An Improved Version of the Volume-Synchronized Probability of Informed Trading (VPIN)[*]

By

Wen-Chyan Ke[a] and Hsiou-Wei William Lin[b]

[*] The authors are indebted to the National Science Council of Taiwan for support under Grant NSC102-2410-H-305-019-. They thank Ivo Welch and two anonymous referees for helpful comments.

[a] Corresponding author; National Taipei University, Department of Finance and Cooperative Management, New Taipei City 23741, Taiwan, phone: 886-2-8674-1111 ext.67712, fax: 886-2-8674-5905, e-mail: wenchyan@gm.ntpu.edu.tw; wenchyan@mail.ntpu.edu.tw

[b] National Taiwan University, Department of International Business, Taipei City 10617, Taiwan, phone: 886-2-3366-4967, fax: 886-2-2362-7203, e-mail: plin@ntu.edu.tw

An Improved Version of the Volume-Synchronized Probability of Informed Trading (VPIN)

Current version: July 2015

Abstract

This study provides a theoretical basis for the transformation of the probability of informed trading (PIN) model to the volume-synchronized PIN (VPIN) setting based on volume buckets. Building on Easley et al. (2011, 2012b), who derive the VPIN metric and provide evidence of its usefulness, we expand the analytical basis of the model and clarify its derivation. We show mathematically that Easley et al.'s VPIN metric becomes unstable for small volume buckets and for infrequent informed trades. In contrast, we use a maximum likelihood estimation to capture the information in volume time, and as a result our improved VPIN mathematical model generates consistent estimates. We also show that the volume time measure helps improve the predictability of VPIN for the flow toxicity.

JEL classification: C52; C13; G14; G12; C51

Keywords: VPIN, PIN, High-Frequency trading, Order flow toxicity

1. Introduction

This study provides a theoretical basis and presents an effective estimation process for the transformation of the probability of informed trading (PIN) model to the volume-synchronized PIN (VPIN) setting based on equal-sized volume buckets. Easley et al. (1996, 2002) develop a microstructure PIN model, and Easley et al. (2011, 2012b) modify the model to create VPIN, which applies to high-frequency markets. VPIN is estimated based on a volume time scale where the basic unit is a fixed-sized volume bucket rather than a constant stretch of calendar time. Easley et al. (2011, 2012b,c) provide empirical evidence that VPIN is useful in monitoring order flow imbalances and conclude that it signals impending market turmoil. Several recent studies also explore the VPIN metric and build on its application (Abad and Yague, 2012; Bethel et al., 2012; Madhavan, 2011; Wei et al., 2013).

However, Andresen and Bondarenko (2014a) conduct mathematical analyses and find that the VPIN metric is imperfect for predicting short-run volatility and the flash crash.[1] Andresen and Bondarenko (2014b) explore the VPIN metric and conclude that the VPIN metric is not theory driven. Given these contradictory findings, we argue, in line with Easley et al. (2011, 2012b), that the VPIN metric is not derived based on a fully specified model but rather serves as a convenient extension of PIN, and therefore a modified procedure of transformation is needed.

[1] Easley et al. (2014) respond to Andresen and Bondarenko's criticism but provide no further analytical evidence.

The goal of this study is not to defend the empirical work regarding VPIN but rather to improve the VPIN theory base by developing a model that illustrates how to transform the PIN model, which is based on calendar time, to the VPIN model, which is based on volume time.

Specifically, this paper demonstrates that PIN and VPIN are different probability measures for informed trading because the volume time, which is the time taken to fill a fixed-sized volume bucket, is stochastic. In Easley et al. (2012b), each equal-sized bucket is equivalent to a random time period of information arrival for VPIN derivation. However, the simplest theory of PIN has information taking place over equal time units, such as a trading day. Namely, PIN is a probability measure for a given time unit whereas VPIN is a probability measure for a given fixed-sized trading volume. To improve on Easley et al.'s (2012b) method of moment estimation (MME) for VPIN, this study proposes incorporating the volume time via maximum likelihood estimation (MLE).

The remainder of this study is organized as follows. Section 2 provides a simple numeric illustration for explaining PIN, VPIN and their difference. Section 3 provides a description of our derived model for VPIN, which is based on the volume bucket. In Section 4, we derive the analytic VPIN metric and show that the VPIN measure is a probability of informed trading in the condition of a fixed trading volume. By contrast, the original PIN estimator generates a probability estimate with respect to a given time interval. Section 5 provides the estimation of

VPIN with volume time via MLE. Section 6 performs a simple empirical test using the trading data of SYP, which are the exchange traded funds of S&P 500 stock index. Finally, Section 7 offers the summary and conclusions.

2. A Simple Numeric Illustration

[TABLE 1 ABOUT HERE]

In Table 1, we construct a numeric example to demonstrate the difference between PIN and VPIN, and the effect of volume time setting on the effectiveness in estimating these measures. In this example, we assume that arrival rates of both uninformed buyers and sellers are equal, that the news type is known, and that one share is one trade. Accordingly, we obtain approximate MLE estimates for PIN and VPIN. Moreover, with no news, we set that buy and sell volumes (V^B and V^S) are equal. Then, Easley et al.'s (2012b) MME, which may generate PIN or VPIN, also perfectly identify the emerging event. Furthermore, the trading imbalance can directly measure the informed trading. Thus, the MME estimates can compare with the MLE estimates.

In Panels A to D of this example, the proportion of event periods is the MLE estimate for the probability of information occurrence. The estimate for uninformed buyers and sellers' arrival rates is the total sum of all buy volumes from no news as well as bad news periods, and

all the sell volumes from no news as well as good news periods, divided by the total sum of corresponding times. The informed traders will be the buyers when the event is good news and will be the sellers when it is bad news. Therefore, the sum of aggregate buy volumes during good news periods and aggregate sell volumes during bad news periods, divided by the total sum of corresponding times, and then minus the estimate for uninformed buyers and sellers' arrival rates, is the estimate of informed traders' arrival rate. Moreover, Easley et al.'s MME is the mean (or sum) of imbalanced trading volumes divided by the mean (or sum) of total trading volumes based on the fixed-length time periods or on the fixed-sized buckets.

For instance, in Panel A of Table 1, these MLE estimates based on time bars are as follows: the probability of information event is 0.4 = 6/15, the uninformed traders' arrival rate is 100 = (18×100+2×(2×75+150))/(18×1+2×(2×1+1)), and the informed traders' arrival rate is 200 = 300 – 100, where 300 = (2×(2×225 +450))/(2×(2×1+1)). Then, by definition, we obtain the PIN MLE estimate 0.286 = (0.4×200)/(0.4×200+2×100), which is the ratio of mean informed trading volume to mean total volume for a given fixed time length. Moreover, the VPIN estimate is 0.2 = 0.4×(200/(200+2×100)), which is the probability of information arrival further multiplied by the ratio of the informed trading volume to a fixed total trading volume conditioned on the event occurrence. Easley et al.'s MME generates the PIN estimate, 0.286 = 80/280 = ((4×|225–75| + 2×|450–150| + 9×|100–100|)/15) / ((4×|225+75| + 2×|450+150| + 9×|100+100|)/15).

Panel A shows that, with the buy and sell volumes measured in a fixed time length, the Easley et al.'s MME of 0.286 serves as a measure for PIN, which is the probability with respect to a given time bar. Panels B and C show that, based on trading volumes from the same time bars in Panel A, the buy and sell volumes are aggregated for the trading days with a fixed three time units and for buckets with a fixed volume of 600 shares and a varied time length. Based on trading days under the Panel B, the Easley et al.'s MME of 0.286 measures the PIN. Based on the buckets in Panel C, Easley et al.'s MME of 0.286 measures VPIN instead of PIN because it is equal to the MLE estimate of VPIN, and is the probability under the condition of a fixed volume of 600 shares. This shows our argument that the varied time length, which is inconsistent with the assumption implied by original PIN, results in different probability measures for informed trading derived via Easley et al.'s MME.

Panels C and D show that the trading imbalance in the short volume time may help improve the MLE estimates of PIN and VPIN. Panel D has shorter volume times in both buckets 1 and 3 than Panel C does. This may mean that the condition shown in Panel D has more informed traders in these buckets than that in Panel C does and should generate a greater PIN or VPIN. Easley et al.'s MME does not change for either panel, but the MLE estimates of PIN and VPIN are greater in Panel D. Namely, the MLE helps improve upon Easley et al.'s VPIN metric, which does not incorporate the information of volume time.

Finally, the Easley et al.'s MME estimates based on time bars, trading days and buckets are not always equal to one another. In this paper, we demonstrate the issue using a special case, in which the news types are known and there are equal buy and sell volumes in no news periods.

