Flow Toxicity and Liquidity in a High-Frequency World

1 Một số khái niệm

1.1 Order Flow Toxicity

Order flow toxicity được định nghĩa là khi các nhà tạo lập thị trường (market makers) bị bất lợi do dòng lệnh từ các nhà giao dịch có thông tin. Họ cung cấp thanh khoản nhưng không ý thức được rằng mình đang giao dịch ở mức bất lợi, dẫn đến thua lỗ.
Tác động: Giảm động lực của market makers, ảnh hưởng tiêu cực đến thanh khoản thị trường và gia tăng biến động giá.

1.2 VPIN: Volume-Synchronized Probability of Informed Trading

VPIN dựa trên sự mất cân bằng của khối lượng giao dịch (imbalance volume) và cường độ giao dịch (trade intensity) được tính trong Volume Time (thời gian dựa trên khối lượng thay vì thời gian thực).
Không yêu cầu ước lượng tham số quan sát, giúp nó phù hợp với các thị trường tần suất cao.
VPIN hoạt động như một chỉ báo hiệu quả cho biến động ngắn hạn do độc tính của dòng lệnh (toxicity-induced volatility) và cải thiện khả năng dự đoán biến động hơn so với các phương pháp truyền thống.

1.3 Bulk Volume Classification (BVC)

Khái niệm: Quy trình phân loại giao dịch thành *buy* hoặc *sell* dựa trên khối lượng lệnh trong một khoảng thời gian cụ thể.
Ưu điểm: Được tối ưu hóa cho các thị trường tần suất cao, nơi các phương pháp phân loại truyền thống như *tick test* hoặc *quote test* có thể không phù hợp.
Bid price: Giá mua - giá cao nhất mà người mua sẵn sàng trả.
Ask price: Giá bán - giá thấp nhất mà người bán sẵn sàng chấp nhận.
Spread: Spread = Ask Price − Bid Price

1.4 High-Frequency Trading Firms

Dù chỉ chiếm khoảng 2% trong tổng số gần 20.000 công ty giao dịch, các HF trading firms lại chiếm:

- Hơn 70% khối lượng giao dịch trên thị trường chứng khoán Mỹ (từ 2009).

- Gần 50% khối lượng giao dịch trên thị trường hợp đồng tương lai.

Hoạt động chính của HF trading firms thường là tạo lập thị trường (market making), cung cấp thanh khoản thông qua lệnh thụ động (passive orders). Họ không đặt cược theo hướng giá cả (no directional bets), mà kiếm lợi nhuận từ biên độ nhỏ thông qua khối lượng giao dịch lớn.

1.5 Volume Bucketing

Dữ liệu được chia thành các khoảng sao cho mỗi *bucket* có khối lượng giao dịch bằng nhau, giúp giảm tác động của thời gian hoặc biến động cụm.
Ví dụ: Nếu tổng khối lượng giao dịch là 1 triệu cổ phiếu và bạn muốn 10 *buckets*, mỗi bucket sẽ chứa 100.000 cổ phiếu, bất kể thời gian khớp lệnh kéo dài bao lâu.
Time Bar: Biểu đồ mà mỗi thanh (bar) đại diện cho một khoảng thời gian cố định (1 phút, 5 phút, 1 giờ, v.v.).
Bulk Bar: Biểu đồ mà mỗi thanh chỉ được tạo ra khi đạt đến một lượng giao dịch nhất định, thay vì dựa trên thời gian cố định.

Bố Cục Bài Báo

- **Phần 1:** Khung lý thuyết và cách PIN ảnh hưởng đến chênh lệch giá mua-bán.
- **Phần 2:** Quy trình ước tính chỉ số VPIN.
- **Phần 3:** Đánh giá độ tin cậy của VPIN.
- **Phần 4:** Ước tính VPIN cho chỉ số cổ phiếu và hợp đồng tương lai dầu.
- **Phần 5:** Khả năng dự báo biến động của VPIN.
- **Phần 6:** Tóm tắt các phát hiện.

2 Nội dung chính của bài báo

2.1 The Model - Mô hình

Trong phần này mô tả mô hình cơ bản để suy ra mức độ độc hại của dòng lệnh. Bắt đầu với một mô hình vi cấu trúc thị trường tiêu chuẩn, từ đó xây dựng thước đo mức độ độc hại của dòng lệnh, PIN, và sau đó điều chỉnh PIN để áp dụng cho thị trường giao dịch tần suất cao.
Một loạt nghiên cứu (Easley và O'Hara 1987, 1992; Easley, Kiefer, O'Hara và Paperman 1996; Easley, Engle, O'Hara và Wu 2008) đã chứng minh cách một mô hình vi cấu trúc thị trường có thể được ước tính cho từng cổ phiếu riêng lẻ bằng cách sử dụng dữ liệu giao dịch để xác định xác suất giao dịch dựa trên thông tin (**PIN**).
Mô hình vi cấu trúc này xem giao dịch như một trò chơi giữa các nhà cung cấp thanh khoản và các nhà giao dịch (position takers), được lặp lại qua các khoảng thời gian giao dịch $i = 1, ..., I$. Ở đầu mỗi khoảng giao dịch, một sự kiện thông tin có thể xảy ra hoặc không, với xác suất α. Nếu thông tin là tin tốt, các nhà giao dịch có thông tin biết rằng vào cuối khoảng thời gian giao dịch, tài sản sẽ có giá $\overline{S}_i$, và nếu thông tin là tin xấu, tài

sản sẽ có giá $\underline{S}_i$, với $\overline{S}_i > \underline{S}_i$. Tin tốt xảy ra với xác suất $(1-\delta)$, trong khi tin xấu xảy ra với xác suất còn lại δ.

