Securities Trading: Principles and Procedures

Joel Hasbrouck

Joel Hasbrouck is the Kenneth G. Langone Professor of Business Administration and Finance at the Stern School of Business, New York University.

Correspondence: Department of Finance, Stern School NYU, 44 West 4th St., New York, NY 10012. Email: jhasbrou@stern.nyu.edu. Web: http://pages.stern.nyu.edu/~jhasbrou.

Disclosures: I have served as a consultant, instructor, and/or advisory board member for numerous private and public institutions.

Version 14a; this draft: January 25, 2024

- Chapter 11 (Trading Halts) updated October 18, 2023
- End-of-chapter problems, exercises and sample exam questions are placed in a separate manuscript (STPPms14aQ).

[Blank Page]

Preface

This manuscript is a set of draft teaching notes for a one-semester course entitled *Principles of Securities Trading*. The target audience is finance students planning careers in trading, investment management, or law, and information technology students seeking to build trading and investment systems. The exposition draws on general economic principles, with an institutional focus on US equity markets.

The high level of institutional content underscores the realism and currency of the material. Given the speed with which markets evolve, however, it is likely (maybe even certain) that some of the details are out of date.

By way of full disclosure, I've taught (for compensation) in the training program of a firm that engages in high frequency trading. I'm presently associated with the US CFTC as an (uncompensated) "special government employee". I've served on various government and industry advisory committees. I give presentations at financial institutions for which I sometimes receive honoraria.

Although these notes draw from the subject generally known as market microstructure, they certainly don't fully cover the field. There are many important areas of academic research that are barely touched upon: the econometrics of high-frequency data; measurement of liquidity; liquidity risk and commonality; liquidity and asset pricing; empirical analysis of price discovery; and so on. These omissions reflect the priority placed on simplifying the foundations of the subject, rather than discussing all the extensions.

The text is organized in parts (broad themes), chapters and sections.

- Part I starts with the basics. It introduces key terms and describes the important players. It explores the floor markets (pre-21st century) and their modern descendants, the continuous electronic limit order markets.
- Part II considers extensions and alternatives to the limit order markets: auctions, dealers and dark trading mechanisms.
- Part III examines informational efficiency. Many readers will have encountered the subject in an earlier finance class. They will have absorbed the idea that the market price of the stock incorporates and fully reflects the split, the takeover announcement, or whatever. The present approach discusses the trading processes that make this incorporation possible. The role of trading procedure is particularly important with respect to private information, which can give rise to bid-ask spread effects, price impacts, market failures and so forth. Part III also discusses some issues of practical and legal importance: securities class action lawsuits and insider trading regulation.
- Part IV introduces algorithmic trading. The approach is incremental, moving from complex order types to statistical models and discussion of the order splitting problem.
- Part V covers current topics in regulation and high frequency trading.

For comments on earlier drafts of these notes I am indebted to Bruce Tuckman.

[Blank page]

Table of Contents

[Blank Page]

Part I. Modern securities markets: the basics

Securities markets rely on highly structured trading procedures and well-defined institutional roles. Part I introduces these institutions and procedures. This part discusses, by way of background, the floor markets. It then goes on to explore the descendants of these floors, our modern limit order markets.

Chapter 1. Introduction

We place strong demands on our securities markets. When we plan our investments or hedge risks, we rely on market prices to tell us the value of what we currently have and the cost of what we might attempt to do. We enter the markets to trade and implement our decisions. As events unfold over time, we return to the markets to monitor our progress and revise our decisions. Finally, when we want to consume the gains from our investments or the hedge is no longer needed, we sell or settle the securities.

In basic economics, supply and demand are usually assumed to play out in an idealized *perfectly frictionless market*. Each buyer and seller are assumed to be *atomistic*. That is, each individual is small relative to the overall market. When acting alone, each is incapable of meaningfully influencing the price. Each trader willingly expresses her true preferences: when she is asked "How much would you buy if the price were x?," for example, she answers honestly. (It does not occur to her to bluff or feign a weaker demand to obtain a lower price.) The buyers collectively define the demand curve (seeking to buy much at low prices, and little at high prices). The sellers define the supply curve. The price at which the total quantity demanded equals the quantity supplied defines the market-clearing price and quantity.

The process of arriving at the market-clearing equilibrium point is (in principle at least) accomplished by an auctioneer. The auctioneer calls out a price, and asks, "Who wants to buy at this price? Who wants to sell?" The auctioneer then adjusts the price until total supply and demand are in balance, and the market clears.

Stock markets are often mentioned as settings that closely approximate this ideal. From one perspective, this is a reasonable conjecture. Stocks are held by thousands of investors, and thousands more might be standing by as potential buyers or sellers.

On closer examination, though, reality breaks from the model. While millions of people might hold a security, only a few might be actively participating in the market when we want to trade. Ultimately the number of market participants might be as low as two: us and our counterparty. From this perspective, the large-number perfect-competition abstraction seems less useful. With few participants, our actions are likely to change the price. Taking this into account,

we behave *strategically*. Most of the time there is no one acting as an "auctioneer". In these interactions, the market procedures and rules matter very much.

These notes are about these rules, the procedures, and the economic principles that shape them. Although we can't avoid talking about the securities (the stocks, bonds, options, and so forth) these notes are not primarily about them, their characteristics, or their uses. The notes attempt to explain instead how they are traded, the details of the market's "plumbing".

A course of study might be organized top-down, starting from a broad conception of a market, the types of markets (floor, auction, limit order, dealer, and so forth), general features of these markets (such as types of participants and varieties of orders), and finally specializing to particular markets (such as the Shenzhen stock exchange). The alternative is bottom-up, an approach that starts with one particular market and its operation, then moves on to alternative modes of trading used in other markets, and then uses comparisons across markets to suggest general principles. These notes are mostly organized on the bottom-up model, and the discussion is usually firmly set in the particulars of some real-world market. It is not really an either/or choice, though. Once a market is described, the questions of how it came to have the form that it does and how we might make it better arise quickly, and the answers are usually determined only by application of general economic principles.

Particularly at the outset, then, it is useful to have one actual functioning market as a central example, and in this respect, the US equity market stands as a good choice. The US equity market is large and active and exhibits an especially wide range of features. More broadly, the economic forces that have converged on it and shaped it are suggestive, for better or worse, of changes that have or played out elsewhere.

This does not imply that the US equity market is the best or that it has always been at the forefront of sound practice and advanced technology. In the 1990s, for example, when the rest of the world had long since adopted decimal prices, US markets were still trading in eighths (of a dollar). Moreover, if the present era can be called the age of electronic markets, the US was in most respects late to the party. Other countries (notably France and Canada) were well ahead of the US in broad adoption of market-unifying technology.

Nevertheless, when the US stock market finally did make the transition to electronic trading, it did so in a flexible and open fashion. The lead regulator, the Securities and Exchange Commission, mostly took the stance that a stock market was not a "natural monopoly", and that there was much to be gained from competition to build better exchanges. This gave rise to rich experimentation with a variety of trading mechanisms and protocols, algorithmic trading, high-frequency trading and other practices that have spread to other markets.

The study of financial markets cuts across many disciplines, spanning almost everything from sociology to physics. The present perspective, though, draws mostly from financial economics. Within financial economics, the area that deals with the study, design, and regulation of trading mechanisms is known as market microstructure.

Market microstructure encompasses diverse lines of thought. Readers looking to supplement these notes might consider the following sources. (Harris, 2003) is a comprehensive review of trading mechanisms, styles and strategies. (O'Hara, 1995) covers the core economic principles. (Hasbrouck, 2006) discusses the empirical implications of these principles, and approaches to working with market data. (Foucault, Pagano and Roell, 2013) provides more depth on economic models and principles. For more extensive analysis of algorithmic trading see (Aldridge, 2013; Bacidore, 2020; Bouchaud, Bonart, Donier and Gould, 2018; Cartea, Jaimungal and Penalva, 2015; Johnson, 2010; Kissell and Glantz, 2003).

Like many other technology-driven sectors of the economy, securities markets have been subjected to many recent changes and disruptions. Discussions that summarize the key change points include (Angel, Harris and Spatt, 2011, 2015; O'Hara, 2015).

The citations in these notes will point the reader to other background sources. Finally, although these notes are primarily focused on the "how" of trading, it is useful to have some sense of the "what" (is being traded), that is, the structure and characteristics of specific securities. In this regard, Bodie, Kane and Marcus (2020) is a useful source to have at hand.

References

Aldridge, Irene, 2013. *High-Frequency Trading* (John Wiley, Hoboken).

Angel, James J., Lawrence E. Harris, and Chester S. Spatt, 2011, Equity Trading in the 21st Century, *Quarterly Journal of Finance* 1, 1-53.

Angel, James J., Lawrence E. Harris, and Chester S. Spatt, 2015, Equity Trading in the 21st Century: An Update, *Quarterly Journal of Finance* 5, 1-39.

Bacidore, Jeffrey M., 2020. *Algorithmic Trading: A Practitioner's Guide* (TBG Press, New York).

Bodie, Zvi, Alex Kane, and Alan J. Marcus, 2020. *Investments, 12th edition* (McGraw Hill, New York).

Bouchaud, Jean-Philippe, Julius Bonart, Jonathan Donier, and Martin Gould, 2018. *Trades, Quotes and Prices* (Cambridge University Press, Cambridge, United Kingdom).

Cartea, Alvaro, Sebastian Jaimungal, and Jose Penalva, 2015. *Algorithmic and High-Frequency Trading* (Cambridge University Press, London).

Foucault, Thierry, Marco Pagano, and Ailsa Roell, 2013. *Market Liquidity: Theory, Evidence and Policy* (Oxford University Press, Oxford).

Harris, Lawrence E., 2003. *Trading and Exchanges* (Oxford University Press, New York).

Hasbrouck, Joel, 2006. *Empirical Market Microstructure* (Oxford University Press, New York).

Johnson, Barry, 2010. *Algorithmic Trading & DMA: An Introduction to Direct Access Trading Strategies* (4Myeloma Press, London).

Kissell, Robert, and Morton Glantz, 2003. *Optimal Trading Strategies* (American Management Association, New York).

O'Hara, Maureen, 1995. *Market Microstructure Theory* (Blackwell Publishers, Cambridge, MA).

O'Hara, Maureen, 2015, High frequency market microstructure, *Journal of Financial Economics* 116, 257-270.

Chapter 2. The Elements of a Securities Market: US Equities

2.1. The larger picture

Financial markets exist worldwide for stocks, bonds, foreign exchange (FX, currency) and a wide range of derivatives (such as options, forwards, futures, and swaps). Among these markets, investors are probably most likely to participate in the FX and stock markets. Most countries have at least one stock exchange, and the workings of stock exchanges are usually well-documented and well-regulated.

The Standard and Poor's Global Broad [Stock] Market Index covers almost 12,000 publicly traded companies around the world. As of June 30, 2020, total market capitalization (market value) of these equities is approximately $55.6 Trillion, USD equivalent.[1] This is comparable in magnitude to the world's annual GDP, $87.752 Trillion USD, (data.worldbank.org).

Market capitalization is one measure of a market's size. Another is the trading volume, the value of securities bought or, equivalently, the value of securities sold, by all market participants (over some period). Alternatively, instead of value, we might use the number of shares, bonds, or contracts traded.

Volume estimates are provided by the stock exchanges, where much of the trade occurs. Table 2.1 summarizes market capitalization and trading volume for some of the world's larger exchanges. The numbers in the first two columns, market capitalization and trading volume are positively related. (Shares in valuable firms are widely held by investors and are also actively traded. To control for this common variation when comparing firms or exchanges, it is useful to look at ratios. The ratio of trading volume to market capitalization is turnover, roughly the

[1] When we multiply the number of shares in a firm by the price per share, we arrive at the firm's equity market, the market value of all the firm's shares. We can total this number for all the firms in a country to get a country's equity capitalization and total all the countries in the world to get a global figure. The total market capitalization reported here is the product of the number of index constituents and the average market capitalization ($11{,}807 \times \$6{,}369.78\ Million \approx \$75\ Trillion$). See (S&P Dow Jones Indices, 2020)

number of times (per year, in this table) that a representative share is traded (2.6, for the NYSE). Alternatively, the reciprocal of turnover (capitalization/volume) is the representative holding period, in this table, the number of years, 2.6 for the NYSE). These ratios are averages, and within each exchange there will be a large variation across trades. Some buyers will hold their newly purchased shares for decades, other buyers will sell them by the end of the day (or even the end of the minute).

A high turnover corresponds to short holding periods. With an average turnover of 3.3 times per year, the average holding period for the Shenzhen Stock Exchange is about 0.3 years (about 3.6 months). On the Euronext and Hong Kong exchanges, the average holding periods are about two and a half years.

Table 2.1. Capitalization, trading volume and Turnover, 2019.

Exchange/Exchange Group	Market cap ($B, US equiv.)	Volume ($B, US equiv.)	Annual turnover	Implied holding period (years)
New York Stock Exchange (NYSE)	24,480	9,305	0.380	2.6
Nasdaq - US	13,002	15,910	1.224	0.8
Japan Exchange Group (Tokyo)	6,191	5,099	0.824	1.2
Shanghai Stock Exchange	5,106	7,790	1.526	0.7
Hong Kong Exchanges and Clearing	4,899	1,877	0.383	2.6
Euronext	4,702	1,920	0.408	2.4
LSE Group (London)	4,183	2,000	0.478	2.1
Shenzhen Stock Exchange	3,410	11,255	3.301	0.3
TMX Group (Toronto)	2,409	1,445	0.600	1.7

Market capitalization is as of the end of 2019. Volume reflects only EOB (electronic order book) trades.

Source: World Federation of Exchanges (http://world-exchanges.org)

2.2. Exchanges

An exchange consists of facilities for trading, such as a trading floor, software that defines the market or connects traders, and so on. An exchange establishes a regularization of the trading process. When we say that a security is exchange-traded, we mean that the trading process is structured, monitored, and standardized.

Most exchange services relate to three areas: listing, trading, and data. Briefly:

- When a firm *lists* on an exchange, the exchange is providing a kind of sponsorship. The firm pays a listing fee. In return the exchange provides trading services and monitors and certifies financial statements and governance procedures.
- The *trading services and facilities* comprise computer systems, standardized trading procedures, and a certain amount of oversight.
- The trading process generates *market data*: reports of trades, quote changes, and so on. These data are valuable for market participants, and their sale generates large revenues.

A firm usually lists with one exchange, or at least designates one as the *primary listing exchange*. The most important US listing venues are the NYSE, NYSE Arca, NYSE MKT, and NASDAQ. They are differentiated by listing fees and listing requirements, but also by public image, investors' perceptions of the "kind" of firms that list there, and other intangibles.

The NYSE (we might call it "NYSE classic" to differentiate the former New York Stock Exchange from other exchanges that carry the NYSE brand) has the highest fees and tightest listing standards. It was historically the dominant US exchange, home to the "blue chip" companies, the largest and oldest industrial and financial companies. An NYSE listing carries associations of seniority and stability. NASDAQ-listed companies tend to be younger, smaller and more concentrated in technology. A NASDAQ listing carries associations of "entrepreneurial" and "growth".

The American Stock Exchange (now NYSE MKT) historically occupied the space between the NYSE and NASDAQ. In the hypothetical corporate life cycle, a firm would first list on NASDAQ, move to the Amex when it grew a little larger, and ultimately step up to the NYSE. From a listing perspective, NYSE Arca represents a NYSE initiative to list companies whose profile comes closer to NASDAQ. In its materials, the NYSE commented, "NYSE Arca is a fully electronic exchange for growth-oriented enterprises. Listed companies can grow on NYSE Arca and transfer seamlessly to the NYSE once they meet the requirements."

For trading purposes, a security is identified by its ticker symbol. Most NYSE and Amex stock listings have ticker symbols of three letters or less, like IBM, GE, or C (Citigroup); most NASDAQ- and ARCA-listed stocks have four-letter symbols, like MSFT (Microsoft), INTC (Intel), and QCOM (Qualcomm). Options and futures have more complicated symbols that encode references to maturity or exercise price.

Until the end of the twentieth century, exchanges tended to be member-owned cooperatives. The members were mostly brokers and traders; the cooperatives were organized as not-for-profit corporations. Memberships (sometimes also called "seats") could be transferred, inherited, bought and sold. A membership comprised partial ownership of the exchange plus trading rights and privileges. Beginning around 1990, exchanges began to reorganize themselves as for-profit-corporations, with publicly traded shares. In this form, ownership and trading rights are separated: owning a share of the exchange does not confer trading privileges, and you can trade without owning any shares. The term "member" now generally refers to the second possibility, someone who has established a relationship with the exchange for purposes of trading.

The US has about twenty-four national securities exchanges (https://www.sec.gov/fast-answers/divisionsmarketregmrexchangesshtml.html). The oldest, the New York Stock Exchange, was established in 1792. Recent additions include: MEMX (the "members' exchange"); the Investors Exchange; the Long-Term Stock Exchange; the Miami International Securities Exchange; MIAX Emerald and MIAX Pearl.

The existence of twenty-four US exchanges might suggest a diverse and competitive industry, but most exchanges are subsidiaries of one of three holding companies: ICE (formed as the Intercontinental Exchange, the owner of the NYSE), Nasdaq, and CBOE (from the Chicago Board Options Exchange). Within each group they coordinate key pricing decisions. The connections extend beyond the US stock market. All three also own markets in futures and options, in the US and beyond.

Non-US exchanges are also grouped into holding companies. The Euronext exchanges include the (formerly national) exchanges of Amsterdam, Brussels, Dublin, Lisbon, Milan, Oslo, and Paris. The Tokyo Stock Exchange Group includes the Tokyo Stock Exchange, but also the Tokyo Commodity Exchange and the Osaka Stock Exchange. The LSE Group operates the London Stock Exchange, AIM, and Turquoise. The TMX Group runs the Toronto Stock Exchange, the Montreal Exchange, TSX Venture, and TSX Alpha.

2.3. Brokers

We can't trade simply by visiting an exchange's web site and flashing a credit-card. For various legal and practical reasons, the exchange requires a more substantial relationship, one that verifies our identity, capability, and authority to trade. Most customers establish this relationship indirectly, by setting up an account with a broker.

A broker conveys or represents customer orders to the market. In this capacity, the brokerage usually provides services directly related to trading: custody of securities purchased, cash loans (for margin purposes), loans of securities (for short-sale purposes), record-keeping and tax reporting.

The process of representing customer orders might be as simple as directly conveying the customer's instructions, for example, "Buy 100 shares of Microsoft." Typically, though, the conveyance requires the broker to make certain determinations and decisions. At an even more involved level, brokers may place at their clients' disposal automated tools known as trading algorithms.

A broker is an *agent* working on behalf of a customer (sometimes called the *principal*). In this capacity, the broker works under a legal obligation to act in accordance with the customer's instructions and in the customer's interest. In broker-customer relations, as in many other principal-agent arrangements, the customer may find it difficult to monitor the broker's effort and actions. Did the broker really make a strong effort to complete our trade at the best possible price? It's often tough to judge. The broker's presumed superior expertise, that is, the very thing that makes the broker's services valuable to us, also makes it more difficult to evaluate his performance. We will encounter in these notes many situations in which the divergence between the customers' and brokers' goals affects market outcomes.

Brokers are sometimes differentiated by clientele and approach. *Prime brokers* provide transaction-related services for large and institutional customers. Individuals go to *retail brokers*. Retail brokers in turn are traditionally divided into "discount" brokers, who focus narrowly on trading services, and "full-service" brokers, who provide more comprehensive investment management and advice.

Brokers have traditionally charged customers commissions on each trade. For many years, commissions were fixed by the NYSE. In 1972, for example, a customer buying 100 shares of a \$50 stock would pay $\$22 + 0.9\%(100 \times 50) = \$67$, by the schedule then in effect (Jones, 2001). Commissions were deregulated on May 1, 1975, under regulatory pressure. Competition among brokers shortly led to lower commissions, and as information technology allowed further efficiencies, commissions continued to drop. By 2012, one discount broker (Scottrade) was advertising 100-share trades for \$7 (WSJ, Feb 27, 2012, Eastern Edition, p. C8).

Around 2015 a startup named Robinhood began to offer commission-free trades (of exchange-traded funds) and an easy-to-use smartphone app. Other discount brokers followed. In October 2019, Charles Schwab, an established full-service broker, adopted zero-commission trading. With that, commissions for retail trades seemed to be a thing of the past. An oft-repeated maxim, though, is "zero commissions does not mean zero costs". Customers pay for trading services in many ways, some of which will be discussed later in these notes.

2.4. Traders and their motives

Trade arises from differences in investment goals, risk exposures, and beliefs about security values. People who are identical in all these respects would want, at any proposed price, to trade in the same direction (buy or sell), and a trade requires both a buyer and a seller. A potential buyer and seller might differ in many ways, large and small. But to get a big picture of the market, it is useful to think about broad groups or clienteles.

Investors are sometimes categorized by their investment horizons. *Long-term investors* include institutions like endowment funds and individuals saving for retirement or a child's education. *Medium-term investors* have holding periods that are on the order of a business cycle (3-5 years). These investors often seek to profit from changes in relative valuations of securities. *Short-term traders* have holding periods ranging from minutes to a few months.

Day traders typically buy and sell within the day and end the day with no net position. They are usually individuals who may have no background or expertise in trading. Trades are often driven by their perceptions of short-term momentum and reversals. One version of a common saying goes, "The trend is your friend / until the end, when it bends."

Although day trading is a long-standing practice, activity surged in the 1990s with lower commissions, order entry via the internet, and a generally rising market. In 2020 day trading again came to the fore. In the pandemic many people became unemployed and/or housebound. With smartphone apps, trading had never been easier; with the elimination of retail trading commissions and lower interest rates on margin borrowing, trading had never been cheaper. A volatile stock market suggested the possibilities of large gains. Less obvious, perhaps, were the dangers of large losses. One case involved the suicide of a young trader (age 20) who believed that he had lost three-quarters of a million dollars. The tragedy was compounded by the likelihood that the actual loss was much smaller. (WSJ, July 28, 2020).

Traders may also be classified by motive. There are many possible motives, of course, but the most important is information concerning the intrinsic value of the security. If our counterparty has superior information (most obviously of the illegal "insider" sort), then we are much more likely to lose. Informational traders usually need to trade quickly (before their information is made fully public) and stealthily (to avoid detection).

Non-informational motives for trade include hedging, arbitrage, and liquidity. *Hedging* trades aim at risk reduction. For example, a farmer who will sell a wheat crop in (say) three months is exposed to price risk, uncertainty in the price of wheat at harvest time. She can eliminate the price risk by taking a short position in wheat futures contracts that mature in three months. This will require only a one-time sale of the contracts. Other strategies (typically involving options or portfolio insurance) require dynamic hedging, in which the underlying is bought and sold repeatedly over the life of the hedge. Even if the hedge is not based on any superior information, the trading may still require speed to maintain an acceptably low level of risk.

Arbitrage involves offsetting trades that lock in a profit. For example, if a stock can be bought on one exchange and sold on another at a higher price, the profit is the difference. From a trading perspective, though, the purchase and sale orders must be submitted with complete certainty that they will occur at the intended prices. If one or the other fails to execute, the arbitrage has "one leg in the air," and the sure profit turns into a likely loss.

Liquidity motives stem from unexpected cash outflows and inflows. A mutual fund's assets under management, for example, change as customers invest or divest shares in the fund. On any given day, these are unlikely to be exactly offsetting, so the fund must sell from its holdings or buy to augment them.

2.5. The price

We often refer to the price of a security as if it were one well-defined number. In fact, the market usually provides us with several alternatives:

- The last sale price (the price of the most recent trade)
- The bid quote (the highest price that someone is publicly willing to pay)
- The ask or offer quote (the lowest price at which someone is publicly willing to sell). "Ask" or "offer" are used interchangeably.

When a price is reported in public media, it is usually the last sale price. The usefulness and validity of this price stems from the fact that the trade actually occurred. The buyer and seller didn't just talk about trade; they really bought and sold. On the other hand, since we see the last sale price after the trade has occurred, the price is not completely current (and in fact might be quite old). The price that we would pay or receive in a trade that we're currently contemplating might be quite different. The bid and ask are hypothetical prices. They are proposals that might or might not lead to a transaction.

The difference between the ask and bid quotes is the *spread*. Assuming that the bid and ask prices don't change, the spread is the cost incurred by someone who buys and immediately sells the security, reversing the initial trade. Often, the bid and offer in a market are posted or set by different traders. Sometimes, however, they are set by one trader who is said to be making a market. If a buyer and seller were to arrive at the same time, the market-maker would buy at her bid price and sell at her ask price. From the viewpoint of the market maker, then, the spread would represent her trading profit.

2.6. Make or take?

The first major decision – to buy, sell, or hold the security – lies in the realm of asset allocation or risk management, beyond the borders of the present discussion. But once this determination is made, we turn to the question of trading tactics. Here, we face a decision that is often simply stated as "make or take". Specifically, when we go into the market to trade, should we *take* the best available price, or should we try to *make* our own price and await the arrival of a counterparty who finds our price acceptable?

We'll start by assuming that we have a stock and a direction (shorthand for "buy" or "sell"). Suppose that we're buying. A buyer entering the market can trade immediately by taking the posted ask price. Or she can put in her own bid, hoping that a seller arrives who is willing to accept her bid. Suppose market in the stock is $100 (per share) bid, offered at $101. She can buy immediately by paying $101, that is, *taking* someone else's price. Or she might *make* a price of her own, for example, by bidding $100.25. If an agreeable seller arrives, she'll buy at 100.25.

The make or take decision is the choice of whether to take someone else's offer and get an immediate execution, or to make a (lower) bid and hopefully buy at the better price. Making a bid entails some risk because a seller might never arrive. The market might move higher, and the buyer might find herself chasing the stock, buying at a price higher than the original $101 offer, and therefore regretting her earlier decision to make.

The specifics of her decision are represented in her order. An order is a request, usually conveyed to the market through a broker. All orders indicate direction (buy or sell) and quantity. Most of the time, an order has a price limit, e.g., "buy 100 shares, limit $102." That is, don't pay above $102 per share. An order with a price limit is usually called a limit order. If the market ask price is $101 when the buy order arrives, the buy order is considered marketable. There is an immediate execution, at $101.

A market order is communicated without a price limit. In the case of a buy order, it says "I will pay the market offer, however high that offer might be." If the market offer price is $101, then someone sending in a market buy order expects to pay 101. But prices can change rapidly, and if the market offer price is $110 when the order arrives, the buyer will pay $110.

Someone putting in a limit order priced at 102 in this situation, also expects to buy at 101. But if the price goes above 102, the order will not be executed. Because market orders can lead to nasty surprises in fast markets, some exchanges do not accept unpriced orders. Similar remarks apply, but in the opposite direction to sell limit and market orders.

A participant in a trade is sometimes called a *side*. A trade has at least one buying side (buyer) and at least one sell side (seller). There may be many sides if there are multiple buyers

and/or sellers. Sides may also be classified as *active* or *passive*. The passive side refers to the trader who is posting the bid or ask/offer and stands willing and available for trade. The passive side is also called the *resting* side.

In any given trade, the active side might be the buyer or the seller. We refer to these situations differently. An active seller *hits the bid*. An active buyer *lifts the offer* (or *lifts the ask*). This distinction might seem unnecessary. In the construction "hit the ask," for example, it seems clear that the seller is passive, and the buyer is active. There are many instances of the expression online and in print. To the traditionalist, though, "hit the ask" sounds wrong, and may even suggest ignorance.[2]

The make/take choice often involves a trade-off between risk and reward. A trader who wishes to buy the stock can execute immediately by paying the offer price. The relative reward to using a limit buy order (a bid priced below the offer) is that the stock might be purchased more cheaply. The risk is that bid won't be hit, and the security won't be purchased. The consequences of this execution failure might be minor (if the trader is only marginally inclined to own the security) but can be major if the desire to own the security (for investment or hedging purposes) is strong. Finally, a limit order usually entails waiting (for the arrival of an order that executes it). Delay causes risk because security prices are constantly in motion and may impose also impose psychological cost from postponed closure (resolution, removal of uncertainty).

2.7. Liquidity (and other terms of the art)

Some terms that we'll encounter are everyday words, but nevertheless possess, in the context of trading and markets, particular meanings or connotations.

Liquidity is a broad term that summarizes the level of cost and difficulty that we encounter when we try to trade. In a liquid market, trading is cheap and easy. Moving beyond this generalization, liquidity is sometimes partially characterized by the attributes of immediacy, tightness, depth, and resiliency:

- *Immediacy* is the ability to trade quickly.
    - Modern electronic securities exchanges that can be accessed instantaneously over the internet or similar network have high immediacy. So-called over-the-counter markets that might require a customer to verbally contact many or more dealers have low immediacy.
- *Tightness* (of the bid-ask spread) implies that a round-trip purchase and sale can be accomplished cheaply.
- *Depth* refers to the existence of substantial buy and sell quantities at prices close to the best bid and offer.
    - Suppose the market in stock *A* is "$10.00 bid for 5,000 shares, and 10,000 shares offered at $10.05", and for stock *B*, "$10.00 bid for 100 shares, and 100 shares offered at $10.05." The tightness for *A* and *B* is the same, but *A* has greater depth.

[2] Why did the hit/lift convention develop? I'm not aware of any authoritative pronouncements, but I suspect that it arose from the need for clarity and consistency. The trading process requires fast and accurate communication. All errors have consequences. Many of the worst errors involve direction: buying when you intended to sell or selling when you wanted to buy. As you read this and contemplate things at leisure, an error of direction might seem unlikely or even preposterous. If you've ever participated in an open-outcry floor market (real or simulated), though, you've probably seen more than a few. The hit/lift construction adds a little more information that helps clarify intent.

- *Resiliency*, in the sense of "bounce back," suggests that any price changes that might accompany large trades are short-lived and quickly dissipate.

Liquidity varies across securities: larger, more widely held securities generally enjoy better liquidity than smaller issues. Liquidity also varies across time. Some of this variation is predictable. The market for a US stock is more liquid during regular trading hours (9:30-16:00, Eastern Time) than after-hours. But some of the time variation is random and unpredictable.

Liquidity is sometimes characterized as a *network effect* or *network externality*. Just as one person's benefit from a telephone depends on how many other people can be reached over the telephone system, liquidity depends on how many other people hold and (by implication) trade the security. If many people are active in a market, it is easier to find a counterparty. [3]

Transparency refers to the amount of information available about the market and trading process. In US equity markets, we generally know the full history of trades (volumes and prices) as well as past and current bids and asks. In currency (FX) markets, trades are not reported and bids and asks are not as freely available. As a relative statement, US equity markets are transparent, and currency markets are opaque. It should be noted that good market transparency doesn't imply that there is full or adequate information about the fundamentals of the security.

Transparency is an attribute of the market, not the security being traded. The term *pre-trade transparency* refers to information available before the trade, such as the bid, the offer, and recent price history. *Post-trade* transparency refers to information available after the trade, such as the trade price, executed volume, and (sometimes) identity of the counterparty.

Latency refers to delays encountered in submitting orders and having them acted upon. Immediacy and latency both refer to speed, but while immediacy is a general attribute that encompasses the whole trading process, latency is more narrowly defined. It is usually measured (in milliseconds or microseconds) as the time that elapses from the receipt of an order at the trading center's computer to the dispatch of a responding message from the computer. It is an attribute of the market's technology.

Manipulation loosely refers to a trading practice that distorts market outcomes (such as prices and traded quantities). This is a working definition, not a precise one, and draws on the principle of "I know it when I see it". Aspects of manipulation include deception, fraud, deliberate intent, and usually a goal of self-enrichment. In a particular context, when a court has found a particular practice to be manipulative, we can be more precise. In the US practices that have been found manipulative in this sense include: pump and dump (Section 10.3); marking the close (6.4); wash sales (13.4); and, spoofing and layering (4.6).

2.8. Ownership and transfers

"I bought it. It's mine. I can do whatever I want with it. I can use it, swap it, sell it, donate it to a charity, lend it to you, or throw it away." These assertions essentially summarize the attributes of ownership for small items (such as a book, backpack or jacket). The purchase of something larger (a car or house, for example) often involves financing provided by a bank or other lender. The "owner" retains most significant rights (to drive the car or occupy the house), but other

[3] "Liquidity" can take on a different meaning in other contexts. In corporate finance and monetary economics, liquidity can refer to how easily something can be converted into cash (either by selling it or borrowing against it). On a corporation's balance sheet, for example, holdings of Treasury bills are considered liquid assets because they can easily be sold if the firm needs cash. Inventories might also be considered liquid under the assumption that the firm could borrow money from a bank using the inventories as collateral. When it is necessary to make the distinction, liquidity in the sense just described is called *funding liquidity*, and liquidity in reference to trading purposes is called *market liquidity* (Brunnermeier and Pedersen, 2009).

attributes of ownership are restricted or redefined. The lender will generally impose obligations (for example, that the car/house be insured and maintained). Moreover, if the car/house is sold, the lender has a claim on the sale proceeds. Ownership is in a sense shared with the lender.

Securities exist within a complex legal framework where the attributes of ownership are many and they may be split apart in various ways. If I own shares in a stock mutual fund, for example, I am a *beneficial owner* of the stock portfolio (along with the fund's other investors). Day-to-day management of the portfolio, though, is delegated to a *portfolio manager* (PM) who makes the buy/sell decisions.

The PM might direct a broker to sell 100,000 shares of MSFT. Can the PM tell the broker to transfer the proceeds of the sale to the PM's personal account? As you might expect, there are safeguards to prevent this from happening. Most importantly, the fund's assets are legally held by a *custodian* (typically a *custody bank*). The custodian receives the proceeds of the sale, and in the case of a purchase receives delivery of the shares. At all times the custodian works for the beneficial owners.

For retail investors these arrangements may be streamlined. Most retail investors make their own investment decisions, acting as their own portfolio managers. Long ago, when securities existed as paper certificates, ownership could be direct. Shares, for example, might be stored on a kitchen shelf. Nowadays, the investor's broker usually serves as the custodian.

A trade results in a transfer of beneficial ownership, usually with securities passing into or out of the custodian's control. These transfers are often of high value, sometimes approaching the beneficial owners' net worth. With so much at stake, the transfers must be accurate. Trading is also fast. A trader might flip a position, buying and selling within a few seconds, and this might happen a hundred times in a day. Are the records of ownership really updated in real time at a frequency that keeps pace with the trading?

Not really. The key events and timings are depicted in Table 2.2. What we usually refer to as "the trade," such as a purchase of stock on an exchange or a sale to a securities dealer, is more accurately viewed as a contract to exchange the security and the payment at some point in the near future. The legal transfer of ownership and payment occurs at *settlement*. In U.S. equity markets, if a trade is executed on day "T", settlement occurs on day T+2. (A trade on Monday settles on Wednesday.) Legal transfer is effective at the close of business on Wednesday, irrespective of Monday's actual trade time.

Trade and settlement are well-defined events. The time interval spanning the two events contains numerous verifications, checks and other processes that altogether comprise *clearing*. In clearing, inconsistencies between buyers' and sellers' understanding of the terms of trade (particularly prices and quantities) should be detected early so that they can resolved quickly. Between trade and settlement, both parties to the trade face risk. This risk can be minimized by shortening the settlement time, regularizing the steps involved, and centralizing clearing and settlement.

Custody, clearing and settlement are sometimes referred to as "back office" functions.

Table 2.2 Events and timings for a trade

	Description/Processes	Timing (days)
Existing ownership	Direct or custodian	
Trade	Execution of a trade (usually involving a securities exchange or dealer)	T
Clearing	Confirmations and verifications (identities of buyer and seller, their brokers, prices, quantities, and so forth). Preparation for settlement (source and destination banks for payment; source and destination for security).	
Settlement	Legal transfer of security ownership and payment.	T+2
New owner-ship	Direct or custodian	

2.9. Regulation

Most countries recognize the crucial role that a well-functioning security market plays in raising capital, allocating capital, and hedging. Due to the broad extent of these markets, the most visible regulation usually exists at the national level, supplemented by efforts at consistency, cooperation, and coordination to manage trans-national concerns.

The pre-eminence of national regulation does not imply, however, that all markets and aspects of trading are closely overseen by federal governments. Rules and procedures are instituted and monitored by participants, industry associations, exchanges, even, in some cases, state governments.

"Securities," in US law, comprise corporate stocks and bonds, state and local bonds, and stock options; they are overseen by the Securities and Exchange Commission (SEC, www.sec.gov). "Commodities", including commodity futures and many financial futures are regulated by the Commodity Futures Trading Commission (CFTC, www.cftc.gov). Other financial derivatives (such as swaps) are regulated jointly by the SEC and CFTC.[4] The markets for US Treasury securities are regulated by the Department of Treasury and the Federal Reserve Bank. Currency (foreign exchange, "FX") markets are regulated indirectly in that the largest participants are banks, which are regulated by multitude of agencies. In addition, since currency forwards and futures have FX as their underlying, the CFTC also possesses derived authority.

The stock, stock option, and (to a lesser extent) bond markets are the most prominent markets. The SEC regulates them under the authorization of several Congressional acts. The 1933 Securities Act mostly applies to the primary market for corporate securities, that is, the initial sale of the securities by a corporate issuer. The 1934 Securities Act regulates secondary trading, that is, transactions where the seller is not the issuer. (Most of these notes are devoted to

[4] In the US Code of Federal Regulations, the most relevant material is found under Title 17, Commodity and Securities Exchange (www.ecfr.gov).

secondary markets.) The 1975 Securities Act updated certain aspects of the 1934 Act, most importantly giving the SEC the power to oversee and facilitate the transition to electronic markets.

The Acts leave most details of rulemaking to the SEC. The SEC in turn delegates some its authority to the exchanges or the Financial Regulatory Authority (FINRA, www.finra.org). FINRA is a non-government, not-for-profit corporation that oversees trading and many aspects of broker-customer relations.[5] The power sharing arrangements are sometimes awkward. If the SEC wishes all exchanges to adopt a rule, it must "request" that each exchange make a rule "proposal," which the SEC then approves.

When we discuss market operations, we'll cover some SEC rules that apply directly to the trading process. The SEC also oversees, however, many aspects of the corporate *disclosure* process and insider trading. These rules affect the information environment in which trading occurs. Information is the primary input to the trading decision, so it's not surprising that almost anything that affects its production, communication and use strongly affects the market.

The CFTC was created by the Commodity Futures Trading Commission Act of 1974. Some of the things that it regulates seem very similar to things regulated by the SEC. A trader seeking broad exposure to the market, for example, might buy an S&P Index ETF (an exchange-traded fund, regulated by the SEC) or go long a stock index futures contract (regulated by the CFTC). The similarities are strong enough that we might expect agreement about how the market should be organized and regulated. In practice, though, the ETF and the futures contract are traded under substantially different rules, and regulatory philosophies differ significantly.

In the European Union, securities overseen by the European Commission's Internal Market and Services Directorate General, Directorate G – Financial Markets. The overarching regulation is the Markets in Financial Services Directives 2 ("MiFID 2"). Much authority still resides with the exchanges and their home countries.

Summary of terms and concepts

Exchanges; listing; brokers (retail, prime, discount, full-service); "make or take"; hit the bid/lift the offer; active vs. passive/resting/standing; liquidity (immediacy, breadth, depth, resiliency); transparency (pre- and post-trade), latency; SEC; 1933 Act; primary market; 1934 Act; secondary market; CFTC; FINRA.

References

Brunnermeier, Markus K., and Lasse Heje Pedersen, 2009, Market liquidity and funding liquidity, *Review of Financial Studies* 22, 2201-2238.

Jones, Charles M., 2001, A century of stock market liquidity and trading costs, Columbia University Graduate School of Business, Available at: https://ssrn.com/abstract=313681.

S&P Dow Jones Indices, 2020, Fact sheet, S&P Global BMI, June 30, 2020.

[5] FINRA administers the examinations that US securities professionals must pass in order to practice. Many employees of securities firms (such as retail stockbrokers) take the "Series 7" exam.

Chapter 3. Floor Markets

Many of today's securities markets started as floor markets. A floor market is simply some central place where people go to trade. The facilities can be modest. The New York Stock Exchange (NYSE) initially operated in the Tontine Coffee House (a sort of precursor to Starbucks). The American Stock Exchange started as the New York Curb Market, operating on the sidewalk outside of the NYSE's building.

On the floor, traders meet face to face. They negotiate, bargain, and attempt to reach agreement on terms of trade. A trade is not inevitable: the attempt at agreement might break down, and then someone walks away. Although most trades are bilateral (one buyer, one seller), the negotiation takes place in a crowd. Everyone can see and hear the proposed terms of trade. Anyone can jump in, perhaps displacing a buyer or seller who has dominated the negotiation up to that point.

As a financial institution, "the floor" reached the zenith of its scope and power in the last half of the twentieth century, when it dominated stocks, futures, and options. At the end of the century, however, most markets transitioned to screen-based electronic trading, and floors closed. The London Stock Exchange closed its floor in 1992; The Chicago Mercantile Exchange closed most of its trading pits in 2015. At this point, the transition is nearly complete and floor markets are largely a thing of the past.

So why study them? There are several reasons. The face-to-face negotiations in a floor market have a logic and an immediacy that may make them more familiar and accessible than their disembodied electronic counterparts. Most importantly, many trading practices, rules, and regulations arose in floor markets, and are best understood in the context of a floor market. An electronic market will sometimes exhibit behavior that at first glance looks like something completely new because it embodies advanced technology (particularly when that technology features speed). Then on closer examination, it becomes more familiar, an adaptation of something we've seen in the trading floors of earlier eras. Simply put, floor markets still provide a useful touchstone in understanding current markets. Throughout these notes we will see many examples.

Although floor markets have faded in importance, there are still some notable survivors. Both the New York Stock Exchange and the Chicago Board Options Exchange maintain trading floors. Both floors closed in March 2020 due to pandemic concerns, and all trading moved to electronic systems. The exchanges might easily have taken the opportunity to make the floor closures permanent. In May 2020, however, both reopened.

Real floor markets are highly structured. Day to day, their trading activities involve the same people, and over time these people have evolved standardized practices and rules. This chapter focuses on these rules and practices. It describes the organization and procedures of a typical floor market, based on the rules of Chicago Mercantile Exchange (CME). This approach reflects a deliberate emphasis here on *operational efficiency*. Viewing the Exchange as a factory, operational efficiency means that trades are "produced" quickly and with minimal effort on the part of the traders. (Later chapters will examine allocational and informational efficiencies.)

The Chicago Mercantile Exchange (CME, "the Merc") started in the 19th century trading agricultural futures. A wheat futures contract, for example, calls for the delivery of a given amount of wheat on a given maturity date. The price that will be paid for the wheat is determined in the market when the contract is traded, but the actual exchange of wheat and money won't occur until maturity. This "deferred settlement" feature means that prior to maturity futures contracts can be traded without actually transferring the underlying commodity. For purposes of speculation and hedging, a position in the futures contract can be similar to, but much more convenient than, direct ownership of the underlying. The CME originally listed contracts in grains (such as wheat, corn, and soybeans) and livestock (such as cattle and hogs).

3.1. Floor procedures

To explore how the floor works, we'll dip into an official CME Rulebook, from a vintage around the turn of the millennium (Chicago Mercantile Exchange, 2004). The rules cover almost all aspects of CME governance and procedures. We focus on Chapter 5, which deals with trading practices. Featured prominently near the beginning is:

Rule 520: TRADING CONFINED TO EXCHANGE FACILITIES:

> All trading … must be confined to transactions made on the Exchange; and … must be confined to the designated trading area during Regular Trading Hours … Any member violating this rule shall be guilty of a major offense.