3. Modeling VPIN

To simplify our modeling, we assume that each trade is perfectly signed. Namely, we do not need any algorithm to sort trades. As mentioned in Easley et al. (2014), the trade classification decision can be made independently from the VPIN metric.

Let B_i and S_i denote the numbers of buyer- and seller-initiated trades (or buys and sells) during a given (unit) trading period i. The notation $\text{Pois}(x; \lambda) \equiv e^{-\lambda}\lambda^x/x!$ denotes the probability density function of Poisson variable x with arrival rate λ.

In the PIN model (Easley et al., 1996, 2002), the joint probability density function (p.d.f.) of Poisson variables B_i and S_i can be specified as

$$\begin{aligned} f(B_i, S_i; \boldsymbol{\theta}) &= \alpha\delta f(B_i, S_i|\text{bad news}; \boldsymbol{\theta}) \\ &+ \alpha(1-\delta) f(B_i, S_i|\text{good news}; \boldsymbol{\theta}) + (1-\alpha) f(B_i, S_i|\text{ no news}; \boldsymbol{\theta}) \qquad (1) \\ &= \alpha\delta \text{Pois}(B_i; \varepsilon_b)\text{Pois}(S_i; \varepsilon_s + \mu) \\ &+ \alpha(1-\delta)\text{Pois}(B_i; \varepsilon_b + \mu)\text{Pois}(S_i; \varepsilon_s) + (1-\alpha)\text{Pois}(B_i; \varepsilon_b)\text{Pois}(S_i; \varepsilon_s), \end{aligned}$$

where α is the probability of an information event occurring during the given trading period i; δ and $(1-\delta)$ are the conditional probabilities of bad and good news types, respectively; ε_b (ε_s) is

the arrival rate of uninformed buys (sells); μ is the arrival rate of informed trades; and vector $\boldsymbol{\theta} = (\alpha, \delta, \mu, \varepsilon_b, \varepsilon_s)$ represents the structural parameters.

Replacing buys and sells, B_i and S_i, respectively, with the buy and sell volumes, V_i^B and V_i^S, Easley et al. (2011, 2012b) modify the PIN model to apply to high-frequency markets. That is, they treat each reported trade as an aggregation of trades of unit size. Accordingly, for example, they treat one trade for five shares at a certain price p the same as five trades of one share, each at price p. The extension implies that V_i^B and V_i^S are also Poisson variables and that their joint p.d.f. may be specified as

$$\begin{aligned} f(V_i^B, V_i^S; \boldsymbol{\theta}) &= \alpha\delta f(V_i^B, V_i^S|\text{bad news}; \boldsymbol{\theta}) \\ &+ \alpha(1-\delta) f(V_i^B, V_i^S|\text{good news}; \boldsymbol{\theta}) + (1-\alpha) f(V_i^B, V_i^S|\text{no news}; \boldsymbol{\theta}) \\ &= \alpha\delta \text{Pois}(V_i^B; \varepsilon_b)\text{Pois}(V_i^S; \varepsilon_s + \mu) \\ &+ \alpha(1-\delta)\text{Pois}(V_i^B; \varepsilon_b + \alpha)\text{Pois}(V_i^S; \varepsilon_s) + (1-\alpha)\text{Pois}(V_i^B; \varepsilon_b)\text{Pois}(V_i^S; \varepsilon_s). \end{aligned} \tag{2}$$

Easley et al. (2011, 2012b) group sequential trades into equal volume buckets of an exogenously defined size V. A volume bucket is, therefore, a collection of trades with volume V. Let $\tau = 1, 2, \ldots, n$ be the index of equal volume buckets. For each volume bucket τ, a corresponding random time interval ω_τ, with a length that equals volume time t_τ, exists. Therefore, the volume in the bucket τ is the sum of V_i^B and V_i^S in a random time interval ω_τ such that $\Sigma_{i \in \omega_\tau}(V_i^B + V_i^S) = V$, where V_i^B and V_i^S may be treated as the buy and sell volumes,

respectively, in a (unit) time bar i. For each volume bucket τ, the volumes of buy and sell $V_\tau^B = \Sigma_{i\in\omega_\tau} V_i^B$ and $V_\tau^S = \Sigma_{i\in\omega_\tau} V_i^S$ may be calculated with $V_\tau = V_\tau^B + V_\tau^S = V$. Easley et al. (2011, 2012b) implicitly assume for the VPIN metric that each information event affects a volume bucket with the probability α.[2] Namely, after an information event occurs, the execution of the informed trader's order may be completed in a volume bucket. This assumption suggests that the PIN and VPIN models are similar but not fully equivalent. In contrast to the PIN setting, the VPIN setting assumes that informed traders submit their buy or sell orders by volume bucket instead of calendar time interval.

Easley et al. (2012b) derive the VPIN estimator based on the argument of two moment conditions, $E[|V_\tau^B - V_\tau^S|] \approx \alpha\mu$ and $E[V_\tau^B + V_\tau^S] = 2\varepsilon + \alpha\mu$, from the Poisson processes. Using our notations, the two moment conditions should instead be expressed as

$$E[|V_\tau^B - V_\tau^S| | t_\tau; \boldsymbol{\theta}] \approx \alpha\mu\, t_\tau \text{ and}$$

$$E[V_\tau^B + V_\tau^S | t_\tau; \boldsymbol{\theta}] = (2\varepsilon + \alpha\mu)\, t_\tau, \tag{3}$$

[2] Easley et al. (2012b, p. 1469) make the following assumption: "We divide the trading day into equal-sized volume buckets and treat each volume bucket as equivalent to a period for information arrival." Following Easley et al. (2011, 2012b), we set V as the one-fiftieth of the mean daily volume, and a user may obtain the VPIN metric with 50 volume buckets on a trading day. If an information event occurs only once each day and does not follow this assumption, these 50 buckets will come from the same (conditional) Poisson distribution and VPIN $\approx \mu/(2\varepsilon+\mu)$ or $0/(2\varepsilon)=0$, of which both do not include α.

where t_τ is a given fixed unit time interval and may be rescaled to 1, and volume V is not equal to the sum of V_τ^B and V_τ^S, which is a random variable. However, according to Easley et al. (2011, 2012b), V_τ^B and V_τ^S do not follow independent Poisson distributions due to the exogenous constraint $V = V_\tau^B + V_\tau^S$. This issue is not relevant for a calendar time interval (such as one trading day) because V_τ^B and V_τ^S can take arbitrary realized values. Namely, the Poisson processes for trade arrivals are not directly implementable on a volume bucket.[3]

After we derive the valid moment conditions, we show that VPIN and PIN measures have different definitions. In line with Easley et al.'s (2011, 2012b) estimation procedure, in the following discussion we accurately model volume time t_τ as randomly determined by an exogenously defined size V and show that the VPIN metric is actually calculated based on $E[|V_\tau^B - V_\tau^S||V; \boldsymbol{\theta}] \approx V \cdot [\alpha\mu/(2\varepsilon + \mu)]$ and $E[V_\tau^B + V_\tau^S|V; \boldsymbol{\theta}] = V$, where volume time t_τ is integrated out in the expectation.

According to the previous descriptions, the joint p.d.f. of V_τ^B, V_τ^S and t_τ given V by rewriting the Eq. (2) is

$$f(V_\tau^B, V_\tau^S, t_\tau|V; \boldsymbol{\theta}) = \alpha\delta f(V_\tau^B, V_\tau^S, t_\tau|V, \text{bad news}; \boldsymbol{\theta})$$

$$+ \alpha(1 - \delta) f(V_\tau^B, V_\tau^S, t_\tau|V, \text{good news}; \boldsymbol{\theta}) + (1 - \alpha) f(V_\tau^B, V_\tau^S, t_\tau|V, \text{no news}; \boldsymbol{\theta}). \quad (4)$$

In Eq. (1) or Eq. (2), an information event affects trades in a given trading period, which is

[3] We appreciate the anonymous reviewer's useful comment.

regarded as a time bar. In contrast, in Eq. (4), an information event affects the trades in a volume bucket, which corresponds to more than one time bar.