α là xác suất mà một sự kiện thông tin (information event) xảy ra ở đầu mỗi khoảng thời gian giao dịch. Nếu một sự kiện thông tin xảy ra, nó có thể là tin tốt (good news) hoặc tin xấu (bad news). ***Xác suất tin tốt:*** $(1-\delta)$ ***Xác suất tin xấu:*** δ*. Nếu không có sự kiện thông tin xảy ra (xác suất $1-\alpha$), thị trường chỉ có giao dịch từ các nhà giao dịch không có thông tin (uninformed traders).*
Sau khi sự kiện thông tin xảy ra hoặc không xảy ra, hoạt động giao dịch trong khoảng thời gian đó bắt đầu, với các nhà giao dịch xuất hiện theo **quá trình Poisson** trong suốt khoảng thời gian giao dịch. Trong những khoảng có sự kiện thông tin, các lệnh từ nhà giao dịch có thông tin đến với tỷ lệ μ. Những nhà giao dịch này sẽ **mua nếu có tin tốt** và **bán nếu có tin xấu**. Trong mỗi khoảng thời gian, các lệnh mua và bán từ những nhà giao dịch không có thông tin đến với tỷ lệ ε cho cả hai phía.
Mô hình cấu trúc liên kết các kết quả quan sát được trên thị trường (tức là các lệnh mua và bán) với các thông tin không quan sát được và các quy trình đặt lệnh cơ bản. Các nghiên cứu trước đây tập trung vào việc ước lượng các tham số xác định những quy trình này bằng phương pháp **hợp lý cực đại** (maximum likelihood method).
Một cách trực quan, mô hình diễn giải mức độ mua và bán bình thường của một cổ phiếu là giao dịch không có thông tin và sử dụng dữ liệu để xác định tỷ lệ dòng lệnh không có thông tin, ε. Các giao dịch mua hoặc bán bất thường được hiểu là giao dịch dựa trên thông tin và được sử dụng để xác định μ. Số khoảng thời gian xuất hiện giao dịch mua hoặc bán bất thường được dùng để xác định α và δ.
Một nhà cung cấp thanh khoản sử dụng kiến thức về các tham số này để xác định mức giá mà anh ta sẵn sàng mua vào (bid) và mức giá mà anh ta sẵn sàng bán ra (ask). Hai mức giá này khác nhau, tạo ra chênh lệch giá mua-bán (bid-ask spread), vì nhà cung cấp thanh khoản không biết liệu đối tác giao dịch của mình có phải là nhà giao dịch có thông tin hay không.
Khoảng chênh lệch này chính là sự khác biệt giữa giá trị kỳ vọng của tài sản khi có người mua từ nhà cung cấp thanh khoản và giá trị kỳ vọng của tài sản khi có người bán cho nhà cung cấp thanh khoản. Hai kỳ vọng có điều kiện này khác nhau do vấn đề lựa chọn bất lợi (adverse selection), được gây ra bởi khả năng xuất hiện các nhà giao dịch có thông tin tốt hơn.
Khi giao dịch diễn ra, các nhà cung cấp thanh khoản quan sát các lệnh và được mô hình hóa theo cách sử dụng quy tắc Bayes để cập nhật niềm tin của họ về mức độ độc hại của dòng lệnh, được mô tả trong mô hình bằng các ước lượng tham số. Gọi $P(t) = (P_n(t), P_b(t), P_g(t))$ là niềm tin của nhà cung cấp thanh khoản về các sự kiện "không có tin tức"(n), "tin xấu"(b), "tin tốt"(g) tại thời điểm t. Niềm tin ban đầu của họ tại thời điểm $t = 0$ là $P(0) = (1-\alpha, \alpha\delta, \alpha(1-\delta))$.
Để xác định giá mua hoặc giá bán tại thời điểm t, nhà cung cấp thanh khoản cập nhật niềm tin của mình dựa trên sự xuất hiện của một lệnh giao dịch có liên quan. Giá trị kỳ vọng của tài sản tại thời điểm t, với điều kiện dựa trên lịch sử giao dịch trước đó, được tính như sau:

$$E[S_i|t] = P_n(t)S_i^* + P_b(t)S_i + P_g(t)\overline{S}_i,$$

trong đó

$$S_i^* = \delta\underline{S}_i + (1-\delta)\overline{S}_i,$$

là giá trị kỳ vọng ban đầu của tài sản.
Giá mua $B(t)$ là giá trị kỳ vọng của tài sản có điều kiện khi có người muốn bán tài sản cho nhà cung cấp thanh khoản. Do đó, ta có:

$$B(t) = E[S_i|t] - \frac{\mu P_b(t)}{\varepsilon + \mu P_b(t)} \left(E[S_i|t] - \underline{S}_i\right).$$

