# Supplemental Material: High-Frequency Trading Synchronizes Prices in Financial Markets

Austin Gerig

austin.gerig@sbs.ox.ac.uk

CABDyN Centre, Saïd Business School, University of Oxford, United Kingdom

**DATA**

The results in the paper are generated using two datasets. The first dataset is provided by NASDAQ and contains all transactions and quotes on the NASDAQ exchange for 120 stocks during the week of February 22-26, 2010. The 120 stocks are a broad representation of the US equity market. Half of the stocks are listed on NASDAQ, half on the NYSE (NYSE listed stocks are also traded on the NASDAQ exchange), and 1/3 are stocks from large companies (large-capitalization or large-cap stocks), 1/3 are medium-cap stocks, and 1/3 are small-cap stocks. Table I contains a list of the stocks. The data is unique for three reasons: (1) the timestamps on the data are precise: they are internally generated by the exchange and are to the millisecond, (2) for every transaction, the initiating order is specified, i.e., I know whether it was the buyer or the seller that caused the transaction, (3) all transactions and quotes are categorized as HFT or non-HFT.

NASDAQ defines the HFT and non-HFT categories as follows: there are 26 firms that specialize in HFT and trade on the NASDAQ exchange; activity that originates from these firms is flagged as HFT, and all other activity is flagged as non-HFT. The 26 firms are primarily independent proprietary trading outfits (although I do not have access to their names, typical examples would include Allston Trading, DRW Holdings, Getco, RGM Advisors, Tradebot, Tradeworx, etc.). The most likely bias in NASDAQ's categorization is under-reporting of HFT. HFT activity that originates from large integrated firms, such as investment banks or large hedge funds, cannot be separated from the other activity of these firms and is therefore not categorized as HFT.

The second dataset is taken from the Thomson Reuters Tick History database (TRTH), which includes records of transactions and quotes from numerous financial markets around the world. From the Reuters database, I collect trade and quote information for the 40 large-cap stocks that are included in the NASDAQ data. I choose the same time period as the NASDAQ data (February 22-26, 2010) but also include data from earlier time periods: February 21-25, 2000 and February 21-25, 2005. 5 of the 40 stocks are not included in the Reuters database during all three periods (GOOG, HPQ, MOS, BIIB, and ISRG), so I disregard the Reuters data for these stocks in all time periods. Furthermore, US markets were closed on February 21 in 2000 and 2005 (both were Mondays), so I disregard the Reuters data for Monday in 2010 as well. The main differences between the Reuters data and the

NASDAQ data are the following: (1) the NASDAQ data flags HFT, (2) both the Reuters and NASDAQ data are timestamped to the nearest millisecond, but the Reuters times are not as precise due to delays in transmitting information from exchanges to Reuters, (3) the Reuters data does not specify whether the buyer or seller initiated the transaction, (4) the Reuters data does not include transactions that are less than 100 shares, (5) the Reuters data includes trades and quotes from all major US exchanges, whereas the NASDAQ data includes information only from the NASDAQ exchange.

## METHODS

To generate the results in Figs. 1(B) and 2, transactions must be classified as buyer or seller initiated. Unfortunately, the Reuters data does not include this information. I estimate the initiator of a transaction for the Reuters data as follows: for each transaction, I determine whether the transaction price is closer to the bid or the ask price. The bid price is the price at which you can immediately sell in the market and the ask price is the price at which you can immediately buy. If closer to the bid, I assume it is seller initiated. If closer to the ask, I assume it is buyer initiated. If the transaction price is at the midpoint of the bid and ask price, I leave the transaction unclassified and ignore it when generating the figures.

In Fig. 1(A), HFT estimates are from the TABB group as reported in "High-frequency trading: Up against a bandsaw", Sept. 2, 2010, *Financial Times*. It is difficult to ascertain how the TABB group calculates these estimates (They did not respond when contacted for clarification). In the NASDAQ data, 49% of share volume is attributable to HFT, which is close to the estimate of 56% in 2010 by TABB. Also, the Aite Group provides similar estimates for HFT ("The fast and the furious"', Feb. 25, 2012, *The Economist*), and both estimates correspond well with the increase in message traffic used as a proxy for algorithmic trading in [1].