To express Eq. (4) in a closed form, we take $f(V_\tau^B,\ V_\tau^S,\ t_\tau|V,$ no news; $\boldsymbol{\theta})$ for an illustration. First, for given t_τ (or, say, ω_τ),

$$f(V_\tau^B,\ V_\tau^S|\ t_\tau, \text{ no news}; \boldsymbol{\theta}) = \text{Pois}(V_\tau^B; t_\tau\varepsilon_b)\text{Pois}(V_\tau^S; t_\tau\varepsilon_s) \tag{5}$$

because $V_\tau^B=\Sigma_{i\in\,\omega_\tau}V_i^B$, $V_\tau^S=\Sigma_{i\in\,\omega_\tau}V_i^S$, V_i^B and V_i^S are Poisson variables (as defined in the PIN model). Let $V_\tau^B = V_\tau^B$ and $V = V_\tau^B + V_\tau^S$; we derive the following joint p.d.f. of V_τ^B and V via the probability theory (see Hogg and Craig, 1995, pp. 165–166):

$$\begin{aligned} f(V_\tau^B, V|t_\tau, \text{ no news}; \boldsymbol{\theta}) &= f(V_\tau^B|V, t_\tau, \text{ no news}; \boldsymbol{\theta})\, f(V|t_\tau, \text{ no news}; \boldsymbol{\theta}) \\ &= \text{B}(V_\tau^B; V, \frac{t_\tau\varepsilon_b}{t_\tau\varepsilon_b + t_\tau\varepsilon_s})\text{Pois}(V; t_\tau\varepsilon_b + t_\tau\varepsilon_s) \\ &= \text{B}(V_\tau^B; V, \frac{\varepsilon_b}{\varepsilon_b + \varepsilon_s})\text{Pois}(V; t_\tau\varepsilon_b + t_\tau\varepsilon_s), \end{aligned} \tag{6}$$

where $\text{B}(x;\ m,\ p) \equiv \frac{m!}{x!(m\text{-}x)!}\ p^x(1\text{-}p)^{m\text{-}x}$ with $x = 1, 2, \ldots, m$ is the p.d.f. of the binomial distribution. Furthermore, because of $V = V_\tau^B + V_\tau^S$,

$$\begin{aligned} &f(V_\tau^B,\ V_\tau^S|t_\tau, V, \text{no news}; \boldsymbol{\theta}) \\ &= f(V_\tau^B|V, t_\tau, \text{ no news}; \boldsymbol{\theta}) = f(V_\tau^S|V, t_\tau, \text{ no news}; \boldsymbol{\theta}) \\ &= \text{B}(V_\tau^B; V, \frac{\varepsilon_b}{\varepsilon_b + \varepsilon_s}) = \text{B}(V_\tau^S; V, \frac{\varepsilon_s}{\varepsilon_b + \varepsilon_s}). \end{aligned} \tag{7}$$

Second, because $f(V|t_\tau,$ no news; $\boldsymbol{\theta})$ is equal to $\text{Pois}(V; t_\tau\varepsilon_b + t_\tau\varepsilon_s)$, the time t_τ needed to reach exactly V cumulative trading volume follows the distribution:

$$f(t_\tau|V, \text{no news}; \boldsymbol{\theta}) = \text{Gamma}(t_\tau; V, \varepsilon_b + \varepsilon_s), \tag{8}$$

where $\text{Gamma}(x; k, \lambda) \equiv \lambda^k x^{k-1} e^{-\lambda x}/\Gamma(k)$ is the p.d.f. of the Gamma distribution with $x > 0$ and $\Gamma(k) \equiv \int_0^\infty y^{k-1} e^{-y} dy$. The p.d.f. $\text{Gamma}(x; k, \lambda)$ is typically adopted to model the wait time until the kth event occurrence for a Poisson process with arrival rate λ (for details, see Hogg and Craig, 1995, pp. 131–133).

According to Eq. (7) and Eq. (8), the joint p.d.f. of V_τ^B, V_τ^S and t_τ given V with respect to no news arrival may be further expressed as

$$\begin{aligned} & f(V_\tau^B, V_\tau^S, t_\tau|V, \text{no news}; \boldsymbol{\theta}) \\ &= f(V_\tau^B, V_\tau^S|t_\tau, V, \text{no news}; \boldsymbol{\theta})\, f(t_\tau|V, \text{no news}; \boldsymbol{\theta}) \\ &= \text{B}(V_\tau^B; V, \frac{\varepsilon_b}{\varepsilon_b + \varepsilon_s}) \cdot \text{Gamma}(t_\tau; V, \varepsilon_b + \varepsilon_s) \\ &= \text{B}(V_\tau^S; V, \frac{\varepsilon_s}{\varepsilon_b + \varepsilon_s}) \cdot \text{Gamma}(t_\tau; V, \varepsilon_b + \varepsilon_s). \end{aligned} \tag{9}$$

Furthermore, the joint p.d.f. of V_τ^S, V_τ^B, and t_τ given V is

$$\begin{aligned} & f(V_\tau^B, V_\tau^S, t_\tau|V; \boldsymbol{\theta}) \\ &= \alpha\delta\, \text{B}(V_\tau^B; V, \frac{\varepsilon_b}{\varepsilon_b + \varepsilon_s + \mu}) \cdot \text{Gamma}(t_\tau; V, \varepsilon_b + \varepsilon_s + \mu) \\ &+ \alpha(1 - \delta)\, \text{B}(V_\tau^B; V, \frac{\varepsilon_b + \mu}{\varepsilon_b + \varepsilon_s + \mu}) \cdot \text{Gamma}(t_\tau; V, \varepsilon_b + \varepsilon_s + \mu) \\ &+ (1 - \alpha) \text{B}(V_\tau^B; V, \frac{\varepsilon_b}{\varepsilon_b + \varepsilon_s}) \cdot \text{Gamma}(t_\tau; V, \varepsilon_b + \varepsilon_s) \\ &= \alpha\delta\, \text{B}(V_\tau^S; V, \frac{\varepsilon_s + \mu}{\varepsilon_b + \varepsilon_s + \mu}) \cdot \text{Gamma}(t_\tau; V, \varepsilon_b + \varepsilon_s + \mu) \\ &+ \alpha(1 - \delta)\, \text{B}(V_\tau^S; V, \frac{\varepsilon_s}{\varepsilon_b + \varepsilon_s + \mu}) \cdot \text{Gamma}(t_\tau; V, \varepsilon_b + \varepsilon_s + \mu) \end{aligned} \tag{10}$$

$$+ (1 - \alpha)\mathrm{B}(V_\tau^S;\, V,\, \frac{\varepsilon_s}{\varepsilon_b + \varepsilon_s})\cdot\mathrm{Gamma}(t_\tau;\, V,\, \varepsilon_b + \varepsilon_s).$$

4. Deriving the VPIN Metric

The estimate in Eq. (10) is more efficient because the complete information set with variables V_τ^B, V_τ^S, and t_τ, is adopted. However, if we integrate out the volume time t_τ, we can derive $f(V_\tau^B,\ V_\tau^S|V;\ \boldsymbol{\theta})$ as

$$\begin{aligned} f(V_\tau^B,\ V_\tau^S|V;\ \boldsymbol{\theta}) &= \int_0^\infty f(V_\tau^B, V_\tau^S,\ t_\tau|V;\ \boldsymbol{\theta})dt_\tau \\ &= \alpha\delta\,\mathrm{B}(V_\tau^B;\, V,\, \frac{\varepsilon_b}{\varepsilon_b + \varepsilon_s + \mu}) \\ &+ \alpha(1 - \delta)\,\mathrm{B}(V_\tau^B;\, V,\, \frac{\varepsilon_b + \mu}{\varepsilon_b + \varepsilon_s + \mu}) + (1 - \alpha)\mathrm{B}(V_\tau^B;\, V,\, \frac{\varepsilon_b}{\varepsilon_b + \varepsilon_s}) \\ &= \alpha\delta\,\mathrm{B}(V_\tau^S;\, V,\, \frac{\varepsilon_s + \mu}{\varepsilon_b + \varepsilon_s + \mu}) \\ &+ \alpha(1 - \delta)\,\mathrm{B}(V_\tau^S;\, V,\, \frac{\varepsilon_s}{\varepsilon_b + \varepsilon_s + \mu}) + (1 - \alpha)\mathrm{B}(V_\tau^S;\, V,\, \frac{\varepsilon_s}{\varepsilon_b + \varepsilon_s}). \end{aligned} \tag{11}$$

VPIN can be estimated using only V_τ^B, and V_τ^S with Eq. (11), but the estimates are less efficient. That is, the VPIN estimates with Eq. (11) have greater standard errors than those of Eq. (10) because Eq. (11) does not include the volume time variable t_τ.