Tương tự, giá bán $A(t)$ là giá trị kỳ vọng của tài sản có điều kiện khi có người muốn mua tài sản từ nhà cung cấp thanh khoản. Do đó, ta có:

$$A(t) = E[S_i|t] + \frac{\mu P_g(t)}{\varepsilon + \mu P_g(t)} \left(\overline{S}_i - E[S_i|t]\right).$$

Những phương trình này minh họa vai trò rõ ràng của sự xuất hiện của các nhà giao dịch có thông tin và không có thông tin trong việc ảnh hưởng đến báo giá. Nếu không có nhà giao dịch có thông tin ($\mu = 0$), thì giao dịch không mang thông tin, và do đó giá mua và giá bán đều bằng với giá trị kỳ vọng ban đầu của tài sản.
Ngược lại, nếu không có nhà giao dịch không có thông tin ($\varepsilon = 0$), thì giá mua và giá bán lần lượt đạt mức giá tối thiểu và tối đa. Ở các mức giá này, không có nhà giao dịch có thông tin nào thực hiện giao dịch, và thị trường về cơ bản sẽ ngừng hoạt động.
Thông thường, cả nhà giao dịch có thông tin và không có thông tin đều có mặt trên thị trường, do đó giá mua sẽ thấp hơn $E[S_i|t]$ và giá bán sẽ cao hơn $E[S_i|t]$.
Chênh lệch giá mua-bán tại thời điểm t được ký hiệu là $\Sigma(t) = A(t) - B(t)$. Chênh lệch này được tính bằng:

$$\Sigma(t) = \frac{\mu P_g(t)}{\varepsilon + \mu P_g(t)} \left(\overline{S}_i - E[S_i|t]\right) + \frac{\mu P_b(t)}{\varepsilon + \mu P_b(t)} \left(E[S_i|t] - \underline{S}_i\right).$$

Thuật ngữ đầu tiên trong phương trình chênh lệch biểu thị xác suất rằng một lệnh mua là một giao dịch dựa trên thông tin, nhân với mức lỗ kỳ vọng khi giao dịch với một người mua có thông tin. Thuật ngữ thứ hai là một biểu thức đối xứng cho các lệnh bán.
Chênh lệch cho các báo giá ban đầu trong khoảng giao dịch, Σ, có một dạng đặc biệt đơn giản trong trường hợp tự nhiên khi các sự kiện tin tốt và tin xấu có xác suất bằng nhau. Nghĩa là, nếu $\delta = 1 - \delta$, thì:

$$\Sigma = \frac{\alpha\mu}{\alpha\mu + 2\varepsilon} \left(\overline{S}_i - \underline{S}_i\right).$$

Thành phần quan trọng của mô hình này là xác suất một lệnh được đặt bởi một nhà giao dịch có thông tin, được gọi là PIN. Có thể dễ dàng chứng minh rằng xác suất giao dịch đầu tiên trong một khoảng thời gian là giao dịch dựa trên thông tin được tính bằng:

$$PIN = \frac{\alpha\mu}{\alpha\mu + 2\varepsilon},$$

trong đó $\alpha\mu + 2\varepsilon$ là tỷ lệ đến của tất cả các lệnh và $\alpha\mu$ là tỷ lệ đến của các lệnh dựa trên thông tin. Do đó, PIN đo lường tỷ lệ lệnh đến từ các nhà giao dịch có thông tin so với tổng dòng lệnh, và phương trình chênh lệch giá chỉ ra rằng đây là yếu tố chính quyết định mức chênh lệch giá mua-bán.
Những phương trình này minh họa rằng các nhà cung cấp thanh khoản cần ước tính chính xác PIN để xác định mức tối ưu để tham gia thị trường. Việc PIN tăng lên bất ngờ sẽ dẫn đến tổn thất cho các nhà cung cấp thanh khoản không điều chỉnh giá của họ.

2.2 Chỉ số VPIN và Ước lượng Tham số

Cách tiếp cận tiêu chuẩn để tính toán mô hình PIN sử dụng phương pháp ước lượng hợp lý cực đại (maximum likelihood) để ước tính các tham số không quan sát được $(\alpha, \mu, \delta, \varepsilon)$, vốn quyết định quá trình ngẫu nhiên của các giao dịch, và sau đó suy ra giá trị PIN từ các ước lượng tham số này.
Trong phần này đề xuất một phương pháp ước lượng trực tiếp về mức độ độc hại (toxicity), không yêu cầu quá trình ước lượng số trung gian đối với các tham số không quan sát được. Chúng tôi cập nhật thước đo này trong thời gian dựa trên khối lượng giao dịch (volume time) nhằm cố gắng điều chỉnh theo tốc độ xuất hiện của thông tin mới trên thị trường.
Cách tiếp cận dựa trên khối lượng này, mà chúng tôi gọi là **VPIN**, cung cấp một chỉ số đơn giản để đo lường mức độ độc hại của dòng lệnh trong môi trường giao dịch tần suất cao. Trước tiên, chúng tôi sẽ thảo luận về vai trò của thông tin và thời gian trong giao dịch tần suất cao.