The wording reflects the CME's organizational structure: it has members, like a club or other association. To participate directly in the trading process, one must be a member. Memberships can be bought, sold or leased, but the total number of memberships is limited. Membership is therefore somewhat exclusionary. I do not have to be a member to buy or sell a wheat contract, but if I am not a member, I must pay someone who is a member to act as my broker. This gives rise to a division between floor traders, who have direct access to the market and its information, and off-floor traders, who necessarily see less of the trading process and must wait a bit longer to see their orders executed.

With Rule 520, the Exchange membership is essentially asserting that there is one market, and that the market is under their control. Furthermore, although the floor is often viewed as an arena of pure competition, anyone who tries to compete by accommodating customers' desire to trade in a place or at a time not approved by the Exchange is "guilty of a major offense." The members agree that while they might compete strenuously against each other within the club and according to its rules, they will not attempt to set up a separate club.

Viewed from this perspective, the rule might be interpreted as an anticompetitive attempt to amass economic power against the interests of those who are not members of the club. But there is another aspect to the rule. The centralization and consolidation of trading, the bringing together of all buyers and sellers at a particular place and time, makes it easier for us to find counterparties and negotiate with them, in full confidence that there are no other secret or hidden markets where we might find better terms of trade. In this sense the rule advances the operational efficiency of the market, facilitating rapid negotiation and high trading volume, ultimately benefiting members and non-members alike.

The tension between these two views of Rule 520 arises in many of the rules that markets devise for their members. Is a market a "natural monopoly"? Should all markets be consolidated, or should we encourage competition and accept the resulting fragmentation? We'll return to the debate later in these notes.

A member on the floor might be trading on his own account, that is, relying on personal funds or those of his employer. Alternatively, he might be acting as a broker, an agent for a customer who is not a member and can't directly participate in the floor trading. An order conveyed by an off-floor customer would specify direction (buy or sell) and quantity (number of contracts). The order might be "at the market" (a market order), which instructs the broker to try to execute the trade as quickly as possible, at the best price currently available in the market. In terms of "make or take," a market order directs the broker to "take". The broker should not delay in hopes of getting a better price.

Alternatively, the customer might submit a limit order. In addition to direction and quantity, a limit order has a limit price. "Buy three December wheat, limit $4.00," instructs the broker to buy ("go long") three wheat futures contracts that mature in (the nearest upcoming) December. The broker can pay as much as $4.00 [per bushel], but the buyer would obviously like to buy the contracts at a lower price, if possible.

So given this order, "Buy three, limit $4.00," how would our floor trader proceed? We turn to …

Rule 521: Pit Trading

> All transactions … shall be by open outcry in the established pit for that transaction …

The CME trading floor is bigger than a soccer field, and it is very crowded. The floor is divided into small areas called pits. Each traded commodity has a designated pit. The pit has a distinctive shape. It is constructed as a set of nested octagons (eight-sided shapes) that slope downwards toward the center, like a sports arena. This ensures maximum visibility for the traders. "Open outcry" means simply that bids to buy and offers to sell must be made orally.

Most of the people in the pit will be traders, but there may also be exchange employees such as reports, who record the prices of trades as they happen, and exchange officials who oversee the activity. Traders, reporters, and officials are distinguished by the color of their jackets. A trader will also be wearing a badge that displays a short code that identifies him for trading purposes. We'll call our broker "ALN". Rule 521 continues:

> A bid shall be made only when it is the best [highest] bid available in the pit.

Why is this necessary? Can't a potential seller hear all bids and simply ignore all bids except for those (there may be more than one) at the top? This might work in principle, but the pit is a crowded and noisy place. If thirty buyers were simultaneously announcing their bids, it would be difficult for a potential seller to hear and keep track of which bid was best. The rule partially shifts the burden of clarifying the best bid to all other bidders. Importantly, the rule does not preclude multiple people making the same bid. If someone is already bidding 100, anyone else can bid 100, joining the established bid.

In addition to indicating a price, anyone bidding or offering must also indicate the quantity:

> A bid is made by stating the price first and quantity next (such as "38.50 on 2," etc.) and by holding a hand outstretched with the palm towards the bidder indicating the quantity by the number of fingers shown.

The price-then-quantity convention for bids is followed in most floor markets. This makes sense when you fill in the missing words: [I'm] bidding $4 for 3 [contracts]. The hand gesture is also significant. With the palm faced inwards, I'm miming the act of pulling the contracts towards myself.

So far, we've been discussing bids. Similar rules apply to offers:

> An offer shall be made only when it is the best [lowest] available offer in the pit. An offer is made by stating quantity first and price next (such as "2 at 38.50") and by holding a hand outstretched with palm away from offeror indicating quantity by the number of fingers shown.

Filling in the missing words, "[I'm offering] two [contracts] at $38.50." With the palm facing away, I'm pushing the contracts away, toward a potential buyer.

So, we follow our broker ALN attempting to buy three contracts limit $4. Suppose that he enters the wheat pit, hears BEV bidding $3.50 for seven, and CAM asking $4.10 for five. He could, within the rules, bid the customer's limit price, "Bidding $4 for three," but it's not a good idea to start any

negotiation by stating your worst acceptable terms. Perhaps ALN can fill the order at a better (lower) price: "Bidding $3.80 for three."

At this point CAM might lower her price: "Asking $3.90 for five." ALN might counterbid closer to CAM's offer, and we might see convergence to the point where agreement (followed by a trade) seems likely.

Now how does a trade actually take place? You might think that a trade would occur automatically whenever there happened to be a match between the bid and offer. Suppose that ALN bids $3.85 and CAM offers at $3.85. We have a buyer and seller who have expressed a mutually agreeable price. This agreement, though, is not sufficient to cause a trade to occur. Rule 521 continues:

> When a trader desires to buy the going offer in the pit, he shall by outcry state "buy it" or "buy them" or "buy" followed by the quantity desired, as the case may be. When selling, the trader shall similarly, by outcry, state "sell it" or "sell them" or "sell" followed by the quantity desired.

Bids and offers are passive. Statements like, "Bidding $3.80 for two" or, "Offering $3.90 for four," simply indicate availability. For a trade to occur, someone must act, shouting "Sell it!" to hit another trader's bid, or "Buy it!" to lift another's offer.

Suppose we have:

> ALN: "[Bidding] $3.85 for three."
>
> CAM: "Five [offered] at $3.85."
>
> A momentary pause, and then,
>
> CAM: "Sell 'em!"

At this point, in the normal course of things, ALN acknowledges CAM, and we'd say that a trade has occurred. A floor reporter overhearing the events would key in the price, and "$3.85" would be broadcast to the world.

For ALN and CAM, though, the dialog continues. Each will report the trade to the Exchange, their respective firms, and, unless they are trading for their own accounts, their customers. The transaction then moves into clearing (mutually confirming the terms of the trade), and settlement (transferring ownership and payment in a manner that is legal and irrevocable). Trades that don't clear, because of some discrepancy in price, quantity or identity of buyer or seller, are bounced back to the traders for resolution.

It all seems quite simple and straightforward. What could go wrong?

3.2. Reputation

In the normal course of events, ALN responds to CAM's "Sell 'em," with a confirmation that they have a deal at $3.85. Failure to confirm a transaction is a trading infraction (Rule 514A.4). Suppose, though, that instead of confirming, ALN shrugs and says, "Okay, but my price is $3.80."

> CAM: "But you were just bidding $3.85!"
>
> ALN: "That's history. My customer just changed his bid. You were too slow."

What do the rules say in a case like this? Once a bid or offer is out of my mouth, for how long can other traders presume that it is available? The formal rules are silent on this point. We have instead a convention, that is, a practical guideline, which says something like "as long as the breath is warm."

That helps a bit. If the elapsed time between ALN's bid and CAM's "Sell 'em" is a half-second, then we side with CAM. If the time is five minutes, then we'd probably say that CAM should have at least confirmed that ALN's bid was still available. But what if the elapsed time were something in between, like two or three seconds? Perhaps there should be a rule that specifies a precise presumed duration, but it is probably not practical for traders in a fast crowded pit to manage stopwatches on top of everything else.

As a practical matter, the situation can be handled by appealing to reputation. On the floor we're face to face. The floor is not an anonymous market. A trader knows, for better or worse, the identities of her counterparties. If ALN backs away from his bid once, that is something that could happen to anyone. If ALN backs away twice, CAM will look for someone else to trade with. If ALN habitually fails to honor his bids and offers, he will find that the floor can be a very lonely place, where it can be very difficult to trade.

Reputation, behavioral expectations based on a presumption of repeated and ongoing interaction, provides a cohesiveness to the market. Small misunderstandings stay small, larger differences are avoided. Reputational enforcement of trading norms arises in a great many situations. Reputation also, on the other hand, can facilitate collusion. We noted earlier the "private club" aspect of Rule 520. This is easier to sustain when members know each other, go along, and get along.

3.3. Priorities

Suppose that there are multiple people bidding 100 for at least one contract. If "SAM" wants to sell a contract, which bidder should he hit? Many real-world transactions rely on time priority: first-come-first-served. An orderly line might work well at, say, an ice cream parlor. In the crush of the pit, however, it is not viewed as practical. There is no time priority at a price. If the bidders are at the same price, SAM can hit the first bid, the last bid, the bid of the trader standing closest to him, the bid of the trader he rides the train with, or a bidder selected at random.

Although the crowd in the CME pit may view time priority as impractical, it may nevertheless be desirable. We want to keep things moving quickly, and time priority encourages this by rewarding the early bidders. The floor at the New York Stock Exchange follows a modified rule in which the first trader to bid or offer at a price has priority, but this doesn't extend to bids or offers after the first. That is, on the NYSE, if ALN, BEV, and CAM bid in that order, SAM would have to sell to ALN first, but after that he (or anyone else) could sell to CAM before BEV.

Besides time, a market might favor other attributes of a bid or offer. One possibility is size (with larger bids or offers at the head of the queue). Among such secondary priorities, rules that favor "customer" (off-floor) orders are common. Electronic markets often allow for hidden (non-displayed) orders that must yield to visible (displayed) orders. We'll examine these and other secondary priorities in later material.

3.4. Trade-throughs

In prohibiting inferior bids and offers (Rule 521), the CME is promoting price competition. This is very important. Customers will send their orders to an exchange only if they believe that the prices are available there are better than those available through search or other alternatives. In a crowded and noisy pit, though, we might bid 100 when someone else bid 101 a split-second earlier. Consider the following sequence:

1. ALN bids 100
2. BEV bids 101
3. CAM hits ALN's bid (trading at 100).

In this situation, CAM has traded through BEV's bid, causing a *trade-through*. The is a direct harm because CAM (or CAM's customer) missed an opportunity to sell at a higher price. There is also an indirect harm, though, in that bidders (like BEV) make aggressive bids hoping to be rewarded by having these bids hit. A trade-through deprives BEV of an execution and discourages further aggressive bids by her and others.

This is considered a trading infraction. It was committed by CAM against BEV. If BEV protests the trade-through (and she probably will), it can be remedied in several ways:

- If they both agree, CAM and ALN can break (cancel) the trade or reset the price of the trade to 101 (matching BEV's bid). Either option deprives ALN of an execution that he would have reasonably assumed to be final.
- CAM can sell to BEV at 101, essentially giving her the execution to which she was entitled.

(See CME Rule 528.) Both remedies impose costs on CAM (and possibly ALN), not to mention the time spent by all parties discussing the situation. These penalties are justified as a deterrent because a market in which trade-throughs are common and tolerated is unlikely to survive.

Looking ahead, price priority and discouraging trade-throughs become important concerns in the transition to electronic markets.

3.5. Crosses

Exchange members are often working on behalf of customers who aren't members and cannot directly participate on the floor. In this capacity, they are working as brokers, that is, as agents for the customers and they're doing agency trading. Their customers send them orders, which they try to execute to the best of their ability.

Often customer orders arrive so quickly that the broker may be responsible for multiple orders at one time. If these orders are all on the same side (buy or sell), the broker can work them using the standard floor procedures described above.

It's not unusual, though, for a broker's working set of customer orders to include both buys and sells (from different customers). Suppose CAT receives two customer orders:

- Belle wants to buy, limit 26.
- Sol wants to sell, limit 20.

Belle would like to buy, not paying above 26, but certainly preferring a lower price. Sol will sell as low as 20 but would obviously prefer a higher price. For simplicity, all quantities in this example are one contract. CAT turns her attention to the pit, where the going bid and offer are from ARI ("bidding 21") and SAM ("asking 25").

CAT's simplest course of action is to turn to SAM and say, "buy it," then quickly turn to ARI, and say "sold." She confirms to Sol that he sold at 21 and confirms to Belle that she bought at 25. Both trades are better than Belle's and Sol's limit prices. CAT collects two commissions and moves on to the next customer order. "Fill 'em and bill 'em!" goes one saying.

There are other possibilities, though. CAT represents her two customers. If Belle and Sol were on the floor face-to-face, they might end up trading directly with each other, negotiating a price somewhere in the middle of the 21 to 25 bid-ask range. Brokers who find good deals for their customers get rewarded with more orders. This motivates CAT (and us) to ask, "Can we just pair off Belle and Sol, letting them trade with each other, at 23 (the bid-ask midpoint)?" Perhaps CAT could just walk over to an Exchange clerk, tell them what she's doing, and let them record the details of the Belle/Sol trade, just to make it official.

This might seem harmless and even benevolent but remember the maxim that all transactions shall be by open outcry. If CAT simply notifies the clerk, there is no bid or offer at 23 that could be seen or heard on the floor and no cry of "buy it" or "sold". The pit is simply informed that a trade has occurred. The open outcry procedures have been bypassed.

There are, as it happens, procedures to handle this situation. Rule 533 states:

> A member who is in possession of both buy and sell orders for different beneficial owners for the same product … may execute [the orders against each other] provided that … In pit trading, a member executing such orders shall first bid and offer by open outcry three times at the same price, stating the number of contracts, and, thereafter, if neither the bid nor the offer is accepted, the orders may be matched in the presence, and with the approval, of a designated Exchange official.

To conform to this rule, before CAT pairs off ("crosses") her customers' orders at 23, she must call out, "Bidding 23, asking 23 … bidding 23, asking 23 … bidding 23, asking 23." The crowd in the pit immediately recognizes this as an attempt at a cross.

CAT's bids and asks are not simply formalities. They are live and active. Anyone may hit CAT's bid of 23 or lift her offer ("accepting," in the words of the rule). Now anyone who does this is *breaking* CAT's cross. This creates more work for CAT, as she is now left with one of her customer orders unfilled. Other traders recognize this, and in the normal course of events might refrain. After all, within a

few minutes, things might be reversed: CAT could be positioned to break one of their crosses. Anyone has the right to break a cross, but sometimes a right is better left unexercised.

Is a formal rule really necessary here? Is the principle of open outcry so important that it must be protected by a complicated ("three times") ritual? If we were visiting the floor, the people to ask would be ARI and SAM. They would probably reply that if crosses happened once a week, bypassing open outcry would be okay. But if crosses became so common that their bids and offers on the floor were rarely being hit or lifted and were simply being used by other traders to compute the midpoints for crosses that they couldn't join, ARI and SAM would have less incentive to bid and offer. They and other potential bidders and offerors might not participate, or at least they might make their bids and offers less aggressive. This would be a problem for everyone. ARI's and SAM's quotes are being used to set the crossing prices. Anything that makes their quotes less authoritative or relevant increases uncertainty even for those who only seek to cross.

Without Rule 533, CAT's cross would simply appear as a trade at 23 with no bid or offer at that price. Trades of this sort, disconnected from the record of visible bids and offers that precedes them, are said to be "dark". They have become more common and significant in modern electronic markets. Chapter 8 provides a deeper discussion.

3.6. The transition to electronic markets

We've now covered enough of the basic floor trading process to set the stage for the modern computerized market. But for the moment, let's put ourselves in the visitors' gallery of a floor market around 1985, wondering as we watch the traders on the floor below, how markets will work in the future.

On my computer, the icon for a "folder" looks like a traditional paper file. My folder icons are arranged for display in a view called a "desktop". The form of these icons intuitively suggests what they represent and how they are used. This visual correspondence is pervasive in our personal electronic devices.

When securities markets were going electronic, the principle of visual correspondence suggested that a computerized market might resemble very closely the floor market that it replaced. In the late 1980's, the Chicago Board of Trade (a rival of the CME) unveiled a system, developed jointly with Apple, called Aurora. The New York Times reported (Berg, 1989):

> Like video games, Aurora will rely on computer graphics to display images of living creatures on the screen of traders' computer terminals. In particular, the Board of Trade has tried with Aurora to replicate the "open outcry system," under which traders shout buy and sell orders to one another in a large, frequently crowded trading ring.
>
> With Aurora, which will operate only when the Board of Trade is closed, traders will be given a computer keyboard and screen. Displayed on that screen will be a large square that is supposed to represent a trading ring. And inside the square will be small circles and squares representing buyers and sellers.
>
> Each of these small circles and squares will have some form of identification so users of the system can know who is who. When a buyer wants to buy futures contracts from someone offering them for sale, he will use a pointing device known as a "mouse" [quotes in the original] to designate a particular seller on the screen. Then he will execute the trade by entering a buy command at his keyboard.

The London International Financial Futures Exchange (LIFFE) developed a similar screen-based representation of the floor called Automated Pit Trading (APT) (Parikh and Lohse, 1995). In an apparent attempt to capture one other aspect of the floor, Parikh and Lohse also note that, "[the APT uses a 10 second limit on all orders after which they must be refreshed. This added feature forces traders to get more involved in the trading process. On the floor, traders must shout and gesticulate continuously in order to place a trade. Thus, on a theoretical system as proposed above, such a feature could be incorporated to keep the traders' orders executable as long as their 'breaths are still warm'," p. 302.

Aurora was never implemented. The APT had a brief period of usage as an after-hours trading system. The future of electronic trading, though, did not belong to the floor, even as a visual representation. At the same time CBT was developing Aurora, the CME was pursuing GLOBEX, an electronic market based on limit orders. Globex and other electronic limit order markets won out, becoming the dominant successors to the floor exchanges. To these we turn next.

3.7. Further reading

MacKenzie (2013) provides a good description of the floor-to-electronic transition at the Chicago Mercantile Exchange, drawn in part from interviews with some of the key participants. Perhaps because of the intense interpersonal interactions, floor markets seem to attract interesting characters. James Allen Smith's documentary film *Floored* shows them telling their stories of success and failure on the floor and beyond (Smith, 2013).

Instructors' note

In introductory economics courses, professors sometimes run double-auction markets in the tradition of the classic experimental markets (starting with (Chamberberlin, 1948)). In most respects these are equivalent to a securities floor market in which brokers represent customer limit orders, but the terminology differs. In an economics class there are typically buyers (customers) with "values" and sellers (producers) with "costs". In an exercise for a finance class, the buyers and sellers simply have limit prices. The objective is identical, though, to maximize the surplus between the trade price and the limit price.

In an economics class, the rules of trade are of lesser importance. Absence of explicit rules may even be a virtue as it helps to demonstrate the robustness of markets. In a securities market, though, operational efficiency (that is, efficiency in the "production" of trades) is generally very important. I model the class exercise based on the CME rules of trade. The players experience the rules and the order that the rules bring to trading. After a few rounds, the students grow very efficient at the process. Moreover, situations arise in class that also occur on a real floor. For example, students will often favor specific counterparties ("preferencing"). So, they get a sense of how a floor community can develop, for better or worse. Trading errors occur, and must be corrected, and so forth.

There is software that implements internet-based double-auction markets. I've used Charles Holt's vEconlab system (http://veconlab.econ.virginia.edu/introduction.php, also see (Holt, 2007)). I also note a recent system, Kiviq, which also runs on smartphones and iPads (Hampton and Johnson, 2020). Both are free for classroom use.

Screen-based instructional markets are useful, and they're certainly compliant with the social distancing requirements that still loom large as of this writing. From a teaching perspective, though, I think that something is missing. Students may feel like they are playing against a machine, or that the system somehow constrains behavior. When they are face to face with human competitors, they learn to recognize and remember the wide variety of behaviors (including bluffing, intimidation, and seemingly random actions) that persist in almost all markets, floor or virtual.

References

Berg, Eric N., 1989, Chicago Board of Trade Challenges Rival Globex System, March 16, 1989., New York Times.

Chamberberlin, E.H., 1948, An experimental imperfect market, *Journal of Political Economy* 56, 95-108.

Chicago Mercantile Exchange, 2004, [Online] CME Rule Book, April 6, 2004.

Hampton, Kyle, and Paul Johnson, 2020, Kiviq. us: A free double auction Internet classroom experiment that runs on any student device, *The Journal of Economic Education* 51, 209-209.

Holt, Charles A., 2007. *Markets, Games and Strategic Behavior* (Pearson (Addison Wesley)).
MacKenzie, Donald, 2013, Mechanizing the Merc: the Chicago Mercantile Exchange and the rise of high-frequency trading, School of Social and Political Science, University of Edinburgh, Available at: http://www.sps.ed.ac.uk/__data/assets/pdf_file/0006/93867/Merc21b.pdf.
Parikh, Satu S, and Gerald L Lohse, 1995, Electronic futures markets versus floor trading: implications for interface design, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.
Smith, James Allen, 2013, Floored.

Chapter 4. Limit order markets

When our markets became computerized, they did not take the form of an app that visually simulated a trading floor, or anything like it. Instead, they drew on a particular component of a floor market: the limit order book. The system that evolved now governs most trading in equities, options, and futures contracts.

Recall that a limit order specifies buy or sell, quantity, and a limit price. "Buy two July wheat, limit $4.40" is an instruction to buy (long) two July wheat futures contracts at a price no higher than four dollars (per bushel). In a floor market, the order would be communicated to a floor broker (exchange member). This member would then act as an agent for the order. He would "work" the order, representing it on the floor, bidding as if he were the actual submitter.

A broker with many customers might accumulate many limit orders, and he would need to maintain a record of those that have not yet been executed (or cancelled). This collection is his limit order book (or simply, "his book"), and it would have originally been maintained on paper.

A broker's book might be quite full, with orders to buy priced "near the market" (that is, close to the current bid) and orders priced "away from the market" (well below the current bid). It might also have orders to sell, near to and away from the market (close to or far from the current offer). In working these orders, the broker has to pay attention to the market activity, standing ready to bid or offer on behalf of any order is his book.

Suppose that the $4.40 limit buy order is given to the broker when July wheat is $4.50 bid, offered at $4.51. The order is priced away from the market, and the customer would not expect an immediate execution. Now suppose that during the trading session the price falls, with bids and offers stepping downwards, so that an hour after the order is submitted, the offer is two contracts at $4.39. At that point, the broker should step forth and shout "buy them," executing his customer's order.[1]

[1] The worst outcome is when an inattentive broker fails to execute the order, and price subsequently moves away. Suppose that the price reaches a low of $4.39, but subsequently rebounds to $4.70. A customer who'd entrusted "buy limit $4.40" to a broker would probably assume that

In a futures market, the broker would typically be handling only wheat limit orders. By remaining in the wheat pit, she'd be positioned to monitor the market. A broker on the floor of a stock exchange, however, might receive limit orders in different stocks. If those stocks are traded in different areas of the floor, following the market might be difficult. When she runs over to listen to stock *A*'s bids and offers, she's too far away to hear stock *B*'s activity.

There are several solutions. She might temporarily give the orders to another broker. Alternatively, the members might collectively decide to rotate the trading. ("Now we're all going to trade Stock *A* … And now we're all going to trade stock *B* …") Finally, the exchange might decide to have one limit order book in the stock, managed by one person who always remains in the trading area.

Each solution has been used effectively, and we'll see some examples, but for now, we'll focus on the last suggestion. At the New York Stock Exchange (say, in the mid to late 20th century), there was one limit order book in each stock, and it was managed by an exchange member called the specialist.[2] On the floor of the Tokyo Stock Exchange, the book was managed by an exchange employee called the saitori.

Traditionally the book was maintained in pencil-and-paper (New York) or on a chalkboard (Tokyo), but the present-day book is computerized, a dynamic data structure that gets revised as orders arrive, get executed, or cancelled. A market can be organized so that all activity flows through the book. We'll now look at how such a market operates.

Some exchanges make their books available in real time via the internet. The CBOE exchange owns several stock exchanges that are very transparent (see http://markets.cboe.com/us/equities/overview/). For the most part, however, book information is not widely and freely disseminated. The NYSE currently charges several thousand dollars per month for its own book (see https://www.nyse.com/publicdocs/nyse/data/NYSE_Market_Data_Pricing.pdf). Building a "consolidated" book covering all US stock exchanges would require subscribing to all exchanges' feeds.

4.1. Basic limit order processing

Let's start with an order, "Buy 200 MSFT limit 25," ("Buy 200 shares of MSFT, but don't pay more than $25 per share.") When this order arrives at a limit order market, there will first be an attempt to match it. A match (also called a trade, execution, or fill) occurs when price of the arriving order meets or crosses the price of a pre-existing ("standing" or "resting") order.

Suppose that it is early in the day and the MSFT book has one resting order: "Sell 200 MSFT limit 25."

- A new arriving order, "buy 200 MSFT limit 25" is marketable. There is a trade at 25. The price is determined by the resting order. An arriving order, "Buy 200 MSFT limit *500,*" would still result in a trade at 25.
- Quantity is determined by the smaller of the buy and sell quantities: "Buy 300 MSFT limit 25" results in a trade of 200 shares (the offer amount). "Buy 50 MSFT limit 25" causes a trade of 50 shares (the buy amount).

For a given incoming marketable order there might be multiple resting limit orders that are valid candidates for execution. Who gets priority? In most markets, the first priority is price. A buy order with a relatively high price is said to be (relatively) aggressive: the buyer is willing to

---

she'd bought at $4.40 a contract that was now worth much more. If the broker failed to execute ("missed the market"), the customer would have a very legitimate grievance.

[2] When the NYSE was founded, each member would have managed his own book. The specialist system evolved in the late 1800s; by the 1960s there was generally one specialist per stock. For more details, see Chapter 7.3.

pay more. A sell order with a relatively low price is aggressive: the seller is willing to accept less. More aggressive limit orders have priority over less aggressive orders.

After price, however, markets vary in their secondary priorities. These priorities are usually established with the intent of promoting the exchange's perceived liquidity, enhancing the attractiveness of the exchange as a desirable trading venue.

The most common secondary priority is time: the order that arrives earlier is executed before the later arrival. This principle rewards a bidder and offeror for stepping in quickly with their best price. An order is usually timestamped when it is received by the market.

Many markets accept hidden orders. These orders reside in the book, and are available for execution, but are not displayed. When such orders are allowed, visible orders usually have priority over hidden orders, even if they arrive later. A displayed limit order is like an advertisement for the market, and so priority for these orders makes sense from the exchange's viewpoint.

Price, visibility, and time, in that order, define the most common priority scheme. To see how these priorities work, consider the following examples.

Suppose that the following buy orders arrived in sequence.

Price	Visibility	Time	Trader
20.05	Hidden	9:30	Amy
20.04		9:31	Brian
20.04		9:32	Chao
20.05		9:33	Dmitri
20.03		9:34	Esteban
20.06	Hidden	9:35	Florio

The hidden orders are in the book, but they are not visible to market participants. In the present discussion this is indicated here by the dimmed (grayed) font. Similarly, the system will know the identity of the traders (or at least, their brokers). We show the names here for our own convenience in referencing the orders, but they aren't visible to market participants.

To build the buy side of the limit order book (the bid book) we first sort by price:

Price	Visibility	Time	Trader
20.06	Hidden	9:35	Florio
20.05	Hidden	9:30	Amy
20.05		9:33	Dmitri
20.04		9:31	Brian
20.04		9:32	Chao
20.03		9:34	Esteban

Next, orders at the same price are ranked by visibility:

Price	Visibility	Time	Trader
20.06	Hidden	9:35	Florio
20.05		9:33	Dmitri
20.05	Hidden	9:30	Amy
20.04		9:31	Brian
20.04		9:32	Chao
20.03		9:34	Esteban

The ranking by visibility changes the relative positions of Dmitri and Amy. At this point we are done. Because the initial set was ordered by arrival sequence, Brian's order and Chao's order, which have the same price and visibility, are correctly ranked. Finally, although we've shown the sorting applied to an already-received collection of unsorted orders, the book is maintained in sorted order and updated whenever a new order arrives.

The highest bid on the book is Florio's 20.06. If any executions in this set were to occur, Florio's order would come first. The market's *bid quote* only reflects displayed orders. The market's bid is therefore Dmitri's 20.05. This is also described as the *top of the market's bid book*.

Organization and display of the ask book

Suppose that we are given the following sequence of sell orders.

Price	Visibility	Time	Trader
20.20		9:30	Gregori
20.18		9:31	Haley
20.18	Hidden	9:32	Inez
20.10	Hidden	9:33	Jing
20.15		9:34	Kala
20.18		9:35	Lou

The ask book is constructed the same way as the bid, with the most aggressive offers at the top. Here is the result:

Price	Visibility	Time	Trader
20.10	Hidden	9:33	Jing
20.15		9:34	Kala
20.18		9:31	Haley
20.18		9:35	Lou
20.18	Hidden	9:32	Inez
20.20		9:30	Gregori

Note that Inez loses time priority to Lou because her offer is hidden. Here, as in the case of the bid book, the highest-ranked order (Jing's 20.10) is hidden. The market's ask quote is Kala's offer at 20.15.

On screens, the bid and ask books are often shown side-by-side, with the bid side on the left. Following this convention, the orders discussed above would be shown as:

Bids					Offers			
Price	Visibility	Time	Trader		Price	Visibility	Time	Trader
20.05		9:33	Dmitri		20.15		9:34	Kala
20.05	Hidden	9:30	Amy		20.18		9:31	Haley
20.04		9:31	Brian		20.18		9:35	Lou
20.04		9:32	Chao		20.18	Hidden	9:32	Inez
20.03		9:34	Esteban		20.20		9:30	Gregori

This is what would be available to an official observer at the market. Traders would not see the hidden orders, or any indication that they existed. An external display would show:

Bids					Offers			
Price	Visibility	Time	Trader		Price	Visibility	Time	Trader
20.05		9:33	Dmitri		20.15		9:34	Kala
20.04		9:31	Brian		20.18		9:31	Haley
20.04		9:32	Chao		20.18		9:35	Lou
20.03		9:34	Esteban		20.20		9:30	Gregori

Alternatively, limit order books are sometimes shown vertically, with offers on the top and bids on the bottom. In this arrangement, the set of orders would be displayed as:

	Price	Visibility	Time	Trader
	20.20		9:30	Gregori
	20.18		9:35	Lou
	20.18		9:31	Haley
Offers	20.15		9:34	Kala
Bids	20.05		9:33	Dmitri
	20.04		9:31	Brian
	20.04		9:32	Chao
	20.03		9:34	Esteban

Note that with vertical arrangement, the positioning of the offers is reversed, to keep the price direction consistent.[3]

4.2. Interactions with incoming orders

The last section briefly described simple executions. When the book has multiple orders, things can get more involved. Consider the following book (presented vertically):

	Visibility	Price	Quantity	Submitted	Trader
		50.12	1,000	9:30	Cathy
		50.11	500	9:32	Bill
	Hidden	50.10	200	9:30	Gina
Offers		50.10	400	9:31	Amy
Bids		50.05	1,000	9:30	David
		50.04	500	9:32	Ellen
		50.03	400	9:31	Fred

Example 1

Now suppose that at 9:40 Hari submits an order, "Buy 200, limit 50.10." This will execute against Amy's order, leaving 200 shares remaining on Amy's order.

[3] The Rotman Interactive Trader market simulator has both types of display. The Book Trader panel is side-by-side; the Ladder Trader panel is vertical.

Example 2

Now suppose Hari's order had been instead, "Sell 1,200 limit 50.04." Because this sell order is priced below the best bid, it will at least partially execute. In this case, 1,000 shares execute against David, at $50.05; 200 shares execute against Ellen at 50.04. Ellen has 300 shares remaining, and 50.04 becomes the market's bid quote. An order that executes at multiple prices is sometimes said to *walk through the book*.

Example 3

Hari (at 9:40): "Sell 1,200, limit 50.05." Now, 1,000 shares will execute at 50.05 against David, but the next available bid, at 50.04, is worse than the Hari's limit price. So, Hari's remaining 200 shares are added to the book on the sell side:

	Visibility	Price	Quantity	Submitted	Trader
		50.12	1,000	9:30	Cathy
		50.11	500	9:32	Bill
	Hidden	50.10	200	9:30	Gina
		50.10	400	9:31	Amy
Offers		50.05	200	9:40	Hari
Bids		50.04	500	9:32	Ellen
		50.03	400	9:31	Fred

Example 4

To the original book, Hari submits: "Buy 500, limit 50.11." Now, 400 shares will execute against Amy at 50.10. Then 100 shares execute against Gina, also at 50.10. The new book is:

	Visibility	Price	Quantity	Submitted	Trader
		50.12	1,000	9:30	Cathy
		50.11	500	9:32	Bill
Offers	Hidden	50.10	100	9:30	Gina
Bids		50.05	1,000	9:30	David
		50.04	500	9:32	Ellen
		50.03	400	9:31	Fred

The market's new ask quote is 50.11.

From the trader's perspective, executions against hidden orders are a bit unexpected. Based on what he could see, Hari would have expected the last 100 shares of his order to execute at 50.11, paying up for those shares. His actual outcome is better: he buys everything at the lower price. Bill is less pleased. Based on what Bill could see, he might have thought that he was next in line after Amy. Gina's hidden order displaces his interest.

4.3. The dynamics of limit order markets

The last section looked at book interactions involving small sets of buyers and sellers, with each interaction completed within a short time window. Studying these interactions in isolation is the best way to learn the basics of book operation. In real markets, though, these scenarios generally flow into each other, with overlapping sets of participants and newly arriving information. This generates important dynamics (changes over time) in the bids, offers and trades.

We'll illustrate some of these dynamics using a widely held and actively traded security, a Standard and Poor's depository receipt (SPDR, or "spider"). The SPDR with the US ticker symbol SPY is an exchange-traded fund (ETF). Its assets are shares in other stocks. The portfolio is designed to mirror the S&P Composite Index of 500 stocks. From an investor's viewpoint, it provides a convenient and low-cost way to hold the index portfolio. It is also attractive from a trader's perspective. Whereas a standard (open-ended) mutual fund can only be purchased or sold at daily closing prices (the net asset value), the SPY can be traded intraday, just like any other stock. The market for the SPY is very liquid: the bid-ask spread is generally $0.01 (the minimum tick); sizes at the bid and ask are large; trading volumes are large.

Figure 4-1 depicts the bid for SPY for a brief time interval (1.5 seconds) on a randomly chosen day. There are some distinctive features. The graph is a "blocky" step function: bids and asks in most US stocks must be in increments of $0.01. The graph is continuous: there is always a bid.[4] Finally, even over this short interval, the bid undergoes frequent changes. Many of these changes persist long enough to be clearly visible in the graph, but some are so quick that they give the appearance of a brief spike. Figure 4-2 depicts the offer (ask) price for SPY over the same time window. The form of the line (a step function on a $0.01 grid) is similar to that of the bid.

Next, to illustrate the comovements of the bid and ask (their joint dynamics), Figure 4-3 plots them together. The bid and offer broadly move together, but they don't move in lock step: it's as if they are loosely tied. The spread between the bid and ask varies. The offer is generally above the bid, but at times it looks like they are almost touching. Within one order book, of course, the bid and ask never touch because any order submission that might cause them to touch would force an execution. With multiple markets (as we'll see in the next chapter), though, touching and even crossing are possibilities.

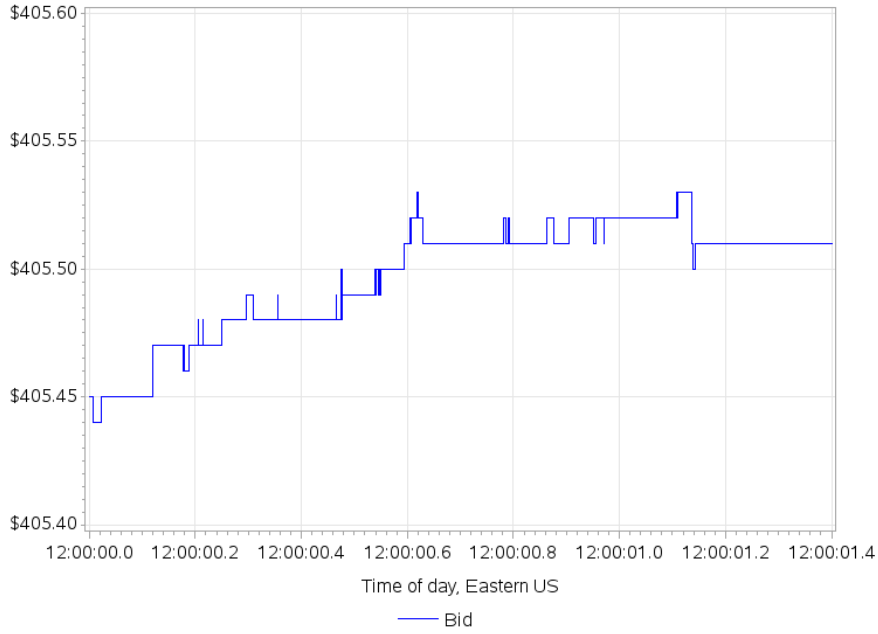
Figure 4-4 plots the trades in SPY. Each trade is depicted as a single dot. Whereas bids and asks are generally continuous and persistent, a trade is a singular event. Each trade occurs at a well-defined time. We might say that the bid *between* 12:00:01.2 and 12:00:01.4 is $405.51, but a trade has no such persistence. It happens and then it's over.

Figure 4-5 pulls everything (bids, asks and trades) into one picture. Now several other features come into focus. First, although there is at any given time one bid and one ask, there might be multiple trades at different prices. Remember that an incoming marketable order might execute against multiple limit orders resting in the book. Also note that one trade (very early in the graph, shortly after 12:00:00) appears to execute between the bid and offer. This might happen if the execution occurs against a hidden order. This early trade also has one other curious feature: it is priced at $405.555 cents, that is, *off* the $0.01 grid that restricts bids and offers. Most "non-penny" trades arise from so-called "dark" trading mechanisms, which we'll examine in Chapter 8.

---

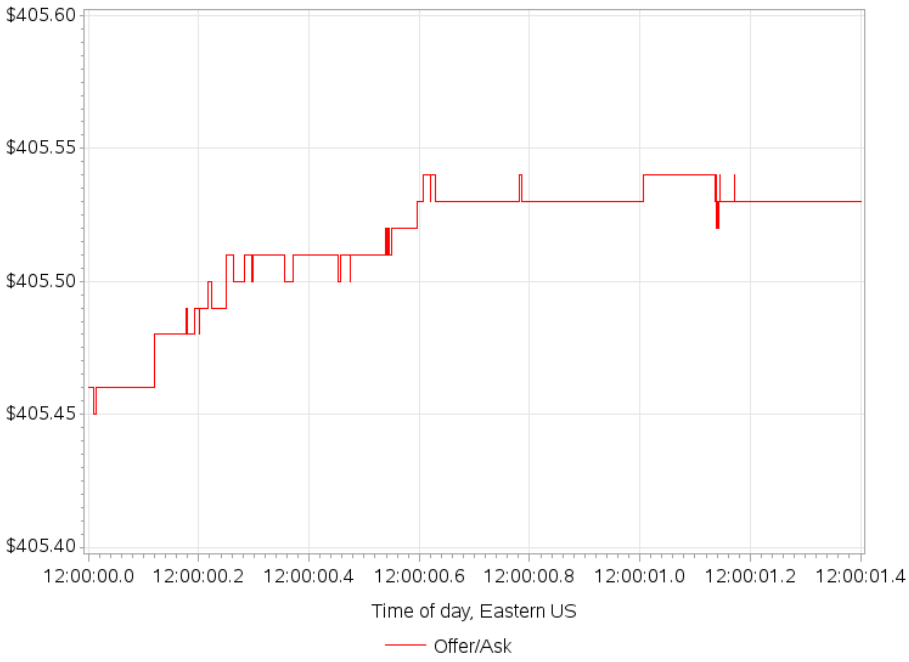
[4] From a formal mathematical perspective, the bid over time is generally said to be right-continuous and left-limited. That is, as we approach a change-point from the left (before the change), the bid function possesses a limit. Approaching a change point from the right (after the change), the function is continuous. In the graph, the change transition is bridged by a solid line, but this is for visual clarity only: at any given time, the bid has only one value.

Figure 4-1 The bid for SPY, starting at 12:00 on Tuesday May 17, 2022



Bid (in dollars per share) versus time (HMS and tenths of seconds). The bid plotted here is the National Best Bid (NBB), the highest bid at the indicated time among all market centers that disseminated quotes in SPY.

Figure 4-2 The ask/offer for SPY, starting at 12:00 on Tuesday May 17, 2022



Ask/offer (in dollars per share) versus time (HMS and tenths of seconds). The offer plotted here is the National Best Offer (NBB), the lowest offer at the indicated time among all market centers that disseminated quotes in SPY.