In Fig. 1(B), the cost of a buyer initiated transaction is measured as the transaction price, $p$, minus the current prevailing midpoint price for the security, $m$, (for a seller initiated transaction, the cost is $m - p$). The midpoint price is the midpoint between the quoted price at which you can immediately buy (called the ask) and the quoted price at which you can immediately sell (called the bid) in the market. The transaction error is measured as the absolute difference between the transaction price, $p$, and the midpoint price 1 minute later, $x$, (see the diagram in the figure). To standardize across securities, costs and errors are measured in basis points (1bps=0.01%) where the cost is divided by $m$ and the error by $p$. To standardize across time periods with different volatilities, costs and errors are divided by the average value of the VIX for the time period (obtained from http://www.cboe.com/micro/vix/historical.aspx). The figure reports the mean and standard error of the mean for the 35 stocks during each time period.

The curves in Fig. 2(A) and 2(B) are calculated as follows: the average price response of stock $i$, conditioned on a price movement in $j$ at time 0, is first determined for all $j \neq i$ (for negative price movements of $j$, the sign of the response is reversed). Price movements are defined as any change in the midpoint prices. Each response curve is normalized by dividing the response by the difference between its maximum and minimum value. The normalized curves are averaged over all $j$ for each $i$, and then the final curve is generated by averaging over all stocks $i$, with the standard error of the mean shown in shaded color.

In Fig. 2(B), I separate out the price response into an amount due to HFT, an amount due to non-HFT, and an amount that was uncategorized. To perform this separation, I determine the cause of each midpoint price change for each security. If the midpoint for a stock increases (decreases) during the same millisecond that a buyer (seller) initiated transaction occurs in that stock, I assume the transaction caused the price change. If the initiator of the transaction was a HFT firm, I assign the midpoint change to HFT. Likewise, if the initiator was non-HFT, I assign the midpoint change to non-HFT. If the midpoint changes without a corresponding transaction, then I first determine whether the bid and/or ask was improved or removed. If improved (i.e., the ask lowered or the bid increased), then I assign the price change to the initiator of the new quote (HFT or non-HFT). If removed (i.e., the ask is increased or the bid is lowered), then I assign the price change to the initiator of the original quote (HFT or non-HFT). I am unable to categorize several instances which are rare in the data: (1) if multiple buyer or seller initiated transactions occur during the same millisecond that the midpoint changed, and if these transactions are mixed between HFT and non-HFT, then I cannot ascertain whether it was HFT or non-HFT that caused the price change, (2) if both the bid and the ask change during the same millisecond, but one change is due to HFT and the other to non-HFT, I cannot ascertain whether the cause was HFT or non-HFT.

I use the NASDAQ data to create the plots in Fig. 4. In Fig. 4(B), the weight between nodes $i$ and $j$ in the full network is $\sqrt{2(1-\rho_{i,j})}$ where $\rho_{i,j}$ is the correlation between 30 second returns for stock $i$ and stock $j$. From this network, I determine the minimum spanning tree using Prim's algorithm. GICS sectors are taken from the CRSP (Center for Research in Security Prices) database.

## DISCUSSION OF MODEL

In the model, liquidity providers are market participants who facilitate trade by transacting with investors' orders. By comparing non-synchronous to synchronous pricing, the model implicitly assumes that liquidity providers were previously unable to synchronize prices, but that HFT (now acting as the de facto liquidity providers) can easily perform this task. Of course, before HFT existed, human liquidity providers would have done their best to keep prices aligned. The point is that computers are much better than humans at performing this task.

In the model, liquidity providers are assumed perfectly competitive so that they set fair prices and make zero profit. In real markets, liquidity providers will require a price concession to transact with an investor's order. Adding a small price concession that allows for liquidity provider profits does not change the results.