2.2.1 Bản chất của thông tin và thời gian

Trong mô hình giao dịch tuần tự tiêu chuẩn, thông tin thường được xem như dữ liệu cung cấp dấu hiệu về giá trị tương lai của tài sản. Trong bối cảnh thị trường cổ phiếu, thông tin có thể liên quan đến các sự kiện tương lai, chẳng hạn như triển vọng của công ty hoặc thị trường đối với sản phẩm của công ty.
Trong một thị trường hiệu quả, giá trị của tài sản sẽ hội tụ về giá trị thông tin đầy đủ của nó, khi các nhà giao dịch có thông tin tìm cách thu lợi nhuận từ thông tin của họ bằng cách giao dịch. Vì các nhà tạo lập thị trường có thể giữ vị thế mua hoặc bán cổ phiếu, nên biến động giá trị tương lai của tài sản ảnh hưởng đến lợi nhuận của họ. Do đó, họ cố gắng suy luận bất kỳ thông tin mới tiềm ẩn nào từ các mô hình giao dịch. Chính niềm tin được cập nhật của họ sẽ được phản ánh vào giá mua và giá bán của họ.
Trong một thế giới giao dịch tần suất cao, các nhà tạo lập thị trường phải đối mặt với cùng một vấn đề cơ bản, mặc dù khung thời gian mà họ hoạt động làm thay đổi một số khía cạnh theo những cách thú vị.
Một nhà tạo lập thị trường tần suất cao, người dự đoán sẽ nắm giữ cổ phiếu trong vài phút, sẽ bị ảnh hưởng bởi thông tin tác động đến giá trị của cổ phiếu trong khoảng thời gian đó. Thông tin này có thể liên quan đến các yếu tố cơ bản của tài sản, nhưng cũng có thể phản ánh các yếu tố liên quan đến bản chất của hoạt động giao dịch trên toàn thị trường hoặc nhu cầu thanh khoản cụ thể trong một khoảng thời gian nhất định. Ví dụ, trong một hợp đồng tương lai, thông tin làm gia tăng nhu cầu phòng ngừa rủi ro (hedging demand) đối với hợp đồng thường sẽ ảnh hưởng đến giá hợp đồng tương lai, và do đó, nó có ý nghĩa đối với một nhà tạo lập thị trường. Định nghĩa rộng hơn về thông tin này có nghĩa là các sự kiện thông tin có thể xảy ra thường xuyên trong ngày và chúng có thể có mức độ quan trọng khác nhau đối với biên độ biến động giá trong tương lai. Tuy nhiên, sự tồn tại của chúng vẫn được phản ánh thông qua bản chất và thời điểm của các giao dịch.
Khía cạnh quan trọng nhất của mô hình giao dịch tần suất cao là các giao dịch không diễn ra cách đều nhau theo thời gian. Các lệnh giao dịch xuất hiện với tần suất không đều, và một số giao dịch quan trọng hơn những giao dịch khác vì chúng tiết lộ lượng thông tin khác nhau.
Ví dụ, như Hình 1 minh họa, hoạt động giao dịch trong hợp đồng tương lai E-mini S&P 500 (đường màu xanh và trục tỷ lệ bên trái của biểu đồ) và hợp đồng tương lai EUR/USD

(đường màu đỏ và trục tỷ lệ bên phải của biểu đồ) thể hiện tính chu kỳ trong ngày khác nhau. Sự xuất hiện của thông tin mới trên thị trường sẽ kích hoạt làn sóng các quyết định được đưa ra, dẫn đến sự bùng nổ về khối lượng giao dịch. Thông tin liên quan đến các sản phẩm khác nhau xuất hiện vào những thời điểm khác nhau, do đó tạo ra sự biến động khối lượng giao dịch theo chu kỳ trong ngày là khác biệt.
Trong nghiên cứu này, thay vì mô hình hóa thời gian theo đồng hồ, chúng tôi làm việc với thời gian theo khối lượng (*volume time*). Easley và O'Hara (1992) đã phát triển ý tưởng rằng khoảng thời gian giữa các giao dịch có mối tương quan với sự xuất hiện của thông tin mới, tạo cơ sở để chúng tôi xem xét thời gian giao dịch thay vì thời gian đồng hồ. Có vẻ hợp lý khi cho rằng một thông tin càng quan trọng thì càng thu hút nhiều khối lượng giao dịch hơn. Bằng cách lấy mẫu mỗi khi thị trường trao đổi một lượng khối lượng cố định, chúng tôi cố gắng mô phỏng sự xuất hiện của thông tin có mức độ quan trọng tương đương trên thị trường. Nếu một thông tin cụ thể tạo ra khối lượng giao dịch gấp đôi so với một thông tin khác, chúng tôi sẽ lấy số quan sát gấp đôi, qua đó nhân đôi trọng số của nó trong mẫu.