Figure 4-3 The bid and ask quotes for SPY, starting at 12:00 on Tuesday May 17, 2022

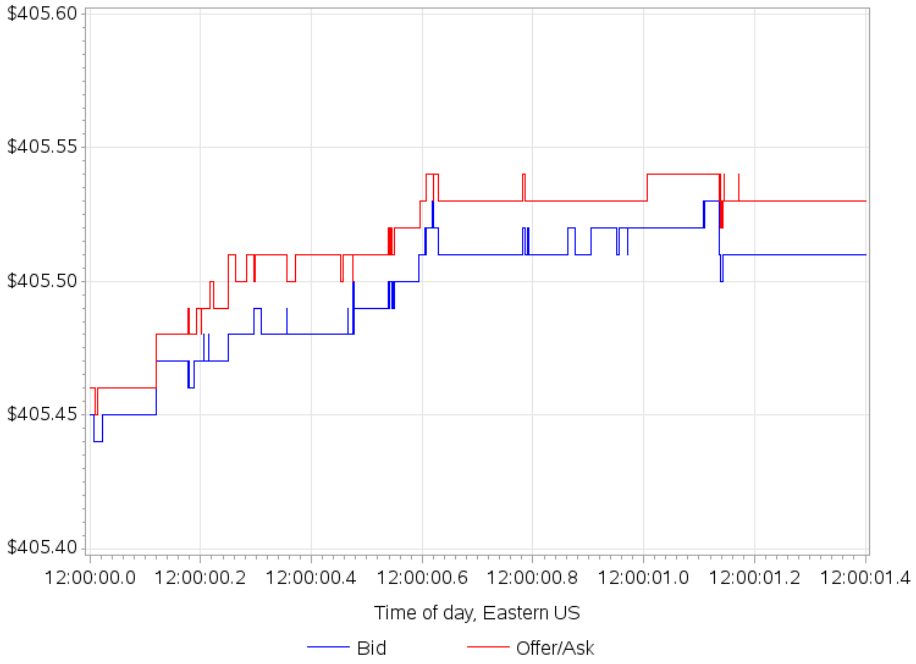


Figure 4-4 Trades for SPY, starting at 12:00 on Tuesday May 17, 2022

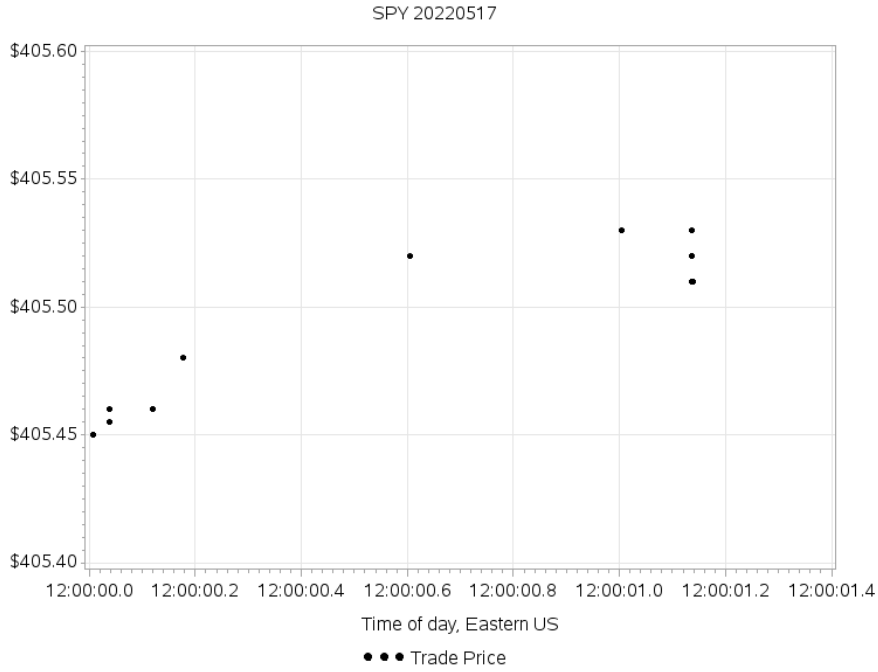
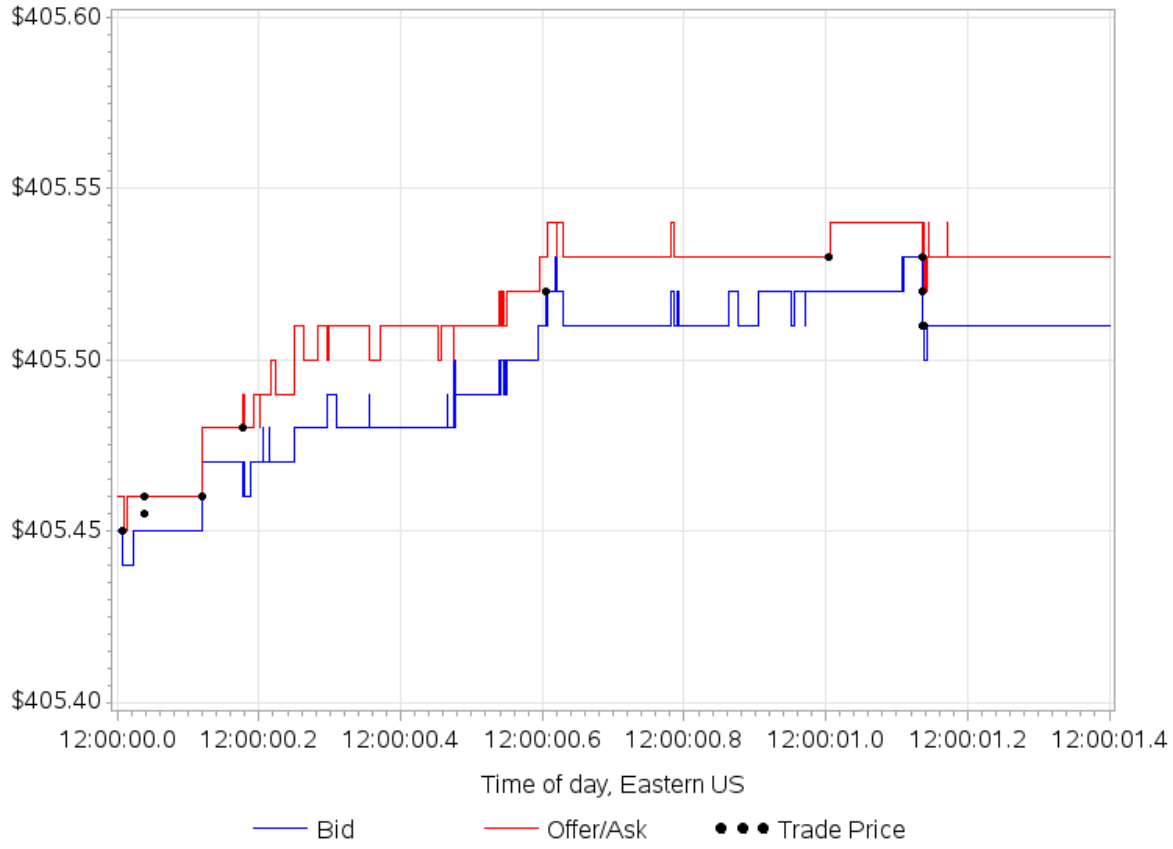


Figure 4-5 Bids, asks, and trades for SPY, starting at 12:00 on Tuesday May 17, 2022



4.4. More complex interactions and order qualifiers

Sometimes qualifications are added to orders. Suppose that the current book is:

	Price	Quantity	Entered at	Trader
	50.12	1,000	9:30	Cathy
	50.11	500	9:32	Bill
Offers	50.10	400	9:31	Amy
Bids	50.05	1,000	9:30	David
	50.04	500	9:32	Ellen
	50.03	400	9:31	Fred

If Gina sends an order, "buy 500 shares, limit 50.10," she gets a *partial fill*. She buys 400 shares @ $50.10 from Amy. Her remaining 100 shares are added to the book. They are added on the bid side (because she's a buyer). The market's new best bid is hers, $50.10. The best offer is now Bill's, $50.11.

One common qualification is *immediate or cancel* (IOC). If Gina's order were stamped IOC, she still would have bought 400 shares, but her unexecuted remainder would not have been added to the bid side. Even if she would be willing to buy 100 more shares at $50.10, she does not want to make her interest visible.

If Gina's order had been stamped *all or nothing* (AON), she would not have bought 400 shares. In most systems, her order would still be considered active, though. If Amy (or some other buyer) adds 100 more shares to the offer side at $50.10, Gina's order would then execute for 500 shares. Gina might use an AON if she were concerned that *any* execution would be perceived as a signal that more shares were sought.

The *fill or kill* (FOK) qualification is a combination of IOC and AON. If Gina's order is marked FOK, nothing happens: the order can't be executed immediately in full and is therefore cancelled.

These qualifiers (IOC, AON, FOK) lie at one end of a continuum that includes conditional orders and complicated algorithms, to which these notes will later return.

4.5. Alternative priority rules

Although price/visibility/time priority is most common, markets sometimes experiment with and adopt alternatives.

Some markets give priority to larger orders. Large orders advertise the exchange as a place that can handle large volumes. This same goal is often accomplished by using pro rata allocations. In a pro rata system, all limit orders at a price split execution quantities in proportion to their sizes, irrespective of when the orders were submitted. For example, suppose that there are two limit orders bidding $100, and that the sizes of these orders are 100 and 900 shares. If a marketable order arrives, say "sell 200 shares limit $100", the 100-share bidder actually receives 20 shares, and the 900-share bidder receives 180 shares.

Some current interest centers on a different kind of time priority. The usual time priority is first-come-first-served, based on when the order was received. Limit orders have another attribute called *time in force* (TIF). If the order is not executed, how long is it considered valid? In the days of floor markets, the default was *good-till-cancelled* (GTC), and orders could sit on the book for many months. Nowadays, the usual convention is end-of-day. At the end of the day, all orders are cancelled, and the book starts the next day empty. But the trader can also submit an explicit TIF, like one minute, or ten seconds. As we'll see later, today's markets receive some limit orders that are cancelled almost immediately. These orders are sometimes considered problematic because their brief duration can confuse other traders. To discourage short-lived orders, some exchanges are considering giving priority to orders based on their commitment to a reasonably long time in force.

4.6. Strategy

Setting the price on a limit order

Despite the simplicity of the book's operation, a trader's actions follow from complex conjectures about how other traders might react. Consider the following book (presented vertically):

	Visibility	Price	Quantity	Submitted	Trader
		50.12	1,000	9:30	Cathy
		50.11	500	9:32	Bill
	Hidden	50.10	200	9:30	Gina
SELL		50.10	400	9:31	Amy
BUY		50.05	1,000	9:30	David
		50.04	500	9:32	Ellen
		50.03	400	9:31	Fred

Suppose that a new trader, Rui, wants to buy 100 shares. First, make or take? Rui can take the best offer (Amy's 50.10), or he can make a bid of his own, using a limit order that will be posted on the book. From Rui's perspective, all sellers in the market are potential counterparties, including those who have not yet entered any sort of order and are simply watching and waiting. All buyers are competitors. The appearance of a new limit order may trigger reactions from competitors and counterparties.

Suppose that Rui submits an order to buy 100, limit 50.06. This will become the market's new best bid. Counterparties (buyers) may react by hitting the bid. Amy might think, "I wasn't willing to sell at 50.05, but 50.06 is acceptable." She hits Rui's bid (selling 100 shares), and she will probably reduce the quantity (from 400 to 300) in her order priced at 50.10. Or Rui's bid might be hit by a seller who hadn't previously submitted anything. A higher bid may induce counterparty sellers to step forward.

Competitors may also react. David might reason, "my order is no longer at the top of the market. I'll reprice it to … 50.07." In formulating these strategies, Rui must weigh the reactions of counterparties and competitors. The analysis is simpler if the Rui's limit order is hidden. Participants can't react to what they can't see. So, no new counterparty sellers will come forth in response to Rui's higher bid, and no competitors (existing bidders) will have cause to reprice their orders.

The modern limit order market is very dynamic. New executions, additions to the book, and cancellations of existing orders may trigger reactions, which trigger further responses, and so forth. The chain of events might also be set off by the arrival of new information.

How would you interpret the following situation? Starting from the book depicted above, there suddenly appears a new large buy order (Stella's):

	Visibility	Price	Quantity	Submitted	Trader
		50.12	1,000	9:30	Cathy
		50.11	500	9:32	Bill
	Hidden	50.10	200	9:30	Gina
SELL		50.10	400	9:31	Amy
BUY		50.05	1,000	9:30	David
		50.04	500	9:32	Ellen
		50.03	400	9:31	Fred
		50.03	20,000	9:40	Stella

Given the prior state of the book, Stella's bid for 20,000 suggests a strong demand: 20,000 is several times as large as the sum of all other orders in the book. The other participants might reason, "Well, right now the 20,000 are priced well behind the current bid, but if that buyer wants to pick up 20,000 in this market, she'll be forced to bid higher." The sellers will contemplate repricing their offers higher. The other buyers might speculate that if she bids higher, she will execute against Amy, Bill, and Cathy. (Gina's order is hidden.) One of them (Ellen?) might wonder, "how much longer will I have the opportunity to buy at 50.10?" Given this additional pressure, some buyer might act. Perhaps Ellen will reprice her 500-share bid at 50.10. At this price, Ellen buys 400 shares from Amy and (in a pleasant surprise) 100 shares from Gina.

Spoofing and Layering

In the last example, things worked out well for Amy. She had been waiting to sell 400 at 50.10, and the arrival of the 20,000-share bid triggered a reaction by one of the other buyers. The next time Amy finds herself in a similar situation, she might think, "The new 20,000 share bid sure moved things in my favor. I was lucky." So far so good, but then, "Maybe I can make my own

'luck'." And Amy *herself* submits the 20,000-share order. To be perfectly clear, the visible book looks the same, since trader's identities aren't disclosed:

	Visibility	Price	Quantity	Submitted	Trader
		50.12	1,000	9:30	Cathy
		50.11	500	9:32	Bill
	Hidden	50.10	200	9:30	Gina
SELL		50.10	400	9:31	Amy
BUY		50.05	1,000	9:30	David
		50.04	500	9:32	Ellen
		50.03	400	9:31	Fred
		50.03	20,000	9:40	Amy

Ellen reacts as before. Amy's sell limit order executes, and she cancels her 20,000-share bid. Other traders are puzzled. "Wait a minute, wasn't someone just trying to buy 20,000 shares a moment ago?" Well, yes, but aren't market prices and quantities always changing? Their suspicions don't last long.

Spoofing is a manipulation that involves the submission of a bid or offer that is not intended to be executed. It is posted only for the purpose of stimulating a reaction by one or more other market participants. It is prohibited in the US under the Dodd-Frank rules. Investigations have led to civil and criminal charges. See, for example, the United States v. Navinder Sarao.[5]

Layering is a related practice. In "Amy's" case, her actions are more likely to be detected because they involve one large order at one price. Layering uses multiple orders submitted at multiple prices, possibly submitted and cancelled at slightly different times to give the appearance of multiple independent agents.

4.7. From floors to books

Trading procedures in modern limit order markets rely on automatic execution ("auto ex"): if an incoming order can be executed, it is executed, immediately. The brokers on most floor markets accepted electronic entry of orders, and computerized maintenance of the book. But they correctly saw that automatic execution threatened their survival, and sensibly resisted it.

In 1992 on the occasion of the New York Stock Exchange's Bicentennial, the Exchange's Chairman William H. Donaldson commented, "It's safe to say that the securities market of the future won't be an unmanned spacecraft, run by computers. It will more closely resemble a Starship Enterprise -- computerized beyond our wildest dreams -- but, in the American tradition, with the intelligence and the soul of human judgment on the bridge," (Widder, 1992).

Floor markets managed the transition to limit order markets with varying degrees of pain. Many non-US stock exchanges seemed to fare well. The Toronto and Tokyo exchanges, and the Paris Bourse readily adopted the electronic limit order books. The US futures exchanges built computerized systems that were initially used off-hours, when the floor was not operating. Although the Chicago Board of Trade's Aurora System did not work as planned, the Chicago Mercantile Exchange's Globex limit order market functioned well. The Board of Trade, the Merc, the New York Mercantile Exchange, and a few other floor-based US futures markets combined, and the Globex system is now used by all of them. Duncan MacKenzie provides an in-depth history of the Merc's transition (MacKenzie, 2013).

[5] A disambiguation: the term "spoofing" is used here with its accepted definition in the context of security market regulation. In networking, it may be used to refer to an illegitimate website that mimics a legitimate one.

The transition for US equity markets was more complicated, involving competing markets and a fair amount of regulatory pressure. At the New York Stock Exchange, fully automatic execution did not occur until the adoption of the SEC's Reg NMS in 2005. Presently, almost all trading at both the NYSE and NASDAQ occurs through their limit order books.

4.8. Further reading

Bessembinder, Panayides and Venkataraman (2009) discuss the usage of hidden orders. The limit order market is usually referred to by economists as a continuous double auction market. Economists usually attempt to characterize the equilibrium behavior of the market. Since the number of potential participants is large, and the set of strategies that any one agent might deploy is rich and varied, however, these models are difficult to solve, and even modest results are hard-won (Glosten, 1994; Goettler, Parlour and Rajan, 2005, 2009; Hollifield, Miller, Sandas and Slive, 2006; Parlour, 1998; Sandas, 2001). Physicists tend to prefer statistical models of order arrival (Bouchaud, Farmer and Lillo, 2008; Bouchaud, Mézard and Potters, 2002; Huang, Lehalle and Rosenbaum, 2015; Smith, Farmer, Gillemot and Krishnamurthy, 2003).

Summary of terms and concepts

Operations of a limit order book; priority (price, visibility, time); undisplayed/hidden orders; matching (execution) procedures; "top of the book"; "walking through the book"; order qualifiers (IOC, AON, FOK); market orders, marketable and non-marketable limit orders.

References

Bessembinder, Hendrik, Marios Panayides, and Kumar Venkataraman, 2009, Hidden liquidity: An analysis of order exposure strategies in electronic stock markets, *Journal of Financial Economics* 94, 361-383.

Bouchaud, Jean-Philippe, J. Doyne Farmer, and Fabrizio Lillo, 2008, How markets slowly digest changes in supply and demand, in Thorsten Hens, and Klaus Schenk-Hoppe, eds.: *Handbook of Financial Markets: Dynamics and Evolution* (Elsevier: Academic Press).

Bouchaud, Jean-Philippe, Marc Mézard, and Marc Potters, 2002, Statistical properties of stock order books: empirical results and models, *Quantitative Finance* 2, 251-256.

Glosten, Lawrence R., 1994, Is the electronic open limit order book inevitable?, *Journal of Finance* 49, 1127-61.

Goettler, Ronald L., Christine A. Parlour, and Uday Rajan, 2005, Equilibrium in a dynamic limit order market, *Journal of Finance* 60, 2149-2192.

Goettler, Ronald L., Christine A. Parlour, and Uday Rajan, 2009, Informed traders and limit order markets, *Journal of Financial Economics* 93, 67-87.

Hollifield, B., R. A. Miller, P. Sandas, and J. Slive, 2006, Estimating the gains from trade in limit-order markets, *Journal of Finance* 61, 2753-2804.

Huang, Weibing, Charles-Albert Lehalle, and Mathieu Rosenbaum, 2015, Simulating and Analyzing Order Book Data: The Queue-Reactive Model, *Journal of the American Statistical Association* 110, 107-122.

MacKenzie, Donald, 2013, Mechanizing the Merc: the Chicago Mercantile Exchange and the rise of high-frequency trading, School of Social and Political Science, University of Edinburgh, Available at: http://www.sps.ed.ac.uk/__data/assets/pdf_file/0006/93867/Merc21b.pdf.

Parlour, Christine A., 1998, Price Dynamics in Limit Order Markets, *The Review of Financial Studies* 11, 789-816.
Sandas, Patrik, 2001, Adverse selection and competitive market making: evidence from a pure limit order book, *Review of Financial Studies* 14, 705-734.
Smith, Eric, J. Doyne Farmer, László Gillemot, and Supriya Krishnamurthy, 2003, Statistical theory of the continuous double auction, *Quantitative Finance* 3, 481-514.
Widder, Pat, 1992, The Party Begins for NYSE, Ends For Dow Streak, Chicago Tribune.

Chapter 5. Multiple markets

A market can be organized so that by law or established custom, all trading in a security is *consolidated*, and occurs through a single exchange. Formerly, this meant that all trading occurred in a single physically convened market (one trading floor). Nowadays, "consolidated" usually means that trading happens in one central computer system usually a consolidated or centralized limit order book (CLOB).

Many present-day regulators, though, are reluctant to give one exchange a monopoly on trading. Allowing multiple exchanges results in a *fragmented* market. A fragmented market can simply result from having multiple limit order books. It may also involve alternative (non-limit-order) mechanisms, in which case the market is also considered to be a *hybrid*.

The US equity market presents an excellent example of a fragmented market.

5.1. US trading venues

Table 5.1 lists the largest venues and their trading volume (in million shares) on a typical recent day. Volume is reported separately for each listing venue. A listing for a US stock used to confer near-exclusive trading rights. If a stock were NYSE-listed, almost all of the trading would occur on the NYSE. Nowadays though, there are many places where a trade might occur. In view of the fact that not all of them are exchanges, these places are called *trading venues* or (in US regulation) *market centers* or *trading centers*. The trading venues differ by fee structure and trading protocols. These differences are important, and we will eventually discuss most of them. For the moment, though, it suffices to note that there are many venues, and they are very competitive. You might recognize some of the names of the trading venues in the table, but a few are probably unfamiliar.[1]

[1] Cboe was originally the Chicago Board Options Exchange. It is now the holding company for exchanges formerly operated by Edge and BATS. BX is the former Boston Exchange, now owned by Nasdaq; PSX was the Philadelphia Stock Exchange. National is the former Cincinnati

Table 5.1. Trading volume (million shares) on Friday, September 11, 2020

	Primary listing venue				
Trading venue	NYSE	Other	Nasdaq	Row totals	Row %
TRF	1,298	560	1,724	3,583	39.94%
Nasdaq	571	249	893	1,713	19.09%
NYSE	827	60	53	939	10.47%
NYSE Arca	205	313	297	815	9.08%
Cboe EDGX	243	117	246	606	6.75%
Cboe BZX	204	152	162	518	5.77%
Cboe EDGA	62	39	52	152	1.69%
IEX	83	15	54	152	1.69%
Cboe BYX	67	41	41	150	1.67%
NYSE National	68	35	28	130	1.45%
Nasdaq BX	30	19	24	73	0.81%
Nasdaq PSX	25	23	16	63	0.71%
NYSE American	12	26	8	47	0.52%
NYSE Chicago	9	11	11	30	0.34%
LTSE	0	0	0	1	0.01%
Col totals	3,704	1,659	3,608	8,971	
Col %	41.29%	18.50%	40.22%		

Notes: TRF=Trade Reporting Facility; LTSE=Long Term Stock Exchange. "Primary listing venue" is a heading of convenience: the columns correspond to CTA Tapes A, B, and C.

Source: IEX website (https://iextrading.com/apps/market/)

On the day sampled in the table (Friday, September 11, 2020), the total volume in NYSE-listed stocks (like IBM, Procter & Gamble, Ford, …) is 3,704 million shares. Yet only 827 million shares, about 22% were traded on the NYSE itself. Similarly, Nasdaq accounted for only 893/3,608 (25%) of trading volume in its own listed companies (like Microsoft, Apple, Amazon, …). For all listing venues the largest entry is "TRF". This refers to the Trade Reporting Facility, a channel for reporting trades that do not take place on an exchange. This category includes most retail trades and most so-called "dark trades" (Chapter 8).

5.2. Trading in fragmented markets.

Traders face great challenges in navigating a fragmented market. Different markets might have different prices. Who, at this moment, is posting the highest bid, or the lowest offer price? If the investor wants to make her own bid or offer (using a limit order), where and how will it be advertised? If someone sees her price, where can they send an order to trade against it?

Around 2005, the US adopted an overarching set of rules governing trading in US equity markets. Labeled "Regulation NMS" (for National Market System) and often simply referred to as "Reg NMS", it establishes a framework that ties the separate markets together in a fashion sometimes called "virtual consolidation" ((Hendershott and Jones, 2005), for example).

Exchange. LTSE (the Long Term Stock Exchange) is one of the newer entrants. IEX (formerly the Investors' Exchange) grew out of a trading unit of the Royal Bank of Canada.

In the language of Reg NMS, any place (or system) where a trade might occur is termed a *market center*. Although the various market centers often trade the same securities, they differentiate themselves according to trading rules and procedures (protocols) or by fees charged.

The centers are linked by *market information systems*, *access systems*, and *routing systems*. Market information systems communicate trade (last sale) reports, current quotes and other information from the market centers to users. Access and routing systems go in the other direction, transmitting users' orders to the market centers.

For New York and American Stock exchange-listed issues, the most important market information systems are the *Consolidated Trade System* (CTS) and the *Consolidated Quote System* (CQS). CTS consolidates reports of trades (wherever they occur); CQS consolidates and broadcasts each market center's best bid and offer (BBO). Both systems are operated by a consortium of the exchanges, the Consolidated Tape Association (https://www.ctaplan.com). Similar (but independent) systems exist for NASDAQ-listed securities. Vendors such as Bloomberg and Reuters purchase these feeds and redistribute the data to their customers. Of particular interest in the quote data are the highest bid at any given time, the *National Best Bid*, and the lowest ask price, the *National Best Offer*. The *National Best Bid and Offer* (NBBO) are important benchmarks for brokers and traders. The US Securities and Exchange Commission designates the operators of most market information systems as SIPs (Securities Information Processors; also see https://polygon.io/blog/understanding-the-sips/).

The market information systems are one-way; they do not provide the means for the investor to send in an order to execute against the NBBO. They are also broadcast systems. A message is not targeted to a specific recipient. They do not enable a market center to report the outcome of a received order back to its originator.

To accomplish these functions, investors rely on *access systems*. Access systems link brokers to market centers and link the market centers to each other. There is little or no consolidation of these systems. Instead, there's a collection of point-to-point communication systems that can convey executable orders and reports.

The systems that guide orders to the market centers where they are likely to be executed at favorable prices are called *routing systems*. Their intelligence comes from combining market information with situation-specific rules and practices. A customer order first arrives at the broker's routing system, which may send it to another broker or a market center or another based on where the stock is listed, who is showing the best bid and offer, and the market center's relationship with the broker. The receiving broker or market center may send it on to another, and in this fashion an order can make multiple hops before it arrives at its final destination. The routing is usually transparent (particularly to the retail customer), but it can take a bit of time. Customers can bypass the routing process by *directing* their orders to a particular destination. The downside is that the decisions made in the routing system often work to the customer's advantage, and the receiving market might not be able to execute the order.

5.3. The NBBO (National Best Bid and Offer)

The current NBBO is generally indicated prominently on our trading screens. It is computed by the Consolidated Quote System and broadly disseminated. In many situations, though, such as transaction cost analysis, we need to determine the NBBO at some precise time in the past. How is this computation performed?

Consolidated quote data consist of a series of time-stamped records each of which contains a bid, an offer, and an exchange identifier. The records usually contain other information as well: the size (number of shares) at the bid, the size of the offer, and various modifiers (condition codes), but the bid, offer and exchange are the most important fields.

An exchange (or similar quoting venue) enters a new record into the stream whenever some feature of its quote changes (or is cancelled). This new record replaces all information on the previous record, and it is presumed valid until the next update. The previous record is valid until the new record arrives.

To determine the NBBO at a given time, we need to determine which exchanges were quoting, and when the most recent update for each exchange occurred. To determine the set of exchanges, we usually need to work forward from the start of the day. To determine the most recent update for each exchange, we work backwards from our reference time.

Table 5.2 Example computation of the NBBO

Quote record					Computations						
					Bid				Offer		
Time	Bid	Offer	Exchange		A	B	C		A	B	C
9:31	70.00	70.10	A		70.00				70.10		
9:32	70.05	70.20	B			70.05				70.20	
9:33	69.90	70.15	C				69.90				70.15
9:34	70.00	70.15	B			70.00				70.15	
9:35					*70.00*	*70.00*	*69.90*		*70.10*	*70.15*	*70.15*
9.50	70.10	70.16	A		70.10				70.16		

Table 5.2 provides, on the left-hand side, a sample record of quotes. Suppose that we want to determine the NBBO at 9:35. First, we scan from the start of the day to determine which exchanges were actively quoting. Next, we set up a table for the bids and a table for the offers, with a heading corresponding to each exchange. Then, for each exchange, starting from 9:35 we scan backwards to determine each exchange's current bid and offer. These are given in the italicized row at 9:35. The NBB is the highest bid, equal to 70.00; the NBO is the lowest offer, equal to 70.10. Both A and B are at the best bid; A is alone at the offer.

Note that in this determination that we are determining the max and min (highest and lowest) across exchanges (horizontally) at a point in time. We are not determining the max and min across time (vertically). The highest value across time for the bid, for example, is 70.05. This is not the NBB as of 9:35, though, because the 70.05 originated from exchange B at 9:32, and B superseded this price at 9:34 with a bid of 70.00.

5.4. Market-wide priority practices

The priority rules in a single limit order book are straightforward: price, visibility, and time. But how do these priority rules play out when there are multiple exchanges – and multiple limit order books. How do priority rules play out among multiple markets?

Most importantly, there is no overall coordination mechanism that consolidates the individual books so that we effectively have a single book where price, time and visibility priorities are observed. The priority rules that hold in a single book do not prevail across markets.

The following situations can happen:

- Violation of price priority. A limit order to buy at a price 100 on exchange *A* might be executed even though there is, at the same time, a limit order to buy at a price of 101 on exchange *B*.
- Violation of visibility priority. An undisplayed order at a price of 100 might be executed on exchange *A* even though there are quantities visible at 100 on exchange *B*.

- Violation of time priority. A limit order to buy at a price of 100 that was entered at 10:00 AM on exchange *A* might be filled before an order to buy at 100 that was entered at 9:30 AM on Exchange *B*.

In a single book, knowing that our order will always get executed before other orders that are priced less aggressively, entered subsequently, or not displayed encourages us to promptly post visible and aggressively priced limit orders. If these priority principles are violated cross-market our incentives are reduced.

The violation of price priority described above is called a *trade-through*. The person bidding at 101 on exchange *B* is traded-through when there's an execution at 100 on exchange A. The seller on exchange *A* is disadvantaged because she sold at 100 when she could have sold at 101. The buyer on exchange *B* is disadvantaged because he is deprived of an execution on terms that would have been acceptable to him. Only the buyer on exchange *A* is better off.

When all trading occurs face-to-face on a floor market, the occurrence of a trade-through, and the identities of the parties to it are usually clearly evident. Suppose that Alice bids 100, Brian bids 101 and Cathy sells to Alice at 100. Brian can observe and protest. One remedy is as follows: Cathy's sale to Alice stands, but Cathy also "owes Brian a fill," that is, Cathy must find shares to sell to Brian at 101.

5.5. Order protection under Reg NMS

The self-policing of trade-throughs that arose naturally in most floor markets did not survive the transition to dispersed electronic markets. Network latencies (delays) coupled with the rapid pace of trading make it difficult to determine exactly what the bids and offers actually were at the precise instant that a trade occurred.

Trade-through protection is nevertheless generally thought to be such a desirable feature of a market that some attempt to preserve and enforce it is warranted. The most important current rule in this respect is the SEC's Regulation National Market System ("Reg NMS"). We'll defer a full discussion of this rule to Chapter 20, but right now we'll examine the *order protection rule* (also called the "trade-through rule") component of the regulation.

The Reg NMS order protection rule defines a class of protected bids and offers, practices that guard against trade-throughs of these orders, and assigns responsibility to market centers. The rule is sometimes described as prohibiting trade-throughs. This would imply, though, that if we observed one, someone would have violated the rule. This is not the case. Trade-throughs continue to occur, for some very good reasons, as we shall see. The rule does not implement "trade-through prohibition," but we believe that with the rule we have fewer of them than we'd experience without a rule.

To be protected, an order must be visible (not hidden), at the market center's best bid or offer (not necessarily the National Best Bid and Offer), and accessible (available for automatic execution). Orders priced away from the market's BBO and hidden orders (even if they are superior to the BBO) are not protected.

But even for protected quotes, trade-throughs are not prohibited. The rule is worded so that markets shall "establish, maintain, and enforce written policies and procedures reasonably designed to prevent" trade-throughs. Instead of a prohibition, the rule calls for a good-faith attempt to avoid them. In practice, this means that a market must route quantities sufficient to avoid trade-throughs, based on what it can see.

As an illustration, Table 5.3 gives the bid books in two markets. Order submission times aren't relevant in this discussion, so we've omitted them.

Table 5.3 Two-market example

	Market A			Market B		
Bid	Trader	Quantity	Visibility	Trader	Quantity	Visibility
50.40	Amy	200	Hidden			
50.39	Alan	300	Display			
50.38	Ava	500	Display	Beth	100	Display
	Anna	200	Hidden	Ben	200	Display
50.37	Arnold	300	Display	Beverly	300	Display
	Alma	100	Display	Belle	100	Display

The top of A's visible book is Alan's bid of 50.39. This is protected. (Amy's bid is higher, but it is hidden, and therefore not protected.) The top of Market B's visible book is 50.38. Both Beth's and Ben's orders (a total of 300 shares) are protected. Assuming that these are the only two markets, the NBB is 50.39. Note that Beth's and Ben's orders are protected even though they aren't priced at the NBB.

In the example, suppose Dana sends to Market B an order to "sell 100, limit 50.37." An execution against B's bid book (against Beth's order at $50.38) would trade through the protected bid on market A. So, B won't allow this to happen.

Beyond attempting to avoid the trade-through, though, Reg NMS imposes no further requirements. Depending on Dana's instructions (and Market B's default practices), Dana's order might be routed to Market A for execution or cancelled. If Dana has the full picture of the market, of course, she (or her broker) will send her order to market A, where she knows that she can get a fill no worse than $50.39. But in a rapidly moving market, she (and her broker) might not be aware of this opportunity.

Executions against unprotected orders can sometimes lead to apparent violations of the rule. This can sometimes be confusing. Suppose that Market B receives an order to sell 600 shares limit 50.37. Any execution against its own book would cause an execution at 50.38. This would trade through A's protected bid at 50.39 (Alan's order), and so would be forbidden. In order to execute even part of the order, B would have to send (route) some portion of it to A. But how much?

In this situation, the rule requires B to route (to A) a quantity equal to the number of A's protected bid, that is, 300 shares. Assume that B does this and executes the remainder against its own book. The resulting outcome is:

- On exchange A, 200 shares execute at 50.40 (against Amy's hidden order), and 100 shares execute at 50.39 (against Alan). Alan has 200 shares remaining to buy, limit 50.39.
- On exchange B, 100 shares execute against Beth (at 50.38), and 200 shares execute against Ben (also at 50.38).

Now this looks like a trade-through. Alan's entire order was considered a protected bid, and since 200 shares remain unexecuted, these too are protected. Didn't B's executions at 50.38 trade through Alan's remaining protected shares? Technically, B's executions might well be considered trade-throughs. The order protection rule, however, only requires market B to make a good faith effort to avoid a trade-through. In this situation, it is considered sufficient for B to attempt to execute against A's protected orders. This is what B has done. From B's perspective,

market A could have executed the 300 shares routed to it against Alan's entire order. From market A's perspective, of course, this is not possible: A must satisfy the hidden bid at a better price. But this is not market B's problem or responsibility.

Here's a related situation. Suppose that B receives an order to sell 1,000 shares limit 50.37. It could, as indicated above, route 300 shares to A, and execute the remaining 700 against its own book. But suppose instead that in a burst of generosity B sends 500 shares to A and executes 500 shares against its own book. The resulting executions would be:

- In market A, 200 shares trade at 50.40 (against Amy), and 300 shares trade at 50.39 (against Alan). The best bid remaining on A's book is 50.38 bid for 500 shares (Ava).
- In market B, 100 shares trade at 50.38 (against Beth), 200 shares at 50.38 (against Ben), and 200 shares at 50.37 (against Beverly).

This is all completely consistent with the rules. Now it could be argued that after A's executions, Ava's bid has become the best, and is therefore considered protected. From this perspective, B's execution at 50.37 is a trade-through. But think for a moment. In a series of executions (like the one on A) that walks through the book, each execution is occurring at a price that is, at the moment of execution, "the best". If this is considered sufficient for protection, then we're really extending protection not just to orders at the top of the book, but to all orders in the book. This is beyond the intent of the rule.

One way of rationalizing the situation is to group the events in two steps. When a market receives an order, it executes the order against its book, possibly against hidden orders, possibly at multiple prices. Then, after any executions have occurred, the market posts its best bid and offer based on its remaining visible orders. Only at this point are these bids and offers considered to be protected. They are not considered protected while the market is processing the original order.

Under the rules, the primary responsibility for avoiding trade-throughs falls on the market centers, who must constantly monitor each other's protected quotes. All this monitoring and checking, of course, can slow down the processing of an order, often to the submitter's disadvantage. For this reason, Reg NMS allows an alternative procedure whereby the order submitter can perform the cross-market check and avoid trade-throughs by routing orders to multiple markets simultaneously. The use of such *intermarket sweep orders* is discussed in Section 20.2.

5.6. Summary

When there are multiple market centers, trading strategies become more complex. The basic make-or-take decision faces the additional question of "where" (the routing problem). Priorities and procedures that are clear in the context of a single market become much less so with multiple destinations. In US equity markets, the trade-through protection afforded under Reg NMS is a partial remedy, but this does not apply to other markets and other securities.

Is this complexity really necessary? Would trading be simpler if all activity was forced into a single well-managed system? The benefits of simplicity are indeed appealing, but there is a cost. The single system would be a monopoly, an entity protected and privileged under law. Experience with near monopolies in trading suggests that they are costly, inefficient and reluctant to innovate. Chapter 20 provides further discussion.

Summary of terms and concepts

Fragmentation and consolidation; types and purposes of linkage systems (market data/information, access, routing); direct (market) access; the Consolidated Trade System; the Consolidated Quote System; the National Best Bid and Offer (NBBO); trade-through; trade-through prevention provisions of Reg NMS.

References

Hendershott, Terrence, and Charles M. Jones, 2005, Island goes dark: transparency, fragmentation and regulation, *Review of Financial Studies* 18, 743-793.

Part II. Alternatives to Limit Order Markets

The electronic limit order book is probably the most widely used trading mechanism. There are nevertheless many securities and settings where it is augmented, or even displaced, by other mechanisms. Here, we look at auctions, dealers, and dark mechanisms: why they are needed and how they work.

Chapter 6. Auctions

Notes, updates and current developments

- The SEC has proposed high-frequency auctions of retail orders (U.S. Securities and Exchange Commission, 2022)

Auctions are widely used in markets for diverse goods and services. The more familiar auctions involve wine, art, and collectibles. In an art auction, buyers compete with each other, bidding higher prices until only one buyer remains. This person buys the item at his last bid price. Christie's and Sotheby's conduct art auctions in which most of the buyers are physically present. Their bidding is coordinated by an auctioneer, who identifies the highest bid (and the person who made it), suggests higher prices, and generally tries to maintain interest, excitement and drama.

Most art auctions are open outcry (also called English) auctions: bidders can hear the bids of others. In other versions, sealed bids are used: bids are collected in sealed envelopes, one bid per buyer; the envelopes are opened; and the highest bid is selected. The art auction is an ascending auction: a bid must be higher than the one that preceded it. Alternatively, in a descending auction, the price is initially set very high, and this price is reduced until one buyer claims it by calling out "mine". The art auction is a seller's auction (one seller and many potential buyers). Many public agencies use buyer's auctions, where suppliers compete by offering lower prices. Klemperer (2004) discusses the alternative formats and their economic principles.

Although auctions are generally viewed as reliable, robust, and fair, they are subject to disruption by manipulative strategies. A seller might use a shill to pose as an aggressive buyer, driving up the bids of others; buyers can collude (agree beforehand) not to bid against each other; and so forth.

Securities auctions differ from art and wine auctions in many ways, but many issues of design and concerns about manipulation are common to both.

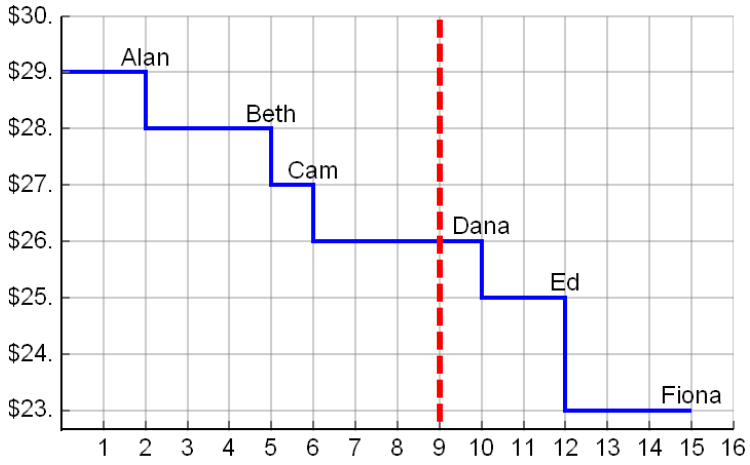
6.1. Primary market auctions

The primary market refers to the process by which a security is initially sold ("placed") by the issuer. It typically involves many identical items (bonds or shares of stock), one seller (the issuer), many potentially buyers, and the transaction is usually accomplished all at once.[1] An auction would seem to be a sensible format.

A bid specifies price and quantity, such as, "$50 per share, for 100 shares." The bids are ranked in price priority (highest bids have highest priority). This defines a demand function that resembles a staircase, descending to the right. For these orders:

Trader	Quantity	Price
Alan	2	$29
Beth	3	$28
Cam	1	$27
Dana	4	$26
Ed	2	$25
Fiona	3	$23

The corresponding demand function is:



The supply is however many units (bonds, shares, whatever) are available for sale. For this sort of auction, the supply does not depend on price, so it is graphed as a vertical line. The figure indicates a supply of nine units.

The supply and demand functions cross at a price of $26. All buyers who bid above this price (Alan, Beth, and Cam) have their orders satisfied in full. Dana gets a partial fill: she gets three of the four units that they wanted.

The US Treasury uses an auction of this type to issue bills, notes, and bonds. The procedures are clearly documented at https://www.treasurydirect.gov/instit/auctfund/work/work.htm. Prior to the auction, the Treasury publishes the specifics of the security being offered and the timing of the auction. The bids are not specified as "dollars per bond".

---

[1] In contrast, most limit order markets, of the sort described in Chapter 3, are *secondary markets*. That is, they involve trade between a buyer and seller, neither of whom is the issuer of the security.

Investors in fixed-income markets usually focus on yield (percent return to the buyer), and all bids are in terms of yield. Bond prices and yields are inversely related, so a low yield corresponds to a high price. Intuitively, someone bidding a low yield is willing to loan the US Treasury dollars at a low rate of interest. Bids may be competitive or non-competitive. Competitive bids specify a desired quantity (in terms of par, or face, value) and a yield. Noncompetitive bids are accepted up to $10M and do not specify a yield. (The limit was raised from $5M in July 2022.) Thus, competitive bids are like limit orders, and noncompetitive bids are like market orders.

At the deadline for bid submission, the noncompetitive bidders are allocated their quantities. Then the competitive bids are ranked and considered from low yields to high, and quantities are accepted until the size of the issue is reached. The last competitive bid accepted determines the yield that will be set on all accepted quantities (noncompetitive and competitive). The auctions are accessible by retail and institutional investors alike, and results are published on the Treasury web site.

The initial sales of most US municipal bonds also use auctions. The Grant Street Group (www.GrantStreet.com) operates several systems. The auctions are open to banks, who then resell the acquired bonds to their end-customers.

Auctions have also been used for equity initial public offerings. From 1999 to 2013, Hambrecht and Quist conducted stock IPOs on its auction system OpenIPO. The largest placement was Interactive Brokers: IBKR, $1.3B in 2007. In 2004 Google (later renamed Alphabet) raised $1.7B in a widely publicized auction (Choo, 2005). Freedman (2021) notes:

> "Google had hoped to sell 25.9 million shares at a price between $108 and $135, but once the bidding process got underway, it ended up selling 19.6 million shares at $85 each [the $1.7B]. 'Obviously, what the founders thought the company was worth didn't exactly match what the public was willing to pay,' Bob Pisani said in a CNBC report."

So despite the broad use of auctions in the issuance of bonds, Jagannathan, Jirnyi and Sherman (2015) comment, "while a number of countries have tried the use of sealed bid share auctions for initial public offerings (IPOs), few continue to use them."

The attraction of auctions for equity IPOs is sufficiently strong (or the memory of past disappointments is sufficiently weak), however, that IPO auctions have returned. They play a notable role in the current capital-raising landscape. We return to these developments below.

6.2. Opening and closing auctions

Regular trading hours for US stocks are 9:30 to 16:00 Eastern Time. There are certainly many ways to trade outside of this period, but even with these opportunities, many investors prefer the regular session.

For a security that normally exhibits low trading activity, it may not be necessary to have a special opening procedure. There may be few orders entered prior to 9:30, and no opportunities for matches until later in the day. It is also possible that as we approach 16:00, the book has few orders and a wide bid-ask spread, with no strong interest on the buy or sell side of the market. Often, though, there is strong trading interest at the beginning and end of the session, with many buyers and sellers who want to trade at the open or at the close. There are sensible reasons for this.

