Algorithmic and High Frequency Trading, Liquidity, Latency and Regulation

Andrei Kirilenko
Imperial College Business School

MFM Summer Camp
Cape Cod
June 2016
The Price Formation Process

Market prices are key to valuation: the valuation of projects, companies, contingent claims, guarantees, risks relies on market prices.

Market prices for the underlying assets – stocks, bonds, currencies, commodities – as well as derivatives – futures, options, swaps – come from trading venues: organized exchanges, bilateral platforms, dark pools and so forth.

Trading venues bring buyers and sellers together to exchange claims over the ownership of assets in return for payment – transaction price.
Finance theory says that there are two fundamental drivers of the price formation process – information and liquidity.

But as transactions in automated trading venues occur milliseconds apart, what information and what liquidity?
London Stock Exchange then and now
The Promise of Automated Markets

Automated markets came with the promise of using faster and cheaper technology to drastically lower execution costs and improve price discovery for fundamental market participants.

For investors who want to buy or sell one hundred shares or a couple of futures contracts, this promise seems to be realized:

They can do it at narrower bid-ask spreads, greater market depth, and prices that can be discovered around the clock.

Then came the Flash Crash.
The Flash Crash - May 6, 2010

Graph showing the time and price movements of the Dow Jones Industrial Average (DJIA), E-mini S&P 500, and S&P 500 Index on May 6, 2010, with a significant drop occurring around 11:00 AM.
HFTs and the Flash Crash

The events of May 6, 2010 were blamed on high frequency traders (HFTs) – a new breed of secretive, hyperactive trading algorithms that take advantage of anyone trying to trade in size.

How did High Frequency Traders trade on May 6, 2010?

What may have triggered the Flash Crash?

What role did HFTs play in the Flash Crash?
Who is Trading?: Classifying 15,000 Traders

**HFTs** (16 traders, 30 percent of trading volume or trades):
High volume, low inventory, end the day flat

**Non-HFT Market Maker (179/10):**
Provide liquidity

**Fundamental (Institutional):**
Take directional positions

**Small (Retail):**
Trade very few contracts

**Opportunistic:**
Trade across multiple markets, against a model, during “events”

---

[Diagram showing trader volumes and positions]
On May 6, 2010, HFTs traded the E-mini futures same way as they did on May 3-5. Small inventory, high trading volume, take more liquidity than provide. High Frequency Traders did not cause the Flash Crash. They did amplify the velocity of the price price move during the Flash Crash.

Kirilenko, Kyle, Samadi and Tuzun (2016, JF forthcoming)
The Flash Crash – The Trigger

Large **Fundamental Seller** initiated an automated execution program to sell E-mini S&P 500 futures minutes before prices collapsed.

Automated sell program – sell the 75,000 E-mini’s with 9% volume participation target.

Largest net position of the year executed in about 20 minutes.

A large imbalance developed between **Fundamental Sellers** and **Fundamental Buyers**.

**Opportunistic Traders** held the imbalance, but for a massive price concession.

Source: CFTC-SEC Report on the Events of May 6, 2010
The Flash Crash – The Cascade

Price decline spread to other assets through cross-market arbitrage – buy E-mini/sell SPY or basket of equities.

System-wide price declines – between 6 and 9 percent across main stock market indices

The Dow Jones Industrial Average experienced the biggest intraday point decline in its entire history.

Source: CFTC-SEC Report on the Events of May 6, 2010
The Flash Crash – The Cascade

Price declines triggered automated alarms and pauses.

Lack of liquidity in individual equities, ETFs, other securities – stop loss orders, systems resets.

E-mini – CME Globex Stop Logic Functionality – an automated 5-second trading pause.

Source: CFTC-SEC Report on the Events of May 6, 2010
The Flash Crash

Stop Logic Functionality
The Flash Crash

Lots of questions:

Can such a systemic event happen again?

Why did the five-second trading pause played such a big role?

Can we engineer automated markets to be more robust? How?
E=mc^2?
Figure 2.1: 10-Year Treasury Yield on October 15
(Cash)

Figure 2.2: Intraday Price and Volume in 10-Year Treasury (Futures)

E=mc^2

Technology and Market Design

Charlie Munger, Vice Chairman, Berkshire Hathaway: “… I think it is very stupid to allow a system to evolve where half of the trading is a bunch of short term people trying to get information one millionth of a nanosecond ahead of somebody else.”
An Automated Market

Order sent to the exchange
Fill information (Order matched)
Price feed
Quote/Order acknowledgment

Risk Manager
- Exchange sets limits
- Trading firm sets limits
- FCM/broker-dealer sets limits

Trading Firm
- Risk limits set at trading server
- Risk limits set on GUI

FCM/Broker-Dealer
- Risk limits set at exchange level

Clearing House
- Internet / VPN access

Drop Server
- Co-location Facility
- Matching Engine
- Proprietary/Vendor system
- Algo/OMS

Settlement Data
- Position
- Reporting
- Risk limits set at exchange level
Latency

Latency is the delay between a signal and a response.

It is measured in units of time: seconds, milliseconds, microseconds, nanoseconds and so forth.

Latency is an essential feature of any nontrivial mechanical or electronic system, e.g., an automated trading platform.

A standard way to measure latency is by determining the time it takes a given data packet to travel from source to destination and back, the so-called *round-trip time* or RTT.
Latency and Messages

The data packet we will use is the so-called message.

A message is a standardized packet of data that enables a trader and an automated trading venue to communicate with each other.

It is a primitive unit of valuable information that enables all pre-trading, trading, and post-trading activity.

Messaging is the only way to affect the price formation process.