In real markets, investors can place aggressive orders that transact immediately at the best available price (called market or marketable orders) or passive orders that specify a price but are not guaranteed to transact (called limit orders). The model does not specifically account for these different order types, but could be adjusted to include either or both. If an investor's order provides a large enough price concession, then regardless of type, liquidity providers will transact with the order. This means the term "liquidity provider" should not be restricted to firms that only transact using limit orders. In fact, in the NASDAQ dataset, HFT firms transact almost equally using market and limit orders, but they can provide liquidity in either case. For example, suppose that identical securities are traded on two different exchanges and that these exchanges have crossing limit orders. An arbitrageur can connect the orders by placing offseting marketable orders in the two markets, which in effect, provides liquidity to both. This example can easily be generalized to two economically related rather than identical securities, and it shows that liquidity providers connect investors not only through time but also across exchanges and securities.

## CALCULATIONS

### Cost and pricing error

As in the main text, assume that the market contains only two securities, $n = 2$, and the initial price of each security is $m_1 = m_2 = 50$ which can increase or decrease by $\delta = 1$. For both securities $\phi_1 = \phi_2 = 0.75$, and their price changes are correlated with $\rho_{1,2} = 0.8$.

I will analyze the transaction cost, $c(B_1) = E[p_1 - m_1|B_1]$, and pricing error, $e(B_1) = E[|x_1 - p_1||B_1]$, for a buy order in security 1, $B_1$, when prices are and are not synchronized (for a sell order, the sign of the cost is reversed). First, if prices are not synchronized, then the buy order transacts at the expected price of the security conditioned on the placement of the buy order

$$\begin{aligned} p_1 &= E[x_1|B_1], \\ &= 49\,\mathcal{P}(x_1 = 49|B_1) + 51\,\mathcal{P}(x_1 = 51|B_1), \\ &= 50.5, \end{aligned}$$

which is calculated by applying Bayes' Rule, $\mathcal{P}(x_1|B_1) = \mathcal{P}(B_1|x_1)\mathcal{P}(x_1)/\mathcal{P}(B_1)$,

$$\begin{aligned} \mathcal{P}(x_1 = 49|B_1) &= 0.25 \\ \mathcal{P}(x_1 = 51|B_1) &= 0.75. \end{aligned}$$

The cost is therefore,

$$c(B_1) = p_1 - m_1 = 0.5,$$

and the average pricing error is,

$$\begin{aligned} e(B_1) &= (51 - p_1)\mathcal{P}(x_1 = 51|B_1) + (p_1 - 49)\mathcal{P}(x_1 = 49|B_1), \\ &= 0.75. \end{aligned}$$

If the market is synchronized, then the transaction price can be one of three values. If the buy order for security 1 arrives before the order for security 2, which occurs 50% of the time, then the cost and pricing error are the same as if prices were not synchronized, 0.5 and 0.75 respectively.

When a sell order for security 2 arrives first, which occurs $\mathcal{P}(S_2|B_1)/2 = 20\%$ of the time,

$$\begin{aligned}
\mathcal{P}(S_2|B_1) &= \frac{\mathcal{P}(S_2, B_1)}{\mathcal{P}(B_1)}, \\
&= \frac{1}{\mathcal{P}(B_1)} \sum_{x_1,x_2} \mathcal{P}(S_2, B_1|x_1, x_2)\mathcal{P}(x_1, x_2), \\
&= \frac{1}{\mathcal{P}(B_1)} \sum_{x_1,x_2} \mathcal{P}(S_2|x_2)\mathcal{P}(B_1|x_1)\mathcal{P}(x_1, x_2), \\
&= 0.40.
\end{aligned}$$