At the open, there are often numerous and large orders that have accumulated overnight. This is particularly true when there has been an overnight or pre-opening news announcement.

Trading interest at the close tends to be even stronger:

- Redemptions and purchases of mutual fund shares are based on net asset values, which generally represent closing prices. Index funds also need to ensure that their composition closely matches that of the index at the close.
- Index futures contracts and other cash-settled derivatives are usually settled at closing index prices (although some use opening prices). Index arbitrageurs seeking to unwind their stock positions will usually need to sell or cover their shorts at closing prices.
- A wide range of index derivative products, such as leveraged and inverse ETFs seek to deliver some multiple of the close-to-close index return.

The need for special opening or closing procedures might not be obvious. Why don't we simply turn on the limit order book at 9:30 and turn it off at 16:00? If we did this, we'd expect to see enormous order flow at these times. Aside from the strains that this might put on the networks, there might be hundreds of trades at different prices occurring nearly simultaneously. This volatility goes contrary to the outcomes we expect from a fair and orderly market.

To ensure that everyone receives the same price, the mechanism must consolidate the total buying and selling interest. The auction used in this context is a *double-sided* auction.

The single-price double auction (SPDA) is simple, in principle. Buyers enter their demands; sellers enter their offers. The system computes the supply and demand curves in real time, as the orders arrive and change. The state of the market can also be communicated back to the buyers and sellers: either the full supply and demand curve, or (more likely) an indication of the clearing price. When everyone is through entering and modifying their orders, the system finds the price and quantity at which supply and demand are equal.

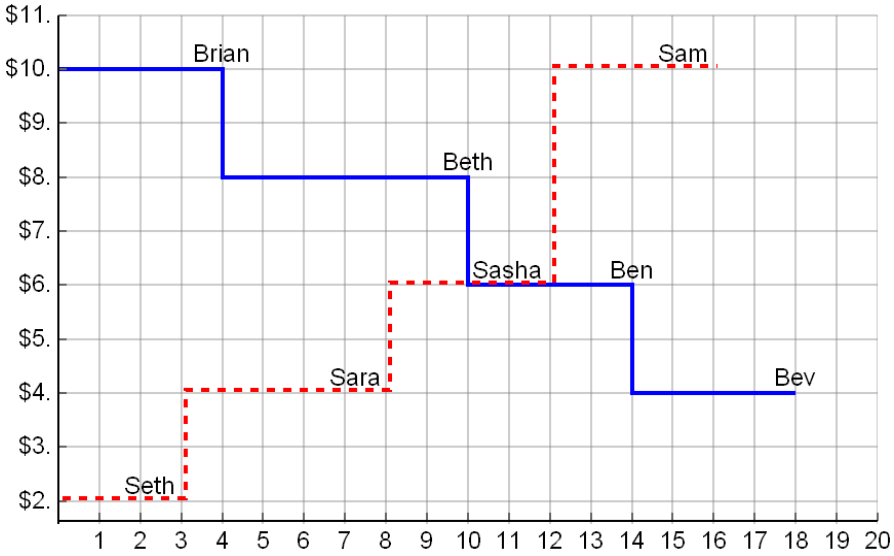
We'll now look a bit closer at this process. First there is a period of order accumulation. Buyers and sellers enter limit (priced orders). Some markets also accept orders without prices, designated as market-on-open (MOO) and market-on-close (MOC) orders.

Buy orders are ranked by price, high to low (essentially willingness to pay). If the rules permit market orders, they are effectively assigned an infinite price. The ranked orders are placed on the horizontal (quantity) axis and cumulated. This defines the demand curve (in reality, a step function).

For example:

Buyers					Sellers			
Buyer	Bid	Quantity	Cumulative Demand	…	Seller	Quantity	Asking	Cumulative Supply
Brian	$10	4	4		Seth	3	$2	3
Beth	$8	6	10		Sara	5	$4	8
Ben	$6	4	14		Sasha	4	$6	12
Bev	$4	4	18		Sam	4	$10	16

These are graphed as:

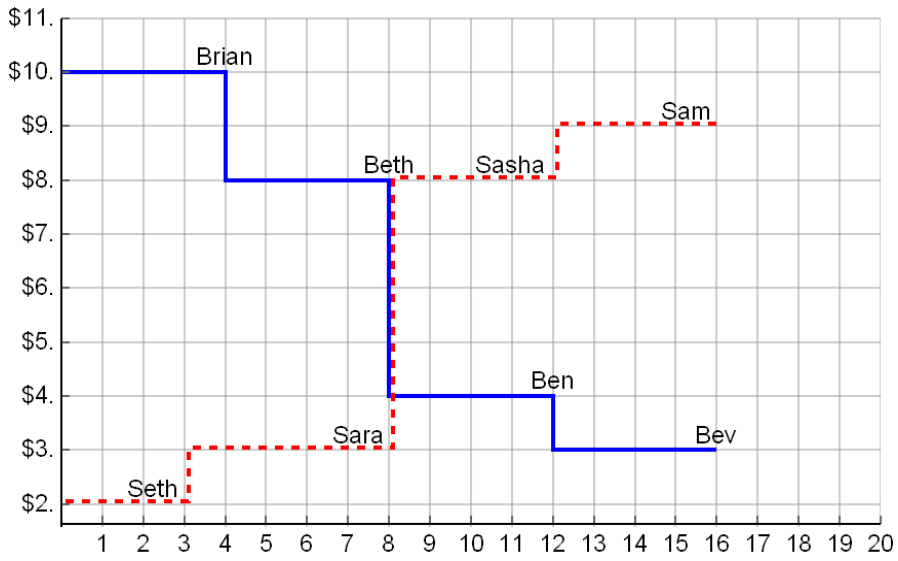


To clear the market, the supply and demand functions are combined, and we search for the point at which the trading volume is maximized: the point that lies farthest to the right, in the set of all intersection points. As usually drawn in introductory economics class, the demand and supply are smoothly sloping curves, and the intersection consists of a single point. With discrete prices and quantities, though, these curves are actually staircase (step) functions, and determination of the clearing price can be a bit more involved.

To start with, at any price, we can read off the supply and demand. The smaller of these quantities defines the *matched volume*. For example, at a price of $9, demand is 4 units (Brian), supply is 12 units (Seth, Sara, and Sasha), and the matched volume is 4 units. The unmatched remainder is the imbalance. The imbalance is signed (as buy or sell). At a price of $9, the imbalance is 8 units, on the sell side. At $6, *demand*=14, *supply*=12, *matched volume*=12, there's a buy *imbalance*=2.

To determine the clearing price, we first find the price that maximizes matched volume. (An exchange, after all, is in the business of trading, and higher volume corresponds to higher revenues.) In this example, that price is $6, for which the matched volume is 12. Immediately above $6, the matched volume drops to 10 units (as Ben drops out). Immediately below $6, matched volume drops to 8 units (as Sasha drops out). In this case, maximization of matched volume defines a unique clearing price.

This need not always be the case, however. Consider:



Here, all prices between $3 and $8 have the same matched volume – 8 units. So, to handle these cases, we need a supplementary rule, which is: minimize the net imbalance.

What if even after applying this supplementary rule, we are still left with multiple possible clearing prices? At this point, there is some variation in the rules. One common rule is to minimize the distance from some reference price. In the case of an opening auction, for example, we might pick the price closest to yesterday's closing price. Minimizing the price change from a previous close is consistent with most investors' preferences for lower price volatility.

Although simple in principle, real-world implementations face certain complexities, to which we turn next.

6.3. Clearing time

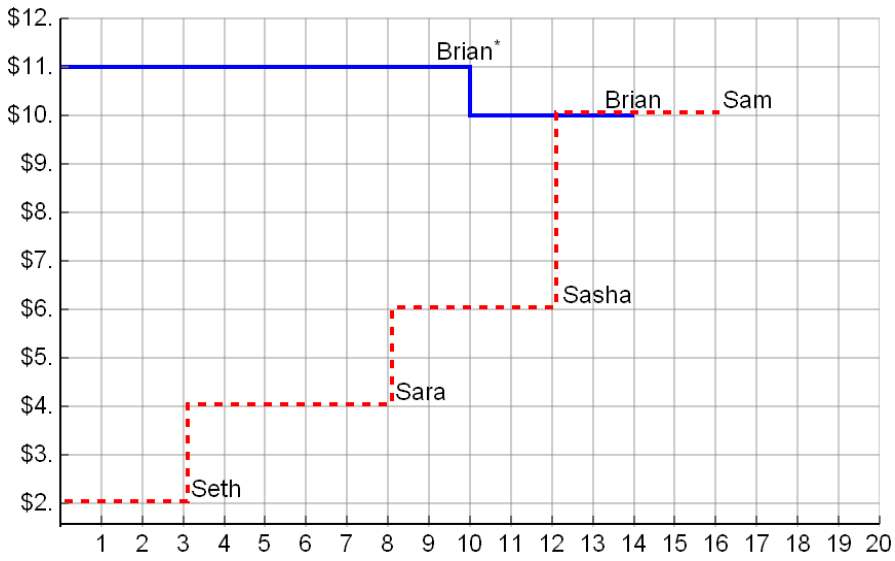
An open-outcry art auction usually doesn't have a fixed stopping time. The bidding simply continues until only one bidder is left. The auctioneer calls out, "Going once, …, going twice, …, going three times, … SOLD!" unless, of course, someone jumps back in with a higher bid.

Online auctions, though, and most opening auctions have a scheduled ending deadline, a fixed time, announced in advance. In many situations, though, a hard deadline simply induces traders to hold off entering their orders until shortly before the deadline.

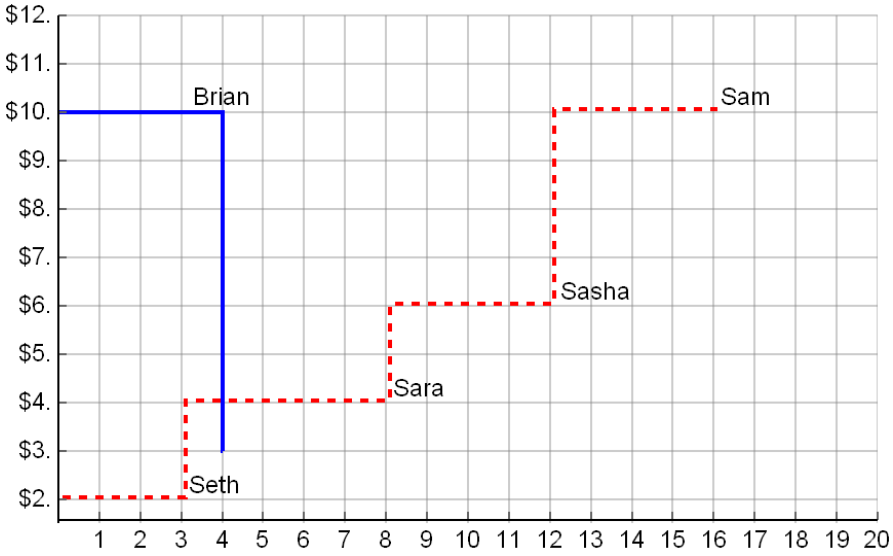
To deal with the last-minute pile-up, online auctions often extend deadlines until a period of time (e.g., ten minutes) has elapsed with no changes in the bid. In an opening security auction, for example, we can imagine a similar procedure: waiting until the orders rest without modification for a decent interval before clearing the market.

In a securities market, though, information, prices, and traders' desires are evolving constantly. On Friday, May 18, 2012, the public offering of Facebook stock was conducted, and the stock began continuous trading. The New York Times reported, "As NASDAQ's systems were setting Facebook's opening price, a wave of order modifications forced the exchange's computers into a loop of constant recalculations. The firm was forced to switch to another system, knocking out some orders and delaying many trade confirmations."

A hard deadline may also make the market susceptible to manipulation. In the main example of this chapter, Brian wanted to buy 4 units, limit $10. Suppose that in addition to this bid, Brian makes a fictitious and deceptive bid of $11 for 10 units. With Brian's two bids, the supply and demand functions look like this:



What happened to the other bidders? Beth, Ben and Bev simply got discouraged. Believing that Brian's aggressive bids were likely to determine a final price of $10 or $11, well above their own limits, they did not bother to bid. Then, one instant before the deadline, Brian cancels his $11 order, leaving the market as:



Brian buys his shares at a lower cost (that is, $4, instead of the $6 he would have paid earlier).

In online auctions, this is known as bid-shielding. When bidding opens on the sterling silver tea set, a bidder enters a price of $10, and then (via an accomplice or a separate online identity) bids $10,000. As the $10,000 establishes the high bid, no lower bids will be accepted. Immediately before the deadline, the $10,000 bid is cancelled, leaving the $10 bid the winner.

To discourage bid shielding and similar tactics, a market may randomize the clearing time. Even a small amount of uncertainty may be enough to discourage manipulations based on last-instant moves. There is some danger that the fictitious, manipulative bid (Brian's $11 bid, for example) might actually be executed, with Brian paying a higher price for much more than he really wanted.

The London Stock Exchange describes their randomization procedure with a particularly thoughtful discussion (London Stock Exchange, 2000). Other markets that employ randomized

clearing times include: the Tel Aviv Stock Exchange (Hauser, Kamara and Shurki, 2012; Tel Aviv Stock Exchange, 2013); and the Euronext markets.

In the US, however, exchanges don't generally use random deadlines. Instead, they attempt to stabilize the price determination process by restricting the information available to auction participants and by restricting order entry and modification immediately prior to the clearing.

In the NASDAQ opening auction ("cross"), the open is timed for 9:30am. From roughly 7:00am onwards, traders can enter orders. These orders may be marked "on open" to ensure that they are held unexecuted until the open. On-open orders must be received before 9:28 and may not be cancelled. Beginning at 9:28, NASDAQ transmits matched volume and imbalance information. Between 9:28 and 9:30, NASDAQ accepts a special kind of order called *imbalance-only*. Imbalance-only orders are only in the direction of minimizing an existing imbalance. Similar procedures are used in the closing auction. Hu and Murphy (2020) compare the NASDAQ and NYSE auction procedures.

6.4. "Marking the close" / "Banging the close"

The closing price of a security is very important. It is used as a reference price for determining mutual funds' net asset values (NAVs), the prices at which new shares are created and old shares are redeemed. It may be a reference price for cash-settled derivative contracts. It may be used to determine the acquisition price in a corporate takeover. Finally, margin calculations are often performed on the basis of closing prices, to determine whether the owner of a margined position has to put up additional cash.

In the usual art or collectible auction, anyone who deliberately overbid would simply be driving up the price he paid, with no obvious benefit. (If I pay $1 Million for painting when the next highest bid is $100,000, can I claim that I own a painting worth $1 Million?)

In the closing auction for a security, though, I may not care very much about overpaying if I have a much larger interest in some related transaction. In less-actively traded securities, for example, there might not be very many orders going into the closing auction and entering one more large buy or sell order might move the closing price by a large amount.

The litigation files of the Securities and Exchange Commission contain numerous instances in which an investor, facing a margin call that would require him to put up cash that he didn't have, placed orders in the final moments of trading to affect the closing price. This is called "marking" or "banging" the close. It is illegal.

6.5. Auctions as an alternative to continuous trading

The opening and closing auctions are used to transition between periods of continuous trading and market closure (or at least off-hours trading). We might wonder if the two auctions would suffice to satisfy all trading demands, dispensing with continuous trading altogether. The Euronext markets (Paris, Brussels, Amsterdam, and others) run call auctions twice a day ("double fixings") in illiquid securities at 11:30am and 4:30pm). These securities do not trade in the continuous market. (At certain times, however, Trading at Last (TAL), that is, at the last auction price is permitted.)

Could the Euronext arrangement be made more flexible? Two auctions per day might be enough for some companies, but others' shareholders might prefer three, four, or more daily auctions. Some authors believe that such arrangements would be superior to continuous limit order books because auctions aggregate multiple orders at a single time and price. The deeper participation results in more stable prices (see (Schwartz, 2001) and the papers by others collected in the same volume).

There are nevertheless considerations that favor continuous trading:

- News arrives continuously. Speculators want to trade on news.

- Hedgers often take a position in a security to offset risk. When they can't trade, the risk must be borne.
- The management of a listed company might like the visibility that comes with belonging to a recognized index. Index membership is usually restricted to stocks that trade continuously.

Clearly, for some traders, auctions that run twice or even ten times a day won't suffice. Suppose, though, that we ran auctions once per minute, say, or even every five seconds. Would this be equivalent to a continuous market? As an approximation, would it be close enough?

This kind of trading mechanism is described as frequent batch auctions (Budish, Cramton and Shim, 2015, BCS). The high frequency weakens one advantage that is claimed for the more traditional auctions, namely that they aggregate trading interest, pulling all potential traders together at one time. (Twice-daily auctions might have hundreds of participants on both sides; a sequence of auctions every five seconds is likely to have many auctions that involve a small number of participants.) BCS make an alternative case, that the continuous limit order market places too much weight on time priority, giving an advantage to high-frequency traders whose orders arrive first. The race for first place leads to overinvestment in technology (in the many millions of dollars) to achieve time improvements (in the milli- or microseconds). These improvements are economically insignificant, and the expenditures are socially wasteful. BCS suggest that auctions might help because during the period of order entry all orders are treated the same, irrespective of arrival time or sequence.[2]

6.6. Auctions for equity IPOs, revisited

Section 6.1 concluded on the generalization that single-price auctions are used in the issuance of US Treasury securities but not corporate stock. Recent developments in stock issuance, though, suggest that auctions are nevertheless important. We examine two current practices.

To set the stage, we first summarize the usual arrangement, an underwritten offering (see, for example, (Bodie, Kane and Marcus, 2020, Chapter 3) or https://en.wikipedia.org/wiki/Initial_public_offering). In this sort of offering, a corporation seeking to raise capital contracts with a group (syndicate) of investment banks that conducts roadshows for potential buyers. In the process, the syndicate learns about the demand for and perceived value of the offering. On the day of the initial sale, the firm and the syndicate set a single offering price for the shares, decide on allocations (which customers? how much to each?), and complete the sales. After the initial sale, attention turns to the secondary market, where the original buyers can sell their shares and new buyers can purchase shares.

The first alternative is a direct listing. Prior to the IPO the firm's owners consist of the early investors, venture capital investors, and employees granted ownership through stock options. The important point here is that the stock already exists. In preparation for the offering the firm registers the stock (with the Securities and Exchange Commission, in the US), releases a prospectus (a legal document that fully describes the firm and its securities), and picks an exchange on which to list. On the day of the IPO, the pre-existing shares can be traded. The listing exchange generally runs an opening auction, and the stock moves into a regular continuous trading phase. Note that unlike the underwritten offering, no new capital is raised. The listing

[2] Related papers include: (Du and Zhu, 2014, 2017; Schwartz and Wu, 2013); (Aquilina, Budish and O'Neill, 2020). There is some practical experience with batch auctions. For a long time they were used as the principal mechanism at the Taiwan Stock Exchange (Lee, Liu, Roll and Subrahmanyam, 2004; Liu, 2016). In March 2020, Taiwan transitioned to a continuous limit order book (a move contrary to the BCS recommendation). Indriawan, Pascual and Shkilko (2021) find that the transition was accompanied by higher trading costs.

simply provides a place for the initial investors to sell their shares. Perhaps the best-known early example is Spotify's 2018 direct listing on the NYSE. The opening auction on the first day of trading is important because it is the first observable public market price. In a sense it substitutes for an IPO auction in which the issue price is determined.

The second alternative is a hybrid auction. In this mechanism, potential buyers enter bids, just as in, for example, the US Treasury auctions. When all bids have been received the seller decodes on the issue price. This price, though, does not have to be the supply-demand crossing. Moreover, the allocations do not have to be in strict price priority. As in an underwritten offering the issuer has discretion in the allocations. Unlike the Treasury auctions, if a buyer bids $12 for 1,000 shares and the issue price is $10, she may not be awarded her full quantity (1,000). This mechanism does not maximize the issuer's proceeds from the sale of the stock, but in the allocation the issuer may favor buyers who appear more likely to hold the stock as long-term investors. This may, in turn, keep the stock out of the hands of short-term traders intent on flipping the stock for a quick profit, which might increase the volatility of the stock price. Robin-Hood used a hybrid auction IPO in 2021 (Driebusch, 2021).

6.7. Further reading

The economic analysis of auctions is well-developed. (Klemperer, 2002, 2004) are excellent introductory sources. Nyborg and Sundaresan (1996) and Das and Sundaram (1996) discuss uniform price vs discriminating auctions in the US Treasury market. Chernov, Gorbenko and Makarov (2013) and Gupta and Sundaram (2011) analyze auctions in CDS settlements. A number of empirical papers analyze opening auctions in stock markets (see Amihud and Mendelson (1991), and references therein). Bogousslavsky and Muravyev (2020) examine current issues with closing prices. This chapter has examined auctions as a supplement to limit order books. Hendershott and Madhavan (2015) discuss their supplemental use in dealer markets.

Summary of terms and concepts

Single-price call auctions, and how they're run; manipulation in auctions; randomization; Nasdaq opening procedures (reference price, matched volume, imbalance); imbalance-only orders; direct listing; hybrid auctions; randomization; stabilization.

References

Amihud, Yakov, and Haim Mendelson, 1991, Volatility, efficiency, and trading: Evidence from the Japanese stock market, *Journal of Finance* 46, 1765-1789.

Aquilina, Matteo, Eric B Budish, and Peter O'Neill, 2020, Quantifying the high-frequency trading "arms race": A simple new methodology and estimates, *Chicago Booth Research Paper*.

Bodie, Zvi, Alex Kane, and Alan J. Marcus, 2020. *Investments, 12th edition* (McGraw Hill, New York).

Bogousslavsky, Vincent, and Dmitri Muravyev, 2020, Should we use closing prices? Institutional price pressure at the close, Available at: https://ssrn.com/abstract=3485840.

Budish, Eric B., Peter Cramton, and John J. Shim, 2015, The high-frequency arms race: frequent batch auctions as a market design response, *Quarterly Journal of Economics* 130, 1547-1621.

Chernov, Mikhail, Alexander S. Gorbenko, and Igor Makarov, 2013, CDS Auctions, *Review of Financial Studies* 26, 768-805.

Choo, Eugene, 2005, Going Dutch: The Google IPO, *Berkeley Technology Law Journal* 20, 405-441.
Das, Sanjiv Ranjan, and Rangarajan K. Sundaram, 1996, Auction theory: a survey with applications to Treasury markets, *Financial Markets, Institutions and Instruments* 5, 1-36.
Driebusch, Corrie, 2021, Robinhood stock sale soured by investor confusion, valuation concern, July 30, 2021, Wall Street Journal.
Du, Songzi, and Haoxiang Zhu, 2014, Welfare and Optimal Trading Frequency in Dynamic Double Auctions, NBER, Available at.
Du, Songzi, and Haoxiang Zhu, 2017, What is the optimal trading frequency in financial markets?, *The Review of Economic Studies* 84, 1606-1651.
Freedman, Robert, 2021, More companies using hybrid auction in lieu of traditional IPO, CFODive.
Gupta, Sudip, and Rangarajan K. Sundaram, 2011, CDS credit event auctions, Stern School of Business, New York University, Available at.
Hauser, Shmuel, Avraham Kamara, and Itzik Shurki, 2012, The effects of randomizing the opening time on the performance of a stock market under stress, *Journal of Financial Markets* 15, 392-415.
Hendershott, Terrence, and Ananth Madhavan, 2015, Click or Call? Auction versus Search in the Over-the-Counter Market, *The Journal of Finance* 70, 419-447.
Hu, Edwin, and Dermot Murphy, 2020, Vestigial Tails: Floor Brokers at the Close in Modern Electronic Markets, Available at.
Indriawan, Ivan, Roberto Pascual, and Andriy Shkilko, 2021, On the effects of continuous trading, Available at: https://ssrn.com/abstract=3707154.
Jagannathan, Ravi, Andrei Jirnyi, and Ann Guenther Sherman, 2015, Share auctions of initial public offerings: Global evidence, *Journal of Financial Intermediation* 24, 283-311.
Klemperer, Paul, 2002, What really matters in auction design, *Journal of Economic Perspectives* 16, 169-189.
Klemperer, Paul, 2004. *Auctions: Theory and Practice* (Princeton University Press, Princeton).
Lee, Y. T., Y. J. Liu, R. Roll, and A. Subrahmanyam, 2004, Order imbalances and market efficiency: Evidence from the Taiwan stock exchange, *Journal of Financial and Quantitative Analysis* 39, 327-341.
Liu, Yu-Jane, 2016, A Simple Alternative: High Frequent Call Auctions, Guanghua School of Peking University, Available at.
London Stock Exchange, 2000, London Stock Exchange Market Enhancements.
Nyborg, Kjell G., and Suresh Sundaresan, 1996, Discriminatory versus uniform treasury auctions: Evidence from when-issued transactions, *Journal of Financial Economics* 42, 63-104.
Schwartz, R. A., and L. R. Wu, 2013, Equity trading in the fast lane: the staccato alternative, *Journal of Portfolio Management* 39, 3-6.
Schwartz, Robert A., 2001, The call auction alternative, in Robert A. Schwartz, ed.: *The Electronic Call Auction: Market Mechanism And Trading* (Springer Science+Business Media, New York).
Tel Aviv Stock Exchange, 2013, Trading Schedule.
U.S. Securities and Exchange Commission, 2022, Order Competition Rule.

Chapter 7. Dealers in public limit-order markets

The last chapter examined the use of auctions in two roles. Auctions can assist limit order markets (in opening and closing continuous trading sessions). They can also serve as the sole trading mechanism, as in the case of the periodic call auctions used for securities with low natural trading interest. The present discussion of dealers takes a similar approach. Like auctions, dealers can both assist limit order markets, and also serve as the sole or dominant trading mechanism. This chapter covers the first case, dealers as supplemental or auxiliary liquidity providers; Chapter 9 discusses dealer markets, where dealers have such central roles in trading that it is difficult for customers to bypass them.

The starting point of our discussion is the definition of a dealer and the basic conflict of interest that informs their activities. We next describe the roles of two kinds of dealers that have historically played central roles in US equities markets: the exchange specialists and the NASDAQ market makers. In the landscape of today's equity markets, though, these traditional dealers play a much smaller part. Technological forces have shifted the dealer functions from specialists and market makers to proprietary ("prop") trading firms.

We are still grappling with the aftermath of this transition. The traditional dealers were heavily monitored and regulated; the newer ones are presently subject to much less oversight. The market-making process has been broadly implicated in several recent market "stress" events (the flash crash of May 10, 2010, and the near-failure of Knight Capital in 2012). These events have raised regulatory concerns, and historical parallels are evident. How did we expect the traditional market makers to behave, and should their replacements assume similar responsibilities?

7.1. What is a dealer?

A dealer is a financial intermediary who specializes in serving as a *counterparty* to customer trades: buying when the customer wants to sell, and vice versa. In this role, a dealer may also be described as a market-maker or liquidity-provider.

The function of a dealer is very different from that of a broker. A broker represents a customer order, conveying it to the market, and acting as its agent. A dealer executes the order against his own account.

Despite this difference, it is often very natural for the same person or firm to serve in both dealer and broker capacities. To see how this might happen, suppose that the NBBO is 20 bid, offered at 21 and the customer sends in an order to buy limit 21. The broker can pass the order to a market center posting the NBO and get a fill at 21.

But suppose that the broker (or someone at his firm, e.g., a proprietary trading desk) wants to sell the stock and is willing to sell at a price better than 21, say 20.90. It would be very natural for the broker to sell directly to the customer.

Although this situation arises quite naturally, it places the broker in a position of conflicted interests. Acting as a broker, his obligation is as agent to his customer buyer, attempting to get the lowest price possible. But acting as a dealer, he wants to get the best terms of trade for himself, selling to the customer at a price as high as possible. In the above example, the broker is willing to accept 20.90, but the customer is willing to pay up to $21. Why not sell to the customer at 20.99? At 21?

Often when we encounter a conflict of interest, we try to assign the conflicting roles or practices to different individuals or institutions. For example, CME Rule 552 states:

> The term "dual trading" shall mean trading or placing an order for one's own account in any contract … in which [the member] previously executed, received or processed a customer order on the Exchange floor during the same Regular Trading Hours session.
>
> - Subject to the following exceptions, dual trading shall be prohibited …:
> - Customer Permission. A member may engage in dual trading … [if the customer grants] prior written permission.
> - Errors. A member may engage in dual trading to offset errors resulting from the execution of customer orders …
> - …
> - Violation of this rule may be a major offense.

Thus, you can't be a broker and a dealer in the same trading session.

Under the separation principle, we'd expect that our securities regulations might make clear distinctions between brokers and dealers. Nevertheless, SEC regulation combines them for many purposes in one category: broker-dealer. In writing the 1934 act that brought the SEC into existence, Congress initially intended separation (segregation) of the two roles, but this was criticized as too disruptive. Instead, further study was indicated. In 1936 the SEC responded with the "Report on the Feasibility and Advisability of the Complete Segregation of the Functions of Dealer and Broker," which noted, "… the great majority of persons engaged in the securities business in the United States *combine the functions of broker and dealer*," (U.S. Securities and Exchange Commission, 1936, p. XIV, italics mine). The report recognized the reality and potential for conflict but concluded that they were best addressed by rules tailored to specific practices and situations, rather than a full segregation.

7.2. Dealers in limit order markets

In our earlier discussion of limit order markets, the bid and offer books were generally assumed to be full of potential buyers and sellers. This is not always the case. As conditions change, limit orders may need to be repriced. This process requires ongoing monitoring and adaptation to new information. If the probability of execution is low, the cost of monitoring and updating may be too high to justify exposing a limit order in the first place. There are no bids or offers.

Potential buyers and sellers checking the market's quotes find nothing, and eventually they stop checking.

If there were even one bid and one offer, a potential trader might enter their own limit order. With sufficiently aggressive limit orders, a marketable order will eventually arrive, and we'll have an execution. In this sense, a bid and offer can encourage other traders to enter and sustain a market going forward. As an exchange's revenue depends on trading volume, it is in the exchange's interest to facilitate this process. The exchange may therefore appoint someone to make the market. This agent is variously called a designated market maker (DMM, the most common term), a liquidity provider or (formerly, on US exchanges) a *specialist*.

When there are insufficient customer limit orders, the DMM is required to post his/her own orders, priced aggressively enough to make a tight spread. The definition of "tight" here depends on the market, the security, and market conditions. This is DMM's primary responsibility, but there may be others.

While there is a consensus that DMMs can usefully augment limit order book, there is less agreement about how they should be incentivized and regulated. The responsibility to make a market is often burdensome (in a volatile market, for example). What do the DMMs get in return? How should they get compensated?

Sometimes the arrangements are direct. In the Euronext markets, DMMs can be paid by the listing company. The argument in favor of this practice is that the beneficiaries of the DMM are the shareholders of the listed company, so they (through the company) are the logical ones to bear the cost. The argument against is that the arrangement links a large trader (the DMM) closely with management (who are likely to possess inside information). Because of this moral hazard problem, US regulators have traditionally opposed subsidized market makers. Recently, however, the practice is being permitted for certain Exchange Traded Products, which are less subject to insider trading concerns.[1]

More typically, however, DMMs are compensated indirectly, through a variety of subsidies, rules and other practices that attempt in some broad fashion to enable them to recover the costs of the services they provide. These mechanisms are rarely isolated features or privileges. They are instead deeply embedded in the specification of the market-maker's role. This will become clearer as we turn to the discussion of US exchange specialists and NASDAQ market makers.

7.3. Exchange specialists and Nasdaq market makers

The New York Stock Exchange was historically a floor market. According to legend, the specialist system began in the 19th century when a member suffering from a broken leg decided to remain in one spot and specialize in trading a small number of stocks. Whatever the truth of this story, the acceptance and adoption of this system must have followed from stronger economic forces at work. In any event, by the advent of federal regulation in the 1930s, the specialist system was well established at the NYSE and other stock exchanges. Around 2005, the role of the specialist was redefined, and they were renamed designated market makers (DMMs).

Our discussion will start with an examination of the specialist system in the last half of the 20th century when floor markets were still widespread. Why start here? The specialist of this era was a single person placed in an extraordinarily powerful position, at the center of trading activity. This power, and the rules designed to monitor and shape it, represented a balance of

[1] Exchange traded products (ETPs) include exchange traded (mutual) funds, like the SPY, and also exchange traded notes (ETFs and ETNs). An ETP usually represents a portfolio of stocks, and its value depends on readily observable market prices. In the case of ETPs, the DMMs are paid by the sponsor of the product, the firm that constructed it and manages it.

public and private interests involving certain principles and trade-offs that have informed regulatory debates well into the era of high-frequency trading. When we think about how today's market makers should be regulated, the rules of the former specialist system are still a relevant touchstone.

The specialist (circa 1989)[2]

The specialist was an independent trader, buying and selling on his own account (not an Exchange employee). Each listed stock generally had one specialist, but a specialist could handle multiple stocks. Since the NYSE accounted for the preponderance of trading volume in its listed companies, the specialist truly stood at the center of the trading process. The specialist was broadly charged with "maintaining a fair and orderly market," but was also subject to a wide range of specific affirmative (positive) and negative obligations.

- *Making a market.* The specialist was expected to always post a bid and offer, not necessarily for large sizes, but with a spread that was reasonably (given the characteristics of the stock) narrow. This is the essential expectation, of course, for all market makers.
- *Avoiding "destabilizing trades."* The specialist was not supposed to trade actively, that is, by hitting/taking public limit orders. The ideal was neutrality, bringing buyers and sellers together in a way that encouraged the emergence of the "natural" price.
- *Public priority.* A specialist was supposed to yield to customer bids and offers at the same price. For example, if the specialist were bidding 50, and a customer were to enter a limit order to buy at 50, the customer's bid would have priority over the specialist's bid. Nowadays we think of the limit order market as a self-contained trading mechanism. But in a floor market, a limit order book is simply a collection of orders written on pieces of paper. Each order needs an advocate, an agent, to represent it in the trading crowd. The specialist was this agent. Public priority follows from the belief that an agent should not trade ahead of those whom he is representing.[3]
- *Crossing public orders.* If market buy and sell orders arrived simultaneously, the specialist was supposed to match them directly, generally at the midpoint of the bid and offer.
- *Price continuity.* Large price swings are sometimes viewed as evidence of a chaotic market. To avoid these jumps, the specialist was required to bridge large changes by a series of trades at intermediate prices. These were often unprofitable, especially when the gap was large.

The specialist was sometimes described as a monopolist. In fact, his market power was more circumscribed. The principle of public priority meant that customer bids and offers were often effective competition against his own. As a result, the specialists' trades did not generally constitute the preponderance of trading activity. Most trades were customer-to-customer, without the direct involvement of the specialist.

The specialist also enjoyed certain advantages. As the agent for the limit order book, he knew the contents of the book. As he was prohibited from disclosing the book to others, this constituted a distinct edge. In his role as agent for incoming market orders, the specialist also

[2] Specialist rules and practices evolved over time. This discussion is based on an NYSE memorandum to members that summarizes the role of the specialist (New York Stock Exchange, 1989).

[3] The practice of agents (brokers) trading in advance of their principals (customers) is called front-running. In many circumstances front running is against market rules. Nowadays, however, the term is more broadly used to describe any situation where someone is trading in advance of another trader's order. This might not be desirable, but if there is no agency duty, it is not usually deemed illegal.

enjoyed a first mover advantage. That is, when a market order arrived, he could decide whether to take the other side of the order or let the order execute against the book. This was not quite a right of first refusal because public priority might compel him to offer a better price. For example, if there were a customer limit order offering stock at $50 and a market buy order arrived, the specialist could not sell to the market order at $50. He would have to offer a better price, say, $49 7/8. (Until roughly the end of the 20th century, the tick size in US equity markets was $\$1/8 = \$0.125$.)

The NYSE's Designated Market Makers

By the 1990s, the specialist system was coming under pressures of technology, competition, and increased regulatory scrutiny. In the early 2000's, the NYSE fundamentally changed its structure. The specialist was replaced by a designated market maker (DMM). As under the specialist system, there is one DMM for each stock. Two prominent responsibilities remain. The DMM must maintain a fair and orderly market and post a bid and an offer. Other rules are quite different. The specialist is no longer the agent for the limit order book, and the principle of public priority has been replaced by *parity* (essentially that the specialist can share executions alongside the limit order book).

NASDAQ Market Makers

NASDAQ's roots were in a dispersed network of loosely linked ("over-the-counter", OTC) brokers and dealers. Over time, the network was computerized, and the role of market makers was formalized.

Some NASDAQ members were classified as order entry firms. An order entry firm could accept customer orders, but it could not trade against them as principal. Instead, the order entry firm would typically pass the orders to a NASDAQ market maker for execution. A NASDAQ market-maker in a particular stock would meet certain minimum capital markets agree to make and maintain a bid and offer in that stock. A NASDAQ-listed company had to have at least two market makers. A large stock (like Microsoft) might have over fifty. A market-making firm might have its own retail arm, but others, variously called "wholesalers" or "OTC market-makers" would specialize in handling the orders sent to them by other firms.

This industrial structure is still a pretty good description of the industry. Most firms familiar to retail investors (like Charles Schwab or E-TRADE) are primarily order entry firms. The wholesalers include the large sell-side banks (such as JP Morgan, Credit Suisse, and Goldman Sachs), and firms that are more focused on market-making (such as Automated Trading Desk, Cantor Fitzgerald, and Knight Capital).

NYSE-listed and NASDAQ-listed stocks trade in a similar fashion, but this was not historically the case. Prior to the 1990s NASDAQ did not have a central limit order book. Customer bids and offers were not displayed, and direct trade between a buyer and seller was rare. The handling of customer limit orders became a point of regulatory conflict. Eventually the SEC's order handling rules compelled NASDAQ market-makers to yield to public orders, display them, and make them available for execution (U.S. Securities and Exchange Commission, 1996).

7.4. De facto market makers ("market makers in fact, if not in name")

Customer bids and offers are competition for a dealer's own quotes. The principles of public priority and the order handling rules attempt to protect the customers' interests. In a historical context, this was sensible. Members of the floor-based exchanges enjoyed a centrality to the market process that often worked to the disadvantage of off-floor customers. These customers

were generally viewed as long-term investors. Who could hope to make short-term proprietary trading profits in competition with the floor?

When orders were communicated verbally or over slow networks, a human market maker on a floor market enjoyed a latency (speed) advantage over public bids and offers. With the advent of low-latency computer systems, though, the competition from off-floor bids and offers became much stronger.

The competition between the (human) market makers and off-floor customers reached a tipping point with the introduction of Reg NMS (2005). As discussed in Section 5.5, Reg NMS mandates trade-through protection of bids and offers, subject to certain restrictions. It only applies to the top of the visible book (a market center's best bid and offer). An additional restriction, though, concerns access. For a bid or offer to be protected, it must be immediately accessible for execution. The bids and offers of non-automated markets or market participants are not covered.

This elevated the relative status of customers who invested in the fastest technology. Their bids and offers largely supplanted those of the traditional market makers. These customers became the new "de facto" market makers. The trend was not limited to the US. (Menkveld, 2013) examines the entry of a new electronic market (CHI-X) into Europe: "[Three] events coincided: [the] new market's take-off, the arrival of a large HFT [high frequency trader], and a 50% drop in the bid-ask spread."

Beginning in 2007, the US entered a financial crisis in which one major investment firm (Lehman) failed, others were merged into stronger partners (sometimes under regulatory pressure), and government guarantees were deemed necessary to restore confidence in the financial system. During this crisis, the markets for some securities (notably mortgage-backed and auction-rate) performed poorly, with greatly reduced liquidity and availability.

The equity markets, however, appeared to function well. In the face of unprecedented volume and volatility, these markets continued to operate smoothly. Spreads widened somewhat in recognition of the higher risks, but limit order books did not suddenly become empty. It was almost always possible to trade.

Despite this generally satisfactory performance, concerns arose. It was observed that the de facto market makers had no formal affiliation or obligations. They generally supplied liquidity, but this provision was opportunistic. There was no penalty for withdrawal in a volatile market. Nor were there prohibitions against destabilizing trades. The 2010 SEC Concept Release on Equity Market Structure noted (U.S. Securities and Exchange Commission, 2010):

> The use of certain strategies by some proprietary firms has, in many trading centers, largely replaced the role of specialists and market makers with affirmative and negative obligations. Has market quality improved or suffered from this development? How important are affirmative and negative obligations to market quality in today's market structure? Are they more important for any particular equity type or during certain periods, such as times of stress? Should some or all proprietary firms be subject to affirmative or negative trading obligations that are designed to promote market quality and prevent harmful conduct? Is there any evidence that proprietary firms increase or reduce the amount of liquidity they provide to the market during times of stress?

The Concept Release was published in January of 2010. A few months later, the flash-crash of May 6, 2010, provided some direct evidence. Around 2:30pm the US broad market indices dropped around 5% in a few minutes, and then rebounded a similar amount also in the span of a few minutes. A Joint CFTC-SEC report analyzes the event (U.S. Commodity Futures Trading Commission and Commission, 2010). "The de facto market makers, it should be emphasized, did not trigger or cause the decline." The report nevertheless notes:

> [At about 2:45pm] based on interviews with a variety of large market participants, automated trading systems used by many liquidity providers temporarily paused in reaction to the sudden price declines observed during the first liquidity crisis. These built-in pauses are designed to prevent automated systems from trading when prices move beyond pre-defined thresholds in order to allow traders and risk managers to fully assess market conditions before trading is resumed.
>
> After their trading systems were automatically paused, individual market participants had to assess the risks associated with continuing their trading …
>
> Based on their respective individual risk assessments, some market makers and other liquidity providers widened their quote spreads, others reduced offered liquidity, and a significant number withdrew completely from the markets. Some fell back to manual trading but had to limit their focus to only a subset of securities as they were not able to keep up with the nearly ten-fold increase in volume that occurred as prices in many securities rapidly declined.
>
> HFTs in the equity markets, who normally both provide and take liquidity as part of their strategies, traded proportionally more as volume increased, and overall were net sellers in the rapidly declining broad market along with most other participants. Some of these firms continued to trade as the broad indices began to recover and individual securities started to experience severe price dislocations, whereas others reduced or halted trading completely.
>
> Many over-the-counter ("OTC") market makers who would otherwise internally execute as principal a significant fraction of the buy and sell orders they receive from retail customers (i.e., "internalizers") began routing most, if not all, of these orders directly to the public exchanges where they competed with other orders for immediately available, but dwindling, liquidity.

In summary, while traditional exchange specialists were expected to maintain a bid and an offer, their successors weren't and didn't. While specialists were discouraged from selling into a declining market (trading in a destabilizing fashion), the newer firms were not so constrained and took advantage of this possibility. Retail orders became like hot potatoes, as firms that would normally give executions re-routed them to other venues.

Current European regulations impose requirements that are more stringent than those in the US. Under MiFiD-II, firms pursuing market making strategies (the de facto market makers) must register as market-makers. Market-makers must commit to ongoing provision liquidity (posting bids and offers). Exchanges must structure arrangements with market makers to incentivize this provision, and they must monitor their market makers.

7.5. Further reading

(Viswanathan and Wang, 2002) contrasts dealer and limit order markets. Interdealer markets are discussed in (Hansch, Naik and Viswanathan, 1998; Reiss and Werner, 1998; Viswanathan and Wang, 1998; Viswanathan and Wang, 2004). (Seppi, 1997) examines strategies of designated market makers (specialists). (Duffie, 2012) discusses many aspects dealer markets, notably transparency.

