extensions of the model that might result in stable estimates for the extension on three markets are suggested in the next section. The arrival rates can be interpreted as arrival rates per hour, as the constant is set to $T = 24$ ($T = 3$ for the interval dataset).

Univariate results are reported in Table 5 to 7 (see Appendix A). These results are obtained using the pinbasic package (Recktenwald, 2018). The estimators are transformed in such a way that they are comparable to the bivariate setting reported in Table 8 to 10 (see Appendix A). Bitfinex denotes a surprisingly high estimate for δ, indicating the estimated probability for a "bad" event day. As the probability for an information event (α) to occur is fairly low to begin with, the PIN is low despite a relatively high arrival rate for informed traders (μ). The sell arrival rate of uninformed traders is lower than the buy arrival rate, indicating that uninformed buyers seem to be the most dominant traders on the market. This differs from the results on Kraken and Poloniex, as uninformed sellers and informed traders seem to be the most prominent agents. When taking into account the substantially lower probability for a "bad" event day δ on Kraken and Poloniex, an interesting thought arises: overall, uninformed traders are more likely to trade in the "wrong" direction when it comes to an information event occurring on the respective markets. The differing δ on Bitfinex compared to Kraken and Poloniex might have interesting implications for the Bitcoin market. Intuitively, the probability should be equal across the markets, since the same asset is traded on all of them. But as the prediction of market sentiments is the main source for information in the Bitcoin market, the differing probabilities imply that traders' sentiments differ across exchanges. The effective direction is ambiguous, but it might indicate that specific traders self-select into different exchanges interpreting market signals and predicting the sentiments differently.

The probability for an information event α overall seems to be fairly similar across markets, as well as the PIN itself. This would support the base model, where arrival rates are allowed to differ while the event probabilities are restricted to be equal. Problematic to this notion is the parameter δ. While it is not significant on Poloniex and only significant on a 5%-Level on Kraken, the values differ substantially across markets, which might be an indicator as of why the multivariate estimation failed, as said parameter was not iterated over. One should also note that Poloniex denotes the highest PIN measure due to a higher event probability α and a high informed trader arrival rate. A possible explanation might be the differing order types across markets. As mentioned earlier, Bitfinex and Kraken offer hidden orders or dark pools, causing informed trading to be less likely to be identified by the model.

Table 8 to 10 report the estimates for the different models when parallel testing the PIN on Kraken and Poloniex. The values are all significant when standard errors are computed with the BHHH algorithm. Note again that these standard errors are imprecise due to non-analytic computation. The test statistics are also reported. The focus is laid on interpreting the results reported in Table 8, as univariate results indicate that Kraken and Poloniex have fairly similar event probabilities while the arrival rates differ. In line with the absolute number of trades on both markets, the arrival rates for informed and uninformed traders are higher for Poloniex. The higher arrival rate on Poloniex for informed traders combined with a low probability of an information event occurring (α) causes the PIN to be below 0.2. The estimates for Kraken report higher arrival rates of uninformed traders, which is in line with the initial intuition based on different order types and causes the PIN_k to be below PIN_p. The values are similar to the univariate results and initial values do not evoke boundary solutions, thus justifying the specification of the model.

Easley and O'Hara (1987) argue that informed traders are more likely to trade larger amounts. This is certainly not in line with the PIN measure observed in this estimation, as Poloniex reports a higher PIN, despite the lower per trade volume. This fact still holds for the aggregated datasets, where estimated PINs are slightly lower on both markets. The initial explanation for the lower volume due to diverse order types is still believed to be true, combined with the idea that the differing order types simply cause informed trades not to be revealed on Kraken to the extent they are on Poloniex.

While the overall arrival rate of informed traders is relatively high, the estimates suggest that there are few information events occurring, such that probability for an informed trade to arise is low. The estimated probability for a "bad" news day (δ) is also fairly low, which would suggest that "good" news days are more common. Overall the PIN can be interpreted as being fairly low, which is in line with the initial intuition.

Here it is worth considering how the PIN measure identifies ”good” or ”bad” days. The model classifies the respective information event days by the order flow imbalance, meaning the difference in buys and sells on active trading days compared to days where overall number of trades is low (no information event day). The inconsistency of the Bitcoin market makes the identification of a low trading activity day difficult. I examine the order flows and its implications for the measure in the next section.