Moreover, if x follows the binomial distribution, of which the p.d.f. is $\mathrm{B}(x;\, m,\, p)$ with $x = 0$, 1, 2, …, m, we can derive the following approximation using Jensen's inequality for a large m and a p diverging from 0.5 (see Figure 1):[4]

[4] With Jensen's inequality, $E[|m - 2x|]/m > |E[m - 2x]|/m = |2p - 1|$ and $E[|m - 2x|]/m = E\{[(m - 2x)^2]^{1/2}\}/m < \{E[(m - 2x)^2]\}^{1/2}/m = \{[(2p - 1)^2 + 4p(1 - p)/m]\}^{1/2} \approx |2p - 1|$ for a large m. Therefore, $E[|m - 2x|] \approx m|2p - 1|$ as m becomes large.

$$E[|m - 2x|] \approx m|2p - 1|. \tag{12}$$

[FIGURE 1 ABOUT HERE]

Then, with Eq. (11) and Eq. (12), we derive the following result:

$$E[|V_\tau^B - V_\tau^S||V; \boldsymbol{\theta}] \approx \alpha\delta V\frac{|\varepsilon_b - (\varepsilon_s + \mu)|}{\varepsilon_b + \varepsilon_s + \mu} + \alpha(1 - \delta)V\frac{|(\varepsilon_b + \mu) - \varepsilon_s|}{\varepsilon_b + \varepsilon_s + \mu} + (1 - \alpha)V\frac{|\varepsilon_b - \varepsilon_s|}{\varepsilon_b + \varepsilon_s}. \tag{13}$$

Following Easley et al. (2012b), we set $\varepsilon_b = \varepsilon_s = \varepsilon$ and rewrite Eq. (13) as

$$E[|V_\tau^B - V_\tau^S||V; \boldsymbol{\theta}] \approx \alpha\delta \cdot V\frac{\mu}{2\varepsilon + \mu} + \alpha(1 - \delta)V\frac{\mu}{2\varepsilon + \mu} = V\frac{\alpha\mu}{2\varepsilon + \mu}. \tag{14}$$

Therefore, we obtain the estimator of Easley et al. (2012b) for VPIN as

$$\text{EVPIN} \equiv \frac{\sum_{\tau=1}^{n}|V_\tau^B - V_\tau^S|/n}{V} = \frac{\sum_{\tau=1}^{n}\left|V_\tau^B - V_\tau^S\right|/n}{\sum_{\tau=1}^{n}V_\tau/n} = \frac{\overline{\left|V_\tau^B - V_\tau^S\right|}}{\overline{V_\tau}} \approx \frac{\alpha\mu}{2\varepsilon + \mu}, \tag{15}$$

which does not approximately measure $\alpha\mu/(2\varepsilon + \alpha\mu)$, the expectation of Easley et al. (2012b).

Hereafter, VPIN is defined as $\alpha\mu/(\varepsilon_b + \varepsilon_s + \mu)$. That is, VPIN $\equiv \alpha\mu/(\varepsilon_b + \varepsilon_s + \mu)$. Based on this definition, the VPIN is a measure from the Bayes' theorem for the probability of informed trading given a fixed-sized trading volume. By contrast, the original PIN $\equiv \alpha\mu/(\varepsilon_b + \varepsilon_s + \alpha\mu)$ measures the probability of the informed trading for a given fixed time interval.

That is, VPIN (PIN) is an expectation conditioned (not conditioned) on the volume bucket. Specifically, VPIN measures the percentage of informed trading for a given fixed number of

transactions, whereas PIN measures the probability of informed trading over a given tiny time interval, during which only a single transaction occurs. VPIN may be explained intuitively as follows. In buckets in which no event occurs, the arrival rate of trades is $\varepsilon_b + \varepsilon_s$, and the relative arrival rate of informed to total trades is zero, as no informed trades occur. In buckets during which an event occurs, the arrival rate is $\varepsilon_b + \varepsilon_s + \mu$. Informed trades, if they exist, arrive at rate μ. Therefore, the relative arrival rate of informed to total trades in buckets in which an event occurs is $\mu/(\varepsilon_b + \varepsilon_s + \mu)$. Information events occur with probability α, so that the expected relative arrival rate of informed to total trades is $\alpha \cdot \mu/(\varepsilon_b + \varepsilon_s + \mu)$.

Analogously, Easley et al.'s VPIN metric (EVPIN) intuitively measures VPIN rather than PIN for $\varepsilon_b = \varepsilon_s = \varepsilon$. Suppose that the volume of trading imbalance measures informed trading. When $\varepsilon_b = \varepsilon_s = \varepsilon$, a trading imbalance may be triggered by a news event such as the announcement of a firm successor. Thus, a large (small) trading imbalance suggests a news (no-new) regime with a high (low) $|V_\tau^B - V_\tau^S|/V \approx \mu/(2\varepsilon + \mu)$ (≈ 0). Easley et al.'s VPIN metric, which may be treated as the mean ratio of the imbalanced volume to the total trading volume, is $\alpha \cdot \mu/(2\varepsilon + \mu)$, that is, the probability of the event times the probability of informed trading within a news regime.

Adopting a Monte Carlo simulation using Eq. (11) with fixed $V = 500$ and $n = 50$ (buckets),

we verify our procedure.[5] We take α from {0, 0.1, 0.2, …, 1} and μ from {0, 50, 100, …, 500} and then set $\varepsilon_b = \varepsilon_s = (V - \alpha\mu)/2$. For each node of (α, μ), we generate 1,000 samples with size n = 50 and calculate the EVPIN and the mean measurement error. Figure 2 plots the results. Panel A (Panel B) shows mean error in measuring the actual VPIN (PIN) value using EVPIN, defined as EVPIN – VPIN (EVPIN – PIN) from 1,000 Monte Carlo simulations. Panel A shows a more stable mean measurement error than that in Panel B. These results suggest that EVPIN captures the VPIN value instead of the PIN value.

[FIGURE 2 ABOUT HERE]

Andersen and Bondarenko (2014a) provide a more precise expression and approximation of $E[|V_\tau^B - V_\tau^S||V; \boldsymbol{\theta}]$ when the trade direction is randomly determined. They express the approximation under $\varepsilon_s = \varepsilon_b = \varepsilon$ and $\mu = 0$ (or $\alpha = 0$) as[6]

[5] According to the pseudo code in the appendix of Easley et al. (2012b), the buy and sell volumes (V_τ^B, V_τ^S) generated by their Monte Carlo simulation do not satisfy the condition $V_\tau^B + V_\tau^S = V$. Therefore, their simulation is inconsistent with the estimation procedure of VPIN.

[6] Andersen and Bondarenko (2014a) derive the approximation under the assumption that the trade classification scheme generates a purely random buy-sell indicator with probability 1/2. However, if the trade classification procedure works well, it will generate the buy–sell indicator with the probability determined by the arrival rates of buy and sell orders. Therefore, the probability 1/2 implies that $\varepsilon_s = \varepsilon_b = \varepsilon$ and $\mu = 0$ (or that $\varepsilon_s = \varepsilon_b = \varepsilon$ and $\alpha = 0$).

$$E[|V_\tau^B - V_\tau^S||V; \boldsymbol{\theta}=(\alpha, \delta, 0, \varepsilon, \varepsilon)] = \frac{(2\nu)!}{2^{2\nu}\nu!\nu!} \approx \sqrt{\frac{2}{\pi V}}, \quad (16)$$

where ν is an integer subject to $V = 2\nu$ or $V = 2\nu + 1$. Eq. (16) implies that EVPIN may be inappropriate for small volume buckets and infrequent informed trades because Eq. (15) generates a greater measurement error under these conditions.[7] Figure 1 supports this intuition. Furthermore, given Eq. (11), EVPIN ignores the information provided by the volume time variable t_τ, and failing to incorporate t_τ leads to inefficiency in estimating VPIN. Accordingly, we propose an MLE procedure for the VPIN metric, which we discuss in the next section.

5. Estimating VPIN with Volume Time via MLE

Based on Eq. (10), Eq. (11), and Eq. (15), EVPIN appears to ignore information regarding volume time t_τ. That is, in practice, EVPIN fails to meet the expectation of Easley et al. (2012b). To solve this problem, we follow the original PIN estimation and incorporate the volume time variable into the VPIN estimation via MLE.

Moreover, without the trade classification, we may simply treat that as if only a trade of one share occurs in a tiny time bar, and thus V is the number of time bars, Q, as defined in Andersen and Bondarenko (2014a).

[7] In our opinion, Andersen and Bondarenko's (2014a, pp. 13–14) Findings 2 and 3 are the results of the worse performance of Easley et al.'s (2012b) VPIN metric under $\varepsilon_s = \varepsilon_b = \varepsilon$ and $\mu = 0$. A large bucket contains a large number of time bars. Then, Easley et al.'s VPIN metric generates a small estimate of imbalance orders and a VPIN estimate of approximately zero due to $\mu = 0$ and the law of large numbers.