Summary of terms and concepts

The broker/dealer conflict of interest; designated market makers; contract market makers; NYSE specialist affirmative obligations (price continuity, maintaining a narrow bid and ask spread, crossing public buyers and sellers, public priority) and negative obligations (trading in a destabilizing fashion); Designated market makers (DMMs); parity.

References

Duffie, Darrell, 2012. *Dark markets: Asset pricing and information transmission in over-the-counter markets* (Princeton University Press).

Hansch, Oliver, N. Naik, and S. Viswanathan, 1998, Do inventories matter in dealership markets? Evidence from the London Stock Exchange, *Journal of Finance* 53.

Menkveld, Albert J., 2013, High-frequency trading and the new market-makers, *Journal of Financial Markets* 16, 712-740.

New York Stock Exchange, 1989, Memorandum to members: New Specialist Job Description, March 3, 1989, in Market Performance Committee, ed.

Reiss, P. C., and I. M. Werner, 1998, Does risk sharing motivate interdealer trading?, *Journal of Finance* 53, 1657-1703.

Seppi, Duane J., 1997, Liquidity provision with limit orders and a strategic specialist, *Review of Financial Studies* 10, 103-150.

U.S. Commodity Futures Trading Commission, and U.S. Securities and Exchange Commission, 2010, Findings Regarding the Market Events of May 6, 2010, (Washington D.C.).

U.S. Securities and Exchange Commission, 1936, Report on the feasibility and advisability of the complete segregation of the functions of dealer and broker, pursuant to Section 11 (e) of the Securities Exchange Act of 1934, (U.S. Government Printing Office, Washington, D.C.).

U.S. Securities and Exchange Commission, 1996, Final rule: Order execution obligations (Release No. 34-37619A; File No. S7-30-95), CFR 17 Part 240.

U.S. Securities and Exchange Commission, 2010, Concept release on equity market structure.

Viswanathan, S., and James Wang, 1998, Why is interdealer trading so pervasive in financial markets, Duke University, Available at.

Viswanathan, S., and James J. D. Wang, 2002, Market Architecture: Limit-Order Books versus Dealership Markets, *Journal of Financial Markets* 5, 127.

Viswanathan, S., and James J.D. Wang, 2004, Inter-dealer trading in financial markets, *Journal of Business* 77, 987-1040.

Chapter 8. Dark Markets

Trades occur when an incoming order executes against a standing bid or ask. Now: did we see the bid or ask (was it visible?) before the trade occurred? If not, then trade is considered dark. Darkness is the absence of pre-trade quote transparency. In US equity markets, the trade is considered dark if the trade price is above the executing market's bid and below the executing market's ask (offer). For example, suppose that the National Best Bid and Offer (NBBO) are $25.00 and $25.10, but the best bid and offer (BBO) on Exchange X are $24.50 and $25.10. If Exchange X executes a trade at $25.00, that trade is dark. Note that the trade still must be reported, like any other trade. Dark trades typically arise from one of following mechanisms:

1. A hidden (undisplayed) limit order in a limit order market (like Nasdaq) that also displays visible orders.
2. A NASDAQ market maker trades against a customer order at the NBBO at a time when the market maker's own quotes are behind (inferior to) the NBBO.
3. The trade occurs in a crossing network or dark pool that posts no visible quotes of its own but matches buyers and sellers at prices determined by the NBBO.

A market that normally has a visible bid and ask is said to be "lit". Case 1 involves a dark trade on a lit market. Case 2 includes the situations where a broker (like Robinhood) sends a customer order to a market maker (like Citadel Securities). Case 3 describes crossing networks and dark pools.

Before considering dark mechanisms in detail, we turn to the question of why they exist in the first place.

8.1. The logic of darkness.

A dark trade is defined by the unwillingness of either party to the trade to post a visible bid or offer. Why should this be? If a buyer is willing to pay, say, $10, why not advertise the bid to attract sellers?

It is useful to consider a more familiar situation. The markets in consumer electronics products are extraordinarily competitive: the product (a given model from a given

manufacturer) is homogeneous, and the internet makes it easy to comparison shop. Many retailers advertise aggressive prices. Other retailers claim, "Bring us any advertised price, and we'll match or better it. Guaranteed." The situation is complex because it is unlikely that any two sellers are truly identical with respect to location, reputation, and many other small differentiating characteristics. We can still draw a few lessons by thinking about their strategies.

Firstly, the price matcher avoids the expense of determining and updating advertised prices. They are letting someone else "do the math," and simply doing as much or as little as necessary to remain competitive. Secondly, in the presence of a price matcher, the advertiser has a reduced incentive to post an aggressive price. When the advertiser lowers the posted price, the additional customers will be split with the price matcher.

Both considerations operate in securities markets. A seller who is willing to match the market's best visible offer can avoid the calculations and judgments necessary to determine their own reservation price. From the viewpoint of a seller posting a visible offer, why post an aggressive (low) offer if we're simply establishing a reference price for someone else's trade? We'll raise or keep the offer above our true reservation selling price .

We now turn to a description of the mechanisms.

8.2. Undisplayed (hidden) limit orders

In a hybrid market (like the US) a limit order has no market-wide time-priority, and only partial market-wide price priority. Not only may a visible limit order lose priority to visible limit orders posted at other market centers, but executions may occur at their prices, via market makers and dark pools who are posting no visible quotes of their own. Visibility has its benefits (in making potential counterparties aware of the price), but in a hybrid market it also carries a cost. Hiding a limit order, or displaying it only for a brief interval, is one way to control this cost. Hidden executions on lit exchanges are normally reported in the same way as lit executions: they are identified by the exchange on which they occur.

8.3. Dealers

When a broker receives a customer market order, he may, if his firm is a market maker in the stock, simply keep the order within the firm. The market-making desk or division can take the other side of the customer order. The main requirement is that the execution occurs at or within the NBB or NBO. It is not necessary that the firm's market-maker's quote be at the NBBO. The execution must be reported, of course. The market maker will probably send the report to FINRA's Trade Reporting Facility. Orders executed within the firm are said to be *internalized*.

The market maker may also be receiving order flow from other firms. Much of this happens by pre-arrangement. An order-entry firm may routinely send (or *preference*) its customer orders to a particular market-maker with the understanding that the orders will receive the NBBO or better.

It was noted in the discussion of private information that less-informed order flow, such as that originating from retail customers, is more profitable for dealers. There is less chance of loss. This is not merely an academic fine point. One aspect of the arrangement between order-entry and market firms involves payment: the order-entry firm receives a small consideration for each order that it sends.

Internalized and preferenced orders never interact with other orders. They aren't generally conveyed to any central limit order book, for example, and so don't fully participate in the market. They receive a price that is set by others. In the price determination process, they represent buyers and sellers who are participating without a vote.

8.4. Point-in-time crossing networks and continuous dark pools

Crossing networks and dark pools are markets that are separate from, but still very reliant on, the lit market. Historically, point-in-time crossing networks came first. Some of them are still active as crossings, but some evolved into continuous crosses, and ultimately, into today's dark pools.

Before exploring the differences, here are the points of similarity.

- The "darkness" is pre-trade only. Trades on dark pools have to be reported (and are visible on the Consolidated Trade System) just like all other trades.
- They don't disseminate quotes or provide price discovery or determination. The prices for all trades are set by reference to prices determined in the lit market (usually the NBBO).
- They are regulated by the SEC as "Alternative Trading Systems" (ATSs).

Across these markets there are many variations in the trading procedures. The following describes typical systems. We'll then discuss modifications and extensions.

Point-in-time crossing networks

Orders are submitted anonymously. They usually specify direction (buy or sell) and quantity (or sometimes a quantity range), but not prices. They are held in the system undisplayed. At some pre-scheduled time, the system goes through the orders and attempts to match (pair off) buyers and sellers. A buyer for 20,000 MSFT and a seller for 30,000 MSFT would be paired off for 20,000 shares, for example. Whenever there is a match, the execution price for the trade is determined by reference to the lit market. Depending on how the dark pool is designed, this price might be the prevailing NBBO midpoint, the day's closing price or the day's volume-weighted average price (VWAP). Of course, with VWAP or closing price, the price is not determined, and the trade is not reported until the end of the day.

Instinet runs VWAP crossings that are marked to full-day and last-hour VWAP. For the full-day VWAP, "Orders are matched during any of the three pre-market matches — 8:35, 8:50 and 9:15 am ET. Fills are sent back immediately after each match. At approximately 4:10 pm ET, fills receive the consolidated full day VWAP and are printed." For the last-hour VWAP cross, "Orders are matched at 3:00 pm ET. Indicative fills are sent back after the match at the stock's NBBO midpoint. At approximately 4:10 pm ET, fills receive the VWAP for the last hour of trading (3:00 pm-4:00 pm) and are printed," (https://www.instinet.com/us-vwap-cross). Instinet also runs closing crosses marked to (variously) the NYSE, NASDAQ or ARCA closing prices: "The cross occurs prior to each exchange's 'on-the-close order submission' cutoff … (NYSE: 3:43PM EST; NASDAQ: 3:48PM EST; ARCA 3:48PM EST)." (https://www.instinet.com/market-close-cross-us).

Trades occurring in crossing networks are sometimes described as "zero impact". When the trade is reported, it can't be determined whether the aggressor was a buyer or seller. Nothing is shown prior to execution. If there is no execution, nothing is shown.

Continuous dark pools

Consider the MSFT example, above. In a crossing network, the timing of order arrival doesn't matter. Everyone must wait for the next scheduled match. In a continuous dark pool, the system looks for a match whenever a new order arrives. If a match can be made, the trade is executed immediately. As in a crossing network, the execution price is set by reference to some price from the lit market.

How is this different from a limit order book in which all orders are hidden? Remember that in a limit order book, the execution price is set to the limit price of the resting order (the

bid or offer that is already in the book). In a continuous dark pool, the trade price is set by reference to the lit market, typically the NBBO midpoint. As with the point-in-time crossing networks, nothing is shown unless there is an execution.

By some counts there are over fifty dark pools. Users often try them sequentially, favoring those that in the past have proven likely to provide fills (the dark pool routing decision).

Variations

Some systems allow traders to specify conditions that an execution must satisfy. The most important of these are:

- Minimum quantity. ("Don't execute the trade unless it is for at least 10,000 shares.")
- Limit price.

Note that when a limit price is specified, it restricts the allowable outcomes, but does not affect the execution price. For example, if "sell 1,000 shares limit $10.50" is an order resting in a conventional limit order book, then (if it is executed) the price will be $10.50. If the same order is sent to a typical dark pool, the execution price will be the NBBO midpoint at the time of the match, and if the NBBO midpoint is below $10.50 there will be no match.

8.5. Crosses on the floor

In our discussion of floor markets (section 3.5), we encountered a broker ("CAT") on the floor of the Chicago Mercantile Exchange (CME) who had a customer order to buy and another customer order to sell. She intended to cross (match) her customers at the average of the bid and ask on the floor, much like a dark pool might execute a midpoint cross. We did not label the intended crossing trade as "dark," but by the definition given at the start of this chapter, it would have been. That is, she would be executing a trade at a price where she had established neither a visible (well, audible) bid or offer.

We noted that a trade of this sort would run counter to the open-outcry requirement. Rule 533 was the Exchange's remedy. It required her to bid (and offer) three times at her intended crossing price before she could execute a trade at that price. The New York Stock Exchange has a similar rule. ""When a member has an order to buy and an order to sell the same stock, he or she must publicly offer at a price higher than his or her bid by the minimum variation," (Rule 76). There is a slight difference between the NYSE and CME procedures. On the CME, the crosser's bid and offer are at the same price; on the NYSE, the crosser's bid and offer are one tick apart. (The NYSE requirement establishes a two-sided market.) This difference, though, should not obscure the fundamental point of agreement: both exchanges require the crossing member to bid and offer, thereby illuminating a trade that would otherwise be dark.

In modern markets, rules like CME 533 and NYSE 76 are generally known as "trade at" rules. In crossing trades, exchanges, dealers, and other parties subject to the rule must trade at their established bid or offer. There are often exceptions for large trades and midpoint trades. Both Australia and Canada have trade-at rules. The US currently does not.

8.6. Enforcement in dark markets

Dark markets are sometimes viewed with suspicion. Confidentiality of customer orders is difficult to monitor and verify; several recent cases raise doubts.

- According to one SEC file (at http://www.sec.gov/litigation/admin/2011/33-9271.pdf):

> Pipeline Trading … operated an alternative system ("ATS"), a private stock-trading platform commonly referred to as a "dark pool." Pipeline held out its ATS as a "crossing network" that anonymously matched customers' interests in trading large amounts of stock. However, Pipeline did not disclose to its customers that the overwhelming

> majority of the shares traded on its ATS were bought or sold by a wholly owned subsidiary of Pipeline [Millstream, the "Affiliate"].

- In the case of a Barclay's dark pool, the New York attorney general alleged that although the firm reassured customers that they were being protected from predatory high frequency traders, the firm was actively soliciting high frequency traders to join the pool.
- UBS paid a $12 Million fine to settle an SEC complaint that in its dark pool UBS allowed high frequency traders and market makers to jump ahead of customer orders.

(Domowitz, Finkelshteyn and Yegerman, 2009) note that:

> Even if the dark pool is operationally secure, other users may still be able to draw inferences via "sniffing" and "sniping". These practices involve using small standing orders to detect larger incoming orders, or small marketable orders to detect larger standing orders.

Any system that matches buyers and sellers at some external reference price gives the users the incentive to manipulate that price. For example, a buyer sending an order to a dark pool knows that any execution will be priced at the NBBO midpoint. The buyer can lower the midpoint by submitting a small aggressive sell limit order to a lit market. After achieving a dark pool execution, the sell limit order is cancelled. This practice, one type of "spoofing," has attracted regulatory and enforcement interest.

8.7. The interplay of dark and lit markets

Dark mechanisms are controversial. Aside from hidden orders, they assign prices using the quotes in the lit market, essentially "free riding" on the lit prices. While they provide liquidity, their contributions aren't visible pre-trade.

An aggressive visible bid or offer is an advertisement that encourages potential counterparties to hit or lift it. If these counterparties can obtain the same price in a dark pool, their orders will migrate away from the visible market. With fewer traders posting orders in the lit market, the bid ask spread will become wider, and (with less participation) more variable. This in turn hurts not only the traders in the lit market, but those in the dark pool as well (since the dark pool relies on the lit market's prices.

Summary of terms and concepts

Dark trades vs. dark markets; lit markets; hidden orders; internalized executions; crossing sessions; continuous dark pools; midpoint matches; spoofing; leakage; pros and cons of dark trades; regulatory concerns; recent cases (Pipeline/Millstream; UBS; Barclays).

References

Domowitz, Ian, Ilya Finkelshteyn, and Henry Yegerman, 2009, Cul de sacs and highways: an optical tour of dark pool trading performance, *Journal of Trading* 4, 16-22.

Chapter 9. Dealer markets

Chapter 7 discusses dealers as supplemental players, providing bids and asks in markets mostly organized about limit order books. In these situations, customers will sometimes trade against dealers, but through the limit order books customers will often trade against other customers. This chapter deals with markets in which the dealer is the defining feature. These markets are sometimes called dealer markets but are also commonly described as over-the-counter (OTC) or quote-driven markets. The important dealer markets include foreign exchange (FX), over-the-counter derivatives, swaps, government bonds, and corporate bonds. Trading practices vary across these markets, but there are often strong similarities. We will focus first on these similarities, and then turn to specifics of the different markets.

In these markets, dealers are central to the trading process. The markets might incorporate limit order books, auctions, and so forth, but they do not usually have the broad coverage and availability that they do in equities markets. Historically, dealer markets evolved without centralized trading floors. Each dealer operated as a separate proprietorship, exchanging securities and payments with incoming customers.

On the surface, dealer markets exhibit a bewildering array of features. In important ways they are all different. Beneath these differences, though, are similarities and commonalities that can help us understand individual markets in relation to the broader collection. The discussion therefore starts with describing the features of typical traditional dealer markets and the evolution of these markets. We then discuss details of foreign exchange and fixed-income markets.

9.1. The traditional dealer market and its evolution

To introduce a traditional dealer market, you might think of a retail currency exchange business at an airport or train station. A traveler arriving in Europe from the US might want to buy euros, paying an ask price stated in dollars. On the return trip, she might want to sell her unspent euros, receiving the bid price (also stated in dollars). In either transaction, an attempt to negotiate for a higher bid or lower ask would probably be summarily dismissed. The dealer's quotes are

take-it-or-leave it, and this rigidity may explain why dealer markets are sometimes described as quote-driven.

The FX trades that take place at retail currency venues are mostly small one-time transactions. Large or frequent transactions will be accomplished through the customer's bank. On a larger scale, an investment fund or corporation (for example) might establish and maintain a relationship with one or more dealing banks. Traditionally, a hedge fund wishing to buy Euros might contact the FX desk (of a large banks) and ask for a two-sided market, that is the dealer's bid and ask quotes. The dealer knows the customer's identity. The quotes are often oral, and actionable only for a brief period (say, a minute). The bid and ask are specific to the customer. Other customers cannot hear these numbers. The dealer may convey different quotes to different customers, possibly considering the profitability of customers' previous trades.

The dealer/customer relationship is sustained by reputation. The dealer will always make a market. The customer must (at least sometimes) trade on the bids and asks she is given. The interaction leading up to a bond or swap trade would be similar.

Traditional dealer markets usually exhibit some common features. Dealers are linked by computers and telecommunications, but there is little centralized coordination. Customers have access only to dealers with whom they have previously established a relationship. This mutual recognition establishes credit limits and trading procedures that minimize the work of processing subsequent trades. A dealer is always the counterparty to a customer trade.

Dealers may disseminate *indicative* bids and asks widely (on Bloomberg, or financial web sites, for example). These are mainly advertising. *Firm* quotes (against which a customer can actually execute an order) are generally given only to customers with whom the firm has a preexisting relationship, and often only in response to inquiry.

Generally, trades are not publicly reported. For investors who are accustomed to the comprehensive last sale reporting available in most equity markets, this may come as a surprise. There is no consolidated feed, for example, that publishes the recent trades in foreign exchange or US Treasury bonds.

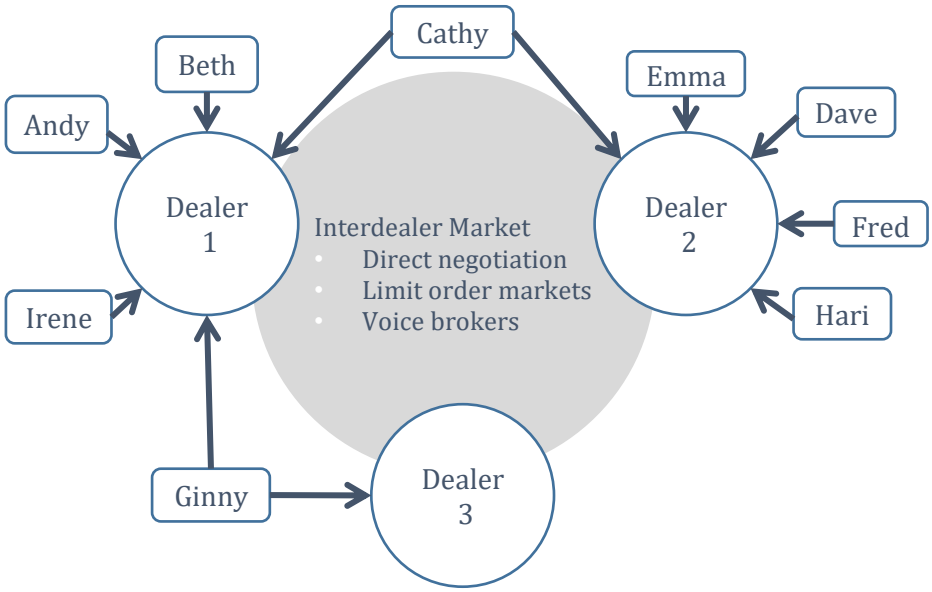
Although the dealers are central and important, they are ultimately intermediaries. The securities (or, in the case of FX, the currencies) ultimately pass from one customer to another. A dealer's long or short position is a source of risk, in the form of exposure to adverse price changes. Ideally, then, the dealer's position is close to zero ("flat"). If the customer were flow were naturally balanced, that is, if a customer selling 100 units were quickly followed by a customer buying 100 units, the dealer would never accumulate a large position. More realistically, though, customer flows are unbalanced, and positions build up.

To some extent, a dealer can change the mix of buyers and sellers by modifying their bids and asks. If the dealer has a large short position, for example, she can increase her bid and ask. This will encourage incoming customer sellers (who will hit the dealer's bid, reducing the short position) and discourage incoming customer buyers (who might lift the dealer's ask, increasing the size of the short position). But this tool has its limits. Some customers can compare prices across dealers, and a dealer who often sets prices far from the market will be dropped from customers' call lists. So, adjusting bids and asks might reduce a dealer's unwanted position, but not entirely eliminate it. The solution to the dealer's problem lies in what is called the interdealer market. This is a venue for dealers to trade against each other to work off the imbalances in their customer order flows.

The trading relationships of customers and dealers in a typical dealer market are illustrated in Figure 9-1. Around the outside of the figure are the named customers who are each connected to one or two dealers. The three dealers are grouped around the interdealer market. The diagram suggests a network, and formal network analysis is often applied to the study of these markets. The networks that arise in many disciplines are often described as core/periphery, with a few highly connected agents (or, in network analysis, nodes) at the center, and a

larger number of peripheral agents with fewer connections. These networks are also called hub-and-spoke, or, in our application, inner and outer markets.

Figure 9-1. Trading relationships in a dealer market



Interdealer markets exhibit a range of trading mechanisms. The simplest is *direct negotiation*: the € dealer at bank X may contact the dealer at Bank Y and ask for a quote. Bank Y will try to accommodate the request because over the course of the next hour or day, it is likely that at some point the situation will be reversed and Bank Y will be making the request. Communication usually takes place via secure instant messaging links. *Voice brokers* are intermediaries who negotiate trades between dealers, but without disclosing either party's identity. Finally, there is extensive use of *order book markets*. These are virtually identical to the systems used in equities, but entry is much more restricted. Very few customers have access privileges, and access fees are high.

The traditional dealer market is structured in a way that discourages customer-to-customer trade. Furthermore, even though some customers (like Ginny and Cathy in Figure 9-1) have relationships with multiple dealers, they can generally only contact the dealers one-at-a-time. This limits the number of competing dealers. To take an extreme example, if the customer were to try to contact ten dealers, the process would be sufficiently long that it would be difficult or at least awkward to return to the first dealer, if that first dealer's bid and ask appeared to be the best.

Although the traditional structure still exerts significant influence, the interplay of technology, increasing customer sophistication and regulatory pressure have introduced changes. Most dealer markets have evolved to accommodate one or more of the following mechanisms:

- Single/multiple-dealer execution platforms
- Request for quote (RFQ) procedures
- Prime brokerage
- Non-bank dealers

These have broadly worked in favor of the customers.

In the traditional interaction, the customer would call the dealer and the dealer would reply with a bid and ask. Nowadays this would more likely occur on a single-dealer [electronic] platform. The customer's screen displays an accessible (actionable) bid and ask, and trades can occur with one click. The bids and asks are updated continually, and they are described as *streaming*. Note, though, that the dealer does not have to display the same bids and asks to all customers. Multiple-dealer platforms function is a similar manner, provide streaming quotes (bids and asks) from a set of dealers. A customer such as Ginny or Cathy in Figure 9-1 will see streaming quotes from all dealers with whom she has a trading relationship. This provides a measure of competition.

In markets operated as continuous limit order books, auctions are sometimes used to concentrate trading at a particular point in time (usually the open or close, as described in section 6.2). On multiple-dealer platforms, request for quote (RFQ) protocols allow customers to initiate an auction. A customer wishing to buy or sell a particular quantity can broadcast the request to multiple dealers. These dealers can respond within a short time window, at the end of which the customer may trade at the best prices. This replaces the traditional one-at-a-time search, and because more dealers may respond, their bids and asks are more competitive. (Hendershott and Madhavan, 2015) describe this mechanism in the context of corporate bonds.

Prime brokerage (PB) provides customers with access to the interdealer market. Under prime brokerage a customer trades through the sponsorship of and in the name of someone (usually a dealer) who has direct access to the system. The terms of access may be more restrictive, though, than in the usual limit order market. A PB arrangement may give a customer the opportunity to submit marketable orders to an interdealer limit order book. The ability to enter a limit order, though, may be more restricted. Displayed customer bids and offers, after all, compete with a dealer's own quotes.

Despite the increasing prevalence of electronic trading platforms, RFQ and prime brokerage, however, customer protections in a dealer market are not as strong as would be typical in an equities market. This might not be immediately apparent. At first glance, the typical screen in a multiple-dealer execution platform might seem almost identical to what the customer sees on their broker's system when they buy or sell a stock. In both cases, the customer gets streaming bids and asks that can be accessed quickly.

There are, however, some major differences. The bid and ask on a broker's screen for MSFT stock will usually be the national best bid and offer (NBBO), that is, the best bids and asks across all exchanges trading the stock. The bids and asks on the multiple-dealer screen are limited to participating dealers and are not necessarily the best across all dealers. In fact, the dealers might be streaming better bids and asks to other customers on other platforms.[1]

Furthermore, in some markets, the dealer retains a right of "last look". That is, once the customer hits a dealer's bid or ask, the dealer has a short period of time in which to reject the trade. Last look calls to mind a floor trader who reneges on an oral bid or ask. As on a floor market, backing away from a quote subjects the dealer to reputational cost. For this reason, it is a right that is exercised sparingly. In posting on its website, JPMorgan provides a disclosure and rationale for last look:

[1] The text discussion simplifies certain points. The bid and ask visible on the stockbroker's screen might not be the current NBBO. The current NBBO would be available, though, by paying a small surcharge. Also, even without knowledge of the NBBO, the order protection rule would ensure that the customer execution price was within NBBO bounds. This applies only in the US.

> The Price Check is intended to protect J.P. Morgan, as a liquidity provider in the electronic FX and commodities markets, against latency inherent in electronic communications or erroneous price formation generated by external systems. (J.P. Morgan, 2018).

The statement suggests that last look is a general response to the limitations of electronic trading systems. Last look, though, is a long-standing practice, while widespread use of electronic systems is a more recent innovation. In fact, anyone who posts a bid or offer is (and always has been) exposed to the risk of being picked-off by someone with more current information.

Historically most dealers were located at major money-center banks. Trading involves the transfer of large sums of money, custody of securities, and (often) credit or financing. Banks are well-positioned to provide all these services. The dealing units of banks, however, have sometimes exposed the banks to substantial risk. In the US, the Dodd-Frank financial reforms were enacted in response to the 2007-2008 financial crisis. The "Volcker Rule," one component of the Dodd-Frank act, imposed limits on banks' dealing activities. Other rules, such as the Basel III accord, increased the capital that banks would have to hold relative to their dealing positions. The Volcker Rule and capital requirements have imposed costs on dealing activities, and banks scaled back. This has created openings for "non-bank liquidity providers", financial institutions such as hedge and other investment funds that act as dealers. These firms are not under the umbrella of banks' government deposit insurance, and so are not subject to the same strict regulations.

We now briefly describe the particulars of some important dealer markets.

9.2. Foreign Exchange (FX)

The FX market is where currencies (the dollar, euro, yen, renminbi, and so forth) are traded. The customers (end-users) are diverse, including individual retail traders, businesses involved in the import and export of goods and services, and investors managing cross-border portfolios of debt and equity. The following discussion relies on: (King, Osler and Rime, 2012) for market structure circa 2010; (Evans and Rime, 2019) and (Schrimpf and Sushko, 2019a, b) for later developments. There are no data sources that provide an ongoing continuous view of trading activity. The data that are most comprehensive across countries and market participants come from the Bank for International Settlement (BIS) Triennial Surveys, of which the most recent was conducted in April 2019 (Bank for International Settlements, 2019).

"Exchange" is an important word in this context. In a typical transaction, hedge fund *F* might contact bank *B*'s FX desk and arrange to pay \$1.1M to the bank and receive €1M. We could say that *F* is buying euros and selling dollars, or that bank *B* is selling euros and buying dollars. The transaction is symmetric. In some cases, it might be natural to emphasize one interpretation or the other. For example, if *F* needed euros to settle a purchase of French stock, the acquisition of these euros is the reason for the trade. The dollar is simply the payment currency, and some other payment currency (like the yen or pound) might have been easily substituted. On the other hand, if *F* had received unneeded dollars from the sale of US stock, it would be natural to think of the trade as a sale of dollars. The purchase and sale stories, though, are background. The trade itself conveys no presumptions.

Market conventions are nevertheless, in one sense, directional. Currencies are identified by three-character codes, the EUR and USD (for example), that are decided by the International Standards Organization. Any exchange involves a pair of currencies, designated with the two codes separated by a slash, as in EUR/USD. The first currency in the pair is the base currency; the second is the quote currency. An exchange rate (price) is given as the number of units of quote currency being exchanged for one unit of the base currency. A price quote in the EUR/USD market is the number of US dollars being exchanged for one euro (currently, as of Jan 2021, around \$1.23). Caution: the "/" is not a fraction bar; the price is dollars per euro, not

euros per dollar. There is no reason why some traders could use EUR/USD prices and others could use USD/EUR prices, but trading is simplified if the market settles on one convention. It is usually "EUR/USD." The convention applies to both bids and asks. On a trader's screen, the exchange rate would be given as, say, $1.23 bid – offered at $1.24. The euro is not always the base currency, of course. In exchanges between the US dollar and the Japanese yen, the convention is USD/JPY, with the dollar as the base currency.

The base/quote convention implies a directionality when trading against quotes from a dealer or limit order book. In the EUR/USD market, for example, we would pay dollars and receive euros when we lift the ask; we would receive dollars and pay euros when we hit the bid.

With around 160 currencies actively used worldwide, there are over 12,000 possible pairs. The number of actively traded pairs, though, is much smaller, and in most of the actively traded pairs the base currency is either the euro, the US dollar, the British pound or the yen. If there's a need to exchange a pair that is not actively traded, it is usually easiest to use two exchanges involving a bridge currency. To exchange Mexican pesos (MXN) and Norwegian Kroner (NOK), for example, we might go through the dollar: a trade in USD/MXN coupled with a trade in USD/NOK.

The FX market has no fixed hours of operation (although activity is light on weekends). Most dealers are found in banks, in all countries. The geographic centers of the market, though, are Tokyo, London, and New York. A plot of volume by time of day shows peaks and overlaps corresponding to regular business hours in these cities.

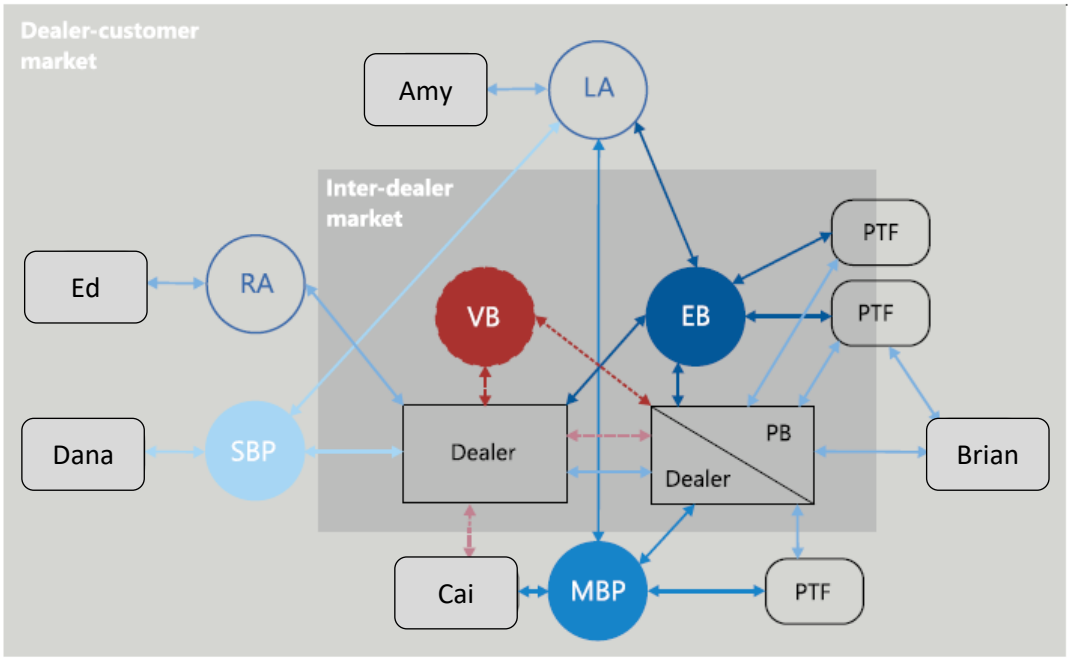
The interdealer market features two prominent limit order systems, informally known as EBS and Reuters. EBS originated in 1990 as Electronic Broking Services, a consortium of dealing banks. Having passed through several owners and name changes, it is presently operated by the NEX unit of the Chicago Mercantile Exchange (CME). The CME has announced plans, however, to merge trading into its Globex system in 2021. Reuters has operated various FX trading platforms aimed at diverse clienteles. The interdealer system is currently named "FX Matching". As of this writing, all of Reuters trading and market data operations have been spun off as a separate firm (Refinitiv) that is owned by the London Stock Exchange Group.

Reuters and EBS do not report trading volumes by currency pair, but the general consensus is that each dominates a different set of currency pairs. Pairs involving Scandinavian currencies or UK commonwealth currencies trade most actively on Reuters. EBS dominates the USD pairs.

The FX market has evolved to reflect many of the changes mentioned earlier in the discussion of dealer markets in general. Figure 9-2 (reproduced from (Schrimpf and Sushko, 2019c), with client/customer names added) depicts the current structure. The organization follows Figure 9-1, with customers at the outer edge of the market and the interdealer market at the core. There are, however, new connections and new players.

The interdealer market is segmented into electronic brokers (EBs, chiefly Reuters and EBS) and voice brokers (VBs). Principal trading firms (PTFs) are hedge funds, sovereign wealth funds, and similar entities trading for their own account. They have direct access to the EBs. Strikingly, they can also act as dealers against their own customer order flow ("Brian"). In this capacity, they are known as non-bank liquidity providers. This is important because they are not subject to regulations that constrain the market-making activities of banks. Prime brokerage (PB) provides a path for Brian to access the EB's directly. Client Dana trades through a single-bank platform (SBP). Cai trades through a multi-bank platform (MBP), seeing quotes derived from a dealer and a PTF.

Figure 9-2 Stylized structure of the FX market (from (Schrimpf and Sushko, 2019b), modified)



EB = electronic broker; LA = liquidity aggregator; MBP = multi-bank platform; PB = prime broker; PTF = principal trading firm; RA = retail aggregator; SBP = single-bank platform; VB = voice broker. Dashed lines indicate voice execution; solid lines indicate electronic execution.

Source: King et al (2012), augmented by adding LA to depict liquidity aggregators and PTFs in their roles as both clients and intermediaries.

9.3. US Fixed-income markets

Fixed-income securities (bills, notes and bonds) represent debt. The initial seller of the security (the issuer) is the borrower; the buyer (the investor) is the lender. The modifier "fixed" is accurate only in a limited sense. The only things that might in fact be considered preset, predetermined or invariant are the payments promised by the issuer. Prices and returns are determined in the market. They will vary depending on the issuer's ability and strength of commitment to make the promised payments, as well as broader forces of supply and demand. The maturity is the initial term of the debt (from issue to repayment). The securities are classified by their initial maturity: anything up to a year is a money-market instrument; one to ten years is a note; and over ten years is a bond. The present discussion focuses mostly on bonds, but also applies to notes. (Money-market practices are significantly different.) Bonds are issued by corporations ("corporates"), municipalities (cities and towns, "munis"), US government-sponsored entities ("agencies"), and the US government itself ("treasuries"). After the initial sale of the securities (in the primary markets), fixed-income securities generally trade in over-the-counter, dealer markets. This section summarizes the operation of that market, drawing on Bessembinder, Spatt and Venkataraman (2020, BSV), an authoritative overview. For a more comprehensive characterization of the securities themselves the reader is referred to one of the introductory financial markets texts listed in the introduction.

By comparison with most other financial markets, the most striking feature of the fixed-income market is the vast number of securities being traded. In the FX market, there are perhaps a few hundred actively traded currencies. In US stocks, the broadest Russell index comprises 3,000 issues. The number of bonds, on the other hand, is much larger. Mizrach (2015)

reports about 35,000 corporate bonds traded in 2015. For municipals at the end of 2017, BSV note around 50,000 issuers and over 1.5 million bonds.

As another point of comparison, stock/equity represents a claim on a firm's ongoing operations. The "lifetime" of an equity claim is indefinite but is often expected to be many decades or more. The maturity of a bond, though, is finite and known. An investor with an expected target horizon (like retirement or a child starting college) might reasonably choose to buy bonds with a corresponding maturity. Many institutional investors also have target dates. These maturity preferences affect bonds' typical trading patterns. At and around a bond's issue, it may be actively traded. Soon, though, trading activity declines as the bonds end up in portfolios where the investors plan to buy and hold to maturity. Most bonds may go for weeks or months without trading.

A few bonds, particularly those recently issued, are actively traded. For these bonds, as in the FX market, dealers may stream bids and asks in these bonds to institutional customers on single- and multiple-dealer platforms. The multiple-dealer systems usually support request for quote (RFQ) trading. Retail customers traditionally faced bids and asks offered by a single dealer, but some larger brokers are aggregating bids and asks from multiple dealers. These parallel the aggregation systems in the FX markets.

Corporate bonds

Among all fixed income markets, the US corporate bond market stands out in one important respect. Under SEC pressure, the market adopted last sale reporting. All bond trades must be reported to the FINRA-operated Trade Reporting and Compliance Engine (TRACE). Data are available through private vendors and on the FINRA website (http://cxa.gtm.idmanagedsolutions.com/finra/BondCenter/Default.aspx). For municipal bonds, the Municipal Securities Rulemaking Board (MSRB) operates a system, Electronic Municipal Market Access (EMMA) that reports last sale prices (https://emma.msrb.org/Home/Index).

The market has long been viewed as one due for an expanded role of limit order platforms. These now appear to be viable and some brokers have started consolidating their bids and offers. Harris (2015) discusses the current trading environment. Sirri (2014) provides a parallel analysis of the municipal bond market.

9.4. US Treasury Securities

Traditionally dominated by dealers, but (relatively) open order books have become more important. These books are not open to retail customers, and there are no publicly available last sale prices. European government bonds are generally traded on the MTS system. US Treasuries trading is concentrated on two systems eSpeed (a subsidiary of NASDAQ) and BrokerTec (a subsidiary of ICAP). (Fleming, Mizrach and Nguyen, 2014) discuss BrokerTec.

On October 15, 2015, the US Treasury market experienced an episode of high volatility. Since this did not seem to be caused by any major news announcement, attention turned to the trading process. The events are discussed in a joint interagency report that drew input from the Federal Reserve Bank, the Treasury Department, the SEC and the CFTC (U.S. Department of the Treasury, 2015). The Treasury Dept. subsequently initiated a request for information (Treasury, 2016).

In July 2021, the Group of Thirty ("G30"), a multinational consortium of academics, industry practitioners, and regulators released a summary of trading arrangements in the US Treasury market, identified areas of potential concern, and made some policy recommendations (Group of Thirty Working Group on Treasury Market Liquidity, 2021). The perceived need for such an analysis arose from many factors, but especially the expected increase in the issuance of US Treasury debt in connection with the COVID-19 stimulus expenditures. In short, trading

volume is likely to grow, but the market-making capacity of established dealers (mostly banks) is unlikely to keep up.

In the present discussion, one of the policy recommendations is particularly notable.

> Recommendation 8: The TRACE reporting system should be expanded to capture all transactions in US Treasury securities and Treasury repos, including those of commercial bank dealers and principal trading firms. Furthermore, subject to a cap on the disclosed size of trades, the data should be publicly disclosed in a manner similar to the way that data on corporate bond transactions are currently disclosed.

This recommendation is aimed at improving market transparency, at least bringing it up to the level of corporate bonds.

Another recommendation is aimed at expanding access to market-making capital.

> Recommendation 1: The Federal Reserve should create a Standing Repo Facility (SRF) that provides very broad access to repo financing for US Treasury securities on terms that discourage use of the facility in normal market conditions without stigmatizing its use under stress …

The "very broad access" refers to inclusion of non-bank dealers. In a sense, this is a traditional part of the Fed's mission. During the stock market "break" of October 1987, for example, the Fed encouraged banks to provide credit to securities firms, and made it clear that it would make financing available to banks for this this purpose. In the Treasury market, this financing is available to banks using repurchase agreements (repos). A bank that might otherwise be forced to sell Treasury bonds to raise capital can instead borrow from the Fed using the bonds as collateral. These repo arrangements are available to banks. The G30 recommendation is that this be expanded to other market participants.

Summary of Terms and Concepts

Dealer/over-the-counter/quote-driven markets; request for quote; prime brokerage; indicative/firm quotes; last-look; Basel III; Volcker Rule; FX pricing conventions (quote and base currencies); TRACE; on/off the run;

References

Bank for International Settlements, 2019, Triennial Central Bank Survey: Foreign exchange turnover in April 2019.

Bessembinder, Hendrik, Chester Spatt, and Kumar Venkataraman, 2020, A survey of the microstructure of fixed-income markets, *Journal of Financial and Quantitative Analysis* forthcoming.

Evans, Martin D.D., and Dagfinn Rime, 2019, Microstructure of Foreign Exchange Markets, July 29, 2019, Oxford Research Encyclopedia of Economics and Finance.

Fleming, Michael J., Bruce Mizrach, and Giang Nguyen, 2014, The Microstructure of a U.S. Treasury ECN: The BrokerTec Platform, Federal Reserve Bank of New York Staff Reports, Available at.

Group of Thirty Working Group on Treasury Market Liquidity, 2021, U.S. Treasury Markets: Steps Toward Increased Resilience, (The Group of Thirty, Washington, DC).

Harris, Lawrence E., 2015, Transaction Costs, Trade Throughs, and Riskless Principal Trading in Corporate Bond Markets, SSRN, Available at.

Hendershott, Terrence, and Ananth Madhavan, 2015, Click or Call? Auction versus Search in the Over-the-Counter Market, *The Journal of Finance* 70, 419-447.

J.P. Morgan, 2018, Trade matching and "last look" in the wholesale electronic foreign exchange and commodities markets, March 30, 2018.

King, Michael R., Carol L. Osler, and Dagfinn Rime, 2012, Foreign exchange market structure, players, and evolution, in Jessica James, Ian W. Marsh, and Lucio Sarno, eds.: *Handbook of Exchange Rates* (John Wiley & Sons, Hoboken, NJ).

Mizrach, Bruce, 2015, Research note: Analysis of corporate bond liquidity, January 18, 2020, (FINRA, Office of the Chief Economist).

Schrimpf, Andreas, and Vladyslav Sushko, 2019a, Beyond LIBOR: a primer on the new reference rates, *BIS Quarterly Review, March*.

Schrimpf, Andreas, and Vladyslav Sushko, 2019b, FX trade execution: complex and highly fragmented, *BIS Quarterly Review* 39-51.