A very peculiar result is fairly similar across all variations of estimation environments (bivariate and multivariate) and datasets (full sample or interval): The third Model (results reported in Table 10), which restricts all parameters to be equal, yields a significantly higher PIN. For all other models and variations the PIN measure seems to lie between close to zero and 0.2, whereas for Model 3 the value hovers around 0.7. When treating the three market specific datasets as basically one market, the PIN is significantly higher, which is due to a higher event probability α. This possibly indicates that when trying to reveal information on Bitcoin one should consider looking at all of the markets, since treating the markets as one causes more informed trading to be revealed. Furthermore, this implicates that informed traders observe the movement on different markets and *then* choose the market to trade at. The earlier notion, that market sentiments differ across markets, is then possibly supported. Differing market sentiments would cause the markets to move dissimilarly and thus attract different traders. This would however give rise to the question about information in Bitcoin markets once again. If market sentiments are a criteria for choosing the market to trade at *conditioned* on the information one might carry, the term information once again becomes ambiguous. The different models are compared using Likelihood Ratio tests in the following subsection.

5.3 Likelihood ratio tests

As mentioned in the beginning of this section, the different models are now tested against each other in the bivariate setting. The results can be found in Table 3. When testing the different models against each other, the fully restricted model (Model 3) is surprisingly found to be the dominant model. Intuitively, this might not seem too surprising when examining the bivariate environment where Kraken and Poloniex are simultaneously tested, as arrival rates and event probabilities seemed to be fairly similar. What makes the result problematic is the fact that the same holds for bivariately testing Bitfinex and Kraken. The arrival rates and the event probability δ substantially differ for the two markets. The Likelihood Ratio test however suggests that Model 3 is the dominant one.

Testing Model 1 (base model) and Model 2 (no restrictions) against Model 3 (fully restricted) yields p-values of zero. This also holds for a downward scaled dataset.

These results indicate that the three markets yield the same arrival rates and event probabilities. While the latter seems reasonable, the former does not. An additionally surprising result is the actual parameter values for the third model. As mentioned in the earlier subsection, the probability of an information event occurring (α) is substantially higher (Table 10), contradicting the results from the first and the second model where the probabilities show similarly low results. The arrival rates of uninformed ε and informed traders μ are also substantially different from the results in Table 8 and 9 (Model 1 and 2). It should be kept in mind that the computational issues mentioned in the beginning of this section cause the ambiguous results in the bivariate setting as well.

Table 3: Likelihood Ratio Test

	Model 1 vs. Model 2	Model 1 vs. Model 3	Model 2 vs. Model 3
Full sample	0.0000	0.0000	0.0000
Scaled sample	0.0000	1.443592e-55	5.080956e-47

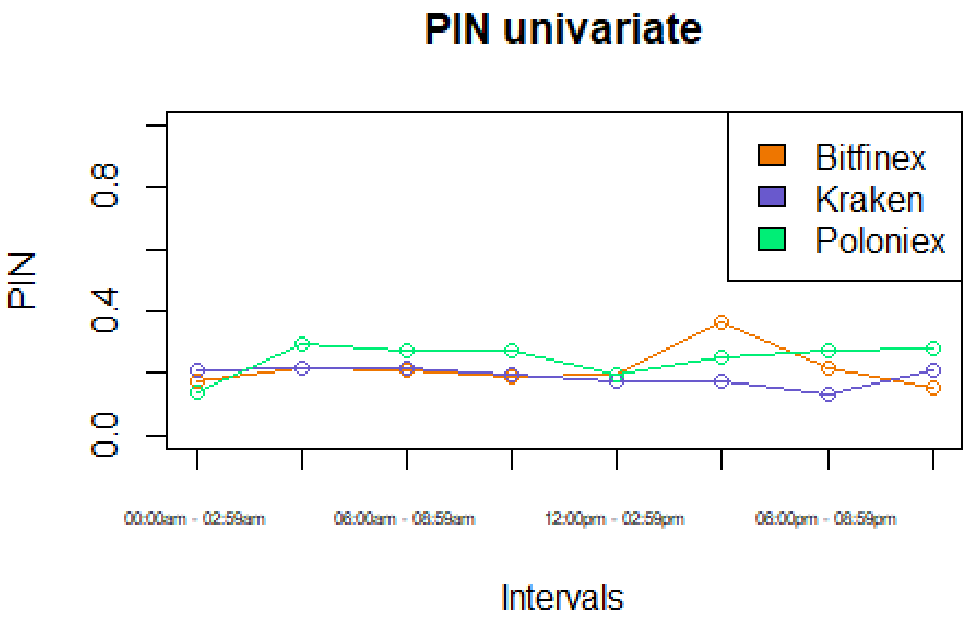
The table depicts results from Likelihood Ratio tests of the bivariate models. Model 1 restricts the event probabilities α and δ to be equal and allows arrival rates ε_n and μ_n to differ across markets. In Model 2 all parameters are unrestricted, while Model 3 is the fully restricted model.
The scaled sample is the original full sample divided by 1000.