2.2.2 Phân chia theo khối lượng (Volume Bucketing)

Trong ví dụ trên, nếu chúng tôi lấy một mẫu hợp đồng tương lai E-mini S&P 500 sau mỗi 200.000 hợp đồng được giao dịch, trung bình chúng tôi sẽ thu được khoảng chín mẫu mỗi ngày. Vào những ngày có hoạt động giao dịch sôi động, số lượng mẫu có thể là bội số lớn của chín, trong khi vào những ngày ít giao dịch hơn, số điểm dữ liệu thu thập được sẽ ít hơn. Vì hợp đồng tương lai EUR/USD có khối lượng giao dịch trung bình hàng ngày chỉ bằng khoảng một phần mười so với hợp đồng tương lai E-mini S&P 500, nên để đạt được mục tiêu lấy chín mẫu mỗi ngày, khoảng cách khối lượng giữa hai lần lấy mẫu sẽ cần được giảm xuống khoảng 20.000 hợp đồng. Do mô hình giao dịch trong ngày khác nhau giữa hai hợp đồng này, khi chúng tôi thu thập mẫu quan sát đầu tiên của ngày đối với hợp đồng tương lai E-mini S&P 500, thì gần như đồng thời chúng tôi đã thu được mẫu quan sát thứ tư của ngày đối với hợp đồng tương lai EUR/USD.
Để thực hiện phương pháp lấy mẫu dựa trên khối lượng này, chúng tôi nhóm các giao dịch liên tiếp vào các khối lượng bằng nhau, gọi là volume buckets, với kích thước V được xác định ngoại sinh. Một volume bucket là tập hợp các giao dịch có tổng khối lượng bằng V. Nếu giao dịch cuối cùng để hoàn thành một bucket có khối lượng lớn hơn mức yêu cầu, phần dư thừa sẽ được chuyển sang bucket tiếp theo. Chúng tôi ký hiệu r là chỉ số của các volume buckets có khối lượng bằng nhau. Một thuật toán chi tiết cho quá trình đóng gói khối lượng này được trình bày trong Phụ lục trực tuyến. Phương pháp lấy mẫu theo volume buckets cho phép chúng tôi chia phiên giao dịch thành các khoảng thời gian có nội dung thông tin tương đương, trong đó sự mất cân bằng giao dịch có tác động kinh tế đáng kể đối với các nhà cung cấp thanh khoản.

2.2.3 Phân loại khối lượng mua và khối lượng bán

Một vấn đề mà chúng tôi chưa đề cập là cách phân biệt giữa khối lượng mua và khối lượng bán. Nhắc lại rằng khối lượng có dấu (*signed volume*) là cần thiết vì khả năng tương quan của nó với mức độ độc hại của dòng lệnh (*order toxicity*). Mặc dù tổng khối lượng giao dịch có thể báo hiệu sự xuất hiện của thông tin mới, nhưng hướng của khối lượng giao dịch lại thể hiện ý nghĩa của nó đối với sự thay đổi giá. Do đó, sự áp đảo của khối lượng mua (bán) có thể cho thấy mức độ độc hại của dòng lệnh phát sinh từ sự xuất hiện của tin tốt (xấu).

Nghiên cứu vi cấu trúc thị trường (*microstructure research*) đã dựa vào các thuật toán dựa trên dấu tick (*tick-based algorithms*) để xác định hướng giao dịch. Tuy nhiên, việc phân loại giao dịch luôn là một vấn đề nan giải.

Một trong những vấn đề là cách báo cáo giao dịch trên các thị trường có thể xử lý lệnh mua và lệnh bán khác nhau. Ví dụ, Sở Giao dịch Chứng khoán New York (*NYSE*) chỉ báo cáo một giao dịch nếu một lệnh bán lớn được khớp với nhiều lệnh mua trên sổ lệnh, nhưng sẽ báo cáo nhiều giao dịch nếu một lệnh mua lớn được khớp với nhiều lệnh bán.

Tương tự, việc chia nhỏ các lệnh lớn thành nhiều lệnh nhỏ hơn có nghĩa là các giao dịch diễn ra trong khoảng thời gian ngắn thực tế không phải là những quan sát độc lập. Để giải quyết những vấn đề này, các nhà nghiên cứu thực nghiệm đã sử dụng phương pháp tổng hợp các giao dịch cùng phía thị trường trong khoảng thời gian ngắn thành một quan sát duy nhất.

Một khó khăn thứ hai là việc xác định hướng giao dịch cũng đòi hỏi phải so sánh giá giao dịch với báo giá hiện hành. Các nhà giao dịch thực hiện lệnh tại giá mua (*bid*) của nhà tạo lập thị trường được coi là người bán, trong khi những người thực hiện lệnh tại giá bán (*ask*) được coi là người mua. Các giao dịch nằm giữa khoảng giá này thường được phân loại bằng thuật toán dựa trên dấu tick (*tick-based algorithm*).

Thuật toán Lee-Ready (1991) cũng đề xuất sử dụng độ trễ năm giây giữa báo giá được ghi nhận và giá giao dịch để phản ánh thực tế rằng cơ chế báo giá lên hệ thống không giống với cơ chế báo cáo giao dịch. Ngay cả trong một môi trường giao dịch đơn giản hơn với các nhà giao dịch chuyên biệt (*specialist trading*), lỗi trong phân loại giao dịch vẫn khá đáng kể.