then the transaction price of the buy order is,

$$\begin{aligned}
p_1 &= E[x_1|B_1, S_2], \\
&= x_1^+ \mathcal{P}(x_1^+|B_1, S_2) + x_1^- \mathcal{P}(x_1^-|B_1, S_2), \\
&= 51 \times 0.5625 + 49 \times 0.4375, \\
&= 50.125
\end{aligned}$$

where again, Bayes' Rule is used, $\mathcal{P}(x_1|B_1, S_2) = \mathcal{P}(B_1, S_2|x_1)\mathcal{P}(x_1)/\mathcal{P}(B_1, S_2)$,

$$\begin{aligned}
\mathcal{P}(x_1 = 49|B_1, S_2) &= 0.4375 \\
\mathcal{P}(x_1 = 51|B_1, S_2) &= 0.5625.
\end{aligned}$$

The cost is therefore,

$$c(B_1) = p_1 - m_1 = 0.525,$$

and the average pricing error is,

$$\begin{aligned}
e(B_1) &= (51 - p_1)\mathcal{P}(x_1 = 51|B_1, S_2) + (p_1 - 49)\mathcal{P}(x_1 = 49|B_1, S_2), \\
&= 0.984375.
\end{aligned}$$

When a buy order for security 2 arrives first (the remaining 30% of the time),

$$\begin{aligned}
\mathcal{P}(B_2|B_1) &= \frac{\mathcal{P}(B_2, B_1)}{\mathcal{P}(B_1)}, \\
&= \frac{1}{\mathcal{P}(B_1)} \sum_{x_1,x_2} \mathcal{P}(B_2, B_1|x_1, x_2)\mathcal{P}(x_1, x_2), \\
&= \frac{1}{\mathcal{P}(B_1)} \sum_{x_1,x_2} \mathcal{P}(B_2|x_2)\mathcal{P}(B_1|x_1)\mathcal{P}(x_1, x_2), \\
&= 0.60.
\end{aligned}$$

similar calculations give $c(B_1) = 0.35$ and $e(B_1) = 0.4375$. Averaging over these three possibilities, the overall expected transaction cost and average pricing error are,

$$c(B_1) = 0.5 \times 0.5 + 0.525 \times 0.2 + 0.35 \times 0.3 = 0.46,$$
$$e(B_1) = 0.75 \times 0.5 + 0.984375 \times 0.2 + 0.4375 \times 0.3 = 0.703125.$$

**Profit of Informed Traders**

In the original Glosten and Milgrom model[2], investors are separated into two groups "informed investors" and "liquidity traders". The former have complete knowledge of the end-of-period price for the security they trade, and they submit buy orders when it is higher and sell orders when it is lower. The latter have no knowledge of the final price and buy or sell randomly (in the paper, I call these investors "average investors"). If the fraction of orders for security $i$ from informed investors is $\gamma_i$, then $\mathcal{P}(B_i|x_i^+) = \phi_i = 1/2(1+\gamma_i)$. In the example given in the paper, $\phi_i = 0.75$ so that $\gamma_i = 1/2$, meaning 1/2 of all orders come from informed investors.

If prices are not synchronized, informed investors make an expected profit of,

$$E[x_1 - p_1|x_1^+] = 0.5,$$

per buy transaction (and also an expected profit of 0.5 per sell transaction). Average investors make an expected profit of,

$$0.5E[x_1 - p_1|x_1^+] + 0.5E[x_1 - p_1|x_1^-] = -0.5,$$

per buy transaction (and the same per sell transaction). Notice that the total amount lost by the average investors is gained by the informed investors; a well-known result of the Glosten and Milgrom model.