5.4 Intra-day PINs

In modern financial markets where information inflow occurs quite frequently, it is reasonable to assume that possible implications concerning the information content remain undetected when only examining data on a daily basis. Thus, the PINs for interval data are computed to evaluate the intra-day development of the measure on Bitcoin markets. The data is divided into 8 intervals, denoting 3 hour periods. The results are depicted in Figure 1. Few fluctuations are observed during the day, indicating that informed trading is not dependent on the time of the day.

Figure 1 depicts the results for the univariate estimation on the three markets. The overall estimates are slightly higher compared to the full sample results. This implies that the division into intervals caused more informed trading to be revealed. As noted earlier, one can observe fairly similar PINs throughout the day and across markets. Poloniex denotes the highest PIN for most intervals, which is in line with the initial estimates derived from the full sample. Bitfinex and Kraken seem to develop similarly and diverge in the 6th interval denoting the period from 03:00pm to 05:59pm, which corresponds to night time in China (CST) and early noon in the US (EST). Bitfinex denotes a high PIN, surpassing Poloniex. Examining the cause of this, I find the probability of a "bad" event to significantly fall during that time period, causing the PIN to rise. The arrival rates differ in their absolute dimension, though the relative relation of μ being high and the arrival rates being lower remains the same in the former and successive interval. The event probability falling indicates that "bad" news are less likely to arrive at said time interval, or traders *interpret* information events more optimistically. Another possible intuition originates from the idea that informed traders act on their information in specific time periods. Admati and Pfleiderer (1988) argue that during periods where the market is thick (usually around the 5th and 6th interval in this dataset, UTC) liquidity traders, as well as informed traders are likely to be active in the market to avoid price impact. They argue that it is ambiguous whether uninformed traders are kept out of the market in said periods due to the higher level of informed trading and overall probability of informed trading will thus increase, or the opposite is the case. Possible evidence is found for the former on the Bitfinex market, explaining the sudden rise in the PIN measure.

Figure 1: Intra-day Probability of Informed Trading



The figure depicts the evolution of the PIN during a trading day. The results are obtained from the univariate testing environment on each market. The day is divided into three hour intervals starting at 00:00am and ending at 11:59 pm.

The shape of the figure is in line with the initial intuition as the model does not imply a gradually increasing probability of informed trading over time, since information is incorporated into the trading process over the whole trading day with informed and uninformed traders arriving at their respective rates.

Chang et al. (2014) suggest a dynamic intra-day measure (DPIN) for the probability of informed trading while accounting for various trade characteristics like trade size. They divide the trading day into 15-minute intervals. When controlling for trade size, they find a U-shaped intra-day pattern for the PIN. Without taking into account specific nuances of trades, the authors find a similar intra-day pattern to the one in this thesis, namely a constant PIN over the day. The constant pattern could have various reasons. One intuition lies behind the arrival of information itself. When the day is split into intervals, each interval is treated as its own trading day. Informed traders are believed to prefer trading as soon as they receive an information signal. Information is believed to arrive randomly, hence informed trades should be allocated randomly throughout the day, causing the PIN to be constant across the day. This is also in line with the sequential trade model idea that traders are chosen probabilistically.