Trong môi trường giao dịch tần suất cao, việc phân loại giao dịch trở nên phức tạp hơn nhiều. Trong các thị trường hợp đồng tương lai mà chúng tôi nghiên cứu, không có nhà giao dịch chuyên biệt (*specialist*), và thanh khoản được cung cấp thông qua một sổ lệnh điện tử chứa các lệnh giới hạn được đặt bởi nhiều loại nhà giao dịch khác nhau. Trong thị trường điện tử này, một nhà giao dịch có thể khớp lệnh tại mức giá giống với giao dịch trước đó hoặc có thể gửi một lệnh giới hạn cải thiện giá giao dịch trước đó, khiến quy tắc tick có thể gán sai hướng của giao dịch. Ngoài ra, việc chia nhỏ lệnh là tiêu chuẩn, việc hủy báo giá và lệnh diễn ra tràn lan, và khối lượng giao dịch lớn đến mức áp đảo. Sử dụng dữ liệu hợp đồng tương lai E-mini S&P 500 từ tháng 5 năm 2010, chúng tôi nhận thấy rằng trung bình một ngày có 2.650.391 thay đổi báo giá tốt nhất (*best-bid-or-offer - BBO*) do việc thêm hoặc hủy lệnh, và 789.676 thay đổi báo giá do các giao dịch. Vì BBO thay đổi nhiều lần giữa các giao dịch, nhiều hợp đồng được giao dịch tại cùng một mức giá thực tế đã diễn ra cả ở giá mua và giá bán. Trong thế giới giao dịch tần suất cao này, việc áp dụng các thuật toán tiêu chuẩn lên từng giao dịch riêng lẻ trở nên có vấn đề.

$$V_\tau^B = \sum_{i=t(\tau-1)+1}^{t(\tau)} V_i \cdot Z\left(\frac{P_i - P_{i-1}}{\sigma_{\Delta P}}\right)$$

$$V_\tau^S = \sum_{i=t(\tau-1)+1}^{t(\tau)} V_i \cdot \left[1 - Z\left(\frac{P_i - P_{i-1}}{\sigma_{\Delta P}}\right)\right] = V - V_\tau^B, \tag{7}$$

trong đó $t(\tau)$ là chỉ số của thanh thời gian cuối cùng được bao gồm trong *volume bucket* thứ τ, Z là hàm phân phối tích lũy (*CDF - Cumulative Distribution Function*) của phân phối chuẩn chuẩn tắc, và $\sigma_{\Delta P}$ là ước lượng độ lệch chuẩn của sự thay đổi giá giữa các thanh thời gian. Quy trình của chúng tôi chia đều khối lượng giao dịch trong một thanh thời gian thành khối lượng mua và khối lượng bán nếu không có sự thay đổi giá từ

đầu đến cuối thanh thời gian. Ngược lại, nếu giá tăng, khối lượng sẽ được phân bổ nhiều hơn về phía mua so với bán, và mức phân bổ này phụ thuộc vào độ lớn của sự thay đổi giá so với phân phối của các biến động giá.
Một điểm khác biệt quan trọng giữa phương pháp phân loại khối lượng (*bulk classification*) và thuật toán Lee-Ready là thuật toán Lee-Ready phân loại toàn bộ khối lượng giao dịch thành hoặc mua hoặc bán, trong khi phương pháp phân loại khối lượng chia một phần khối lượng là mua và phần còn lại là bán. Nói cách khác, thuật toán Lee-Ready cung cấp một phân loại rời rạc, trong khi thuật toán phân loại khối lượng là liên tục. Điều này có nghĩa là ngay cả trong trường hợp cực đoan khi một thanh thời gian duy nhất lấp đầy một *volume bucket*, khối lượng vẫn có thể được cân bằng hoàn toàn theo phương pháp phân loại khối lượng, phụ thuộc vào $\frac{P_i - P_{i-1}}{\sigma_{\Delta P}}$.
Mục đích chính của chúng tôi khi sử dụng khối lượng giao dịch là để tính toán sự mất cân bằng lệnh (*order imbalance*). Gọi $OI_\tau = |V_\tau^B - V_\tau^S|$ là sự mất cân bằng lệnh trong *volume bucket* τ. Thước đo này, tất nhiên, chỉ là một xấp xỉ của mất cân bằng lệnh thực tế vì nó dựa trên phương pháp phân loại khối lượng giao dịch theo xác suất.
Chúng tôi trước tiên xem xét cách $E[OI_\tau]$ liên quan đến tốc độ giao dịch bằng cách chứng minh rằng nó không bị ảnh hưởng bởi một phép co giãn đơn giản của khối lượng giao dịch. Giả sử rằng khối lượng của mỗi thanh thời gian được co giãn theo một hệ số $\beta > 0$, tức là $V_i^* = \beta V_i$, và sự mất cân bằng khối lượng được phân phối đồng đều trong *bucket*. Khi đó, số lượng thanh thời gian cần thiết để lấp đầy một *bucket* sẽ tỷ lệ nghịch với β, tức là $\frac{t(\tau) - t(\tau-1)}{\beta}$. Từ phương trình (7), giá trị kỳ vọng của mất cân bằng lệnh vẫn không đổi:

$$E[OI_\tau^*] = E\left[\left|V_\tau^{*B} - V_\tau^{*S}\right|\right] = \frac{1}{\beta} E\left[\left|\beta V_\tau^B - \beta V_\tau^S\right|\right] = E[OI_\tau]. \tag{8}$$