Each message is time-stamped when it passes through a specific subsystem of an automated trading platform.
Trading Platform Latency

In automated financial markets, there are three main types of latency that affect the trading process:

- communication latency (can be reduced by buying co-location services),
- market feed latency (can be reduced by buying proprietary data feed),
- trading platform latency (cannot be reduced by a trader; taken as given).

Trading platform latency is the time it takes for a message to travel inside an automated platform.
Trading Platform Latency is a Random Variable

Kirilenko and Lamacie (2015)
A Mixture of Lognormal and Power Law

WINJ14 - RTT - All Types

Data < trunc
Log-Normal Fit
mean = 1174.79
std = 20865.68
trunc = 1500

WINJ14 - RTT - All Types

Data > trunc
Power Law Fit
alpha = 3.19
trunc = 1748
Time Series of Trading Platform Latency
Time Series of the Volatility of Volatility
Intraday dynamics of median trading platform latency and the dispersion of latency add significant explanatory power to changes in volatility and the volatility of volatility.

These effects arise over and above changes in the number of messages (a control variable).

What seems to matter is not HOW MANY messages a trading platform needs to process, but HOW LONG a trading platform takes to process certain types of messages.

Why?
Latency and the Matching Engine

A possible explanation:

Messages that result in changes of the best bid and offer take more time for a platform to process than messages that change the lengths of existing queues.

That’s because the matching engine needs to adjust the “locations” of many queues.

But, the very messages that result in the changes of the best bid and offer are certainly much more valuable in terms of the price formation process than those that add a bit of depth here and there, but don’t move the midpoint price.
Latency as a Technological Risk Factor?

One possibility is that some technologically-advanced traders monitor trading platform latency on an ongoing basis as a proxy for “the collective wisdom about the state of the market” AND THEN actively adjust their trading behavior to changes in latency and in the variability of latency which they view as predictive about future imbalances or orders.

Another possibility is that both trading platform latency and asset prices are driven by changes in one or several latent variables (e.g., “information”, “liquidity”, “sentiment”) that first manifest themselves in the characteristics of latency and then in asset prices simply due to trading technology.
CME Group Inc., the world’s largest exchange operator, just completed an upgrade traders said would eliminate a shortcoming that gave some participants an advantage.

Under the old system, data connections that linked customers to CME -- where key products like Treasury futures and contracts tied to the Standard & Poor’s 500 Index trade -- had noticeably different speeds, opening up the potential for gaming, according to traders and other experts. Those who knew how to gain faster access could increase their odds of being first in line to trade.

The new design appears to stamp that out.
Regulating Latency?

Many regulators and policy makers decided to focus on latency measures: to “slow things down”, to put in “speed bumps”, or to remove the “speed advantage” of HFTs.

It’s not so simple though. Latency is a random variable. It has jitter.

Automated trading lies in the intersection of four highly specialized fields:

- Regulation,
- Finance,
- Technology, and
- Data processing.
Rule Implementation

From an e-mail to me:

“I have now been at [a BIG automated trading platform] for over a year and have seen us struggle with trying to bridge the gap between rules that are written by lawyers and then translated into code by our development team.

In nearly all cases, the rules, design of the solution and the development and testing are all happening in parallel which opens us up to the risk that a lawyer changes a sentence late in the process that has a meaningful impact on our actual code.

As you can imagine, when found before we go live with the software change, this results in delays, rework and extra testing.

When found after we have gone live in production, this can result in regulatory action including monetary fines.”
Lost in Translation

Modern financial markets are complex adaptive systems of computer algorithms. Yet, the regulatory framework is still fully human-centric.

Human regulators are writing rules with the idea that they would be interpreted and implemented by a human market operator.

There is, indeed, a human operator, but it is no longer a trader or a regulatory compliance officer of the past.

Instead, it is a software architect or a computer programmer who has to translate regulations into lines of computer code.

This creates a possibility of a significant amount of intended regulatory content to be “lost in translation.”
What Can Be Done?: Go From Analog to Digital

Each regulation is a circuit: it allows some things, prohibits others and is neutral over a range of values.
Solution: A “Blueprint” Schema for Each Regulation

To enhance regulatory efficiency and reduce implementation costs, regulations written for automated financial markets need to include a separate section.

The Technological Implementation and Data (TID) section – a schema of sorts – should describe the circuit architecture of each rule in a form of standardized logic gates, truth tables or other alternatives.

Circuit architecture can be efficiently implemented by software architects and developers.

It could also include scenarios and data required to validate implementation.
Circuit Architecture
Safe Harbor

The TID section would be akin to the cost-benefit analysis (CBA) section currently required for each regulation, but with one big exception – it should be a “safe harbor” section, not a litigation target.

Otherwise, it would quickly evolve into a “litigation risk management” tool that would further add to implementation errors.

Safe harbor means that specific implementation protocol as provided in the TID section should be deemed sufficient.

Deals with the issue of implementation of principles-based regulations: A trade off then between being more prescriptive and improving regulatory efficiency and being less prescriptive and fostering innovation.
Financial Regulation 2.0

*Systems-Engineered.* Regulate automated markets as complex systems composed of software, hardware, and human personnel.

*Safeguards-Heavy.* Make risk safeguards consistent with the machine-readable communication protocols and operational speeds.

*Transparency-Rich.* Mandate that versions and modifications of the source code that implement each rule are made available to the regulators and potentially the public.

*Cyber-centric.* Change regulatory surveillance and enforcement practices to be more cyber-centric rather than human-centric.

*Platform-Neutral.* Make regulations neutral with respect to computing technologies.

Kirilenko and Lo (2013, JEP)