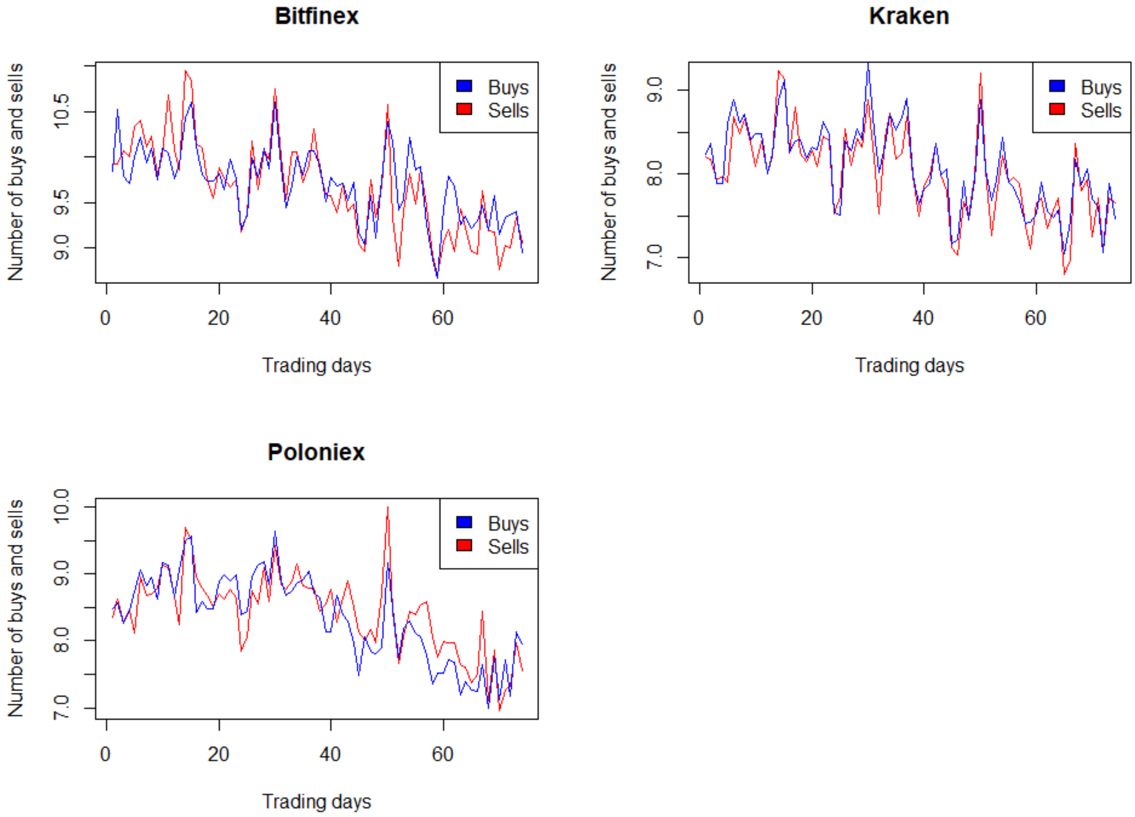
6 Empirical fit of Bitcoin data and the PIN

As stated in the introductory section, the PIN has been subject to various extensions and alterations to fit the model to data or purposes. The appeal of the PIN measure is mainly its low cost of data, as the number of daily buys and sells suffices to estimate all parameters. Another perk is the fact that it is an information asymmetry measure that is not reliant on returns, which makes it applicable for regression analysis trying to estimate returns. However, its simplicity bears various sources for misspecification and poor data fit.

Gan et al. (2017) analyse the fit of the PIN measure to empirical data. They construct a goodness of fit measure for different PIN models and find that estimated PIN models generally have poor fit for empirical data. They suggest that the model should be applied with caution. Often the PIN is used as an indicator for information asymmetry in regression analysis without being investigated further. It is important to note that poor empirical fit does not necessarily cause the measure to be inapplicable, but that it should instead be applied with caution perhaps by comparing it to other information asymmetry measures which do not solely rely on order flow.

I tries to identify traits in the dataset to suggest model extensions that might help fitting the empirical data to an extended estimated PIN model. Figure 2 depicts the logarithmized number of daily number of buys and sells for our sampling period. As already noted, Bitfinex denotes a higher overall daily number of trades followed by Poloniex. The overall volatility of trading intensity stands out. The daily number of trades decreases slightly over the sampling period. Buys and sells show a positive correlation, indicating that trading activity increases on both sides of the market similarly. This gives insights about the informational value of said trades. One would expect the number of sells to increase on a ”bad” day, while number of buys should remain either constant or decrease, depending on whether and which information event occurred the day before. This is a key assumption in the PIN model, as the number of buys and sells are used to identify ”good” and ”bad” news days. These patterns cannot be observed in this thesis’ dataset, giving us a possible explanation as of why the model does not yield stable results.