Schrimpf, Andreas, and Vladyslav Sushko, 2019c, Sizing up global foreign exchange markets, *BIS Quarterly Review* 21-38.

Sirri, Erik R., 2014, Report on Secondary Market Trading in the Municipal Securities Market, (Municipal Securities Rulemaking Board).

Treasury, U.S. Department of the, 2016, Notice Seeking Public Comment on the Evolution of the Treasury Market Structure, (Federal Register).

U.S. Department of the Treasury, 2015, Joint staff report: The U.S. treasury market on October 15, 2014, in U.S. Department of the Treasury, Board of Governors of the Federal Reserve System, Federal Reserve Bank of New York, U.S. Securities and Exchange Commission, and U.S. Commodities Futures Trading Commission, eds.

Part III. Information and efficiency

Changes in securities prices are driven mainly by new information. The competition to act on the new information leads to market prices that incorporate the information, a desirable property known as market efficiency. The trading process is critical in the transition to the new price. It matters very much whether the information is public (known to all) or private (known to a select few). Insufficient public information impairs one of the key benefits of a financial market, its ability to produce price signals that can guide real investment. Too much private information can result in a market so overshadowed by mutual distrust that nobody will trade. The nexus of financial markets and information is an important concern for law and regulation. These laws determine the type and quantity of information that is produced, how that information is disseminated, and what kinds of information are deemed for purposes of trade to be illegal.

Chapter 10. Public Information

Investors form beliefs about securities' values largely based on public "common knowledge" information. This information set is extremely broad, ranging from public fundamental information of obvious relevance (such as the firm's financial statements) to more diffuse information that might affect investor sentiment (such as a political development in a distant country). Prices generally adjust to reflect changes in this information, and trades are often a part of the adjustment process.

The doctrine that a security price fully (and, on average, accurately) reflects all available public information is one form of the efficient market hypothesis. Violations of market efficiency raise the possibility of "incorrect" market valuations, and trading strategies that can profit by exploiting these errors. Therefore, an alternative (and more provocative) form of the efficient market hypothesis conjectures the impossibility of *consistently* outperforming the average investor ("beating the market").

The economic force driving market efficiency is competition among investors. They must accurately assess and interpret the available information, but ultimately, they must also trade, to capture the profits left by the valuation mistakes of others. Trade, or at least the potential to trade, is therefore an important part of the process. Impediments to trade can interfere with efficiency. Following on Shleifer and Vishny (1997), the limits-to-arbitrage literature has identified various imperfections. Many of these imperfections are broadly related to liquidity. The knowledgeable investors must be able to access the market, to trade at minimal cost, and to do so at a scale that can recover the costs of finding the inefficiency the first place. Trade is therefore very important to generating and sustaining market efficiency.[1]

[1] Market efficiency is so central to the practice and study of finance that most finance texts discuss it at some length (Bodie, Kane and Marcus, 2020, Chapter 11, for example) Lasse Pedersen's *Efficiently Inefficient* is a modern and accessible discussion of why markets are almost, but not quite, completely efficient (Pedersen, 2015).

This chapter begins with a discussion and example of what happens when public information is released in a scheduled announcement. Because the time of the announcement (but not the content of the announcement) is known, market participants can prepare to focus on the announcement and its implications. Then we turn to unscheduled news, in which the announcement is unforeseen and takes market participants by surprise. This case illustrates the potential of news to cause extreme volatility in prices. The sudden convergence on the market of large buy and sell flows can lead to a chaotic disorderly market. To deal with these situations we consider approaches to shutting down the market altogether, such as trading halts and price limits.

10.1. Scheduled public announcements.

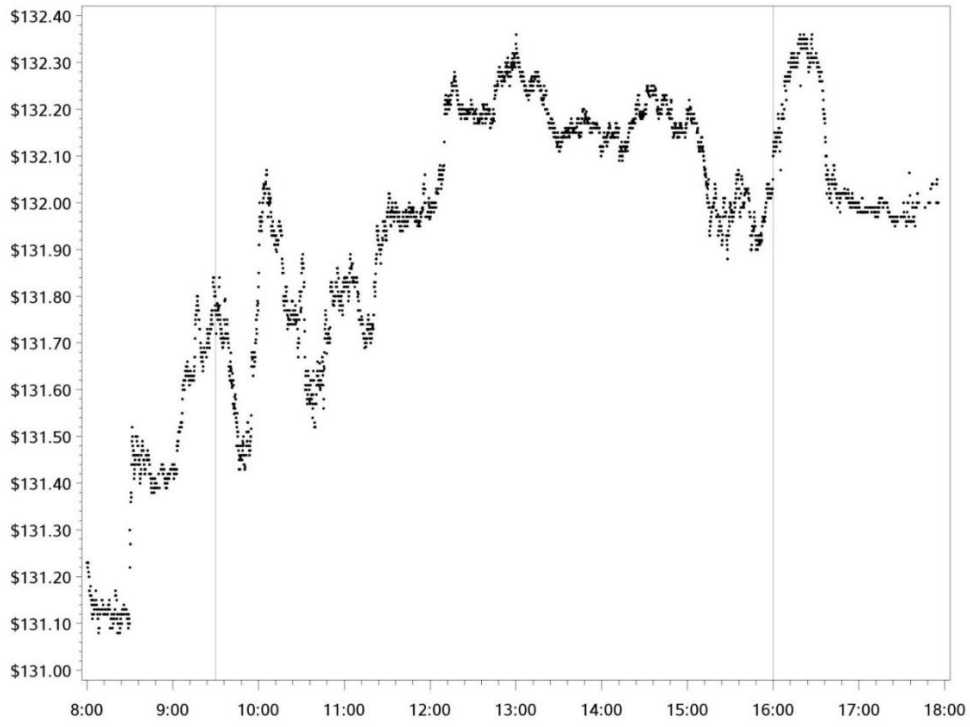
The security with US ticker symbol SPY is an exchange-traded fund (ETF). Its assets are shares in other stocks. The portfolio is designed to mirror the S&P Composite Index of 500 stocks. From an investor's viewpoint, it provides a convenient and low-cost way to hold the index portfolio. It is also attractive from a trader's perspective. Whereas a standard (open-ended) mutual fund can only be purchased or sold at daily closing prices (the net asset value), the SPY can be traded intraday, just like any other stock. The market for the SPY is very liquid: the bid-ask spread is generally $0.01 (the minimum tick); sizes at the bid and ask are large; trading volumes are large.

SPY on April 15, 2011

Figure 10-1 depicts trade prices for SPY on April 15, 2011. During many periods throughout the day, successive price changes are small and random (13:30 to about 15:00, for example). In the figure, at 8:30am, there is a sudden jump in the price. What happened?

Figure 10-1 Trades in the composite index SPDR on April 15, 2011.

The data are "thinned" for clarity. Each point marks the price of the last trade in a ten-second interval. Vertical lines demarcate the traditional trading hours (9:30am to 16:00pm).



Economic analysts in the US government and elsewhere measure economic quantities that reveal the current state of the economy. Inflation figures, industrial production, housing starts and the like are among the useful indicators. Estimates are usually released on a well-publicized schedule. "Economic calendars" that list upcoming announcements are featured on many financial web sites, including the Wall Street Journal online and www.briefing.com. April 15, 2011, was a particularly busy day. Table 10.1 is a partial record.

Table 10.1 Public announcements on April 15, 2011.

Scheduled release time	Release statistic	For	Actual	Briefing.com Consensus
8:30 (Eastern)	Consumer price index (CPI)	Mar	0.50%	0.50%
8:30	Empire Manufacturing Survey	Apr	21.7	15
9:15	Industrial Production	Mar	0.80%	0.60%
9:15	Capacity Utilization	Mar	77.40%	77.40%
9:55	Michigan Consumer Sentiment	Apr	69.6	66.5

Source: briefing.com

To connect the news to the price movements, note first that the announcement may incorporate information previously known. The information content of the announcement must be assessed relative to what market participants already knew and their previously formed beliefs. "At $1.50 per share, earnings rose 10% from the same period last year." This sounds positive, but if market participants were expecting earnings of $2.00, the announcement conveys bad news. An old Wall Street adage holds that "the baby is born." (Birth is the predictable outcome of a previously observed condition.)

For each statistic, the table gives the released estimate, and the pre-release consensus of surveyed analysts. At 8:30, there were two announcements: the consumer price index and the Empire Manufacturing Survey. The CPI figures came in on the consensus value – no surprise there. The Empire release, though, was substantially higher than the consensus value.

The Empire Manufacturing Survey is a monthly product of the Federal Reserve Bank of New York. Manufacturing firms in New York (the "Empire State") are surveyed about general business conditions. The direction of the surprise suggests that business conditions will be better than previously thought. Production, sales, and presumably earnings will be higher.

Figure 10-2 provides more detail around the 8:30 announcement. The entire time span covered by the plot is thirty seconds. Prior to the announcement, there are a few trades at the bid, and a few at the ask. Neither the bid nor the ask makes a major move, but since the bid drops by a few pennies, the bid-ask spread widens.

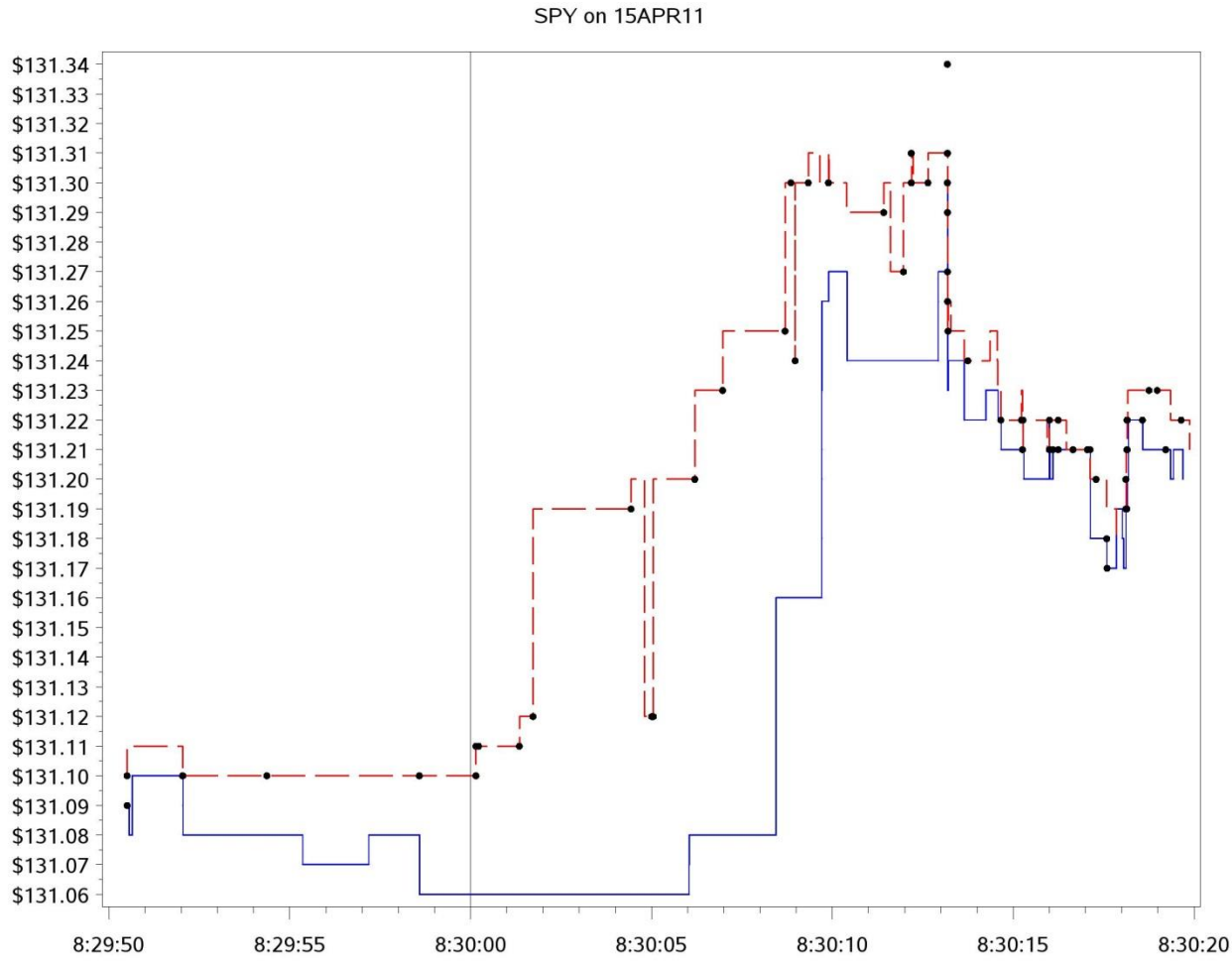
Immediately after 8:30, more marketable buy orders arrive. They take out the standing sell limit orders at $131.10, leaving $131.11 as the newly exposed ask. Additional marketable buy orders exhaust the quantities at $131.11, and those at $131.12. The next most aggressive sell orders are priced at $131.19, a jump of seven cents. (It turns out that there was a "hole" or "air pocket" in the book.) A buy at $131.19 leaves the ask at $131.20. While these orders were walking through the ask side of the book, the bid remained steady at $131.06.

The development around 8:30:05 is interesting. Suppose that someone wants a quick sale. They could hit the bid at $131.06, but pattern of recent trades suggests that the market is moving up. So, they try a limit sell order priced at $131.12. Sometimes limit orders can sit unexecuted for hours. But this one is priced so aggressively that it is hit within a fraction of a second.

After this, the flow of marketable buy orders continues, pushing prices (on bids, asks, and trades) higher. Then some sellers enter the market, and prices drop a bit. The whole set of events has played out in about fifteen seconds, and the net price change is about ten cents. All of this analysis, of course, refers to a particular example. Which features generalize?

Generally, prior to a scheduled public announcement, trading volume drops, and the bid-ask spread widens. There are many reasons for this, some of which we'll investigate later. But for the moment, it suffices to note that the period immediately subsequent to the announcement is likely to have high volatility. Someone who buys or sells prior to the announcement is bearing high risk.

Figure 10-2. Quotes and trades for SPY around 8:30 on April 15, 2011



In the adjustment to the new information, trades (executions) are not a necessary feature. The bid and ask may rise or fall together, bracketing the market's new estimation of the security's value. Usually, though, trades do occur following a news announcement, and volume is often high. There are several reasons for this.

- Traders might disagree about the importance of the information.
- Traders who established a position with the intent of betting on the impact of the announcement will unwind.
- Any kind of announcement brings the stock to people's attention.

The process of arriving at the new price, involving a complex interplay of bids, asks, and trades, is called *price discovery*. The term "discovery" emphasizes that the outcome is unknown. Although everyone might agree that the news is positive, no individual trader knows the economic value of the news. This value can only be established collectively. In economic terms, the market aggregates the heterogeneous beliefs of the participants.[2]

[2] Lee, Mucklow and Ready (1993) discuss the market dynamics around earnings announcements.

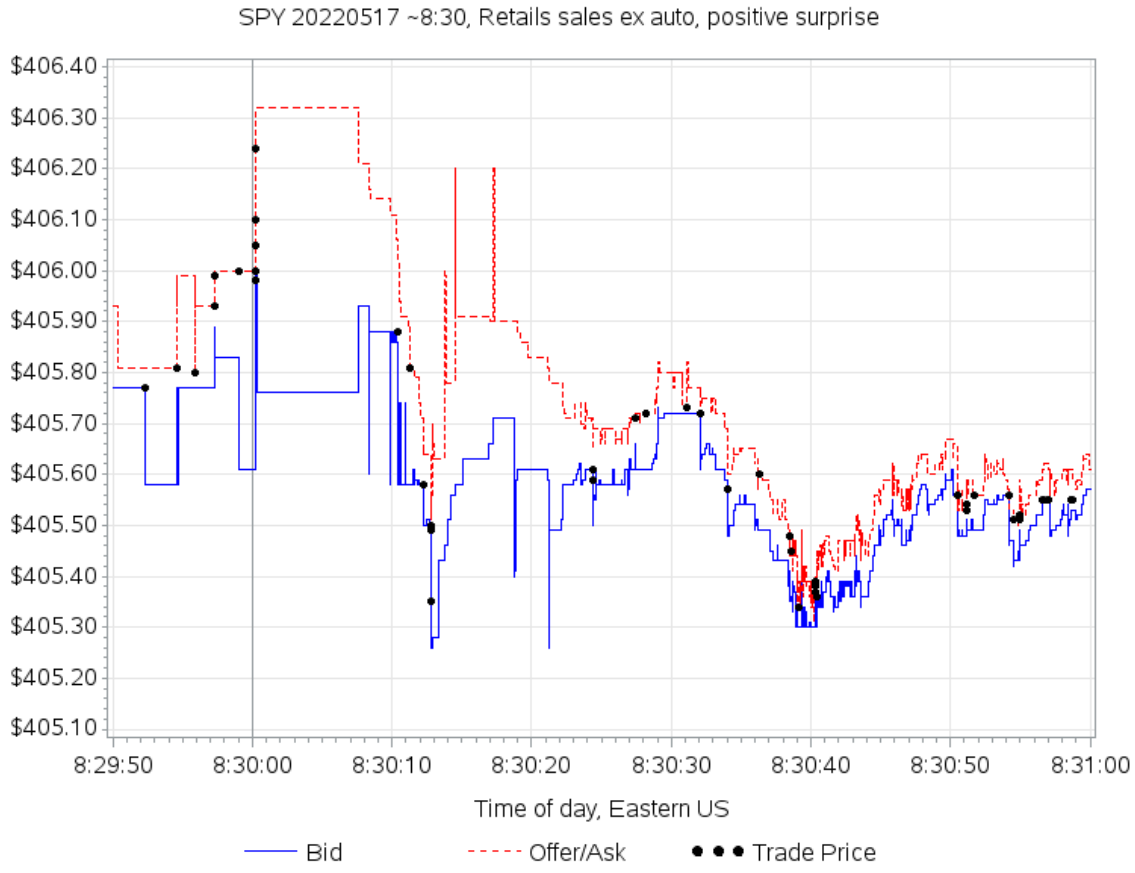
SPY on May 17, 2022

The US Census Bureau publishes a monthly estimate of retail sales growth (% relative to previous month). Two figures are released: an overall estimate and an estimate that excludes automobiles. For May 17, 2022, the briefing.com entry for that day was:

08:30 ET: **Retail Sales** For: **Apr** | Trading Impact: **High** | Actual: **0.9%** | B.com Forecast: **1.1%** | B.com Cons: **0.9%** | Prior: **1.4%** | Revised From: **0.5%** --	»
08:30 ET: **Retail Sales ex-auto** For: **Apr** | Trading Impact: **High** | Actual: **0.6%** | B.com Forecast: **0.6%** | B.com Cons: **0.3%** | Prior: **2.1%** | Revised From: **1.1%** --	»

For each release, the calendar notes the Briefing.com forecast (from some unspecified model) and a consensus forecast (based on a survey), prior to the announcement. The "actual" refers to the announced figure. For retail sales, the actual, forecast and consensus are in good agreement. For retail sales ex-auto, the actual is above the consensus, implying a positive surprise. The market activity in SPY around this time is shown in Figure 10-3. Note that: The bid-ask spread widens prior to the announcement; within milliseconds after the 8:30 announcement; the ask side of the book is repeated lifted by buyers; the spread widens more; there are many bid and ask changes, but few trades. The spread gradually narrows and there are many trades.

Figure 10-3 Market dynamics in SPY around a US Census Bureau release.



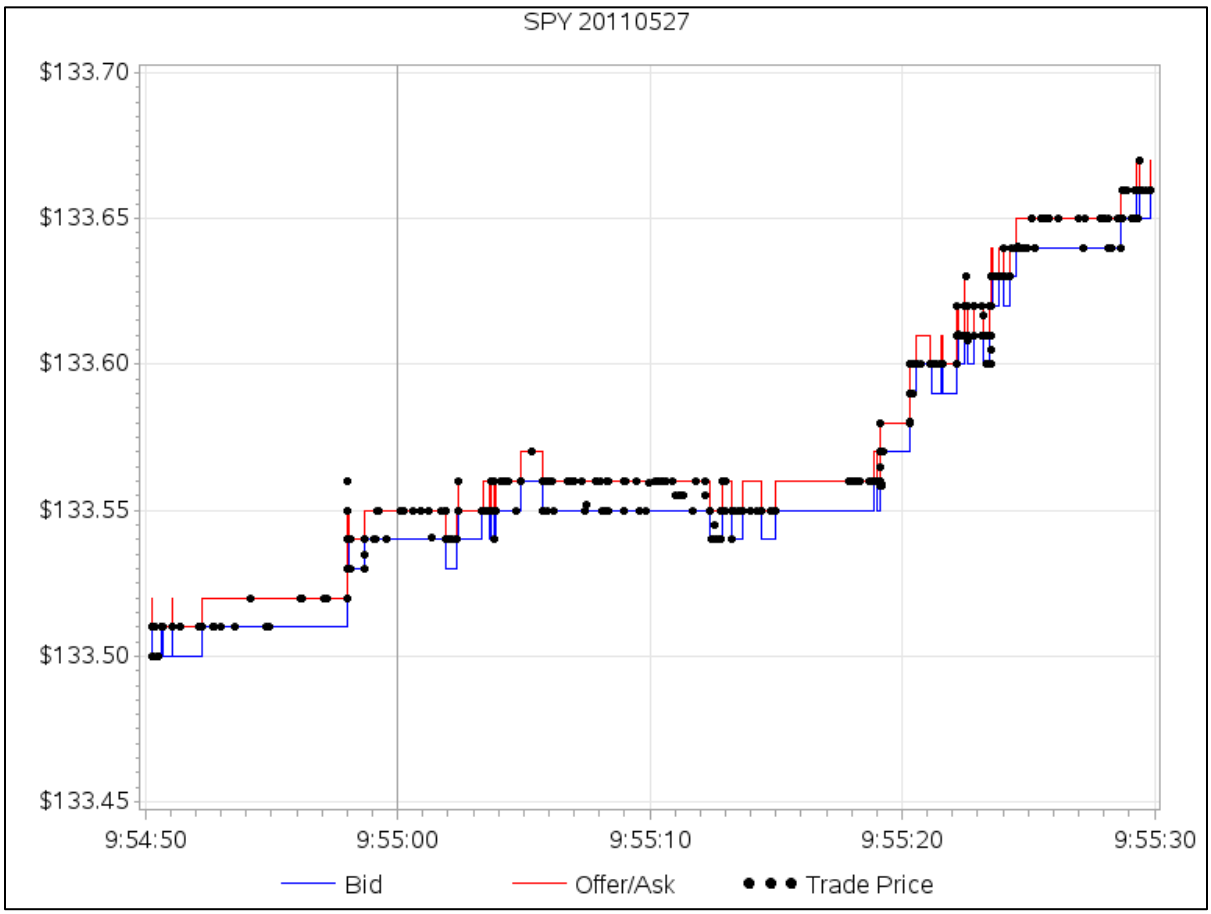
SPY on May 27, 2011. Release of the Michigan Consumer Sentiment Index

The University of Michigan compiles the Michigan Consumer Sentiment Index, a widely followed statistic that is based on surveys of consumers' planned purchases. On May 27, 2011, the Seeking Alpha website reported that

> The … final report for May [2011] came in at 74.3, an unexpected improvement over both the April final of 69.8 and the May preliminary reading of 72.4. The Briefing.com consensus expectation was had been for 72.4 and Briefing.com's own forecast was for 72.6.

Figure 10-4 depicts the price reaction. Despite the importance of the news, it is not as sharply defined as the retail order release. There are more trades lifting the offer, but there are still many hitting the bid. Nevertheless, by thirty seconds after the announcement the price has increased by around $0.10 per share.

Figure 10-4 Market dynamics in SPY around a Michigan Consumer Sentiment release.



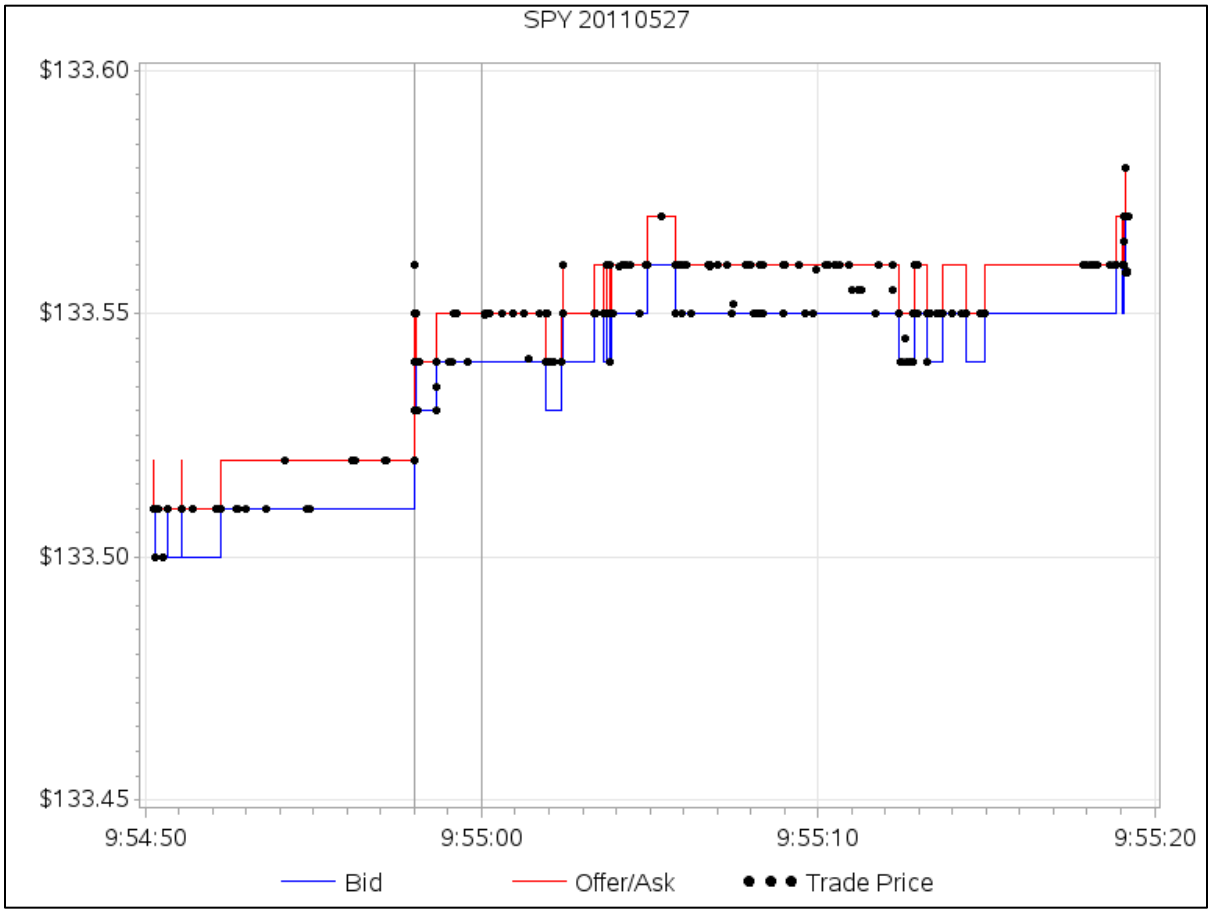
The event is notable in one other respect. Figure 10-5 shows a more detailed picture of the announcement window at a finer time scale. The first wave of buy orders lifting the offer started to arrive at 9:54:58, that is, two seconds prior to the public announcement. According to a Wall Street Journal article (Mullins, Rothfeld, McGinty and Strasburg, 2013):

> [An] early look at the consumer-sentiment findings comes from Thomson Reuters Corp. The company will pay the University of Michigan $1.1 million this year for rights to distribute the findings, according to the university. … In turn, Thomson Reuters's marketing materials say the firm offers paying clients an 'exclusive 2-second advanced feed of results…designed specifically for algorithmic trading.' Clients who pay a subscription fee to Thomson Reuters, which for some is $5,000 a month plus a $1,025 monthly connection charge, get the high-speed feed at 9:54:58 a.m. Eastern time. Those who pay for Thomson Reuters's regular news services get the report two seconds later.

Shortly after publication of the WSJ article, the University of Michigan and Reuters ceased this arrangement. Hu, Pan and Wang (2017) examine this case at length. Although the circumstances might suggest unfair advantage, Hu et al find that the "early peek" improves market efficiency and reduces volatility.

This incident illustrates the gray area between public information and private information. Two seconds does not seem like a large time span when compared to the time scales for production, consumption, and other macroeconomic processes. It can nevertheless shift the realization of trading profits and the incidence of the related costs.

Figure 10-5 Market dynamics in SPY around a Michigan Consumer Sentiment release, detail.



Because major *company* announcements often induce volatility, they are generally scheduled outside of regular trading hours (before the official market open or after the close). Of course, the force of this practice has declined over time, as trading activity has spread beyond regular hours, but the timing persists. If a company decides that an announcement must be made during regular trading hours, the company will usually notify the listing exchange. If the news is major, the listing exchange will halt trading immediately prior to the announcement.

When US companies release information, they are for the most part constrained by the Securities and Exchange Commission's Regulation FD ("Reg FD;" "Full Disclosure"). The force of this rule is that disclosures must be made public in a way that ensures that the information is available to everyone at the same time. For example, it typically prohibits management from holding private conversations with favored shareholders to give them advance knowledge of important developments.

Reg FD does not apply to the release of information by people or entities that aren't connected to the company (if any) that's the subject of their comments. This encompasses independent research firms, doing company-specific or market-wide research.

10.2. Unscheduled announcements

With an unscheduled information shock the event itself comes as a surprise to most market participants (an earthquake, for example). Most of these shocks will be broad, affecting large portions of the world economy. But unscheduled surprises can also happen in ways that are very specific to a particular company.

Acorda Therapeutics (ticker symbol: ACOR) is a NASDAQ-listed pharmaceutical firm. Figure 10-6 describes trade prices on April 14, 2011. Until about 13:10 the stock trades in a narrow range around $21.20 per share. Shortly after 13:10, the price suddenly jumps. Figure 10-7 provides detail around the initial period of transition.

Figure 10-6 Trading in Acorda, April 14, 2011

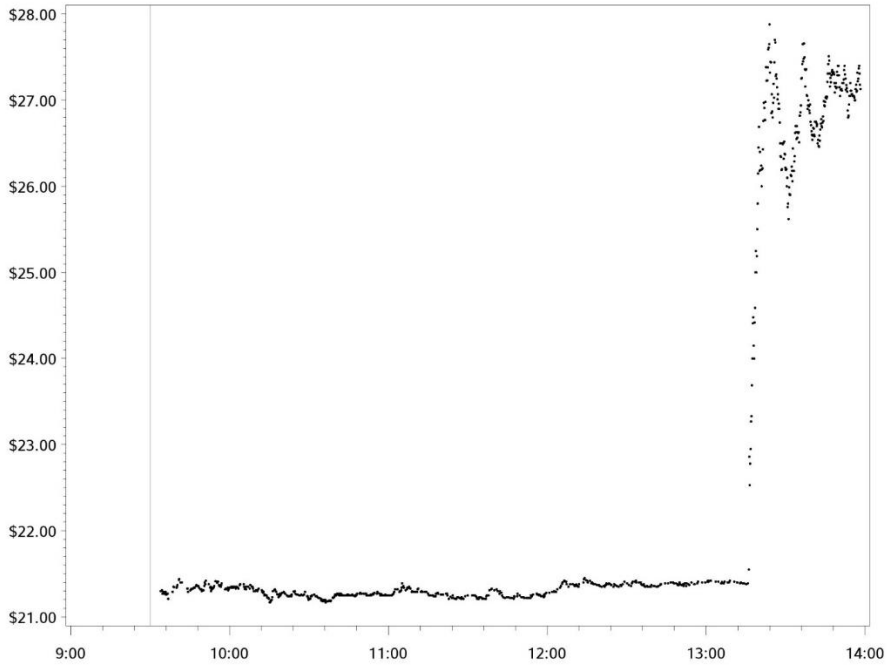


Figure 10-7 Detail

In this case, the news came not from the company itself, but from an analyst following the company. RBC Capital Markets is an investment bank affiliated with RBC (the Royal Bank of Canada). According to a posting on Xconomy (www.xconomy.com), "RBC Capital Markets speculated in a report that the patent on the company's multiple sclerosis (MS) drug, dalfampridine (Ampyra), might extend for longer than initially expected." The news is unambiguously positive, and it appears to have caught traders completely by surprise.

The price discovery process is much rougher than that for the scheduled Empire announcement. The Empire announcement moved the price of the SPY by about ten cents, and the adjustment process was over in fifteen seconds. Here, the price oscillates wildly between about $26 and $28. The price chart stops at about 14:00. At this point, NASDAQ halted trading in the security.

A pre-scheduled news release allows all potential buyers and sellers of security to coordinate their trading activities. It focuses attention, saying in effect, "pay attention to your news feeds". Absent a scheduled news release, the number of people following the market in any given stock might be very low. Without strong participation, the price discovery process may be extremely volatile.

A surprise information event usually sets up a race. Limit orders pre-existing in the book are, relative to the new information, mispriced. Alert traders will try to hit the mispriced side of the book (in the ACOR case, the ask) before the limit orders can be cancelled or repriced. This is sometimes called "picking off stale limit orders." Traders pursuing pick-off strategies typically work off of "low-latency" news feeds, augmented by text analytics. One offering claims:

> "The Dow Jones Elementized News Feeds are an ultra-low latency, XML-tagged, machine-readable news data feeds that deliver economic indicators and corporate news, with a corresponding elementized archive, into quantitative models and electronic trading programs. These innovative feeds revolutionize how news flow can be interpreted and give firms an enhanced news source for analyzing and identifying trading, investing and hedging opportunities—while moving on information in milliseconds."

(Groß-Klußmann and Hautsch, 2011) discuss stock price reactions to signals extracted from text analytics. (Aquilina, Budish and O'Neill, 2021) examine a sample of high-frequency races on the London Stock Exchange. They find that uneven access to markets imposes significant costs on relatively slow traders and suggest that these costs might be mitigated by frequent batch auctions.

10.3. Public *mis*information

The principle of market efficiency suggests that the price fully reflects the information held and believed by the market, even if that information is incorrect. This can arise from honest confusion. The stock of Zoom Video Communications, the well-known operator of videoconferencing systems (ticker symbol ZM), soared through the initial stages of the COVID-19 pandemic. The stock of Zoom Technologies (ticker symbol ZOOM) also soared, however, despite an absence of required regulatory filings. After it had risen 240%, the SEC suspended trading. Wieczner (2020) describes this and other similar instances. (Some investors apparently bought Snap Interactive, thinking that it was Snapchat. Ticker symbol FACE turned out to be Physicians Formula Holdings, not Facebook.)

The power of misinformation, though, also supports deliberately manipulative strategies. The SEC website defines one of the most common:

> "Pump and dump" schemes have two parts. In the first, promoters try to boost the price of a stock with false or misleading statements about the company. Once the stock price has been pumped up, fraudsters move on to the second part, where they seek to profit by selling their own holdings of the stock, dumping shares into the market.
>
> These schemes often occur on the Internet where it is common to see messages urging readers to buy a stock quickly. Often, the promoters will claim to have "inside" information about a development that will be positive for the stock. After these fraudsters dump their shares and stop hyping the stock, the price typically falls, and investors lose their money.

The misinformation might be appealingly topical. In a June 2020 release:

> The Securities and Exchange Commission today charged a penny stock trader in Santa Cruz, California, with conducting a fraudulent pump-and-dump scheme in the stock of a biotechnology company by making hundreds of misleading statements in an online investment forum, including a false assertion that the company had developed an "approved" COVID-19 blood test.

Answers

Aquilina, Matteo, Eric B Budish, and Peter O'Neill, 2021, Quantifying the high-frequency trading "arms race": a simple new methodology and estimates, Financial Conduct Authority, University of Chicago, Financial Conduct Authority, Available at.

Bodie, Zvi, Alex Kane, and Alan J. Marcus, 2020. *Investments, 12th edition* (McGraw Hill, New York).

Groß-Klußmann, Axel, and Nikolaus Hautsch, 2011, When machines read the news: Using automated text analytics to quantify high frequency news-implied market reactions, *Journal of Empirical Finance* 18, 321-340.

Hu, Grace Xing, Jun Pan, and Jiang Wang, 2017, Early peek advantage? Efficient price discovery with tiered information disclosure, *Journal of Financial Economics* 126, 399-421.

Lee, Charles M., Brenda Mucklow, and Mark J. Ready, 1993, Spreads, depths and the impact of earnings information: an intraday analysis, *Review of Financial Studies* 6, 345-374.

Mullins, Brody, Michael Rothfeld, Tom McGinty, and Jenny Strasburg, 2013, Traders pay for an early peek at key data, June 13, 2013, Wall Street Journal (Dow Jones).

Pedersen, Lasse Heje, 2015. *Efficiently inefficient : how smart money invests and market prices are determined* (Princeton University Press, Princeton, NJ).

Shleifer, Andrei, and Robert W. Vishny, 1997, The Limits of Arbitrage, *The Journal of finance* 52, 35-55.

Wieczner, Jen, 2020, 'ZOOM' stock halted after investors confuse it with Zoom Video stock, March 26, 2020, Fortune.

Answer to problem 10.1

Chapter 11. Circuit breakers, trading halts, and price limits

At its best the trading process involves many participants reacting thoughtfully and deliberately to the unfolding of public information. The ideal of a "fair and orderly" market comes to mind. While the meaning of these two words might be debated without end, it must be admitted that situations arise which are by common agreement anything but. Such instances make a case for temporarily closing the market. This discussion describes company-specific halts, market-wide circuit breakers, price limits, and the US limit-up limit-down procedures.

This can happen by a variety of mechanisms. Trading halts are generally news-related, reflecting information originating from or in relation to a specific company. Circuit-breakers and price limits are triggered by market price movements.

11.1. Trading halts

In the US stock market, the primary listing exchange has the responsibility of declaring a halt, which is communicated to market participants by a message sent over the quote stream. Trading is typically halted when a news announcement is pending or in process. Halts are common. Table 11.1 gives a sample.

A halt should only last long enough to ensure widespread dissemination of accurate information. Once this has occurred, the market can be reopened. This typically happens using a single-price auction similar to the daily opening auction that is run by the primary listing exchange (see Chapter 6).

Table 11.1. Current Trading Halts. Jan 25, 2012. Halt times displayed are Eastern Time (ET).

Halt Date	Halt Time	Issue Symbol	Issue Name	Reason Code	Resume Quoting	Resume Trading
01/25/2012	15:04:42	NUVA	NuVasive Inc	T1		
01/25/2012	09:19:08	GLRE	Greenlight Capital Re, Ltd.	T3	12:45:00	12:50:00
01/25/2012	09:00:38	PNNW	Pennichuck Corporation	T12		
01/25/2012	08:10:27	INCB	Indiana Community Bancorp	T3	08:40:00	08:45:00
12/19/2011	13:29:39	FEED	AgFeed Industries, Inc.	T12		
11/29/2011	07:01:23	BQI	Oilsands Quest Inc	T2		

Notes: The reason codes are: T1 (a news release is pending); T2 (a news release is in process); T3 (trading will resume shortly); T12 (NASDAQ is requesting additional information).
Source: www.Nasdaqtrader.com

11.2. Market-Wide Circuit Breakers (MWCBs)

Circuit breakers are market-wide trading halts triggered by declines in the S&P 500 Index relative to the prior day's close. The current SEC information bulletin states (U.S. Securities and Exchange Commission, 2016):

> A cross-market trading halt can be triggered at three circuit breaker thresholds—7% (Level 1), 13% (Level 2), and 20% (Level 3). These triggers are set by the markets at point levels that are calculated daily based on the prior day's closing price of the S&P 500 Index.
>
> A market decline that triggers a Level 1 or Level 2 circuit breaker before 3:25 p.m. will halt market-wide trading for 15 minutes, while a similar market decline "at or after" 3:25 p.m. will not halt market-wide trading. A market decline that triggers a Level 3 circuit breaker, at any time during the trading day, will halt market-wide trading for the remainder of the trading day.

The index levels that would trigger the MWCBs are updated daily at https://www.nasdaqtrader.com/trader.aspx?id=CircuitBreaker. Once triggered, if the halt does not extend through the end of the trading day, the individual stocks reopen with the usual auction procedures.

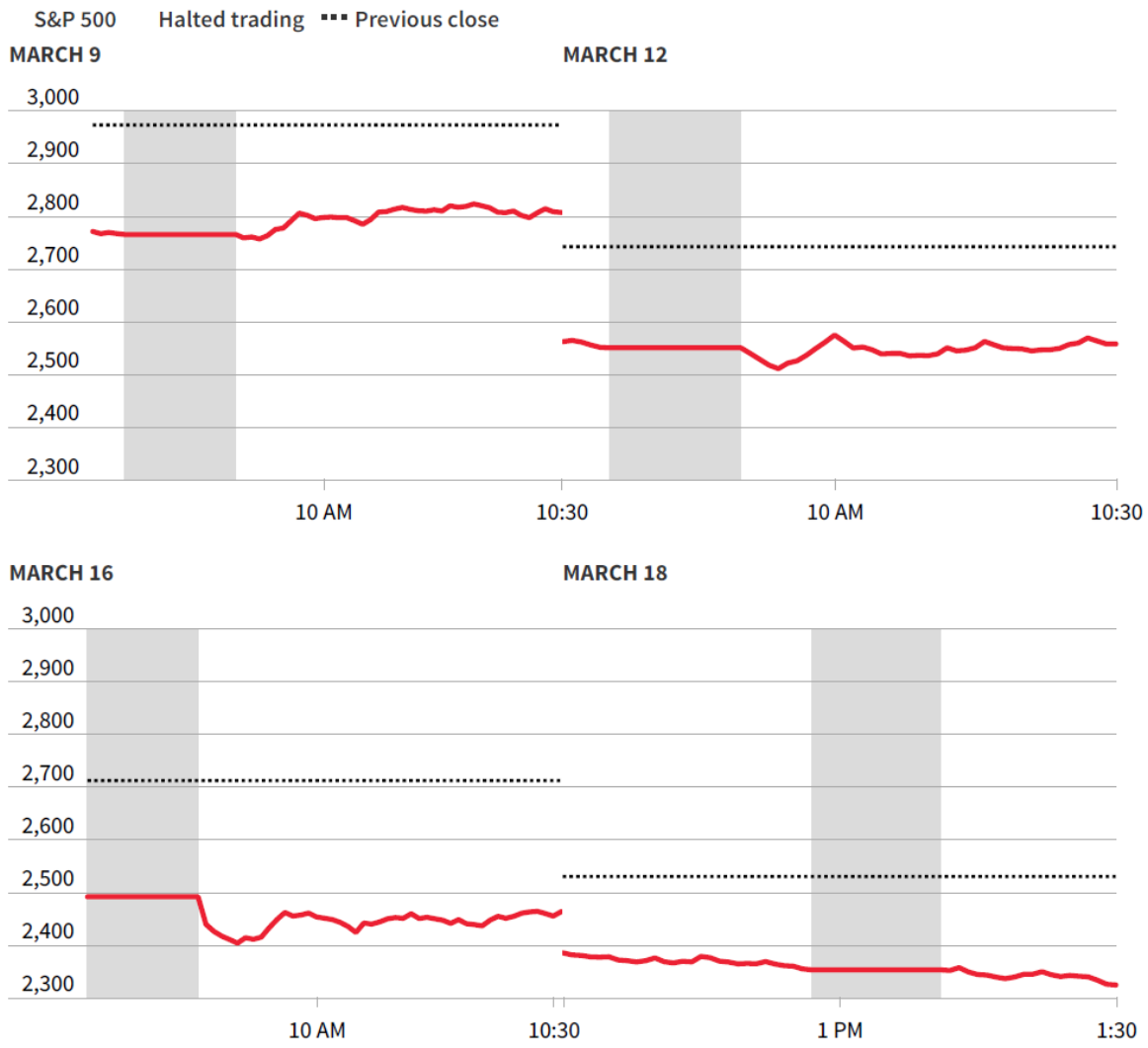
The market-wide circuit-breakers were first instituted in response to the October 1987 "market break," a sudden price decline across virtually all stocks. They were triggered for the first time about ten years later, on October 27, 1997. At that time the consensus was that they had been triggered prematurely, and the trigger levels were lowered, requiring a more extreme drop before being triggered.