With the assumption that the volume buckets $\tau = 1, 2, \ldots, n$ are independent of one another, the (log-)likelihood $L_I(\boldsymbol{\theta}|\mathbf{D})$ of observing a series of $(V_\tau^B, V_\tau^S, t_\tau)$ with $V_\tau^B + V_\tau^S = V$ over the n volume buckets is the sum of the log of p.d.f. $L_I(\boldsymbol{\theta}|V_\tau^B, V_\tau^S, t_\tau) \equiv \log(f(V_\tau^B, V_\tau^S, t_\tau| V; \boldsymbol{\theta}))$:[8]

$$L_I(\boldsymbol{\theta}|\mathbf{D}) \equiv \sum_{\tau=1}^{n} L_I\left(\boldsymbol{\theta}|V_\tau^B, V_\tau^S, t_\tau\right) = \sum_{\tau=1}^{n} \log\left(f(V_\tau^B, V_\tau^S, t_\tau| V; \boldsymbol{\theta})\right), \tag{17}$$

where $\mathbf{D} \equiv \left\{\left(V_\tau^B, V_\tau^S, t_\tau\right)\right\}_{\tau=1\ldots n}$ represents the buy and sell volumes and the volume times for the volume buckets $\tau = 1, 2, \ldots, n$. V_τ^B and V_τ^S may be determined by the signed trading data or the trade classification algorithms (see Easley et al., 2012a). With an appropriate numerical method, $\boldsymbol{\theta}_I \equiv (\alpha_I, \delta_I, \mu_I, \varepsilon_{bI}, \varepsilon_{sI})$ is the consistent estimate of $\boldsymbol{\theta}$ by MLE using Eq. (17). We denote the estimates from (17) by the subscript *I*. Then, the improved PIN estimate, given $\boldsymbol{\theta}_I$, is

$$\text{IPIN} = \frac{\alpha_I \mu_I}{\varepsilon_{bI} + \varepsilon_{sI} + \alpha_I \mu_I}, \tag{18}$$

and the improved VPIN estimate is

$$\text{IVPIN} = \frac{\alpha_I \mu_I}{\varepsilon_{bI} + \varepsilon_{sI} + \mu_I}. \tag{19}$$

In contrast, under $\varepsilon_b \neq \varepsilon_s$, an investor cannot adopt the observed imbalance order flow directly to estimate VPIN because she cannot exclude the possibility that the imbalance results from

[8] The $L_I(\boldsymbol{\theta}|V_\tau^B, V_\tau^S, t_\tau)$ may be simplified for the stable numerical computing stability via a similar means adopted by Lin and Ke (2011). The reformulated $L_I(\boldsymbol{\theta}|V_\tau^B, V_\tau^S, t_\tau)$ is $\log(\exp(e_{\tau,1}-e_{\tau,\max}) + \exp(e_{\tau,2}-e_{\tau,\max}) + \exp(e_{\tau,3}-e_{\tau,\max}))$, where $e_{\tau,1} = \log(\alpha\delta) + V_\tau^B\log(\varepsilon_b) + V_\tau^S\log(\varepsilon_s+\mu) - t_\tau(\varepsilon_b+\varepsilon_s+\mu)$, $e_{\tau,2} = \log(\alpha(1-\delta)) + V_\tau^B\log(\varepsilon_b+\mu) + V_\tau^S\log(\varepsilon_s) - t_\tau(\varepsilon_b+\varepsilon_s+\mu)$, $e_{\tau,3} = \log(1-\alpha) + V_\tau^B\log(\varepsilon_b) + V_\tau^S\log(\varepsilon_s) - t_\tau(\varepsilon_b+\varepsilon_s)$, and $e_{\tau,\max} = \max(e_{\tau,1}, e_{\tau,2}, e_{\tau,3})$.

liquidity trading during a no-news period.

Unlike VPIN, the PIN metric measures the probability of informed trading within a fixed time interval. The number of trades occurring within each fixed time interval is unknown. PIN identifies the number of informed trades among the trades that occur during each fixed time interval. The VPIN metric, in contrast, measures the probability of informed trading for a given trading volume. Therefore, VPIN is the ratio of informed trades for a given fixed number of trades. The time needed to accumulate the given number of trades is a priori unknown. Namely, PIN and VPIN measure the probabilities based on different data constructions from the same raw data set.

Adopting a simulation using Eq. (10) with a fixed small $V = 50$ and $n = 50$ (buckets), we verify the performance for small volume buckets. We take α from $\{0, 0.1, 0.2, \ldots, 1\}$ and μ from $\{0, 5, 10, \ldots, 50\}$, and set $\varepsilon_b = \varepsilon_s = (V - \alpha\mu)/2$. For each node of (α, μ), we generate 50 samples with $n = 50$ and calculate IVPIN and EVPIN and their mean measurement errors. Figure 3 plots the results. Panel A (Panel B) presents mean error of estimating the actual VPIN value using IVPIN (EVPIN) defined as IVPIN – VPIN (EVPIN –VPIN) from 50 simulations (for each node).[9] Panel A shows a more stable mean measurement error than that in Panel B. These results

[9] As we increase the number of simulations that we conduct, the plot becomes smoother; however, the process also becomes more time consuming. Therefore, we perform 50 simulations for each node.

suggest that IVPIN is more suitable for small volume buckets than EVPIN.

Figure 3 also implies the significant marginal predicting power of volume time in estimating VPIN. If μ is small, volume time is likely to be large according to Eq. (10). Therefore, when volume time is large, μ may be small, and thus the difference between IVPIN and EVPIN may become large. However, when μ is large, the difference between IVPIN and EVPIN is small regardless of volume time. Namely, the marginal effect of volume time becomes more pronounced as μ decreases. Thus, volume time may be more beneficial to the identification of informed trading for less frequently traded stocks.

[FIGURE 3 ABOUT HERE]

6. Empirical Tests

Using the trading data of SYP, the ETF of S&P 500 stock index, from the Trade and Quote (TAQ) database, we calculate EVPIN and estimate VPIN via MLE both with and without volume time t_τ on May 6, 2010. We use the one-fiftieth of the mean daily trading volume in 2010 to set the bucket size, which is approximately 3,694,243.315 shares. We use the bulk classification algorithm (Easley et al., 2012a) with a one-minute time bar to calculate the volumes of buy and sell, V_τ^B and V_τ^S, respectively. For a time bar, if the cumulative volume exceeds V, we assume

that the trades are uniformly distributed and then calculate V_τ^B, V_τ^S, and t_τ proportionally.

For each rolling window of 50 buckets, we calculate both EVPIN and IVPIN. To improve the computational efficiency of IVPIN, we set the solution of the previous rolling window as the initial value of the subsequent one. If the previous solution fails to optimize the MLE problem, we generate another 25 initial values and rerun the optimization. We obtain IVPIN using the log-likelihood function Eq. (17), which is based on Eq. (10) with volume times. We also calculate VPIN via MLE based on Eq. (11), which does not include volume times, with the constraint of $\varepsilon_s + \varepsilon_b = 1$ and the following log-likelihood function:

$$L_M(\boldsymbol{\theta}|\mathbf{D'}) \equiv \sum_{\tau=1}^{n} L\left(\boldsymbol{\theta}|V_\tau^B, V_\tau^S\right) = \sum_{\tau=1}^{n} \log\left(f(V_\tau^B, V_\tau^S|\, V; \boldsymbol{\theta})\right), \tag{20}$$

where $\mathbf{D'} \equiv \left\{\left(V_\tau^B, V_\tau^S\right)\right\}_{\tau=1\ldots n}$. We denote the estimates from Eq. (20) by the subscript *M*. Moreover, the VPIN (PIN) estimate from Eq. (20) is MVPIN (MPIN), which does not incorporate any volume time information. The MVPIN is the MLE version of EVPIN.

[FIGURE 4 ABOUT HERE]

Figures 4A and 4B show the price of SYP and EVPIN and volume time, respectively. The price drops substantially during the flash crash between 14:30 and 14:48. EVPIN keeps increasing before, during, and after the flash crash. The phenomenon is consistent with the

observation of Andersen and Bondarenko (2013a) that EVPIN does not peak during the crash. On the other hand, the volume time appears to decrease before the crash, reach the bottom during the crash, and then increase after the crash. This information is useful for estimating VPIN.

[FIGURE 5 ABOUT HERE]

Figure 5 shows the estimates from MLE with and without volume time t_τ. After the crash, these figures show different patterns. In Panel B, the MVPIN is closer to EVPIN than MPIN. This result is consistent with that of our simulation test in Figure 2. This difference is from the non-extreme α_M (Panel D) [10] and the increasing μ_M/ε_M (Panel H), where $\varepsilon_M = \varepsilon_{bM} + \varepsilon_{sM} = 1$. Moreover, the three estimates, EVPIN, MVPIN, and MPIN, keep increasing after the crash. In contrast, Panel A shows that IPIN (IVPIN) moves within a narrow (a somewhat tighter) band and declines slightly. Therefore, Panels A and B show the incremental effect of incorporating volume time, which improves the reasonability of VPIN estimates after the crash.