Figure 2: Daily number of buys and sells



The graphics depict the logarithmized number of daily buys and sells over the full sampling period on the respective exchanges. Each panel depicts one market.

Duarte and Young (2009) suggest a model extension to account for positive covariance of buys and sells, and large buy and sell volatility by introducing a positive liquidity shock parameter that simultaneously impacts buy and sell order flows (from here on DY-PIN). I highly suggest further research in this direction when computing the PIN using Bitcoin data, as it accounts for symmetric order flow shocks. These shocks are believed to shape Bitcoin data (see Figure 2), as market sentiments play an important role in trading Bitcoin. Once the market displays more activity, this activity rises on both sides of the market. The drawback of this method is the introduction of various new parameters, possibly further restricting the domain of definition. One should thus consider using grid search (Yan and Zhang, 2010) or hierarchical agglomerative clustering (Gan et al., 2015) to obtain starting values [5].

Duarte, Hu, and Young (2018) show that the DY-PIN outperforms the traditional PIN model with respect to fit to the actual data. It should also be noted, that the DY-PIN implied variances still do not reflect the data accurately. The authors therefore set up a generalized PIN model (GPIN), which is also based solely on order flow, hence keeping the low cost of data. The basic idea is not to mix two discrete models as Duarte and Young (2009) do, but to rather let the noise trade intensity vary continuously instead of differentiating between high and low trading intensity. This approach also circumvents the restricted domain of definition due to the mixture of Poisson distributions. The trade intensity follows a Gamma distribution while the arrival rates are still Poisson-distributed with the trade intensity as their parameter. Informed traders are believed to follow noise traders and to trade a specific additional portion of the uninformed traders' intensity. For further explanations see Duarte et al. (2018). The GPIN seems to outperform the DY-PIN and the traditional PIN model in terms of explaining variability of noise trade in the data itself. It seems to be a possible match to this thesis' data, where noise trade is in fact quite variable. Furthermore, the issue of the restricted domain of definition could be resolved, as arrival rates are more dynamic, which is better fitting to this model. The main issue of the restricted domain of definition is avoided as arrival rates are not constant as before, but are allowed to vary.

Easley et al. (2012) suggest another extension where they create a volume synchronized probability of informed trading (VPIN) to estimate order flow toxicity.[6] They intend to overcome the issues that accompany high frequency data by measuring stochastic time in volume units. They incorporate volume as an important indicator for information in a trade, and the estimation does not require numerical estimation. For more intuition see Easley et al. (2012). The VPIN however, does not explicitly estimate the probability of informed trading as the extensions investigated before, but it rather can be used as a measurement for risk of subsequent large price movements. Therefore it does not pose as a viable extension in this case.

[5] Implementing the idea of symmetric shock into this model, I find that starting values still seem to be an issue, hence the recommendation of implementing one of the above algorithms.

[6] "Order flow is regarded as toxic when it adversely selects market makers who are unaware that they are providing liquidity at a loss." (Easley et al., 2012)

7 Conclusion

The PIN measure has been a core part of research with respect to information asymmetry. It is a widely used measure and the paper by Easley et al. (1996) has been cited almost 1,800 times.[7] It was thus attempted to use this measure to apply it to a highly discussed asset: Bitcoin. The estimation was intended to be applied simultaneously on three markets, namely Bitfinex, Kraken and Poloniex. The extended model I applied here revealed various issues. The extension on three markets significantly restricted the domain of definition for the estimation equation, causing the estimation application to be problematic. Various algorithms from literature which possibly solve this issue were referred to.

I thus made use of a bivariate and univariate setting to estimate the PIN measure. Fairly low PIN values for the univariate case were found, which is in line with the initial intuition of the PIN being low as Bitcoin is believed to not carry information. A surprising result was found for the bivariate case: The model restricting all parameters to be equal seemed to be the fitting specification, while the PIN for said model yielded high values, providing evidence against my initial intuition. This possibly indicates a characteristic of the Bitcoin market, where various exchanges have to be investigated to draw implications for information incorporation.

Intra-day PINs revealed a fairly constant structure in line with previous research and supporting the sequential trade model, where traders are chosen probabilistically and informational advantage is then traded on immediately.