Thứ hai, chúng tôi đặt câu hỏi liệu trong một giới hạn hợp lý, khoảng thời gian của một thanh thời gian có ảnh hưởng đến thước đo mất cân bằng lệnh hay không. Để xác định điều này, chúng tôi tính toán mất cân bằng lệnh cho hợp đồng tương lai E-mini S&P 500 trong giai đoạn từ ngày 1 tháng 1 năm 2008 đến ngày 1 tháng 8 năm 2011, sử dụng các thanh thời gian có độ dài từ 1 đến 240 phút. Với mỗi quy định về thanh thời gian, chúng tôi sử dụng 50 *volume buckets* mỗi ngày và tính toán tỷ lệ giữa mất cân bằng lệnh so với kích thước *bucket*, được đo lường bằng khối lượng trong mỗi *bucket*. Chúng tôi nhận thấy rằng tỷ lệ mất cân bằng lệnh so với kích thước *bucket* (như một hàm của số lượng trung bình các thanh thời gian trên mỗi *bucket*) tăng dần khi số thanh thời gian trong mỗi *bucket* giảm, nhưng không bao giờ tiến gần đến giá trị 1, và cuối cùng ổn định khi sử dụng các thanh thời gian quá dài một cách phi thực tế. Do đó, đối với các thanh thời gian có độ dài hợp lý, thời lượng của một thanh thời gian có rất ít tác động đến việc đo lường sự mất cân bằng lệnh.
Phương pháp này sẽ dẫn đến một số trường hợp phân loại sai khối lượng giao dịch. Tuy nhiên, mục tiêu của chúng tôi không phải là phân loại chính xác từng giao dịch riêng lẻ (một nhiệm vụ gần như bất khả thi), mà là phát triển một chỉ báo về sự mất cân bằng tổng thể của giao dịch, giúp xây dựng một thước đo mức độ độc hại của dòng lệnh (*toxicity*). Chúng tôi sử dụng các thanh thời gian (*time bars*) để tạo khoảng thời gian cho giá thị trường điều chỉnh theo hướng giao dịch, thông tin mà chúng tôi cố gắng khôi phục thông qua phương pháp phân loại khối lượng (*bulk classification*). Sau đó, trong bài báo này, chúng tôi sẽ trình bày bằng chứng cho thấy phương pháp phân loại khối lượng của chúng tôi mang lại kết quả hữu ích hơn trong việc ước tính mức độ độc hại của dòng lệnh so với phương pháp phân loại chi tiết dựa trên dữ liệu giao dịch thô.

2.2.4 Xác suất giao dịch có thông tin đồng bộ theo khối lượng (Chỉ số độc hại dòng lệnh VPIN)

Mô hình PIN tiêu chuẩn chỉ xem xét số lượng lệnh mua và bán để suy ra thông tin về cấu trúc thông tin cơ bản; trong đó không có vai trò rõ ràng của khối lượng giao dịch. Trong các thị trường tần suất cao mà chúng tôi phân tích, số lượng giao dịch là một vấn đề phức tạp. Quay lại nền tảng lý thuyết của PIN, điều chúng tôi thực sự muốn là thông tin về ý định giao dịch phát sinh từ các nhà giao dịch có hoặc không có thông tin. Mối liên kết giữa các ý định giao dịch này và dữ liệu giao dịch thực tế có rất nhiều nhiễu, vì ý định giao dịch có thể bị chia nhỏ thành nhiều phần để giảm tác động thị trường, một lệnh có thể tạo ra nhiều giao dịch khớp lệnh, và các giao dịch dựa trên thông tin có thể được thực hiện dưới nhiều hình thức lệnh khác nhau. Vì những lý do này, chúng tôi xem mỗi giao dịch được báo cáo như một sự tổng hợp của các giao dịch có kích thước đơn vị (tức là một giao dịch gồm năm hợp đồng ở một mức giá p sẽ được xem như năm giao dịch, mỗi giao dịch một hợp đồng tại mức giá p). Quy ước này đưa cường độ giao dịch vào phân tích một cách rõ ràng.

Chúng tôi biết từ nghiên cứu của Easley, Engle, O'Hara và Wu (2008) rằng đối với mỗi khoảng thời gian, giá trị kỳ vọng của mất cân bằng lệnh là

$$E\left[\left|V_\tau^S - V_\tau^B\right|\right] \approx \alpha\mu$$

và tổng số lệnh kỳ vọng là

$$E[V_\tau^B + V_\tau^S] = \alpha\mu + 2\varepsilon.$$

Phương pháp phân chia theo khối lượng (*volume bucketing*) giúp chúng tôi ước lượng mô hình này một cách đơn giản. Cụ thể, chúng tôi chia ngày giao dịch thành các *volume buckets* có kích thước bằng nhau và xem mỗi *bucket* là tương đương với một khoảng thời gian tiếp nhận thông tin. Điều này có nghĩa là $V_\tau^B + V_\tau^S$ là một hằng số và bằng V cho mọi τ. Chúng tôi sau đó xấp xỉ mất cân bằng lệnh kỳ vọng bằng cách tính trung bình mất cân bằng lệnh qua n buckets.