After the crash the MLE generates increasing μ_M/ε_M (Figure 5, Panel H). In contrast, μ_I/ε_I decreases (Panel G), where $\varepsilon_I = \varepsilon_{bI} + \varepsilon_{sI}$. Along with increasing α_I (Panel C), these results create a flatter pattern in Panel A than in Panel B after the crash. Namely, after the crash, uninformed

[10] When α=1 or 0, VPIN = PIN.

traders keep selling their shares and thus add to the order flow imbalance. Without any information regarding t_τ, the market maker may perceive that the extent of informed trading increases. However, with an increasing t_τ, the market maker can recognize that sell orders are in fact submitted by uninformed traders.

Panel A of Figure 5 shows that IVPIN responds earlier than EVPIN. Specifically, IVPIN (EVPIN) starts to rise at approximately 13:30 (14:00).[11] The difference between IVPIN and EVPIN appears to be based on whether information regarding volume time is used in the estimation. Panel A shows that before the crash, when volume time is large, the difference is large. In contrast, during the crash period, when volume time is small, the difference is small. Moreover, α_I and δ_I differ significantly from α_M and δ_M.[12]

In summary, incorporating volume time variable (t_τ) helps improve the predictability of VPIN for the flow toxicity. Furthermore, the estimates of VPIN parameter sets generated by the estimation procedure that includes t_τ differ significantly from the estimates generated by the procedure that excludes t_τ.

7. Conclusion

[11] The cumulative probability of IVPIN from the empirical cumulative distribution appears to start rising at approximately 13:30, and the one of EVPIN begins to rise at approximately 14:00.

[12] Variables needed to explain their differences include (i) the extent of market efficiency and (ii) whether the informed traders divide the trade into a large number of smaller orders. Specifically, if the market is efficient and the informed trader chooses not to divide her trades, we expect a zigzag pattern of the α_I and δ_I. In contrast, if the market is inefficient and the informed trader divides her trades into many orders, we expect a smooth pattern for α_I and δ_I.

We model the VPIN and the derived VPIN metric from the PIN model, demonstrating that VPIN and PIN metrics are different measures. VPIN measures the percentage of informed trading for a given fixed number of transactions, whereas PIN measures the probability of the informed trading over a given tiny time interval, during which only a single transaction occurs. Building on Easley et al. (2011, 2012b), who derive the VPIN metric and provide evidence of its usefulness, we then expand theoretical basis of the model and clarify its derivation.

The logical inconsistency between Easley et al.'s VPIN and PIN metrics primarily results from determining the bucket by fixed-sized trading volume. If the bucket is determined by certain conditions, such as price duration as in Engle and Russell (1997, 1998), a revised Easley et al.'s VPIN metric (EVPIN) may serve as an estimator of PIN, but the other revised VPIN metrics may not.[13] In contrast, our VPIN estimator (IVPIN) may be viewed as a Bayesian estimator of VPIN given *V*.

Our proposed generalized VPIN estimation measure is broadly applicable. It extends the usefulness of the original VPIN metric for small volume bucket and infrequent informed trading conditions. Specifically, small bucket volume may result in overestimating order imbalance and thus a large VPIN estimate even when no informed trading occurs.[14] Thus, market

[13] A revised EVPIN, $\left(\Sigma_{\tau=1}^{n}\left|V_{\tau}^{B}-V_{\tau}^{S}\right|/t_{\tau}\right)/\left(\Sigma_{\tau=1}^{n}V_{\tau}/t_{\tau}\right)$, may serve as the estimator of PIN ($\equiv \alpha\mu/(\varepsilon_b+\varepsilon_s+\alpha\mu)$), but the other revised EVPIN, $\Sigma_{\tau=1}^{n}\left|V_{\tau}^{B}-V_{\tau}^{S}\right|/V_{\tau}$, may still be the estimator of VPIN ($\equiv \alpha\mu/(\varepsilon_b+\varepsilon_s+\mu)$).

[14] This result is consistent with Findings 2 and 3 of Andersen and Bondarenko (2014a, pp. 13–14).

microstructure studies that apply Easley et al.'s VPIN metric need to establish a clear criterion for a high-frequency market to avoid including in their samples securities with infrequent trades.

From a practical perspective, even in a high-frequency market, Easley et al.'s VPIN metric may mislead market makers. For example, an imbalance may exist when an overwhelming number of liquidity trades accompanies fairly few (or no) informed trades. Also, market makers may base their bid and ask quotes on a small volume bucket in an attempt to reduce or cover inventory costs or to shorten the reaction time. The VPIN metric may be overstated in these situations and lead to a wide bid–ask spread, thereby jeopardizing operational efficiency in the market.

Despite the simplicity in calculating VPIN, Easley et al.'s (2012b) estimation procedure ignores volume time information (t_τ), which may be crucial. Accordingly, we provide the MLE procedure for the VPIN and PIN with volume time, which remains valid in an infrequent market. We also show that volume time helps to improve the predictability of VPIN for flow toxicity. The MLE of VPIN may be useful for assets pricing studies that choose to replace PIN with VPIN. Both PIN ($\equiv \alpha\mu/(\varepsilon_b + \varepsilon_s + \alpha\mu)$) and VPIN ($\equiv \alpha\mu/(\varepsilon_b + \varepsilon_s + \mu)$) can be calculated based on buckets via MLE: Which metric is empirically advantageous is another issue. Furthermore, MLE can be applied with various bucket sizes. Determining the right bucket size is not a trivial matter, and one-fiftieth of mean daily volume, as reported in prior studies, serves merely as one alternative.

We incorporate volume time in a microstructure model based on the PIN setting of Easley et al. (1996; 2002). With this model, market makers of thinly traded stocks may update their bid and ask quotes with long rolling windows to infer the degree of information asymmetry using cumulative trading volume. With this proposed procedure for modeling VPIN, extended PIN models can potentially incorporate the volume bucket variable to estimate information parameters (e.g., Duarte and Young, 2009.)

References

Abad, D. and J. Yague. 2012. From PIN to VPIN: an introduction to order flow toxicity. *The Spanish Review of Financial Economics* 10, 74–83.

Andersen, T.G. and O. Bondarenko. 2014a. VPIN and the flash crash. *Journal of Financial Markets* 17, 1–46.

Andersen, T.G. and O. Bondarenko. 2014b. Reflecting on the VPIN dispute. J*ournal of Financial Markets* 17, 53–64.

Bethel, E.W., D. Leinweber, O. Rubel, and K. Wu. 2012. Federal market information technology in the post-flash crash era: Roles for supercomputing. *Journal of Trading* 7(2), 9–25.

Duarte, J. and L. Young. 2009. Why is PIN priced? *Journal of Financial Economics* 91, 119–138.

Easley, D., N. M. Kiefer, M. O'Hara, and J. Paperman. 1996. Liquidity, information, and infrequently traded stocks. *Journal of Finance* 51, 1405–1436.

Easley, D., S. Hvidkjaer, and M. O'Hara. 2002. Is information risk a determinant of asset returns? *Journal of Finance* 57, 2185–2221.

Easley, D., M. López de Prado, and M. O'Hara. 2011. The microstructure of the "flash crash." *Journal of Portfolio Management* 37(2), 118–128.

Easley, D., M. López de Prado, and M. O'Hara. 2012a. Bulk classification of trading activity.

Available at: http://ssrn.com/abstract=1989555.

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

Easley, D., M. López de Prado, and M. O'Hara. 2012c. The volume clock: insights into the high frequency paradigm. *Journal of Portfolio Management* 39(1), 19–29.

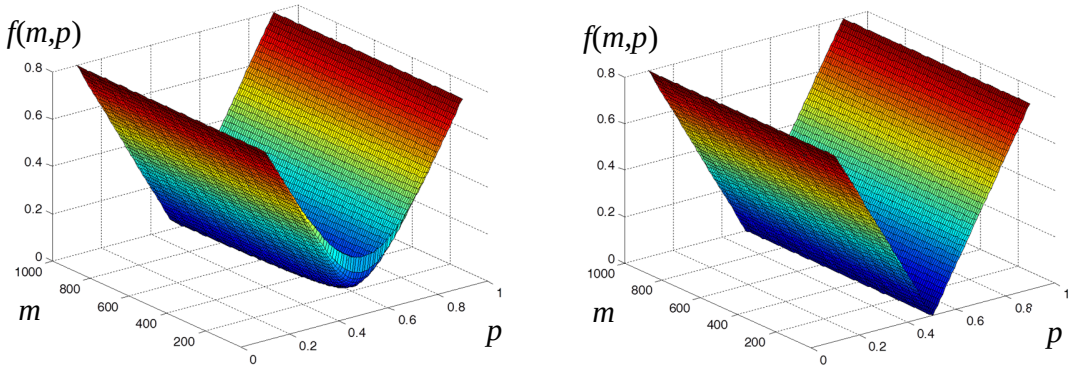
Easley, D., M. López de Prado, and M. O'Hara. 2014. VPIN and the flash crash: A rejoinder. *Journal of Financial Markets* 17, 47–52.