Furthermore, various extensions of the PIN measure were investigated to find a possibly better fitting theoretical model for our data. Duarte and Young (2009) suggest the DY-PIN accounting for a positive relation between buys and sells, which in fact could be useful for this dataset. A better performing extension of the DY-PIN is the GPIN (Duarte et al., 2018). Based on theoretical reasoning I find the GPIN to be a fitting model for Bitcoin datasets. It accounts for volatility, positive covariance in the order flow and possibly solves computational issues related to the restricted domain of definition. Further research on Bitcoin data using the Probability of Informed Trading is suggested to be conducted with the GPIN as it accounts for various issues the traditional PIN model does not.

This thesis revealed weaknesses of the extension of the PIN model on three Bitcoin markets. Multiple extensions from the literature were introduced and linked to said weaknesses. Implications for the microstructure of Bitcoin were unfolded and confirmed.

[7] according to Google Scholar [03-01-2020]

References

Anat R Admati and Paul Pfleiderer. A theory of intraday patterns: Volume and price variability. *The Review of Financial Studies*, 1(1):3–40, 1988.

Nihat Aktas, Eric De Bodt, Fany Declerck, and Herve Van Oppens. The pin anomaly around m&a announcements. *Journal of Financial Markets*, 10(2):169–191, 2007.

Elli Androulaki, Ghassan O Karame, Marc Roeschlin, Tobias Scherer, and Srdjan Capkun. Evaluating user privacy in bitcoin. In *International Conference on Financial Cryptography and Data Security*, pages 34–51. Springer, 2013.

Dirk G Baur, Kihoon Hong, and Adrian D Lee. Bitcoin: Medium of exchange or speculative assets? *Journal of International Financial Markets, Institutions and Money*, 54:177–189, 2018.

bitfinex.com. Features. URL `https://www.bitfinex.com/features`. accessed [2019-12-27].

blockchain.com, 2020. URL `https://www.blockchain.com/de/charts`. accessed [2020-01-03].

Ekkehart Boehmer, Joachim Grammig, and Erik Theissen. Estimating the probability of informed trading - does trade misclassification matter? *Journal of Financial Markets*, 10(1):26–47, 2007.

Rainer Böhme, Nicolas Christin, Benjamin Edelman, and Tyler Moore. Bitcoin: Economics, technology, and governance. *Journal of Economic Perspectives*, 29(2):213–38, May 2015. URL `http://www.aeaweb.org/articles?id=10.1257/jep.29.2.213`.

Sanders S Chang, Lenisa V Chang, and F Albert Wang. A dynamic intraday measure of the probability of informed trading and firm-specific return variation. *Journal of Empirical Finance*, 29:80–94, 2014.

Eng-Tuck Cheah and John Fry. Speculative bubbles in bitcoin markets? an empirical investigation into the fundamental value of bitcoin. *Economics Letters*, 130:32–36, 2015.

Thomas Dimpfl. Bitcoin market microstructure. *Available at SSRN 2949807*, 2017.

Jefferson Duarte and Lance Young. Why is pin priced? *Journal of Financial Economics*, 91(2):119–138, 2009.

Jefferson Duarte, Edwin Hu, and Lance A. Young. A comparison of some structural models of private information arrival. *Available at SSRN 2564369*, December 2018. URL `https://ssrn.com/abstract=2564369`.

Gerald P Dwyer. The economics of bitcoin and similar private digital currencies. *Journal of Financial Stability*, 17:81–91, 2015.

David Easley and Maureen O'Hara. Price, trade size, and information in securities markets. *Journal of Financial economics*, 19(1):69–90, 1987.

David Easley, Nicholas M. Kiefer, Maureen O'Hara, and Joseph B. Paperman. Liquidity, information, and infrequently traded stocks. *The Journal of Finance*, 51(4): 1405–1436, 1996. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1996.tb04074.x`.

David Easley, Soeren Hvidkjaer, and Maureen O'Hara. Is information risk a determinant of asset returns? *The journal of finance*, 57(5):2185–2221, 2002.

David Easley, Soeren Hvidkjaer, and Maureen O'Hara. Factoring information into returns. *Journal of Financial and Quantitative Analysis*, 45(2):293–309, 2010.

David Easley, Marcos M Lopez De Prado, and Maureen O'Hara. The microstructure of the "flash crash": flow toxicity, liquidity crashes, and the probability of informed trading. *The Journal of Portfolio Management*, 37(2):118–128, 2011.

