February 19, 2015

Mr. Brent Fields
Secretary
U.S. Securities and Exchange Commission
100 F Street, N.E.
Washington, D.C. 20549-1090

Re: Equity Market Structure

Thank you for the chance to comment on the SEC's Market Structure microsite. My comments are about a staff white paper called "Equity Market Structure Literature Review Part II: High Frequency Trading," a survey of empirical high frequency trading research posted there ("Paper").

More than one-third of the studies surveyed in the Paper - a dozen of them - analyze data the staff describe as "data for equity trading on NASDAQ that NASDAQ has made available to researchers ('NASDAQ Datasets')." The NASDAQ Datasets are unique in U.S. stock market research because, unlike other data available to academic researchers, trades and orders are grouped into participant categories. Each trade or order is marked as either high frequency trading activity or not, and because of those classifications research based on that data is influential, widely known, and often cited.

The NASDAQ Datasets might be unique but they are far from perfect. The Paper aptly describes some of their limitations, and suggests that those limitations could affect research based on the data. But problems with the data are even more troubling than the Paper sets out. I'll explore all that here.

Miscoded

According to researchers, the NASDAQ Datasets include all trades from 2008 and 2009 in 120 stocks. Nasdaq labeled a trade participant in the data as "HFT" if it were one of 26 firms designated by Nasdaq as a high frequency trading firm. All other participants went into a non-HFT bucket. HFT and non-HFT are the only two participant categories in the data. Certainly there were many more than 26 HFT firms active on Nasdaq during 2008-2009, so the rough sort of HFT and non-HFT participants in the NASDAQ Datasets should cause some worries.[1] SEC staff do worry about it, and the Paper discusses the problem:

> Another limitation of the NASDAQ Datasets is that they do not cover all HFT activity. For example, they do not include HFT at firms that also act as brokers for customers because this activity cannot be clearly identified. The NASDAQ Datasets thereby exclude the proprietary trading desks of large integrated broker-dealer firms. The NASDAQ Datasets also do not include HFT at firms that route their orders through integrated firms because this activity cannot be clearly identified, which may exclude smaller HFT firms that rely on other firms for market access.

Said another way - and more bluntly - any research on the NASDAQ Datasets that characterizes HFT and non-HFT behavior could be fatally mistaken because of obvious sampling bias.

The problem isn't only that the NASDAQ Datasets don't cover all HFT activity and "exclude the proprietary trading desks of large integrated broker-dealer firms," as the Paper says. And the problem isn't just that the

[1] Hirschey (2013) uses Nasdaq data similar to that in the NASDAQ Datasets. Hirschey's data includes "unique trade and trader-level data" that distinguish it, however.

NASDAQ Datasets "may exclude HFT firms that rely on other firms for market access" (direct market access or sponsored access firms). The problem is that all this data - as we'll see, an extraordinary amount - is not excluded at all. The data is included but *miscoded* as non-HFT in the NASDAQ Datasets, and researchers use that miscoded data to sort and characterize HFT and non-HFT behavior. Millions of trades and billions of traded shares might be coded non-HFT even though they're HFT activity. Compounding that already ugly sampling problem, and as an artifact of how and why they were systemically miscoded non-HFT, miscoded firms could have very different strategies and very different effects on market quality than the two dozen or so firms Nasdaq identified as HFT. No one knows.

An early draft of one research paper based on this data described the HFT/non-HFT labeling this way:

> Firms that others may define as HFT are not labeled as HFT firms here [in the NASDAQ Datasets] if they satisfy one of the following: firms like Lime Brokerage and Swift Trade who provide direct market access and other powerful trading tools to its customers, who are likely engaging in HFT and thus are likely HFT traders but are not labeled so; proprietary trading firms that are a desk of a larger, integrated firm, like Goldman Sachs or JP Morgan; an independent firm that is engaged in HFT activities, but who routes its trades through a MPID of a non-HFT type firm; firms that engage in HFT activities but because they are small are not considered in the study as being labeled a HFT firm.

Let's parse this. For a start, in the time period covered by the NASDAQ Datasets, Lime Brokerage was one of the largest sponsored access brokers in the market. In 2010, Lime described itself as "a leading provider of low-latency, high-throughput trading technologies to the professional trading community." It also said the firm "has been active in this space for over 10 years and a significant percentage of our clients are High Frequency Trading firms." In 2009, Forbes wrote that Lime "handles trades for 200 high-frequency trading firms and individuals." And yet, according to the description quoted here, not one of Lime's customer trades was marked HFT in the NASDAQ Datasets.

Wedbush Securities has described itself as "a leading provider of clearing services and sponsored access solutions for registered broker-dealers and non-registered entities." A regulatory action last summer said that "During the relevant period [which includes 2008 and 2009], Wedbush was one of the largest volume market access providers, including through its provision of market access to overseas high-frequency, high-volume, algorithmic day-trading firms and anonymous foreign traders" and estimated that "Wedbush market access customers traded on NASDAQ over 695 million shares daily in 2009," or perhaps as much as *one-third* of Nasdaq's volume that year. Was any of this activity coded HFT in the NASDAQ Datasets? It likely wasn't.

An enforcement action against Athena Capital Research last autumn accused the high frequency trading firm of "placing a large number of aggressive, rapid-fire trades in the final two seconds of almost every trading day during a six-month period to manipulate the closing prices of thousands of NASDAQ-listed stocks." The SEC said that Athena's activity lasted from "June to December 2009 and made up more than 70 percent of the total NASDAQ trading volume of the affected stocks in the seconds before the market close." Are any of Athena's trades marked HFT in the NASDAQ Datasets, or was Athena a sponsored access customer, with all its trades coded non-HFT?