The MWCBs stood by, unused and largely unremarked, until March 2020. In that month they were triggered four times. Figure 11-1 depicts the level of the S&P index around the closures. Did the MWCBs work to prevent, or limit the extent of, further declines? It is impossible to say with certainty. There are only four events. We don't know what would have happened without the MWCBs. With the possible exception of March 9, though, it does not appear that the halts provided "support" to the market.

Notice, though, that three of the four episodes occurred at or near the market open. In this connection Connault makes an important point (Connault, 2020a, b). The S&P index is a weighted average of recent trade prices. Before a stock has opened, the most recent trade is the prior day's close. Since this is the point of reference, until the stock has opened there is no indication of a price change, even though all market participants might agree that the value has dropped. As the individual stocks open, their lower prices are incorporated into the index, and the index falls, eventually triggering the MWCB.

Now how do we know that prices are falling if the individual stocks haven't opened? Connault suggests that we look at the index securities, the S&P 500 Exchange Traded Fund (ETF, ticker symbol SPY) or the stock index futures contract. Both of these are actively trading as of 8:30, that is, an hour before the individual stocks open. Both were down significantly from their previous day's close. Connault accordingly suggests that these securities be considered as indicators of when to trigger the halt.

Figure 11-1 Market-wide circuit breakers in March 2020



Source: (Funakoshi and Hartman, 2020)

11.3. Price limits

Halts and market-wide circuit breakers pull the plug: trading does not occur. Price limits specify a range of prices. Within this range, trades are permitted; outside of the range, trades are prohibited. Price limits are temporary. The market will periodically (typically daily) allow the limits to adjust.

For example, the CME lists a lumber futures contract. On August 3, 2021, the near contract (maturing in September) closed at a price of $606.50. The price limits for the following day are set as $606.50 $\pm$ $7.00, that is, $599.50 to $613.50. In April 2021, as the economy emerged from the pandemic, the daily price limits were hit in nine trading sessions (Dezember, 2021).[1]

[1] The size of the contract is 110,000 board feet. One board foot is equivalent to a piece of lumber that is $(one\ foot) \times (one\ foot) \times (one\ inch)$. Contract prices are dollars per 1,000 board feet. At a price of $606.50, the value of the lumber underlying the contract is about $66,715.

Futures price limits are symmetric. Whereas the stock market's MWCBs are triggered only on declines, price limits can bind on the upside as well. One might rationalize the difference as follows. The stock market is operated to enhance capital formation and long-term investment. Rising stock prices encourage both activities. The futures market, though, exists to facilitate hedging, the transfer of risk between oppositely situated parties. A home builder is exposed to the risk that the price of lumber rises; a sawmill suffers if the price of lumber falls. A one-sided price limit might hurt one or the other. Although this reasoning has some appeal, it is somewhat misleading. Hedging is not unique to the futures market; it occurs in the stock market as well. In fact, we will shortly encounter a type of price limit used in the stock market that is symmetric.

When a futures contract nears maturity, some trades simply can't be postponed. Traders who had no intention of making or taking delivery of the underlying need to close their long positions (sell) and cover their short positions (buy). A long-term investor needs to roll over the expiring contract, by selling it and simultaneously buying the next near contract. For these reasons, the CME removes price limits for trades occurring in the maturity month.

The CME futures exchange uses daily price limits for most of their agricultural and currency contracts (lumber and Japanese yen, for example). Energy, metals, and interest rate contracts use *Dynamic Circuit Breakers*, in which the limits are reset within the day. The upper and lower limits are constructed from trade, bid, and ask prices over a one-hour lookback window. A full description (and video) are posted at https://www.cmegroup.com/globex/trade-on-cme-globex/frequently-asked-questions-dynamic-circuit-breakers.html.

11.4. Limit Up Limit Down (LULD)

The US stock market's MWCBs and the US futures markets' regular price limits are set once per day, at the beginning of the day, and relative to the previous day's final price. The approaches discussed in this section are more reactive: the limits can be revised within the day. These revised limits are pegged to an average of trade prices over some short prior interval.

LULD applies to individual US stocks. The procedures are implemented as coordinated rules across US market centers, not (as one might expect) as an SEC Rule. It is very much a group effort, and it is referred to as the LULD Plan (www.luldplan.com).

The procedures are somewhat complex. A simplified description follows. The price limits are set around a quantity known as the *reference price*. At the start of the day the reference price is the previous day's closing price. After that, the reference price is the average of trades reported in the preceding five minutes. If there are no trades in this interval, the previous reference price is used.

The procedures are designed to prevent trades from occurring outside of a range centered around the reference price. The upper and lower bands (limits) of this range are set as the reference price $\pm$ 5%. Normally, the National Best Bid and Offer (NBBO) are within the reference band (as in Figure 11-2).

Although the reference price is an average of trade prices, most of the procedures relate to bid and ask quotes. Specifically, for example, if the National Best Bid (NBB) drops below the lower band, it is flagged as *unexecutable* (Figure 11-3). In the figure, the National Best Offer (NBO) is within the reference band. It is therefore executable. This is a *straddle state*.

If the NBO then drops to the lower limit, it is said to be a *limit quote* and we enter a *limit state* (Figure 11-4). Entry into a limit state triggers the start of a 15-second clock. If within this

---

Alternatively, the $\pm$\$7.00 price limit corresponds to a maximum daily gain or loss of \$770. This statement, though, seems to imply that price limit controls volatility in some fundamental way. In fact, when positions are marked to the true (unconstrained) price, gains and losses may be much larger. Full contract specifications are given on the CME website and the CME Rulebook.

period the limit quote is (fully) executed or withdrawn, we return to the straddle state. If the clock runs out, the market enters a five-minute trading pause. If the Limit State offer persists, the pause is extended for another five minutes. At the end of the second pause the primary listing exchange may reopen the stock, using the normal opening procedures. If the second five-minute pause extends beyond the normal close of the market, the primary listing exchange will operate its normal closing auction procedures.

Note that a limit state is triggered by the NBO dropping *to* the lower band, not "to or through." The NBO can't drop through the lower limit because coordinated rules across market centers prohibit display of an offer below the lower band. (See, for example, NASDAQ LULD FAQ, NYSE Rule 7.11.)

In a sense, the bid and ask quotes are being used as indicative prices. If, in the limit state described above, the ask quote rises above the lower bound, this suggests that at least some sellers believe the stock is worth more, and the limit state is terminated. The rise in the ask quote may result from execution or cancellation of asks at the lower bound. The trading halt/pause is to be avoided, if possible, and the fifteen-second window gives an opportunity for the price decline to reverse. If, on the other hand, the asks that triggered the limit state persist through the two five-minute pauses, this suggests that the price decline is unlikely to reverse. In this case a stronger remedy (a reopening) should occur. In a sense, the jaws of the price limits do not snap shut, but close gradually.

Figure 11-2. Reference band in normal conditions

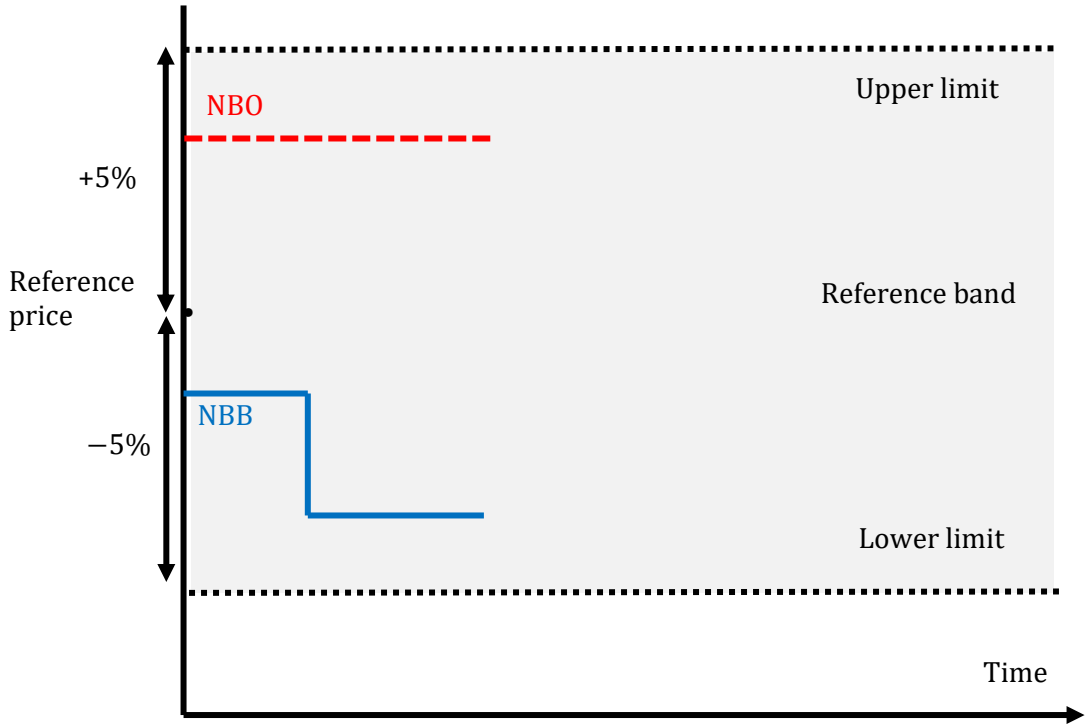


Figure 11-3. An unexecutable bid in a straddle state

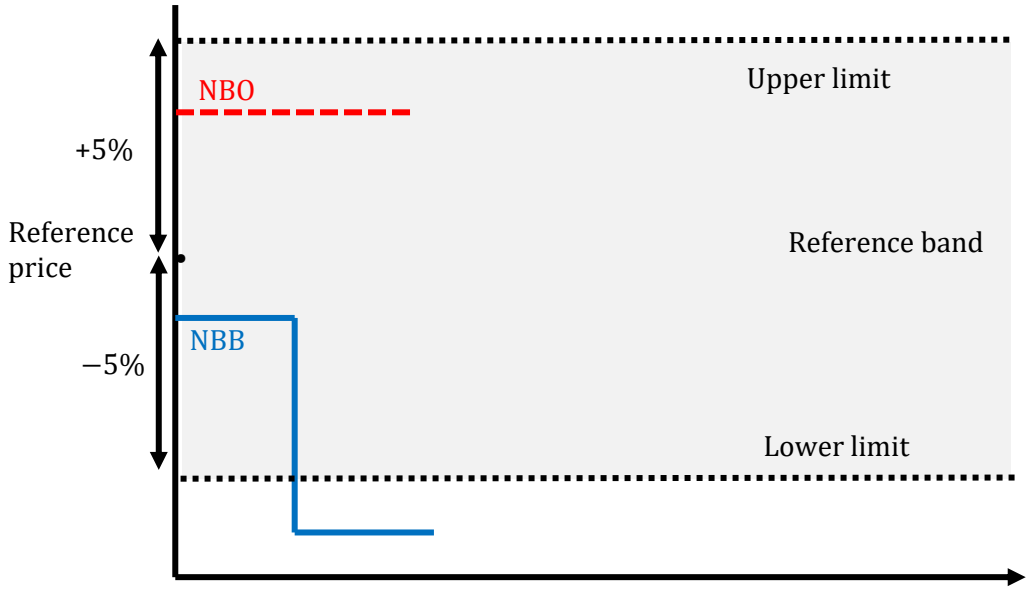
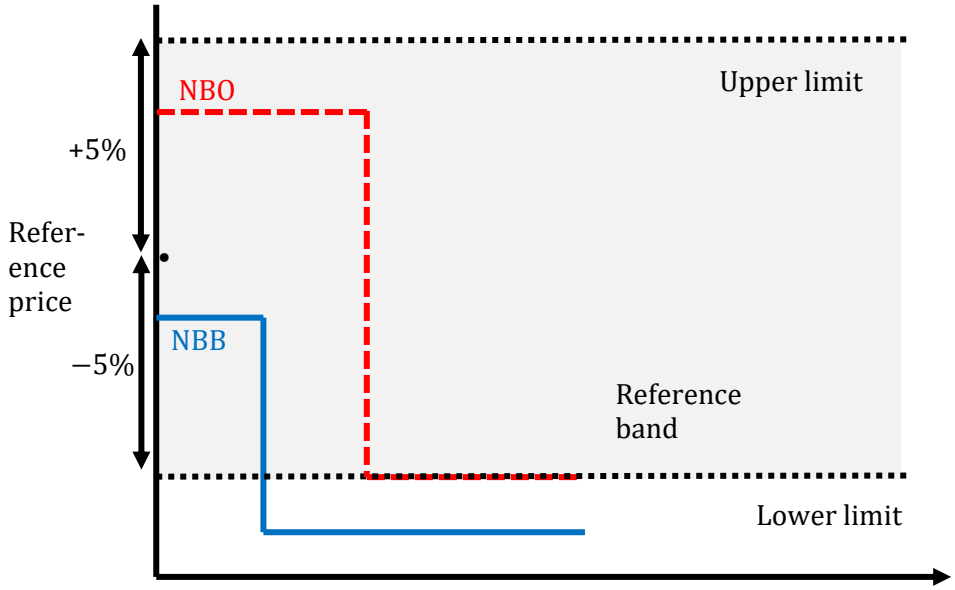


Figure 11-4. A limit state



The example underlying this discussion involves a price decline. The procedures are completely symmetric. In the case of a price rise, the treatment would have started, "If the National Best Offer (NBO) rises above the upper band, it is flagged as unexecutable," and so on.

There are some simplifications invoked in the discussion. The ±5% offsets apply to Tier-1 ("actively traded") stocks and previous day closing prices above three dollars. Below three dollars, the offsets are larger. A fuller description is posted at luldplan.com. Moise (2021) also documents the procedures and evaluates their effectiveness.

11.5. Discussion

Trading halts and price limits are controversial.
The arguments for:

- A sudden news announcement is unfair to limit-order traders. It can make what was a relatively level playing field very uneven. It is better to halt trading, at least for a few minutes, to allow everyone the chance to see and react to the new information.
- The loss in efficiency from a short trading halt might be very small. (Does it really matter if the information reflected in prices is a few minutes old?)
- The market is more liquid if the bid-ask spread is tight, i.e., if buyers and sellers are posting aggressive limit orders. If a trader thinks he'll be picked off, he'll price his order less aggressively. If everyone does this the bid-ask spread widens.

The arguments against:

- A trade is a voluntary act between two consenting parties, with no effects on anyone else. From a Libertarian perspective, a trading halt is an arbitrary prohibition.
- Markets are most efficient when prices reflect a free flow of information. Trading halts impede the market's reaction and therefore impair efficiency.

The effects on market volatility are complex and ambiguous. Obviously, while a trading halt is in effect reported market prices don't change. Perceptions of fundamental values, however, can't be prevented from changing, and during a halt they are based on less information than usual. Moreover, if traders believe that a trading halt is imminent, they might accelerate their trading plans, aggravating volatility. For example, if the price has dropped to the neighborhood of a lower price limit, someone contemplating a sale might act quickly out of concern that the ability to sell will be shortly taken away. This leads to accelerated declines when the price nears the limit, a phenomenon that is called "gravitational pull." Wong, Kong and Li (2020) find evidence of this dynamic for price limits in the Chinese stock market.

In recent years, however, market volatility caused by wayward algorithms has emerged as the overriding concern. In the absence of any other practical way to contain these programs, price-based halts remain regulators' first choice.

Summary of terms and concepts

Limit-up/limit-down procedures; market-wide circuit breakers (MWCBs); timing of public announcements; trading halts; re-openings.

References

Connault, Benjamin, 2020a, Price Discovery During Market Wide Circuit Breakers (Part 1), IEX, Available at: https://medium.com/boxes-and-lines/price-discovery-during-market-wide-circuit-breakers-part-1-88587efb0957.

Connault, Benjamin, 2020b, Price Discovery During Market Wide Circuit Breakers (Part 2), IEX, Available at: https://medium.com/boxes-and-lines/price-discovery-during-market-wide-circuit-breakers-part-2-98d3194fa4f0.

Dezember, Ryan, 2021, Lumber Prices Reach High, Boosting Sawmill Profits, 05/04/2021 May 04, Wall Street Journal (New York, N.Y.).

Funakoshi, Minami, and Travis Hartman, 2020, March Madness, March 18, 2020, (Reuters Graphics).

Moise, Claudia E., 2021, Circuit breakers and the COVID-19 crisis, Fuqua School of Business, Duke University, Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3768937.

U.S. Securities and Exchange Commission, 2016, Investor Bulletin: Measures to Address Market Volatility, January 4, 2016.

Wong, Kin Ming, Xiao Wei Kong, and Min Li, 2020, The magnet effect of circuit breakers and its interactions with price limits, *Pacific-Basin Finance Journal* 61, 101325.

Chapter 12. Securities Class Action Lawsuits.

The last chapter looked at how markets react to newly arriving information. But the importance of public informational efficiency goes beyond the market adjustment process, a dynamic that mostly plays out over a few minutes. The connection between market prices and information is crucial to investors who use the information and those (usually corporate managers and accountants) who produce it. This chapter takes a brief look at the role of market efficiency in the regulation of information and the resolution of disputes involving information. The subject is a complex one, however, and this overview will just touch on a few of largest concepts.

In the early 20th century, there was little national securities regulation. Most US rules were adopted by the individual states. Then, as now, government was greatly concerned with protecting investors from unscrupulous promoters of dubious investment schemes. Because they sought to eliminate the sale of securities based on empty promises, the state laws that emerged were (and still are) known as "Blue Sky Laws". Protecting the public from bad investments, though, turned out to be much more complicated than it seemed at first. Could government really evaluate the merits of new securities? By the time that the first federal securities laws were being framed in the 1930s, it seemed better to settle for a more modest aim: full disclosure. The difference between merit and disclosure is important. When Apple Computer first sold stock, they did not have to convince a regulator that the securities would be a good investment. They only had to establish that they were openly disclosing the risks and uncertainties.

The principle of disclosure arises in many places in US regulation, but perhaps the most well-known statement occurs in the 1934 Securities Act (§240.10b-5):

> It shall be unlawful for any person, directly or indirectly, by the use of any means or instrumentality of interstate commerce, or of the mails or of any facility of any national securities exchange,
>
> (a) To employ any device, scheme, or artifice to defraud,
>
> (b) To make any untrue statement of a material fact or to omit to state a material fact necessary in order to make the statements made, in the light of the circumstances under which they were made, not misleading, or
>
> (c) To engage in any act, practice, or course of business which operates or would operate as a fraud or deceit upon any person,
>
> in connection with the purchase or sale of any security.

This law has far-reaching implications. Although it seems at first to deal only with public information (what we say or don't say), it in fact underlies a lot of the private information (insider trading) laws we'll be considering later. It is also worth pointing out that the 1934 Act applies mostly to stocks that are already trading (that is, in the secondary market). A similar and generally stronger provision of the 1933 Act applies to the initial public offering (the primary market).

The wording of this law has been subjected to extremely careful examination and interpretation by lawyers, experts, and courts, and the volume of commentary is substantial. For the moment, though, it suffices to draw out one phrase, "It shall be unlawful … to make any untrue statement of a material fact or to omit to state a material fact …" Simply put management has to tell the truth. What makes a fact "material"? One common formulation is that a reasonable investor would consider it important in considering the purchase or sale of the securities. That is, the difference between knowing it or not might cause an investor to view the stock very differently.

The disclosure obligation has some important exceptions. A firm need not publicly disclose its legitimate business secrets, such as technical information that might be useful to competitors. Activities that are questionable or even illegal are not necessarily exempt from the requirement. In September 2015, a suit was filed against Volkswagen AG, alleging failure to disclose facts connected to the emission-testing defeat devices. The alleged failure to disclose (to investors) the activity is an offense quite distinct from the activity itself.

As with many securities laws, the primary responsibility for disclosure enforcement lies with the SEC. The SEC, for example, recently announced an investigation of whether Chrysler Fiat was inflating its monthly sales figures, (Boudette, 2016). Rule 10b-5, however, also has a *private right of action*. People who bought or sold the securities can bring lawsuits on their own, seeking to recover *damages* from the firm, without relying on the SEC. This has given rise to many lawsuits, sometimes known simply as "10b-5" cases.

To simplify matters somewhat, a 10b-5 situation arises when a corporation makes some misstatement (a lie or a lie of omission). An investor buys stock in the company. At some later time, the company admits the lie, and the stock price drops, leaving the investor with a loss.

The investor then sues the company. Crucially, at this point, the investor is suing not just to recover her own losses, but also on behalf of a class of all other purchasers in the same time period. This *class action* greatly magnifies the damages sought in the suit. The suit is brought in a Federal court. The investor is the plaintiff, and the company is the defendant. Both sides retain economic and accounting experts. The case then proceeds through the court system. Typically, though, the final outcome is not a judicial verdict. Instead, both sides reach a settlement. In the

settlement, the defendant company pays the plaintiff's lawyers, the plaintiff's experts, and damages to the investors.[1] The sums involved can be large: Enron settled for over $7 Billion. The Securities Class Action Clearinghouse at Stanford University (in collaboration with Cornerstone Research) maintains a database of the filings (http://securities.stanford.edu).

12.1. Informational efficiency

The importance of efficiency in these cases can be seen by comparing things before and after the concept of efficiency took hold. Although many ideas related to market efficiency go back a hundred years or more, it was formalized as an economic principle in the 1960's and 1970's. Shortly thereafter the idea entered the legal arena. In arguing the legal case for market efficiency, (Fischel, 1982) captures the essence of the transition to our present view.

By the time of Fischel's paper, the courts had adopted a set of criteria that 10b-5 plaintiffs had to satisfy:

- Materiality. Would a reasonable investor consider misstated or omitted fact important in making an investment decision?
- Reliance. Did the plaintiff actually rely on the misstated fact (or would have relied on the omitted fact) in making the purchase decision?
- Causation. Did the misstated/omitted fact cause the economic loss suffered by the plaintiff?
- Damages. What are the losses (dollar amounts) that can be attributed to the misstated/omitted fact?

These could be difficult things to establish. An investor might have, for example, relied on a thorough financial analysis of the company, perhaps using a valuation model in which the misstatement could be reduced to a single input. The damages could be calculated as the difference between the intrinsic values computed with the wrong and right inputs. But valuation models can be subjective, and it is likely that the defendant could find an expert who would determine the misvaluation to be negligible. Would every plaintiff have to produce her own financial analysis? Even if the analysis could be rigorously documented, would the damages for a purchaser of 100 shares, or even 100,000 shares, be enough to cover the attorney's costs? These considerations established high barriers to the pursuit of these cases.

But if the market is efficient, reliance can be established by simply appealing to the market price. A purchaser of the stock might not have used any sort of valuation model at all, but he certainly would have considered the market price of stock, at least at the moment when he traded. Assuming efficiency, this price would have reflected all public information, including the misinformation alleged by the plaintiff. It is not necessary for the investor to demonstrate direct reliance on the misstatement. It suffices to establish indirect reliance, via the market price. In the new view, someone making a misstatement is causing a *fraud on the market*. This presumption of efficiency, and indirect reliance on the price, defines a broad class of investors, essentially anyone who purchases the stock while the misstatement is operative.

Market efficiency dramatically simplified the establishment of reliance. The concept also had implications for materiality, causation, and damages, but these results are not as sweeping. An investor cannot simply claim, "I bought at the efficient price; I sold at the efficient price; the company lied; the company should reimburse my losses." To establish materiality, causation and damages, the presumption of efficiency certainly helps, but it is not enough.

The following discussion is drawn from Fischel (1982), Dunbar and Tabak (1999), and Feitzinger (2014).

[1] By way of disclosure, I was retained and paid as a consultant in one such case.

12.2. The Framework

There is usually little argument about what the company told its investors: accounting statements, SEC filings, press releases and so forth are a matter of public record. The questions of what management knew and what should have been told to investors, on the other hand, are frequently open to disagreement. Management's information can sometimes be reconstructed from emails and other internal communications, so these things are often requested during the discovery phase of the legal process. Defining exactly what management should have disclosed and when is often a question of accounting practice, and expert opinions here can be clarifying. In any case, the information available to investors vs. the information that should have been available must be established at the outset. We'll assume that the misstatement occurred on day 100, with the release of an erroneous accounting statement and ended on day 400, when the company admitted the error (the *corrective disclosure*).

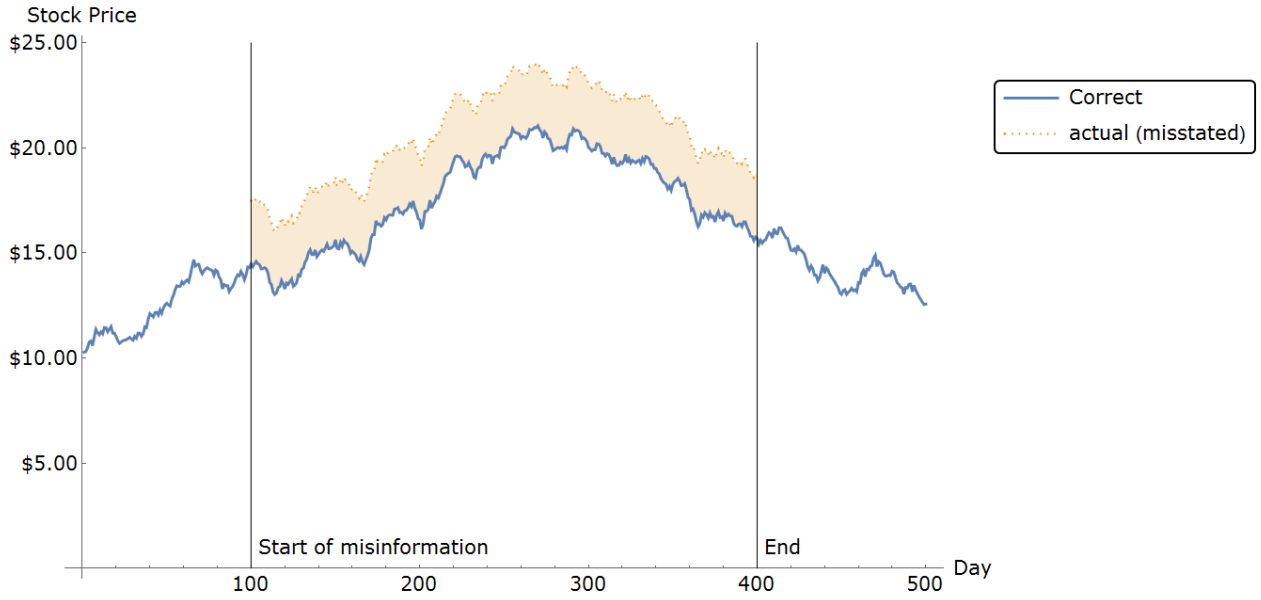
The plaintiff *class* is all investors who purchased the stock over the *class period*, the span of time covering the misstatement. When the suit is initially filed it must name a real purchaser, identified by name. Other purchasers are included by reference, as in, "Joel Hasbrouck, and all others similarly situated." The defendants comprise the corporation, and often managers and directors who are named individually. The defense usually immediately files a *motion for dismissal*, and the plaintiffs indicate how they will pursue their case. If the motion for dismissal is rejected, the case proceeds to the *discovery* phase. In discovery, the plaintiffs may obtain access to internal corporate documents such as emails and memoranda relevant to the case. The plaintiffs use this information to build an informational timeline of what management knew, and what they told (or didn't tell) the investors.

The principles of market efficiency come into play when we try to connect this information to prices. Ideally, we'd like to assemble a picture like Figure 12-1. This graph shows the actual stock price and a price that we think would have reflected the correct information. Appealing to market efficiency, we might then assert that the actual stock price includes the effect of the misinformation, and the shaded difference between the two lines is a measure of the inflation due to misinformation. Suppose that the misinformation caused the value of the stock to be inflated by $3 per share, beginning on day 100 and ending on day 400.

If we could construct such a picture, we could determine the effect of the misstatements on any investor, based on when the investor bought and sold. Let's look at some of possibilities. Suppose that Amy bought a share on day 290 for $23.61 and sold on day 450 for $13.03. She lost $10.58 on her investment, and since she bought at a price that reflected the error and sold after the error was corrected, she might well believe that she deserves to recover her entire loss. In fact, though, only $3 of her loss is caused by the error (and its subsequent correction). Her total loss was larger due to other factors besides the error.

Suppose instead that Amy had sold on day 350 for $21.27, for a loss of $2.34. This is a poor outcome, but none of the loss can be considered to have been caused by the reporting error. Essentially, she bought at an inflated price and sold at an inflated price.

Figure 12-1



Suppose that Brian purchased the stock on day 80 (prior to the error) for $14.12 and sold on day 450 for $13.03. He, too, has a loss (of $1.09), but since neither his purchase nor his sale occurred on a day when the stock price was inflated, he is not entitled to any damages.

Some investors might have benefitted from the error. Suppose that Cathy purchased stock on day 99 (before the error) for $14.30 and sold at on day 100 for $17.49. Her profit is $3.19, of which $3 is directly attributable to the error. Does she have to give up the $3? Usually, no: she was an *innocent beneficiary* of the error.

These examples are simplified because the price inflation attributable to the accounting error is a constant $3 and the start and finish times are precisely defined. In practice, though, the inflation might not be constant over time. The start and stop times might also be blurry. The accounting error might have arisen from an incorrect process that was adopted gradually, and the corrective disclosure might have occurred in the form of several pronouncements, spread out over days or weeks. Both the start of the error and the corrective disclosure are likely to be entangled with other announcements or market developments. In attempting to unravel these effects, where do we begin?

12.3. Event studies: an overview

Suppose that the corrective disclosure occurs in the form of a single announcement, released after the close of regular trading hours. Daily stock data usually comes in the form of closing prices, so the close-to-close return spanning the disclosure will run from the market close on day 399 to the closing price on day 400.

We can't simply attribute the entire day's return to the disclosure, however. Other information will have affected the price of the stock. Most stocks, for example, have returns that are driven in large part by the returns on the broader market. To separate the effect of the corrective disclosure from the market return, a good place to the start is the market model, also known as the single-index model.

The market model is a statistical model that is based on simple linear regression. We'll label the stock that we're trying to model as stock i, and let m refer to the overall market (in

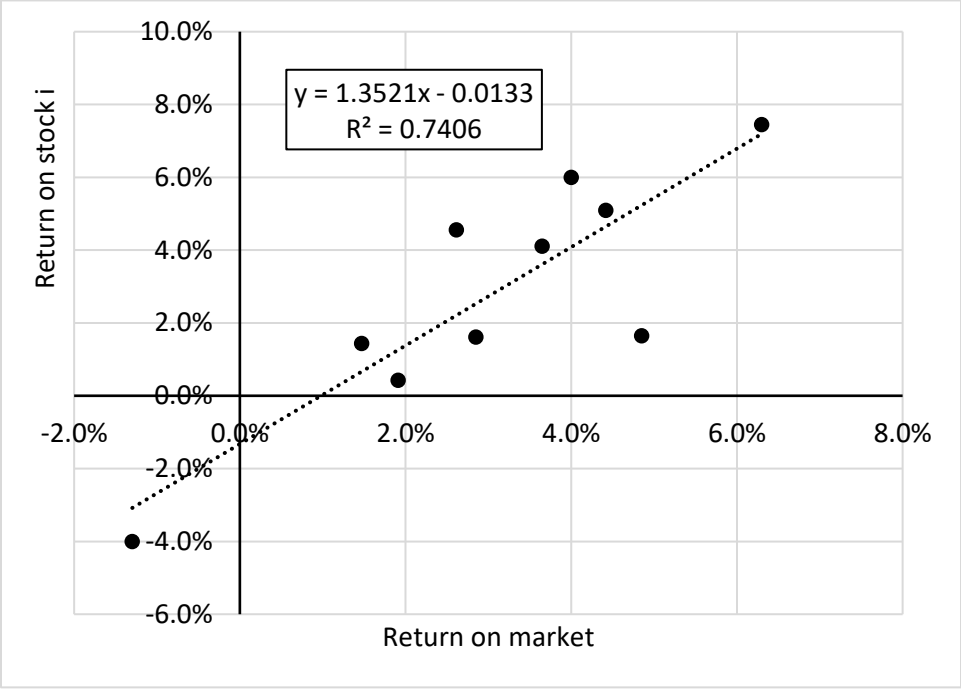
practice, a market index, like the S&P Composite). Then the return for stock i on day t is r_{it}. It is related to the market return for the day by:

$$r_{it} = \alpha_i + \beta_i \times r_{Mt} + e_{it}$$

This breaks r_{it} into three pieces. α_i is a constant "starting point" for the return. $\beta_i \times r_{Mt}$ is the market-related piece: β_i is a multiplier, usually a number close to one. Although β_i is constant for a given stock, it multiples r_{Mt}, which is random and different every day. The final piece, e_{it}, captures the part of the stocks return that is *not* related to the market. Company-specific news (like a corrective disclosure) shows up in e_{it}.

The market model is essentially a best-fit line through a scatterplot of returns. To construct the scatterplot, we first compile a sample of returns, observations for r_{it} and r_{Mt} for a number of days. The pair of returns on day t defines a data point on a graph in which r_{Mt} is on the horizontal axis and r_{it} is on the vertical. The alpha is the intercept of the best-fit line; the beta is the slope. Figure 12-2 depicts an example.

Figure 12-2



The single-index model is usually introduced in connection with portfolio theory, the science of combining many risky securities (like stocks) to achieve the best possible trade-off between risk and the return. In portfolio applications, we're almost always primarily interested in the stock's beta, which measures risk relative to the market.

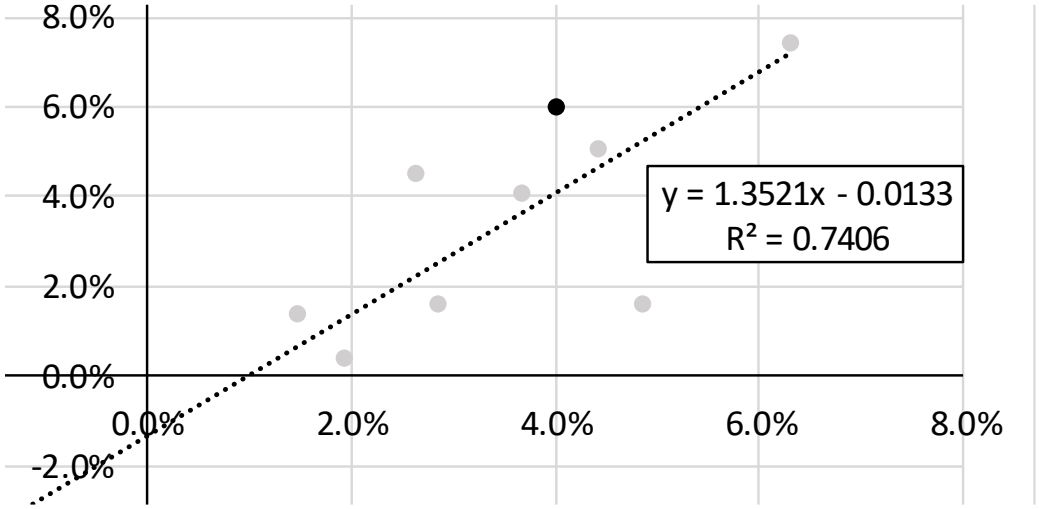
For present purposes, though, it is more important to consider how the model can help attribute returns to information. In Figure 12-3 we drill down on one observation. All days have been grayed out, except for the one in which $r_M = 4.0\%$ and $r_i = 6.0\%$. The estimated

regression equation essentially says that if the market return is four percent, then we'd expect the return on the stock to be

$$Er_i = -0.013 + 1.352 \times .04 = .041 = 4.1\%.$$

The actual return is $r_i = 6.0\%$, so for that day $e_i = r_i - Er_i = 1.9\%$. This is the unexplained return for the day. Or, more properly, it is unexplained by the market. It is therefore attributed to other influences, such as company-specific news announcements.[2]

Figure 12-3.



12.4. Building the event study

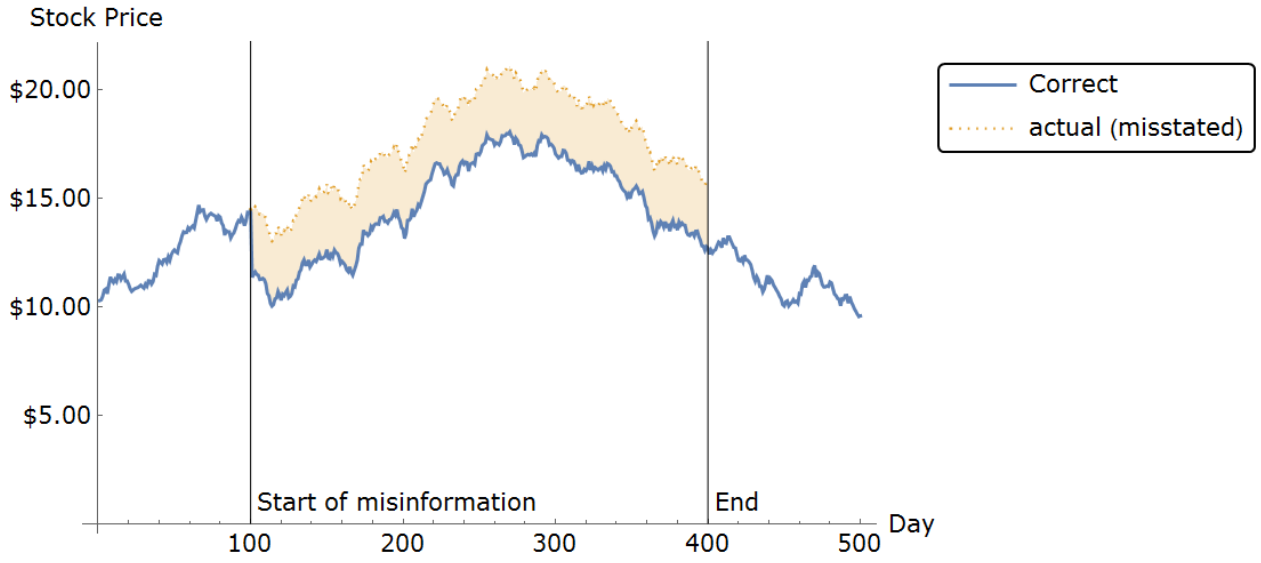
See "Performing the basic event study" (Dunbar and Tabak, 1999).

The first step in estimating the value of the information is to determine an *event window*, a short period of time that corresponds to when the information would have affected the value of the stock. In the situation illustrated in Figure 12-1, the timing of the information is clear. The alleged misinformation started on day 100 and ended on day 400. Looking at the plot, there are two candidates. We might examine the return on day 100, when the price jumped in reaction to the misinformation, or on day 400, when the price dropped at the time of the corrective disclosure.

[2] Now if we are ultimately interested in modeling the stock's price (as in Figure 12-1), why don't we start with a statistical model that is based on prices rather than returns? Analogously to the return model, for example, we might specify something like $p_{it} = a_i + b_i p_{Mt} + z_{it}$ where p_{it} is the closing price of the stock and p_{Mt} is the level of the market index. The error in this regression, z_{it} might be interpreted as the price component that contains non-market information (or misinformation). It turns out in most cases that the relation between a stock's return and the market return is much more stable than the relation between the stock's and market's prices. When the regression is specified in terms of prices, the estimates of the parameters (and residuals) are subject to large errors.

Typically, though, the misinformation involves a lie by omission. Negative information is withheld from investors, and then belatedly disclosed. Often the first indication of a problem is the price drop at the time of the corrective disclosure. There is no corresponding price jump at the time of the omission because there is nothing that would have caused the market to react. This changes the picture of the value inflation.

Figure 12-4. Withholding of negative information



In Figure 12-4, bad news should have been disclosed on day 100. The correct valuation should have dropped on that day. On day 400 we have a belated corrective disclosure. On day 100, though, there was no news, so the stock price simply continued on its path. Only on day 400 do we have a significant drop. In this case the event window would only include day 400. In order to compute the residual on the day of corrective disclosure, e_{400}, we need estimates of α and β. These will be based on data compiled over an *estimation window*. The behavior over the estimation window should be representative of the stock's normal co-movement with the market. Estimates are usually more precise when the sample is large, so there is a case for including all the available data. Companies change over time, however, as new products are introduced, new competitors emerge, and so on. These real changes are often mirrored by changes in the stock price dynamics. This argues in favor of a sample that is restricted to data close in time to the event period, usually a smaller sample. In practice, these two conflicting goals must be balanced. It is not usually necessary to exclude the entire period of misinformation. After all, the market could not have been reacting to what wasn't known.

In addition to the estimates of α, β, and e_{400} we must also allow for errors in these estimates. To judge how well the model fits the data, we first consider the regression coefficient of determination, the R^2. This number indicates the explanatory power of the model, and ranges from zero to one. An R^2 close to zero implies that the stock return and the market return are essentially uncorrelated, while an R^2 near one suggests that the stock and market are moving in near lockstep. We then construct confidence intervals around the parameters. A confidence interval for β, for example, is a range, centered around the estimated β that contains the unknown true value for β. The regression model may be extended to incorporate industry effects and other types of announcements.

Summary of terms and concepts

Rule 10b-5 (not exact wording, but the content); material fact/materiality; reliance; causation; inflation (Feitzinger); damages under the "traditional" view (Fischel); material … damages under the market-efficiency view; fraud on the market; of materiality and reliance; corrective disclosure; the single-index stock return model and its use in damage calculation; *event window; class period; how we identify the class; identification of injured buyers and sellers*.

References

Boudette, Neal E., 2016, U.S. investigates Chrysler Fiat over sales figures, July 19, 2016, New York Times.

Dunbar, Frederick, and David Tabak, 1999, Materiality and magnitude: event studies in the courtroom, National Economic Research Associates (NERA), Available at: http://www.nera.com/publications/archive/1999/materiality-and-magnitude-event-studies-in-the-courtroom.html.

Feitzinger, Kristin, 2014, Estimating recoverable damages in rule 10b-5 securities class actions, (Cornerstone Research).

Fischel, Daniel R., 1982, Use of modern finance theory in securities fraud cases involving actively traded securities, *Business Lawyer* 38, 1-20.

Chapter 13. Private Information

13.1. Overview

The public-information version of the efficient market hypothesis usually strikes people as a reasonable principle. After all, if public facts support a unanimous public consensus on the value of a stock, then any deviations from this value will represent clear profit opportunities. Our belief in this principle is sufficiently strong that if we were to see Apple stock (generally trading in the neighborhood of $500 a share) offered at $100 a share (right this instant, take it leave it), we might well pass. After all, which of the following scenarios is more likely?