Hogg, R.V., and A.T. Craig. 1995. *Introduction to Mathematical Statistics*. 5th ed. New Jersey: Prentice Hall.

Lin, H.W. and W.C. Ke, W.C. 2011. A computing bias in estimating the probability of informed trading. *Journal of Financial Markets* 14 (4), 625–640.

Madhavan, A. 2011. Exchange-traded funds, market structure and the flash crash. *Technical Report, Black Rock*. Available at: http://ssrn.com/abstract=1932925.

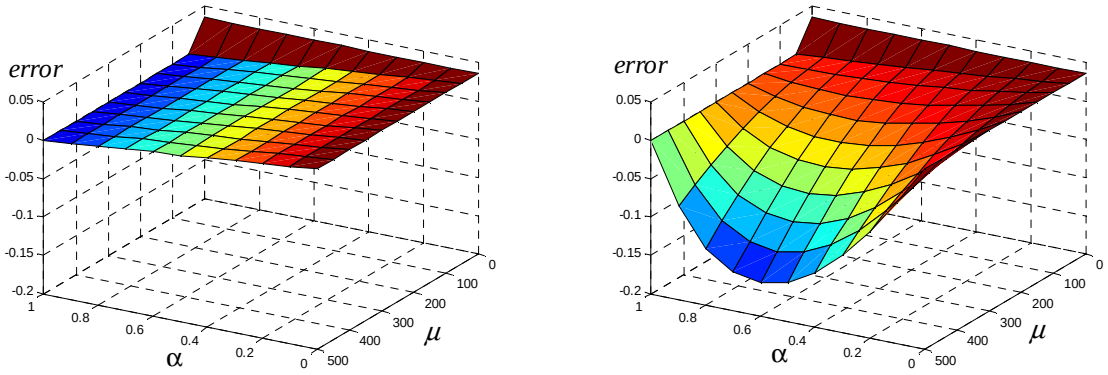
Wei, W. C., D. Gerace, and A. Frino. 2013. Informed trading, flow toxicity and the impact on intraday trading factors. *Australian Accounting Business and Finance Journal* 7, 3–24.



Panel A Panel B

Figure 1: Panel A depicts the function $f(m, p) = E[|m - 2x|]$, where x follows the binomial distribution $B(x; m, p)$. Panel B depicts the function $f(m, p) = m|2p - 1|$.

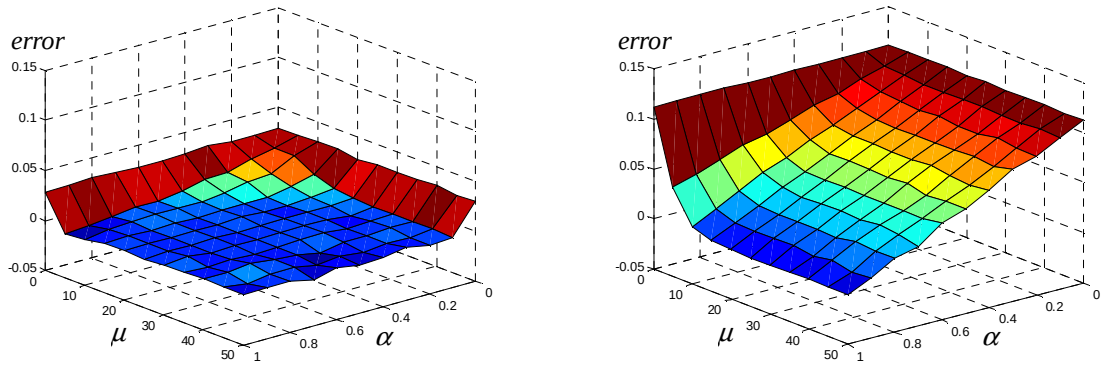
Interpretation: This figure shows that $E[|m - 2x|] \approx m|2p - 1|$ when x follows the binomial distribution, of which the p.d.f. is $B(x; m, p)$ with $x = 0, 1, 2, \ldots, m$. Moreover, when m is small or p is approaching 0.5, the approximation is inaccurate.



Panel A Panel B

Figure 2: Panel A shows the mean error in measuring the actual VPIN value using EVPIN. In contrast, Panel B presents the mean error in measuring the actual PIN value using EVPIN. Both panels result from 1,000 Monte Carlo simulations with V=500 and n = 50.

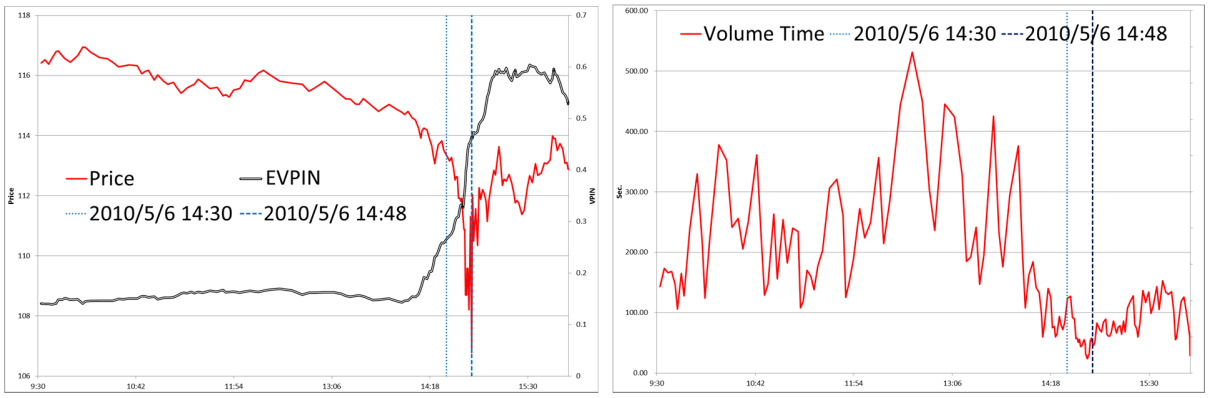
Interpretation: This figure shows that EVPIN is the estimator of VPIN $\equiv \alpha\mu/(\varepsilon_b + \varepsilon_s + \mu)$ rather than PIN $\equiv \alpha\mu/(\varepsilon_b + \varepsilon_s + \alpha\mu)$ when the simulation itself fits the assumption of fixed-sized volume bucket.



Panel A Panel B

Figure 3: Panel A shows the mean error in measuring the actual VPIN value using IVPIN. In contrast, Panel B presents the mean error in measuring the actual VPIN value using EVPIN. Both panels result from 50 Monte Carlo simulations with V=50 and n = 50.

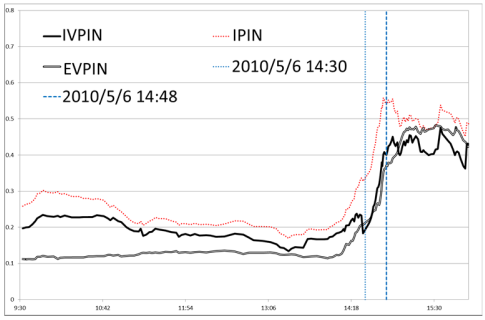
Interpretation: This figure suggests that IVPIN is more suitable for small volume buckets than EVPIN. In addition, it implies the significant marginal predicting power of volume time in estimating VPIN because IVPIN incorporates the information of volume time.



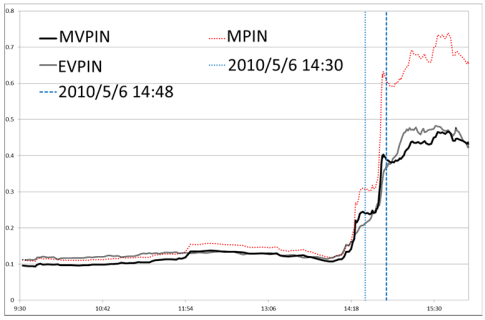
Panel A: ETF Price and EVPIN Panel B: Volume time

Figure 4: Panel A shows the (ending) stock price of SPY and EVPIN at the end of each bucket. Panel B shows the volume time at the end of each bucket.