Từ các giá trị được tính toán ở trên, chúng tôi có thể viết xác suất giao dịch có thông tin đồng bộ theo khối lượng (*volume-synchronized probability of informed trading*), hay chỉ số độc hại dòng lệnh VPIN, như sau:

$$VPIN = \frac{\alpha\mu}{\alpha\mu + 2\varepsilon} \approx \frac{\sum_{\tau=1}^{n}\left|V_\tau^S - V_\tau^B\right|}{nV}. \tag{9}$$

Ước lượng chỉ số VPIN yêu cầu lựa chọn V, tức là khối lượng trong mỗi *bucket*, và n, số lượng *buckets* được sử dụng để xấp xỉ mất cân bằng lệnh kỳ vọng. Trong thiết lập ban đầu, chúng tôi tập trung vào giá trị V bằng một phần năm mươi của khối lượng giao dịch trung bình hàng ngày. Nếu sau đó chúng tôi chọn $n = 50$, thì chỉ số VPIN sẽ được tính toán trên 50 *buckets*, và trong một ngày có khối lượng giao dịch trung bình, điều này tương ứng với việc tìm chỉ số VPIN hàng ngày. Kết quả của chúng tôi vẫn nhất quán với nhiều lựa chọn khác nhau của V và n, như sẽ được thảo luận trong Mục 5. Chỉ số VPIN được cập nhật sau mỗi *volume bucket*. Do đó, khi *bucket* thứ 51 được lấp đầy, chúng tôi loại bỏ *bucket* thứ 1 và tính toán VPIN mới dựa trên các *buckets* từ 2 đến 51. Chúng tôi cập nhật chỉ số VPIN theo thời gian khối lượng (*volume time*) vì hai lý do. Thứ nhất, chúng tôi muốn tốc độ cập nhật VPIN phản ánh tốc độ thông tin đến thị trường. Chúng tôi sử dụng khối lượng giao dịch như một đại diện cho tốc độ

tiếp nhận thông tin để đạt được mục tiêu này. Thứ hai, chúng tôi mong muốn mỗi lần cập nhật dựa trên một lượng thông tin tương đương. Trong các giai đoạn giao dịch có ít sự tham gia, khối lượng giao dịch có thể rất mất cân bằng, và trong những giai đoạn có khối lượng thấp như vậy, dường như ít có khả năng xuất hiện thông tin mới. Vì vậy, nếu cập nhật chỉ số VPIN theo thời gian đồng hồ (*clock time*), các lần cập nhật có thể dựa trên các lượng thông tin không đồng nhất.

Ví dụ, hãy xem xét giao dịch hợp đồng tương lai E-mini S&P 500 vào ngày 6 tháng 5 năm 2010. Khối lượng giao dịch trong ngày này (được nhớ đến vì sự kiện *flash crash*) cực kỳ cao, do đó quy trình của chúng tôi tạo ra 137 ước lượng của chỉ số VPIN, so với mức trung bình 50 ước lượng hàng ngày. Vì độ dài mẫu (n) cũng là 50, nên khoảng thời gian được sử dụng để tính toán một số ước lượng của chỉ số VPIN vào ngày 6 tháng 5 năm 2010 chỉ kéo dài vài giờ, so với mức trung bình là 24 giờ.

Hình 2 minh họa cách phạm vi thời gian trở nên "co giãn"tùy thuộc vào cường độ giao dịch, vốn được xem như một đại diện cho tốc độ tiếp nhận thông tin. Vào lúc 9:30 sáng (EST), dữ liệu được sử dụng để tính toán VPIN bao phủ gần như toàn bộ một ngày.

Tuy nhiên, khi Sở Giao dịch Chứng khoán New York (*NYSE*) mở cửa vào ngày 6 tháng 5 năm 2010, thuật toán của chúng tôi cập nhật chỉ số VPIN thường xuyên hơn và dựa trên một khoảng thời gian ngắn hơn theo thời gian đồng hồ. Đến 12:17 chiều, VPIN được tính toán dựa trên khoảng thời gian chỉ bằng một nửa ngày. Lưu ý rằng việc giảm phạm vi thời gian của mẫu dữ liệu không dẫn đến các ước lượng nhiễu hơn. Ngược lại, chỉ số VPIN tiếp tục thay đổi theo một xu hướng liên tục. Lý do là vì các khoảng thời gian không chứa cùng một lượng thông tin có thể so sánh được. Thay vào đó, chính các khoảng khối lượng (*volume ranges*) mới tạo ra lượng thông tin tương đương cho mỗi lần cập nhật.

Các mô hình GARCH cung cấp một cách tiếp cận thay thế để xử lý hiện tượng cụm biến động (*volatility clustering*), đặc trưng của dữ liệu tần suất cao khi lấy mẫu theo thời gian đồng hồ (*clock time*). Làm việc theo thời gian khối lượng (*volume time*) giúp giảm tác động của hiện tượng cụm biến động, vì chúng tôi tạo ra các ước lượng dựa trên các mẫu có khối lượng bằng nhau. Vì những biến động giá lớn thường đi kèm với khối lượng giao dịch lớn, nên việc lấy mẫu theo khối lượng có thể được xem như một đại diện cho việc lấy mẫu theo biến động giá. Kết quả thu được là một tập hợp các quan sát có phân phối gần với phân phối chuẩn hơn và ít có tính phương sai không đồng nhất (*heteroscedasticity*) hơn so với khi lấy mẫu đồng đều theo thời gian đồng hồ. Do đó, làm việc theo thời gian khối lượng có thể được xem như một phương pháp thay thế đơn giản cho việc sử dụng mô hình GARCH.