David Easley, Marcos M López de Prado, and Maureen O'Hara. Flow toxicity and liquidity in a high-frequency world. *The Review of Financial Studies*, 25(5):1457–1493, 2012.

Oguz Ersan and Aslı Alıcı. An unbiased computation methodology for estimating the probability of informed trading (pin). *Journal of International Financial Markets, Institutions and Money*, 43:74–94, 2016.

Quan Gan, Wang Chun Wei, and David Johnstone. A faster estimation method for the probability of informed trading using hierarchical agglomerative clustering. *Quantitative Finance*, 15(11):1805–1821, 2015.

Quan Gan, Wang Chun Wei, and David Johnstone. Does the probability of informed trading model fit empirical data? *Financial Review*, 52(1):5–35, 2017.

Lawrence R Glosten and Paul R Milgrom. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of financial economics*, 14 (1):71–100, 1985.

Joachim Grammig, Dirk Schiereck, and Erik Theissen. Knowing me, knowing you: Trader anonymity and informed trading in parallel markets. *Journal of Financial Markets*, 4(4):385 – 412, 2001. ISSN 1386-4181. URL `http://www.sciencedirect.com/science/article/pii/S1386418101000180`.

William H. Greene. *Econometric Analysis*. Pearson Education LTD., 7th edition, 2012.

Arne Henningsen and Ott Toomet. maxlik: A package for maximum likelihood estimation in R. *Computational Statistics*, 26(3):443–458, 2011. URL `http://dx.doi.org/10.1007/s00180-010-0217-1`.

Moonsoo Kang. Probability of information-based trading and the january effect. *Journal of Banking & Finance*, 34(12):2985–2994, 2010.

kraken.com. Introducing the kraken dark pool, 2015. URL `https://blog.kraken.com/post/259/introducing-the-kraken-dark-pool/`. accessed [2019-12-26].

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

Hsiou-Wei William Lin and Wen-Chyan Ke. A computing bias in estimating the probability of informed trading. *Journal of Financial Markets*, 14(4):625 – 640, 2011. ISSN 1386-4181. URL `http://www.sciencedirect.com/science/article/pii/S1386418111000176`.

Satoshi Nakamoto et al. Bitcoin: A peer-to-peer electronic cash system. 2008.

David Pan. Ecb official says digital currency could be an alternative to cash, November 2019. URL `https://www.coindesk.com/ecb-official-says-digital-currency-could-be-an-alternative-to-cash`. accessed [2020-01-03].

Andreas Recktenwald. *pinbasic: Fast and Stable Estimation of the Probability of Informed Trading (PIN)*, 2018. URL `https://CRAN.R-project.org/package=pinbasic`. R package version 1.2.2.

Clara Vega. Stock price reaction to public and private information. *Journal of Financial Economics*, 82(1):103–133, 2006.

Yuxing Yan and Shaojun Zhang. An improved estimation method and empirical properties of the probability of informed trading. *SSRN Electronic Journal*, 03 2010. doi: 10.2139/ssrn.890486.

David Yermack. Is bitcoin a real currency? an economic appraisal. In *Handbook of digital currency*, pages 31–43. Elsevier, 2015.

A Additional Tables

Table 4: Full sample, Model 1 (multivariate)

Parameter	Estimate	Std. error (BFGS)	Std. error (BHHH)	t-Stat (BHHH)
α	0.036	Inf	NA	NA
δ	***0.5	Inf	NA	NA
ε_b	753.752	Inf	NA	NA
μ_b	4999.988	Inf	NA	NA
ε_k	156.238	Inf	NA	NA
μ_k	3000.002	Inf	NA	NA
ε_p	250.145	Inf	NA	NA
μ_p	1999.989	Inf	NA	NA

Maximum Likelihood Value: 33036208

$PIN_b = 0.1068494$ $PIN_k = 0.2572182$ $PIN_p = 0.126021$

The table reports estimates for the probability of an information event day α, the probability for a "bad" information event day δ, uninformed trader arrival rates ε and informed trader arrival rates μ. The indices denote the respective markets for which the estimation is conducted (b for Bitfinex, k for Kraken and p for Poloniex).

*** denotes no reported iteration over the parameter, causing the Hessian to have zeros in the respective column and row. BHHH standard errors are thus not reportable as the determinant of the Hessian is zero.