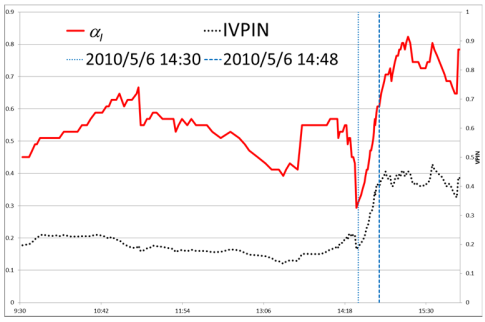
Interpretation: This figure shows that EVPIN keeps increasing before, during, and after the flash crash. Namely, EVPIN does not peak during the crash. On the other hand, the volume time appears to decrease before the crash, reach the bottom during the crash, and then increase after the crash. This information is useful for estimating VPIN.



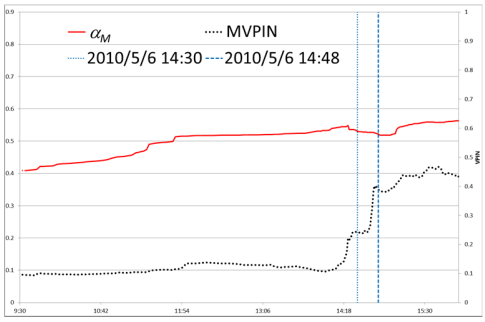
Panel A: IVPIN



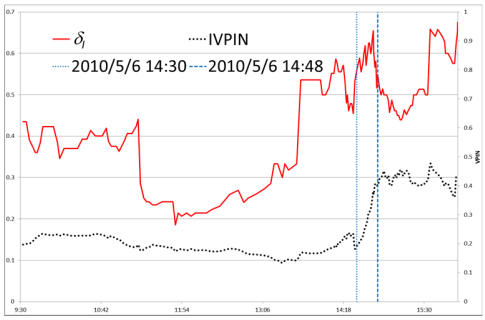
Panel B: MVPIN



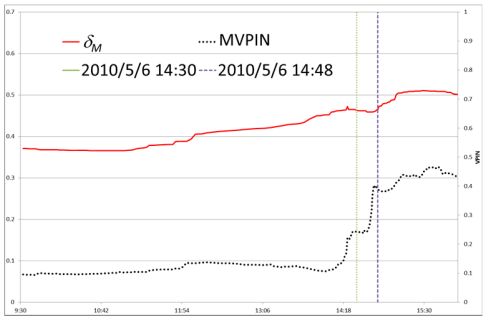
Panel C: α_I



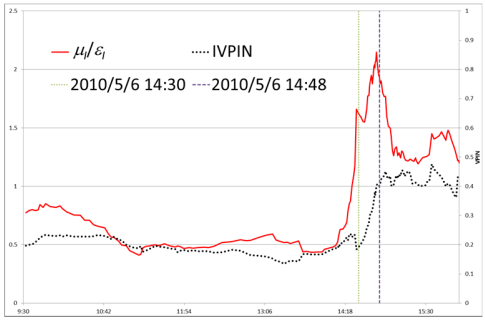
Panel D: α_M



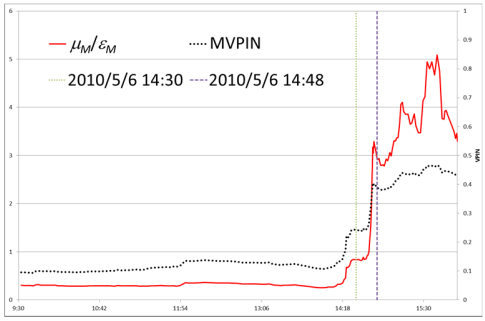
Panel E: δ_I



Panel F: δ_M



Panel G: μ_I/ε_I



Panel H: μ_M/ε_M

Figure 5: Panels A, C, E and G show the estimates from Eq. (17) with t_τ. Panels B, D, F and H show the estimates from Eq. (20) without t_τ. For MLE of Eq. (20), we set the constraint $\varepsilon = \varepsilon_b + \varepsilon_s = 1$ to obtain the unique solution. For each panel, between the two vertical lines we depict the flash crash between 14:30 and 14:48. We set the ending time to be 14:48 because at which there is a lowest bucket end price.

Interpretation: Incorporating volume time variable (t_τ) helps improve the predictability of VPIN for the flow toxicity. Specifically, after the crash, the estimates of VPIN parameter sets generated by the estimation procedure that includes t_τ differ significantly from the estimates generated by the procedure that excludes t_τ.

Panel A: Estimates based on time bars with one time unit.

News Type	**Good News**			**Bad News**			**No News**									**MME**		**MLE(α_i = 0.4)**	
Time Bar i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	$\overline{\lvert V_i^B - V_i^S \rvert}$	80	μ_i	200
Buy vol. V_i^B	225	225	450	75	75	150	100	100	100	100	100	100	100	100	100	$\overline{(V_i^B + V_i^S)}$	280	ε_i	100
Sell vol. V_i^S	75	75	150	225	225	450	100	100	100	100	100	100	100	100	100	PIN_i	**0.286**	PIN_i	**0.286**
Time Unit	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	$VPIN_i$	-	$VPIN_i$	0.2

Panel B: Estimates based on trading days with three time units.

News Type	**Good News**		**Bad News**		**No News**			**MME**		**MLE(α_t = 0.4)**	
Day t	1		2		3	4	5	$\overline{\lvert V_t^B - V_t^S \rvert}$	240	μ_t	200
Buy vol. V_t^B	900		300		300	300	300	$\overline{(V_t^B + V_t^S)}$	840	ε_t	100
Sell vol. V_t^S	300		900		300	300	300	PIN_t	**0.286**	PIN_t	**0.286**
Trading Time	3		3		3	3	3	$VPIN_t$	-	$VPIN_t$	0.2

Panel C: Estimates based on buckets with fixed trading volume of 600 shares.

News Type	**Good News**		**Bad News**		**No News**			**MME**		**MLE(α_τ=0.571)**	
Bucket τ	1	2	3	4	5	6	7	$\overline{\lvert V_\tau^B - V_\tau^S \rvert}$	171.4	μ_τ	200
Buy vol. V_τ^B	450	450	150	150	300	300	300	$\overline{(V_\tau^B + V_\tau^S)}$	600	ε_τ	100
Sell vol. V_τ^S	150	150	450	450	300	300	300	PIN_τ	-	PIN_τ	0.364
Volume Time	2	1	2	1	3	3	3	$VPIN_\tau$	**0.286**	$VPIN_\tau$	**0.286**

Panel D: Estimates based on buckets with fixed trading volume of 600 shares and short volume time

News Type	**Good News**		**Bad News**		**No News**			**MME**		**MLE(α_τ=0.571)**	
Bucket τ	1	2	3	4	5	6	7	$\overline{\lvert V_\tau^B - V_\tau^S \rvert}$	171.4	μ_τ	340.90
Buy vol. V_τ^B	450	450	150	150	300	300	300	$\overline{(V_\tau^B + V_\tau^S)}$	600	ε_τ	109.09
Sell vol. V_τ^S	150	150	450	450	300	300	300	PIN_τ	-	PIN_τ	0.472
Volume Time	1	1	1	1	3	3	3	$VPIN_\tau$	**0.286**	$VPIN_\tau$	**0.348**

Table 1: A simple numeric illustration for estimating PIN and VPIN.

Assuming equal arrival rates for unformed buyers and sellers ($\varepsilon_b = \varepsilon_s = \varepsilon$) and that the news type is known, we can obtain approximate MLE estimates for the probability of information event occurrence α, the informed traders' arrival rate μ and the uninformed traders' arrival rate ε. For instance, in Panel C, these MLE estimates are derived as follows: α_τ = 4/7 ≈ 0.571, ε_τ = (6×300+2×(150+150))/(6×1+2×(2+1)) = 100, $\varepsilon_\tau + \mu_\tau$ = (2×(450+450))/(2×(2+1)) = 300, and μ_τ = 300 – 100 = 200. Then, PIN MLE estimate is 0.364 = $(\alpha_\tau\mu_\tau)/(2\varepsilon_\tau + \alpha_\tau\mu_\tau)$ and VPIN estimate is 0.286 = $\alpha_\tau \times (\mu_\tau/(2\varepsilon_\tau + \mu_\tau))$. Moreover, Easley et al.'s (2012b) MME generates a VPIN estimate, 0.286 = $\overline{\lvert V_\tau^B - V_\tau^S \rvert} \Big/ \overline{(V_\tau^B + V_\tau^S)}$ = ((4×|450−150|+3×|300−300|)/7)/((4×|450+150|+3×|300+300|)/7) = 171.429/600, based on fixed-sized buckets.

Interpretation: Panels A and B show that, when the buy and sell volumes are measured in a fixed time length, the Easley et al.'s MME serves to measure the PIN. Panels B and C show that, based on buckets, the Easley et al.'s MME measures the VPIN but not the PIN. Panels C and D show that Easley et al.'s VPIN metric (MME) does not incorporate the information of volume time.