In 2009 the Aite Group estimated that as much as half of all U.S. equity market activity came from sponsoring firms like Lime or Wedbush. The vast majority of that flowed through naked access services. At the time these services were particularly appealing to high frequency trading firms. Naked access was as fast or faster than any other route into an exchange and let high frequency firms leverage the sponsoring firm's volume discounts. It also camouflaged high frequency firms because they got to markets under the sponsoring firm's credentials instead of their own. Aite estimated that as much as 38% of all U.S. equity market activity came through naked access.

As a guess, then, so far the NASDAQ Datasets overlook close to 40% of Nasdaq's trade activity as "HFT" though it came from what almost everyone, including the firms themselves, would agree were high frequency trading firms. And we haven't even talked about Goldman Sachs or JP Morgan.

Late last year the European Securities and Markets Authority ("ESMA") published a report that estimated HFT activity in Europe using two different methods. The first method estimated HFT activity "based on the identification of HFT firms according to their primary business or the types of algorithms they use," much as Nasdaq did when it prepared the NASDAQ Datasets, and found that HFT accounted for 24% of value traded. The second method estimated HFT "based on statistics such as lifetime of orders or order-to-trade ratio" and found that HFT accounted for 43% of value traded. ESMA said "The results based on the primary business of firms provide a lower bound for HFT activity, as they do not capture HFT activity by investment banks, whereas the results based on the lifetime of orders are likely to be an upper bound for HFT activity." No doubt the second method also captured some non-HFT algorithmic trading, nevertheless the direct identification method of classifying HFT found only 55% of presumed HFT activity the "lifetime of orders" method found.

ESMA's study gives even more scale for what's been overlooked in the NASDAQ Datasets, and it's remarkable. Altogether, adding sponsored activity and investment bank activity, it could be that the majority of high frequency trades in the data were miscoded as non-HFT. Miscoding investment bank trades is an especially sour note since it was news of the 2009 arrest of a former Goldman Sachs employee, Sergei Aleynikov, on suspicion of stealing Goldman's high frequency trading code, that first introduced millions of people to the business. It doesn't matter. None of Goldman's trades were coded HFT. They're all non-HFT, the same category as any one-lot trade from an ordinary retail investor like Mrs. Betty Johanssen of Red Lake, Minnesota.

Dressed and perfumed

Here's where we are:

- The Aite Group estimated nearly 40% of equity market activity in 2009 came from naked sponsored access services; the vast majority of that almost certainly came from high frequency firms. Based on published descriptions, likely all this data was coded non-HFT in the NASDAQ Datasets;
- All activity from large integrated broker-dealers or investment banks was coded non-HFT, according to published descriptions. To get a sense of how much data that might be, we can look to ESMA's work, where, as a group, these large firms could account for nearly half of HFT activity;
- According to a published account, Lime Brokerage had somewhere near 200 "high frequency trading firms and individuals" as customers in 2009; researchers say all Lime's data was coded non-HFT. And what about firms like Wedbush or Newedge?

Questions about the NASDAQ Datasets aren't new. Recent regulatory actions and ESMA's study turn up the heat. SEC staff correctly point out that their review of HFT research "must deal with the various metrics researchers used to define HFT and how their definitions may affect their conclusions about HFT activity." For example, based on the NASDAQ Datasets, some academic researchers conclude that HFT firms reduce volatility, results at odds with, among other foundational studies of HFT behavior, key points in the SEC and

CFTC's joint report on the Flash Crash and in Kirilenko's landmark paper about the crash. Perhaps that's because, as the staff say in the Paper, "the particular metrics used to classify HFT can greatly affect findings about key factual characteristics of HFT activity," and Kirilenko and the SEC and CFTC at the time had much better data.

According to the Paper, the NASDAQ Datasets are "the only datasets available to academic researchers that directly classify HFT activity in U.S. equities," but even so, however tempting it might be, we can't let hope trump wisdom to imagine we're much better off because of them. Suppose a large number of men were miscoded as women, or women as men, in a national health database, and then researchers rushed to report that women were taller and heavier than before, with a surprising incidence of prostate problems, or men were shorter and lighter than ever and were giving birth in record numbers. Are we better off believing any of that?

There's an old joke about economists. The joke makes fun of a common fault of economic researchers. The fault is that they use whatever data's available, no matter how flawed or incomplete the data might be, to draw conclusions about the world. Wikipedia has a nice version of the joke:

> A policeman sees a drunk man searching for something under a streetlight and asks what the drunk has lost. He says he lost his keys and they both look under the streetlight together. After a few minutes the policeman asks if he is sure he lost them here, and the drunk replies, no, and that he lost them in the park. The policeman asks why he is searching here, and the drunk replies, "this is where the light is."

Despite miscoding a giant portion - even half or more - of all high frequency trading activity as non-HFT, the NASDAQ Datasets are all we have detailing U.S. equity HFT activity. It's where the light is, and like that drunk under the streetlamp, researchers are looking for truth in the wrong place.

Why is that? The CFTC has collected and supplied researchers high quality data for years. The CFTC's futures data identifies firms more accurately than almost any available equity market dataset. Along the way the CFTC endured nasty letters from law firms about what it was doing and endured a lengthy investigation into its research program, but the CFTC stayed with it, opened the data to researchers, and we've seen excellent research based on that data.

The SEC has the authority to get good data on almost any equity market question. Now a half-decade after the Flash Crash, why hasn't the SEC done it? Instead, SEC staff walk through research on deeply flawed data, and despite grim reservations about that research it's all they have. Too many researchers - and so, the public - have no choice but to rely on data dressed and perfumed by firms with a lot of money at stake in the debate over HFT. This can't be what the SEC wants. It can't be what Chair White wants for her new "data-driven" market structure advisory committee.

Sincerely,

R. T. Leuchtkafer