When prices are synchronized, the informed make an expected profit of,

$$E[x_1 - p_1|x_1^+] = 51 - E[p_1|x_1^+].$$

If the buy order for security 1 arrives first, which occurs 50% of the time, then $E[p_1|x_1^+] = 50.5$. When a sell order in security 2 arrives first, which occurs $\mathcal{P}(S_2|B_1, x_1^+)/2 = 15\%$ of the time,

$$\begin{aligned}
\mathcal{P}(S_2|B_1,x_1^+) &= \frac{\mathcal{P}(S_2,B_1,x_1^+)}{\mathcal{P}(B_1,x_1^+)}, \\
&= \frac{1}{\mathcal{P}(B_1,x_1^+)}\sum_{x_2}\mathcal{P}(S_2,B_1|x_1^+,x_2)\mathcal{P}(x_1^+,x_2), \\
&= \frac{1}{\mathcal{P}(B_1,x_1^+)}\sum_{x_2}\mathcal{P}(S_2|x_2)\mathcal{P}(B_1|x_1^+)\mathcal{P}(x_1^+,x_2), \\
&= 0.30,
\end{aligned}$$

then $E[p_1|B_1,S_2,x_1^+] = 50.125$. When a buy order in security 2 arrives first, which occurs the remaining $\mathcal{P}(B_2|B_1,x_1^+)/2 = 35\%$ of the time,

$$\begin{aligned}
\mathcal{P}(B_2|B_1,x_1^+) &= \frac{\mathcal{P}(B_2,B_1,x_1^+)}{\mathcal{P}(B_1,x_1^+)}, \\
&= \frac{1}{\mathcal{P}(B_1,x_1^+)}\sum_{x_2}\mathcal{P}(B_2,B_1|x_1^+,x_2)\mathcal{P}(x_1^+,x_2), \\
&= \frac{1}{\mathcal{P}(B_1,x_1^+)}\sum_{x_2}\mathcal{P}(B_2|x_2)\mathcal{P}(B_1|x_1^+)\mathcal{P}(x_1^+,x_2), \\
&= 0.70,
\end{aligned}$$

then $E[p_1|B_1,B_2,x_1^+] = 50.75$. Putting it all together, the expected profit per buy transaction for an informed investor when prices are synchronized is,

$$\begin{aligned}
E[x_1 - p_1|x_1^+] &= 51 - (0.5 \times 50.5 + 0.15 \times 50.125 + 0.35 \times 50.75) \\
&= 0.46875.
\end{aligned}$$

Because of the symmetry of the example, the expected profit per sell transaction for informed investors is the same (as well as the expected profit per transaction in security 2). The informed therefore make less profits (0.46875 per transaction vs. 0.5 per transaction) when prices are synchronized.

## PROOFS

Consider a market with $n$ securities as described in the main text. For the proofs below, I will consider a buy order in security $i$. The same results hold if considering a sell order. If prices are not synchronized, then the transaction cost, $c(\cdot)$, of a buy order for security $i$ is the following:

$$c(B_i) = E[x_i|B_i] - E[x_i], \tag{1}$$

and the average pricing error, $e(\cdot)$, of a buy order for security $i$ is,

$$
\begin{aligned}
e(B_i) &= E\left[|x_i - E[x_i|B_i]||B_i\right], && (2)\\
&= E\left[s_i(x_i - E[x_i|B_i])|B_i\right], && (3)\\
&= E[s_i x_i|B_i] - E[s_i E[x_i|B_i]|B_i], && (4)
\end{aligned}
$$

where $s_i = +1$ if $x_i = x_i^+$ and $s_i = -1$ if $x_i = x_i^-$.

If prices are synchronized, then individual transactions cause price updates in all securities as they occur. Therefore, the transaction cost and pricing error of an order depend on the set of transactions that have occurred prior to the order's arrival. The expected transaction cost, $c'(\cdot)$, of a buy order for security $i$ when prices are synchronized is the following:

$$
\begin{aligned}
c'(B_i) &= E\left[\,(E[x_i|\omega, B_i] - E[x_i|\omega])\,|B_i\right], && (5)\\
&= E[x_i|B_i] - E\left[E[x_i|\omega]|B_i\right], && (6)
\end{aligned}
$$

where $\omega$ is the set of transactions that occur before $B_i$ and $\omega \in \Omega$ where $\Omega$ is the set of all sets of transactions in securities $j \neq i$ that can occur before $B_i$. The average pricing error, $e'(\cdot)$, of a buy order for security $i$ when prices are synchronized is,

$$
\begin{aligned}
e'(B_i) &= E\left[|x_i - E[x_i|\omega, B_i]||B_i\right], && (7)\\
&= E\left[s_i(x_i - E[x_i|\omega, B_i])|B_i\right], && (8)\\
&= E[s_i x_i|B_i] - E[s_i E[x_i|\omega, B_i]|B_i]. && (9)
\end{aligned}
$$