Table 5: Full sample, Bitfinex

Parameter	Estimate	Std. error	t-Stat	Pr(>t)
α	0.365	0.056	6.520	0.000
δ	0.778	0.100	7.791	0.000
ε $(Buys)$	716.959	3.144	228.043	0.000
ε $(Sells)$	533.764	3.258	163.808	0.000
μ	676.403	7.380	91.651	0.000

Maximum Likelihood Value: 627815.8 $PIN_b = 0.1648076$

The table reports estimates for the probability of an information event day α, the probability for a "bad" information event day δ, uninformed trader arrival rates ε and informed trader arrival rates μ.

Table 6: Full sample, Kraken

Parameter	Estimate	Std. error	t-Stat	Pr(>t)
α	0.360	0.057	6.304	0.000
δ	0.151	0.070	2.154	0.031
ε $(Buys)$	114.789	1.860	61.716	0.000
ε $(Sells)$	133.900	1.410	94.985	0.000
μ	131.832	3.688	35.743	0.000

Maximum Likelihood Value: 88594.57 $PIN_k = 0.1602786$

The table reports estimates for the probability of an information event day α, the probability for a "bad" information event day δ, uninformed trader arrival rates ε and informed trader arrival rates μ.

Table 7: Full sample, Poloniex

Parameter	Estimate	Std. error	t-Stat	Pr(>t)
α	0.444	0.058	7.660	0.000
δ	0.060	0.042	1.441	0.150
ε $(Buys)$	126.468	1.842	68.661	0.000
ε $(Sells)$	214.874	1.757	122.280	0.000
μ	215.867	3.708	58.210	0.000

Maximum Likelihood value: 144036.4 $PIN_p = 0.2193409$

The table reports estimates for the probability of an information event day α, the probability for a "bad" information event day δ, uninformed trader arrival rates ε and informed trader arrival rates μ.

Table 8: Full sample, Model 1 (bivariate)

Parameter	Estimate	Std. error (BFGS)	Std. error (BHHH)	t-Stat (BHHH)
α	0.3784	Inf	0.0828	4.5706
δ	0.2501	Inf	0.0859	2.9097
ε_k	125.9479	Inf	0.0137	9184.4923
μ_k	116.9752	Inf	0.0432	2705.8022
ε_p	180.4543	Inf	0.0117	15486.3152
μ_p	201.7569	Inf	0.0370	5457.8149

Maximum Likelihood Value: 9715023

$PIN_k = 0.1494615$ $PIN_p = 0.1746046$

The table reports estimates for the probability of an information event day α, the probability for a "bad" information event day δ, uninformed trader arrival rates ε and informed trader arrival rates μ. The indices denote the respective markets for which the estimation is conducted (k for Kraken and p for Poloniex).

Table 9: Full sample, Model 2 (bivariate)

Parameter	Estimate	Std. error (BFGS)	Std. error (BHHH)	t-Stat (BHHH)
α_k	0.3379	Inf	0.0926	3.6510
δ_k	0.2806	Inf	0.1149	2.4413
α_p	0.4058	Inf	0.0963	4.2130
δ_p	0.3679	Inf	0.1029	3.5741
ε_k	126.1395	Inf	0.01441	8752.5147
μ_k	129.8796	Inf	0.05245	2476.3060
ε_p	178.4472	Inf	0.0126	14117.1954
μ_p	198.2244	Inf	0.04090	4846.9977

Maximum Likelihood Value: 9718411

$PIN_k = 0.1481966$ $PIN_p = 0.1839346$

The table reports estimates for the probability of an information event day α, the probability for a "bad" information event day δ, uninformed trader arrival rates ε and informed trader arrival rates μ. The indices denote the respective markets for which the estimation is conducted (k for Kraken and p for Poloniex).

Table 10: Full sample, Model 3 (bivariate)

Parameter	Estimate	Std. error (BFGS)	Std. error (BHHH)	t-Stat (BHHH)
α	0.7838	0.0000	0.0663	11.8226
δ	0.4308	0.0000	0.0675	6.3782
ε	70.0367	0.0000	0.0055	12797.5600
μ	399.5634	0.0232	0.0324	12328.9142

Maximum Likelihood Value: 9615260 $PIN = 0.690959$

The table reports estimates for the probability of an information event day α, the probability for a "bad" information event day δ, uninformed trader arrival rates ε and informed trader arrival rates μ.