The difference in transaction costs when prices are synchronized is,

$$
\begin{aligned}
c(B_i) - c'(B_i) &= E\left[E[x_i|\omega]|B_i\right] - E[x_i], && (10)\\
&= E\left[E[x_i|\omega] - E[x_i]|B_i\right], && (11)\\
&= E[2\delta_i(\mathcal{P}(x_i^+|\omega) - \mathcal{P}(x_i^+))|B_i], && (12)\\
&= 4\delta_i\left(E[\mathcal{P}(x_i^+|\omega)\mathcal{P}(B_i|\omega)] - \frac{1}{4}\right), && (13)\\
&= 4\delta_i cov[\mathcal{P}(x_i^+|\omega), \mathcal{P}(B_i|\omega)]. && (14)
\end{aligned}
$$

Because the covariance is positive, transaction costs are lower when prices are synchronized.

The difference in average pricing error when prices are synchronized is:

$$
\begin{aligned}
e(B_i) - e'(B_i) &= -E[s_i E[x_i|B_i]|B_i] + E[s_i E[x_i|\omega, B_i]|B_i], && (15)\\
& && (16)
\end{aligned}
$$

which after some algebra is,

$$e(B_i) - e'(B_i) = 4\delta_i \mathcal{P}^2(x_i^+|B_i) \left( E\left[ \left. \frac{\mathcal{P}(\omega|x_i^+)}{\mathcal{P}(\omega|B_i)} \right| x_i^+ \right] - 1 \right) \tag{17}$$

Because the expectation of $\mathcal{P}(\omega|x_i^+)/\mathcal{P}(\omega|B_i)$ is larger than 1, the average pricing error is lower when prices are synchronized.


[1] T. Hendershott, C. M. Jones, and A. J. Menkveld, Journal of Finance **66**, 1 (2011).

[2] L. R. Glosten and P. R. Milgrom, Journal of Financial Economics **14**, 71 (1985).

| Large-cap | | | | Medium-cap | | | | Small-cap | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| AA | AAPL | ADBE | AGN | AINV | AMED | ARCC | AYI | ABD | ANGO | APOG | AZZ |
| AMAT | AMGN | AMZN | AXP | BARE | BRE | BXS | CBT | BAS | BW | BZ | CBEY |
| BHI | BIIB | BRCM | CB | CETV | CHTT | CKH | CNQR | CBZ | CCO | CDR | CPSI |
| CELG | CMCSA | COST | CSCO | COO | CPWR | CR | CRI | CRVL | CTRN | DCOM | DK |
| CTSH | DELL | DIS | DOW | CSE | CSL | ERIE | EWBC | EBF | FFIC | FPO | FRED |
| EBAY | ESRX | GE | GENZ | FCN | FL | FMER | FULT | IMGN | IPAR | KNOL | KTII |
| GILD | GLW | GOOG | GPS | GAS | ISIL | JKHY | LANC | MAKO | MDCO | MFB | MIG |
| HON | HPQ | INTC | ISRG | LECO | LPNT | LSTR | MANT | MOD | MRTN | MXWL | NC |
| KMB | KR | MMM | MOS | MELI | NSR | NUS | PNY | NXTM | PBH | PPD | RIGL |
| PFE | PG | PNC | SWN | PTP | ROC | SF | SFG | ROCK | ROG | RVI | SJW |

TABLE I: Table of stocks in NASDAQ HFT dataset separated by market capitalization.