- There has been a sudden news announcement that has devastated the value of Apple stock, and we haven't yet checked the news websites, or
- Sellers of Apple stock have offered us an opportunity to exploit an obvious valuation error.

The second possibility is very unlikely. Yes, one does encounter obvious valuation errors from time to time, but they are very rare and usually for small amounts. The market forces causing prices to reflect public information are powerful.

But why should the price of a security reflect private, *non-public* information? The simplest answer is that anyone who possesses such information has a strong incentive to trade until the public price hits their private estimate of value. For example, suppose that the offer price for a share of a mining stock is $10 per share. One person ("Clarence") determines that the company's reserves suggest a value of $11. Clarence will buy, depleting limit orders on the offer side of the book, stopping only when the offer reaches $11 per share. Those setting the offer prices may be completely ignorant of the fundamental reason for the price runup (the higher reserves). They only see that a buyer (or buyers) is putting a value of $11 on the stock.

This explanation is a mechanical one, though, and it is far from satisfying. Once all offers on the book at $10 are taken, why don't new ones come in to "refresh" the book? If the company has made public no new information, why should sellers be reluctant or unwilling to offer shares at $10? Clarence makes a profit on all shares that he buys for less than $11, but the profit on a share purchased at $10 is $1, while a share purchased at $10.10 nets him only $0.90. Why

doesn't he wait until the offers come back in at $10? Below, we analyze the market's reaction in detail.

For the moment, though, it is useful to emphasize two important things about the process. Firstly, it necessarily involves trading. Without the opportunity to purchase the shares, there would be no reason for the market prices to move. Secondly, for the trader to realize the profits, the information must be made public. Once the reserves are known to justify $11 per share, buyers will be willing to pay that amount (or something close), the informed trader will be able to unload his shares and realize a cash profit. A patient buy-and-hold investor with favorable private information can capture her advantage gradually over time, eventually realizing the benefit of the information in the stream of dividends that turn out to be higher than what would generally be predicted. But to a buy-and-sell trader, valuable private information is advance knowledge of public information.

The necessity of trade explains the framing of our insider laws and regulations. In the normal course of a firm's activities, many people inside and outside of the company will be in possession of non-public information. They are simply restricted in the extent to which they can trade. The SEC states (U.S. Securities and Exchange Commission, 2013):

> Illegal insider trading refers generally to buying or selling a security, in breach of a fiduciary duty or other relationship of trust and confidence, while in possession of material, nonpublic information about the security.

To the extent that the law is followed, we would not expect prices to reflect illegal private information. Full compliance with the law is not, however, something that can be simply assumed by other traders in the market. We have many instances of violations, and presumably many instances of violations that weren't detected. Furthermore, even when violations are detected and prosecuted, the counterparty to the insider's trade has dim prospects of recovering losses.

The following discussion will show that private information affects markets in profound ways. Among the most important:

- Private information causes a spread between the bid and ask prices.
- Private information induces an order impact: orders that lift the ask ("buys") cause all subsequent prices, including bids to rise; orders that hit the bid ("sells") cause prices to fall.
- The order impact mechanism opens the market to manipulation.
- Trading against people who have superior information will usually lead to losses. Avoiding these losses motives market participants to expend effort to identify their counterparties.

To explain these effects, we first consider the person most directly affected by the private information: the dealer.

13.2. The dealer's perspective.

When market participants differ in the amount and quality of their information, the market is said to exhibit asymmetric information. Asymmetric information is a feature of many current economic analyses, but one of the earliest descriptions of the problem actually focused on a dealer in the stock market (Bagehot, 1971). "Bagehot" was a pen name adopted by Jack Treynor, a successful professional money manager. (The real Walter Bagehot was a 19th century British writer.)

Treynor noted that many investors (retail and professional) appeared to lose money. He suggested that these people lost money because they incurred trading costs, that these costs arose when investors bought at the ask and sold at the bid, and that the reason for the bid-ask spread was the dealers' dilemma caused by a few people with better (private) information.

Suppose, following Treynor, that the dealer faces two sorts of customers. Liquidity traders are driven by idiosyncratic factors unrelated to the fundamental value of the security. They buy because other activities left them with surplus funds to invest; they sell because there arose an unexpected need for cash. Any one of them is equally likely to buy or sell. A dealer's trades against such customers can create uncertainty because on any given day liquidity buyers or liquidity sellers might be more numerous. But this inventory risk (or position risk) can be managed by prudently arranging offsetting trades. If a liquidity seller were to arrive, followed a short time later by a liquidity buyer, the dealer would capture the bid-spread. On average, therefore, the dealer's profit on each liquidity trader is one-half of the spread.

The other group of customers, however, consists of informed traders. By whatever mechanism, legal or not, they simply possess superior information. Unlike the liquidity traders, they have no idiosyncratic trading motives. They will only buy, paying the dealer's ask price, if they believe that the security is worth more than the ask price. They will only sell, receiving the dealer's bid, if that bid is above their estimate of the security's value. They always trade in the direction of their information, never selling when their news is good or buying when the news is bad. In a trade against an informed customer, the dealer loses on average.

The dealer, of course, would like to identify the customers. But this might be very difficult. Orders might arrive anonymously. Even if there is a name attached to the order, the customer might be informed for one set of securities but not another or be informed in one type of situation. We'll later encounter situations that might help a dealer make a better guess as to the type of customer, but ultimately the quality of the customer's type is unknown. The dealer must act as if he faces a random mix of informed and uninformed customers.

In facing this random mix, the dealer's expected losses depend on the likelihood that the next trader is informed, and the quality of that trader's information. A population in which nine out of ten traders are informed imposes costs higher than if there were only one informed trader out of a hundred. The losses will also be higher if the trader's information concerns an upcoming takeover announcement, instead of, say, a promotion of the third assistant sales manager at a regional office.

For the dealer to survive the losses to informed traders must be at least offset by the profits realized from the uninformed traders. The dealer accomplishes this by setting his bid-ask spread sufficiently wide: a high offer price and a low bid price. With a narrow spread the dealer can't recover enough revenue on the uninformed trades to cover his losses to the informed.

Why should the dealer exercise any restraint at all? If the "efficient" price of the stock is $20, why not bid $15 and offer at $25, making a full $10 profit on each pair of uninformed buyers and sellers? Recall from the earlier discussion of dealers that they are constrained by competition and refusal. Competition comes from other potential dealers who offer their customers better terms of trade. Refusal refers to the ability of customers to decline a dealers' bids and offers. There is a balance here, in that some customers will be slow to locate and switch to a more competitive dealer, and some customers may have such strong trading needs that they are inclined to take almost any price their dealer offers. But over time, competition and refusal impose substantive limits on dealers' behavior.

13.3. A formal model of insider trading

What's the simplest picture of insider thinking we might compose? We'd need a random outcome; an insider who knows the outcome in advance; and at least one other trader (a passive one) who is going to post the bid and offer prices that the insider will hit or lift.

For the outcome, let's start with and end-of-day share value V that will be $V=Low$ or $High$ with equal probability. V will be determined at the start of day, prior to trading, but whether it is *Low* or *High* is not generally known.

Preparing for the first trade of the day

Giving ourselves a role as the dealer or passive trader, we have to post a bid and an offer. If the insider is the only other trader in the market, we won't even bother to show up. If $V = Low$, the insider will hit any bid above *Low*, making a profit for himself, and a loss for us. If $V = High$, the insider will lift any offer below *High*, and we lose again. The withdrawal of the passive trader is not only a modeling inconvenience, but also a practical problem as well. A market with "too much" private information is not sustainable.

If we consistently lose when trading with the insider, there has to be another trader (or traders) against which we'll generally profit. So, we introduce a group of traders that are generally described as liquidity, uninformed, or noise traders.

Although there are informed and uninformed traders, the first one of the day to arrive and maybe trade at the passive trader's quotes is either one or the other, not some sort of blend. The selection device is a random draw. Let's say that there is a 20% chance that the first arriving trader is informed.

Events occur in the following order.

1. Nature (the force of fate and chance) flips a coin: $V = Low$ or $V = High$. This outcome is known only to informed traders.
2. We post a bid and an offer (for one share).
3. The first trader arrives.
   a) With 20% probability, the trader is informed. If $V = High$, he'll buy at any ask price below *High*; if $V = Low$, he'll sell at any bid above *Low*.
   b) With 80% probability, the trader is uninformed. Uninformed traders buy or sell with equal probability.

Figure 13-1 describes the sequencing of the random events in the model. (The setting of bid and ask quotes is not indicated.) Now let's think about our bid. All of the paths that end in "Sell" will hit our bid, but we can't tell either before or after the trade which path brought us to that point.

We'll pick some numerical values. Suppose that *V* is equally likely to be *Low* ($100) or *High* ($150). (Think of a stock that is presently at $100 per share. At the close of trading, the board will announce whether the firm has won a contract worth $50 per share.) And it's given that the probability of an informed trader arrival is 0.20. These probabilities are shown in italics in Figure 13-1. They are *transitional* probabilities, reflecting the likelihood of moving from one box ("node") to another.

From the transitional probabilities, we can compute the *joint probability* of a path of events that winds up at a particular node. For example, the probability of $V = High$ and the arrival of an informed trader is:

$$Pr(V = High, Informed) = 0.5 \times 0.2 = 0.10$$

The probability that $V = High$, *and* the trader is informed, *and* the trader hits our bid is:

$$Pr(V = High, Informed, Sell) = 0.5 \times 0.2 \times 0 = 0$$

These joint probabilities are shown in bold in Figure 13-2. Note that the *total probability* of a buy is 0.5, as is the total probability of a sell.

Figure 13-1 Sequencing of random events (transitional probabilities are in italics).

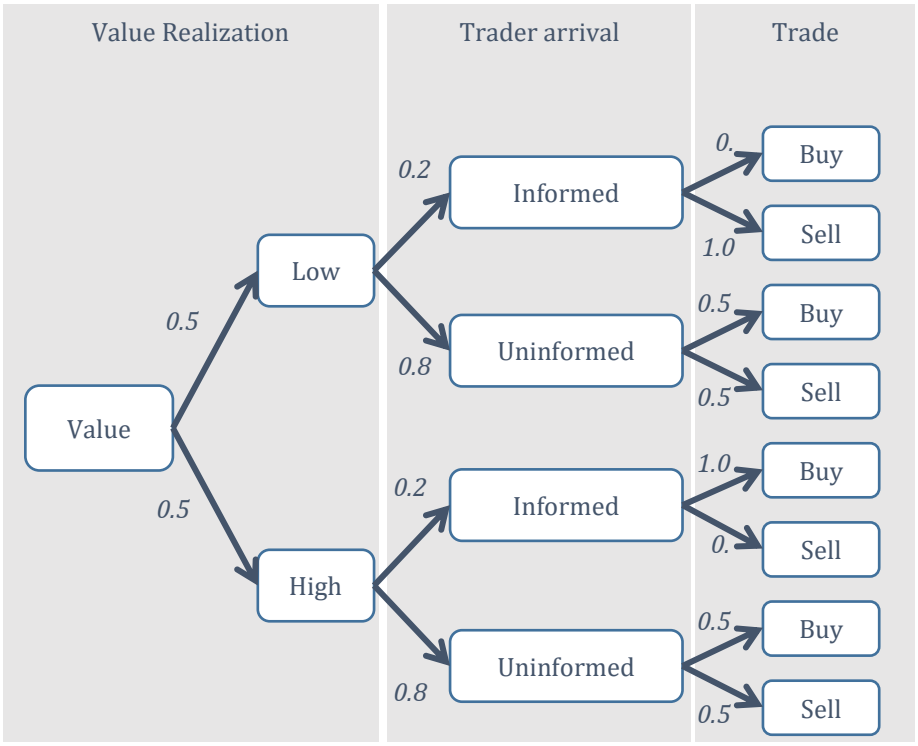
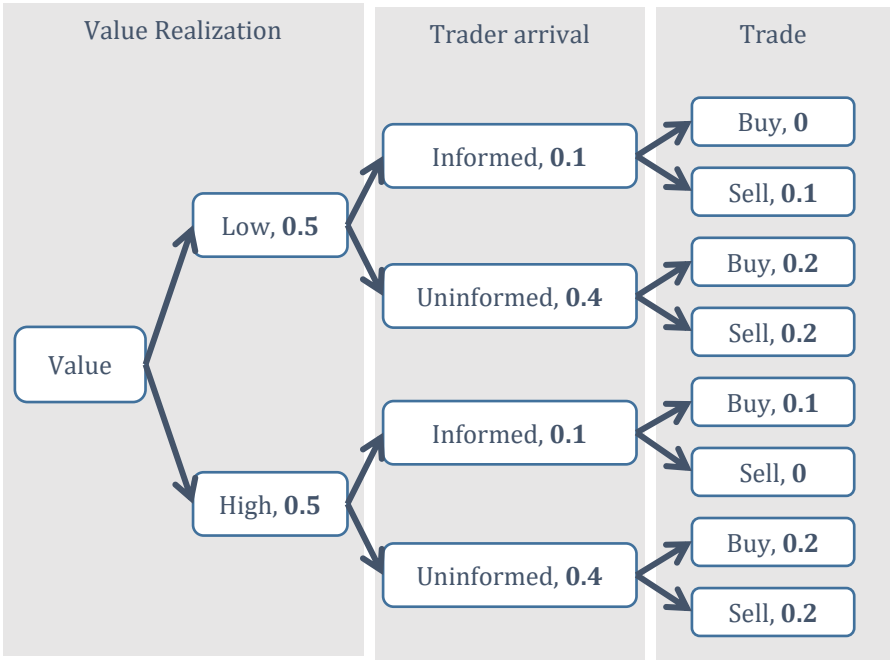


Figure 13-2. Joint probabilities



Although a buy and a sell are equally likely, the paths leading to these events are not identical. Informed traders never buy on bad news, and never sell on good news. As a result, the trade direction tells us something. If our bid is hit, we learn something. The relevant conditional probability here is:

$$P(V = Low|Sell) = \frac{P(V = Low, Sell)}{P(Sell)} = \frac{0.1 + 0.2}{0.5} = 0.6$$

You can check that $P(V = Low|Buy) = 0.4$, $P(V = High|Buy) = 0.6$, and $P(V = High|Sell) = 0.4$.

Our revised beliefs affect our expectation of the firm's value. Before observing the trade, we thought the firm was worth the unconditional expected value

$$\begin{aligned} E[V] &= 0.5 \times Low + 0.5 \times High \\ &= 0.5 \times 100 + 0.5 \times 150 = 125 \end{aligned}$$

The expected value conditional on a sell is,

$$\begin{aligned} E[V|Sell] &= 0.6 \times Low + 0.4 \times High \\ &= 0.6 \times 100 + 0.4 \times 150 = 120 \end{aligned}$$

In this way, we (and other uninformed market participants) can "read" the order flow, to draw inferences about what the informed participants already know.

How should we set our bid price? If we are unconstrained, we'll set it low, to make as much profit as possible from the uninformed traders. But in reality, when we bid we are competing against other actual or potential bidders. The question then involves how high we should be willing to bid.

Suppose that before the first order arrives, I'm bidding \$121, and that's the best bid in the market. If someone hits my bid, I've bought the stock at \$121. The problem is, given that someone hit my bid, I (and the rest of the market) think the stock is worth (on average) \$120. I have a one-dollar loss.

If I lower my bid to \$119, then if I'm hit, I'll have (on average) a \$1 profit. But why should other traders let me capture that profit? Someone will bid \$119.10, someone else will raise to \$119.20, and so on, at least until the bid is \$119.99 (one tick below \$120). Anyone hit at that price will make an average \$0.01 profit.

If anyone overbids, above \$120, they'll be facing an expected loss. Would someone bother to bid exactly \$120? Would someone go to the trouble of putting in a zero-expected-profit order just for the psychic joy of doing a trade? We probably shouldn't rule out the possibility. The point is that the bid is set close to $E[V|Sell]$.

In this view, on average, bidders break even. Informed sellers make money, and uninformed sellers lose. Informed trading works like a tax on uninformed traders. The analysis on the offer side of the market is similar, leading to an ask that is close to $E[V|Buy]$.

When we vary the assumptions that go into this model, we get some interesting results.

- Increasing the relative proportion of informed traders in the population (from 20% to, say, 30%) causes the competitive bid to fall. The ask rises, and the bid-ask spread increases.
- We can raise the volatility of the security by increasing the distance between the low and high values. (For example, instead of $Low = 100 \; and \; High = 150$, let $Low = 90 \; and \; High = 160$.) If we do this, the bid-ask spread increases.

Real markets seem to show both effects. Immediately prior to a scheduled news announcement the spread tends to widen. This may also be the time when, in the process of releasing the news, there is the largest chance of a pre-announcement leak. We also find that when market volatility increases due to an intense flow of public news, spreads are higher.

The second trade and order impact

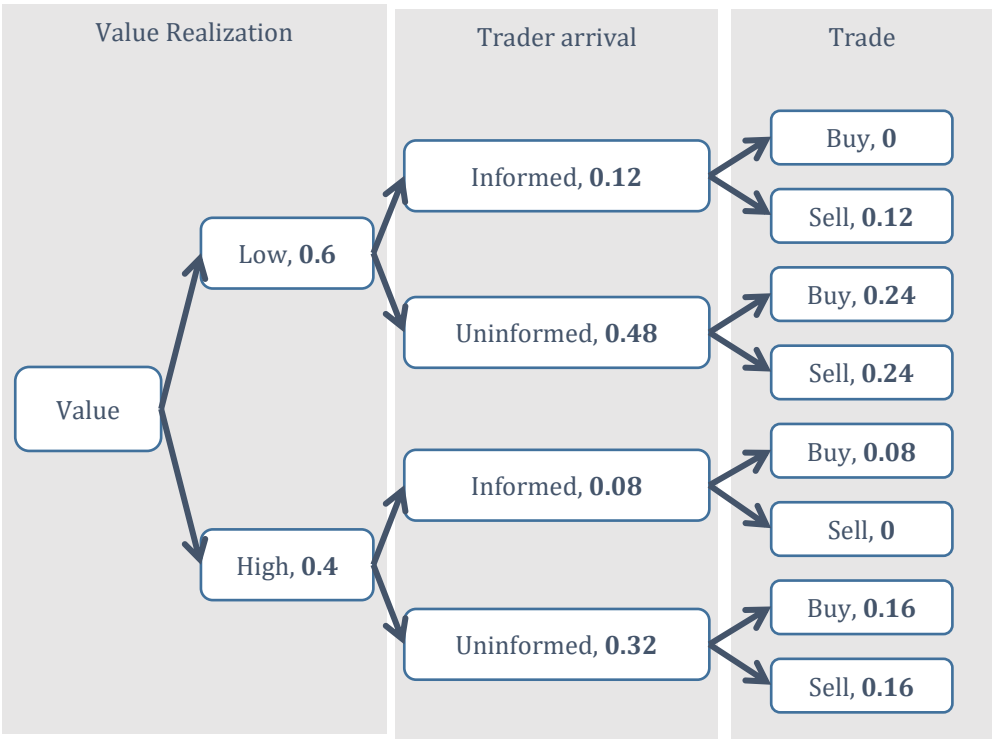
One of the mystifying things about the mining stock example was why limit order sellers did not refresh the orders depleted by the informed trader. With no news announcements from the company, and no apparent change in the company's fundamental economic prospects, why were they suddenly reluctant to sell at a price ($10) that seemed fair and reasonable only a few minutes earlier.

Almost all security markets exhibit something called *order impact* (or *price impact*, depending on whether we want to highlight the cause or the effect). This refers to a movement in market prices that is driven by and apparently caused by incoming marketable orders. For example, a sequence of incoming buy orders that successively lift the ask will move all market prices (bid, ask, and subsequent trade prices) upwards. In the case of the ask quote the reaction is almost mechanical: execution of the buy orders reduces and eventually exhausts the quantity at the prevailing ask price, exposing the next higher sell limit order in the book, and so on. But the bid and subsequent trade prices (even those resulting from marketable sell orders) will also generally be higher. Often the price reverts, but this reversion is usually only partial. A portion of the order impact appears to be permanent.

Increased demand in any market, from candy bars to cars, is also generally associated with a price increase. So why is order impact in security markets remarkable? For one thing, there is an apparent disconnect between the relative size of the order and the price impact. For example, a firm might have 10 million shares of stock outstanding. At a price of, say, $20 per share, the firm's equity market capitalization is $200 Million. A 1,000-share marketable buy order might move the market share price upwards by $0.01. The total value of the order is about $20,000. Yet the market capitalization of the firm has increased by 10 million x $0.01 = $100,000, about five times the total value of the order. It is as if an over-eager car buyer, accepting the dealer's high initial "list price" offer, caused the price of that model to increase for everyone else in the world.

To answer these questions, we return to the simple model, and look at what happens next. Once the first trader arrives and bid is hit or the offer is taken, the probabilities of low and high values are revised. In setting the next bid and offer, the market looks ahead to the second trade. Suppose that the first trade was a "sell" (the bid was hit). The analysis of the second trade is formally identical to that of the first trade, but our initial assessment of the probability of a low value is 60%. The event tree is given in Figure 13-3.

Figure 13-3. The tree for the second trade



By summing the joint probabilities, you can see that $P(Sell) = 0.12 + 0.24 + 0.16 = 0.52$, and $P(Buy) = 0.24 + 0.08 + 0.16 = 0.48$. So, on the second trade we don't expect the incoming order flow to be symmetric. The probability of a low value conditional on a *second* sell is

$$P(V = Low|Sell, Sell) = \frac{P(Low, Sell, Sell)}{P(Sell, Sell)} = \frac{0.12 + 0.24}{0.52} = 0.692$$

So, although we started believing that the probability of a low value was 0.5, once two sellers in a row have arrived, that probability is up to 0.692.

Conditional on observing the first sell, and then the second,

$$\begin{aligned} E[V|Sell, Sell] &= 0.692 \times Low + 0.308 \times High \\ &= 0.692 \times 100 + 0.308 \times 150 = 115.38 \end{aligned}$$

Repeating the argument that we used above; this should equal the next (second round) bid price.

On the offer side of the second market, if the incoming trade following the first sell is a buy, $P[V = Low|Sell, Buy = 50\%]$, and $E[V|Sell, Buy] = 0.5 \times Low + 0.5 \times High = 0.5 \times 100 + 0.5 \times 150 = 125$. With competitive sellers, this will be the new ask price.

This is what we believed before there was any trade. This is reasonable, if the initial seller is followed by a buyer, the order flow appears balanced and symmetric. It is one-sidedness in the order flow that reveals the private information.

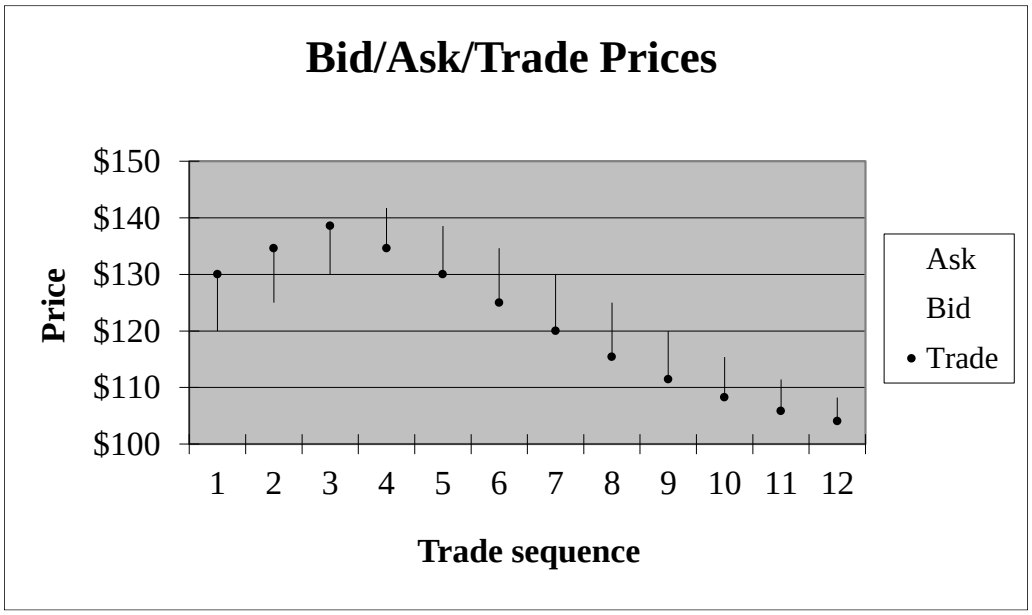
This analysis also clarifies why bids and offers are not generally refreshed at the same level, once acted upon. Although there may be no new fundamental information released by or

concerning the company, the fact of the trade itself (and the fact of its direction, buy or sell) enters the public information set. After the trade, the market has learned something.

13.4. The dynamics of prices and orders

We can use the model to study any sequence of bids and offers over any horizon. SimpleSequentialTradeModel.xlsx (on the class website) is set up to perform these calculations. Figure 13-4 depicts a graph for a sample sequence of twelve trades: three buys followed by nine sells. At each trade, the top of the vertical line is the ask immediately prior to the trade (and the bottom of the line is the bid). A dot marks the trade.

Figure 13-4. Sample price path



Although this is a graph for one particular sequence of trades, it illustrates some general features of the model and, we think, of reality.

Market prices (in the general sense of bids, asks and trades) follow the order flow. Buys drive prices up; sells drive prices down. This order (or price) impact is an important consideration in trading strategies. *Order-splitting strategies* divide a large order into multiple *child* orders, which are executed over time. It is an empirical fact that the child orders executed early in the sequence move the price to the detriment of the later child orders.

Since the order flow reflects on average the true value outcome, the private information is revealed in the market prices over time. We'd be more likely to observe a sequence of nine sells when the value is in fact *Low*.

Not only does the order flow give us a directional signal about the true value, but it also reduces (over time) the uncertainty about the true value. After a period of one-sided order flow, the spread narrows.

13.5. Causation and detection

A customer lifts the ask (buys) and subsequent prices rise. If the customer had hit the bid (sold), subsequent prices would be lower. In this sense, incoming orders cause price movements. But the ultimate value of the security does not depend on the order flow. In our stylized model the value is randomly set at the start of the day. In real-world equity markets, the value is determined by the production, sales and costs of the underlying firm. Most changes in the firm's ownership do not affect these things. (Exceptions involve activist investing, takeovers, and so on.)

The apparent impact of order flow on prices and the ultimate irrelevance of order flow for firm value seem completely inconsistent. On a deeper level, though, they agree. This is because the effect of order flow comes about because it reveals, through the trading process, what has already occurred. Order flow is a signal, a window onto things that aren't yet public.

This also explains the apparent paradox of scale that opened this chapter. Why can a trade cause a change in market capitalization that is much larger than the value of the trade? The answer is that the trade is a signal of value, not just for the shares involved in the trade, but for all shares held by everyone.

In our simple model, the change in bids and asks after the execution of (say) a buy order do not depend on whether the buyer is in fact informed. The maker makers setting the bids and asks can guess, but they don't know. The buyer's knowledge of her own type (informed or uninformed) sets up an additional information imbalance. If she is uninformed, she knows that the bid and ask after her trade are erroneously high. They reflect the market's estimate of the likelihood that she was informed. Only she knows that this likelihood is zero.

Can an uninformed investor move the price on a sequence of orders in a way that yields a trading profit? Huberman and Stanzl (2004) describe this as quasi-arbitrage and show that its absence imposes some structure on the market. It is apparently difficult to execute, however. The SEC litigation releases on pump-and-dump manipulations generally involve cases where the price is run up by the release of misleading positive information, rather than a series of buy orders (see chapter 10.3).

The last section examined sequences of trades implied by the simple model. To someone who is charting prices, a sequence of trades at higher prices suggests convergence to the "good news" outcome; a sequence at lower prices suggests a "bad news" outcome. In a manipulation, trading to establish this pattern involves buying at progressively higher prices (or selling at lower prices), at the going market bids and asks. Alternatively, two floor traders ("Alice" and "Bob") might conspire to establish the trend as follows. Alice sells to Bob at 10; Bob sells back to Alice at 11; Alice sells to Bob at 12; and so forth. The same shares are passed back and forth at higher (or lower) prices. This practice (sometimes called "painting the tape") relies on wash sales. Wash sales are trades in which the buyer and seller are the same party (or, as with Bob and Alice, colluding). In most markets, wash sales are considered a serious infraction in most markets.

Can we identify insider trading after the fact by examining the trade-price record immediately prior to an information release? Perhaps. Some illegal insider trading has been detected in situations where the orders are unusually large, of localized origin (a particular broker or city), and overwhelmingly one-sided (on, of course, the profitable side). It is well to remember, though, that the dynamics that arise in the simple multiperiod model are a consequence of the market's *beliefs* about the incidence of informed trading – not the *actual* incidence. If the market believes that the proportion of informed traders in the population is zero, an insider will enjoy an extremely deep market, and may trade extensively with little impact.

Statistical models of impact play a large role in the formation of dynamic trading strategies. Their estimation and uses are considered in **Error! Reference source not found.**.

13.6. Further reading

The academic literature on asymmetric information in securities markets is largely divided into two approaches. The sequential trade models (like the one considered in this chapter) view the market as a series of moves where at any given time one player is facing a decision. Some of the early key papers are: (Easley and O'Hara, 1987, 1991, 1992; Glosten and Harris, 1988; Glosten and Milgrom, 1985). Strategic trade models form the other main branch. In these models, one or more informed traders interact over time with market makers and uninformed traders. Mostly in these models it is the informed traders who are strategic, trading enough to make profits, but not so intensely as to move the market prices any more than necessary (as in the order splitting models of Chapter 17). Representative papers would include (Foster and Viswanathan, 1993; Holden and Subrahmanyam, 1992; Kyle, 1985, 1989; Spiegel and Subrahmanyam, 1992; Subrahmanyam, 1991). (Back and Baruch, 2004) provide an elegant synthesis that bridges the two approaches.

Summary of terms and concepts

Private vs. common values; symmetric and asymmetric information; Treynor/"Bagehot" analysis of the bid-ask spread. The sequential trade model of the bid-ask spread and price impact, market failure.

References

Back, Kerry, and Shmuel Baruch, 2004, Information in securities markets: Kyle meets Glosten-Milgrom, *Econometrica* 72, 433-465.

Bagehot, Walter, 1971, The Only Game in Town, *Financial Analysts Journal* 27, 12-22.

Easley, David, and Maureen O'Hara, 1987, Price, trade size, and information in securities markets, *Journal of Financial Economics* 19, 69-90.

Easley, David, and Maureen O'Hara, 1991, Order form and information in securities markets, *Journal of Finance* 46, 905-27.

Easley, David, and Maureen O'Hara, 1992, Time and the process of security price adjustment, *Journal of Finance* 47, 576-605.

Foster, F. Douglas, and S. Viswanathan, 1993, The effect of public information and competition on trading volume and price volatility, *Review of Financial Studies* 6, 23-56.

Glosten, Lawrence R., and Lawrence E. Harris, 1988, Estimating the components of the bid/ask spread, *Journal of Financial Economics* 21, 123-42.

Glosten, Lawrence R., and Paul R. Milgrom, 1985, Bid, ask, and transaction prices in a specialist market with heterogeneously informed traders, *Journal of Financial Economics* 14, 71-100.

Holden, Craig W., and Avanidhar Subrahmanyam, 1992, Long-lived private information and imperfect competition, *Journal of finance* 47, 247-270.

Huberman, Gur, and Werner Stanzl, 2004, Price Manipulation and Quasi-Arbitrage, *Econometrica* 72, 1247-1275.

Kyle, Albert S., 1985, Continuous auctions and insider trading, *Econometrica* 53, 1315-1336.

Kyle, Albert S., 1989, Informed speculation with imperfect competition, *Review of Economic Studies* 56, 317-355.

Spiegel, Matthew, and Avanidhar Subrahmanyam, 1992, Informed Speculation and Hedging in a Noncompetitive Securities Market, *Review of Financial Studies* 5, 307-329.

Subrahmanyam, Avanidhar, 1991, A Theory of Trading in Stock Index Futures, *The Review of Financial Studies* 4, 17-51.
U.S. Securities and Exchange Commission, 2013, Insider trading.

Chapter 14. Insider Trading

Inside information is private information that has crossed the line into illegality. But where and how should that line be drawn?

14.1. The social value of informed trading

Informed traders live under a cloud of negative associations. The clearest examples of private information are the most extreme cases of illegal inside information. Furthermore, as in the model, the profits earned by the informed traders often come at the expense of the uninformed. Are the informed traders simply parasitic? Should the insider trading laws be broadened to include all private information, no matter how obtained?

Unfortunately, information is costly to produce. Securities analysts are highly educated professionals, and they rely on comprehensive (and expensive) data. Were they to cease their research, the market would be reliant on information produced by the listed company. If a modest profit at the expense of the uninformed is the price of independent research, perhaps private information is not as bad as we might have initially thought.

14.2. The social cost of informed trading

In our simple model of the last chapter, the profits of the informed traders come at the expense of the uninformed traders, that is, other investors. Market makers on average break even. In real world markets, of course, market makers may well fare better than simply breaking even, but there is no question that informed traders profit at the expense of the uninformed. So, if, in a sense, the market is "rigged" against them, why do uninformed traders participate?

The uninformed might participate out of ignorance, of course. They might also believe themselves smarter than they in fact are. But even limiting our attention to the rational and sober, it is a fact that uninformed traders participate because the pursuit of their long-term investment goals makes the stock market difficult to avoid.

Difficult, perhaps, but not impossible. Could the process become so stacked against public investors that they simply refuse to invest? Looking about the globe, we must admit this possibility. More than a few countries have no public capital markets. Investment is made either by the government, or by family firms.

Private information raises the possibility of market failure. Once the investing public has "lost confidence" in the securities market, it is very difficult to restart the market. Capital is not raised, risks are not hedged, and the entire economy suffers. The consequences of crossing this line are so costly that regulators often try to ensure that markets stay well to the safe side of it. This takes the form of required disclosure, and prohibitions against insider trading.

14.3. The boundary between public and private information

In the section on public information, we noted that information can only be considered public if it is widely available. An importance aspect of this availability is the delay or, in current jargon, the *latency* of the communication process. A trader who receives the information first, and has the opportunity to trade on it, is enjoying a brief interval where the public information is for all intents and purposes, private. The justification suggested above in favor of private information is most convincing when the private information arises from the analysis of fundamental information. It is less attractive if the private information is being produced by expenditure on a computer network that will allow the user to trade on the basis of a press release ten milliseconds before everyone else.

14.4. Manipulation

Trade not only facilitates the incorporation of private information into prices. It almost inevitably creates new private information. If the first trader in the market buys, the price goes up by the same amount whether or not the trader is informed. The market reaction is the same because nobody in the market can tell what group the incoming trader belongs to. The trader herself, though, does know (barring self-delusion) her group.

Suppose that the first trader of the day is uninformed, and she buys. The prices rise. From the viewpoint of all uninformed traders in the market, this is a fair and appropriate reaction given the possibility of informed trader. All traders, that is, except one. The first trader knows she's uninformed and given that knowledge the price reaction is a mistake.

In general, uninformed traders can move prices – not because they actually know anything, but because the market can't tell whether they're informed or not. Are there trading strategies where they can use this to their advantage? Can they drive up the price by buying, for example, and unload all of their shares at the top? In the simple model of this section, this strategy doesn't work, but slight changes in the model can make it profitable.

The term "manipulation" does not have a universally agreed-upon definition, either in economics or in law. But from a rough "I-know-it-when-I-see-it" viewpoint, manipulations often involve market activity that moves the price to an "artificial" level, and trades that attempt to profit from this movement.

How far can an uninformed trader push the price, and for how long? Remember that the order impact occurs because the market reads the order as a signal of information that is private for the moment but is soon expected to become public. An urgent flurry of purchases may appear to originate from buyers anxious to establish their positions prior to a favorable news announcement. But if no public announcement is forthcoming, and particularly if company management denies the existence of any development that might explain the purchases, the share price will revert.

So, suppose that events in fact play out as we suggest. An uninformed trader throws in a string of buy orders, at increasing prices. Then, after it is clear that no favorable developments

are pending, the price falls back to where it was originally. Doesn't this mean that the shares purchased at high prices are now worth less, and that that trader has incurred a loss? This might indeed be the case for the series of trades considered in isolation for everything else.

But market prices are sometimes used as reference prices, for determining the value of "cash-settled" derivative contracts or for computation of margin requirements. The trader might therefore have a strong interest in a higher or lower price at a particular instant, but not at other times. Suppose, for example, that a trader has established a highly levered long-margin position (with borrowed funds). The account is typically valued at closing prices for purposes of determining how much actual cash the trader must put up. By pushing up the price at the close, the trader can increase the apparent net worth in the position and minimizing the additional cash contribution.

It is not the purpose of this discussion to certify whether this practice, a form of "marking the close," constitutes manipulation in a formal economic or legal sense. Regulatory records, however, contain numerous instances where traders have settled the charges and paid penalties, in preference to asserting innocence at a trial.

14.5. US insider trading laws

Under US law, persons who trade while in possession of superior information might be subject to civil or criminal penalties. The relevant statutes (written laws) do not precisely define what constitutes illegal insider trading, however, so our practical sense of the term is the outcome of principles and precedents that have emerged over the years as the courts have decided actual cases. This evolution is still in process. In many areas technological advances have shaped and reshaped our notions of information, its economic value, and its property rights (who owns it). This is certainly true in how we think about insider trading.

Rule 10b-5 states that, "It shall be unlawful for any person … to engage in any act … that operates as a fraud or deceit upon any person in connection with the purchase or sale of any security." Insider trading is subject to regulation because it is potentially fraudulent.

Who exactly is an "insider"? With respect to a particular firm's securities, a corporate *insider* is a manager, officer or employee of the firm. In the normal course of business, we'd expect them to know more about the firm than most others who hold or trade the firm's stocks or bonds. But the reach of the law extends beyond this group. Accountants, attorneys, consultants, and others who work with the firm may also learn about the firm's activities. These *outsiders* may be considered *constructive insiders*, that is, those whose relations with the firm make them, at least temporarily, in a position similar to that of a manager, officer or employee. A *tipper* is someone who passes along information; a *tippee* is someone who receives it.

Information can be *material* in the everyday sense of "significant". In this context, though, the courts have adopted a criterion that the information would likely be considered important by an investor buying or selling the stock. A considerable body of law has evolved to refine a more precise definition of the term.

Of course, small bits of information can add up to strong conclusions. This is sometimes called the *mosaic* principle. Takeover announcements, for example, are usually very material to the value of a target, and many insider trading cases have involved direct leaks of target identities and takeover timing. But public news releases, speeches, and so forth can also signal an acquirer's intentions. Defendants sometimes advance the mosaic defense, that argument that a takeover was in fact deduced from disparate facts, hearsay and rumor. In commenting on a series of 2013 cases, however, (Morrison-Foerster, 2015) note that, "The 'mosaic theory' defense, while still viable in theory, met with no success in court because defendants who have asserted that they legitimately pieced together a cogent investment thesis from bits of immaterial

nonpublic information faced direct evidence – from wiretaps or former co-conspirators – that provided far less innocent explanations for their purchases or sales of securities."

The following discussion first discusses the rules for corporate insiders, and then considers the broader rules that might apply to anyone buying to selling security.

14.6. Special provisions for corporate insiders

Corporate insiders are prohibited from trading on the basis of non-public material information. They can't sell short (thereby profiting from declines in the value of the firm). Profits realized from buying and selling within six months are considered short-swing profits and must be given to the firm. They must report their trades to the SEC within two days.

Managers have all sorts of legitimate reasons to hold their stock. An ownership stake aligns their interests with those of the other shareholders. On the other hand, a stake that represents a large proportion of a manager's wealth is undiversified, and therefore especially risky. Selling part of this investment will reduce the risk and may be desirable even if the manager has an overall favorable view of the stock's expected performance. To allow for these sorts of transactions, the law contains a safe harbor, a course of action that is presumed to protect a manager against prosecution. Specifically:

> … [A] person's purchase or sale is not "on the basis of" material nonpublic information if the person making the purchase or sale demonstrates that … Before becoming aware of the information, the person had:
>
> (1) Entered into a binding contract to purchase or sell the security,
>
> (2) Instructed another person to purchase or sell the security for the instructing person's account, or
>
> (3) Adopted a written plan for trading securities;"

These so-called "10b5-1" plans can be used to achieve diversification over time. A manager can commit to selling a given number of shares at regular intervals, for example.

There is nevertheless evidence of misuse. The plans do not have to be disclosed or filed with the SEC. They can be modified or cancelled at any time. One study finds that the plans are set up on short notice, sometimes on the same day as the trade; about half the plans consist of a single trade; and many trades occur immediately prior to earnings announcements (Larcker, Lynch, Quinn, Tayan and Taylor, 2021; Shifflett, 2021). The SEC is reviewing changes in the rules.

14.7. Some key principles of US insider trading law

Although insider trading is not clearly defined in any rule, principles have emerged over time as cases have been argued. The following material is summarized from Bainbridge (2000, 2001, 2012).

Disclose or abstain

The principle of "disclose or abstain" holds that anyone in possession of non-public material information must disclose that information to a counterparty or simply refrain from trading. Since it is not often practical in an anonymous decentralized security market even to identify one's counterparty (let alone disclose the information), the principle is effectively a prohibition.

In 1964, Texas Gulf Sulfur (a US mining company) determined that an area of Ontario where they'd been engaged in geological explorations was likely to be particularly valuable.

Company personnel bought stock well before the news was made public. The SEC brought an insider trading case and won.

Fiduciary duty of confidentiality

A "fiduciary duty" is a legal obligation to act on behalf of someone else. The courts have held that a duty to disclose only exists if the information was obtained through a relationship of trust. This principle narrows the applicability of disclose or abstain.

In Chiarella v. US (1980), Chiarella worked for a financial printer. In printing documents for corporate acquirers, he determined the identities of target firms. He purchased the stock prior to the announcement. Chiarella (an outsider) was convicted, but the Supreme Court reversed the conviction on the grounds that he had no fiduciary relationships to the target companies or their stockholders.

Misappropriation of information

Misappropriation involves taking something (information, in this case) and using it for trading purposes.

O'Hagan was a partner in a law firm, Dorsey and Whitney. The firm was advising Grand Metropolitan on a takeover of Pillsbury. O'Hagan was not working on this case but became aware of the deal by seeing papers on another partner's desk. He bought shares of Pillsbury, and profited when the takeover was later announced. Since Grand Met was not his client, there was no violation of attorney-client confidentiality. The information that he used in trading, though, was not his to use.

R. Foster Winans (ca. 1983) wrote a column "Heard on the Street" for the Wall Street Journal. Companies mentioned often experienced price moves after the column appeared. Prior to publication, Winans identified the stock others (including a stockbroker), who went on to generate large trading profits. The SEC asserted that the information was the property of Winans' employer (the Journal), and in tipping others he had misappropriated that information from the Wall Street Journal. The Court of Appeals agreed. Interestingly, if we grant that the Wall Street Journal owned the information, presumably it would have been okay for them to use the information in trading.

In using the Journal's information for personal profit, Winans was breaching an employee's *duty of confidentiality.* The duty of confidentiality can be critical in ascertaining misappropriation, so it's important to note that it can arise in all sorts of business and personal relationships. From the Code of Federal Regulations, §240.10b5-2 Duties of trust or confidence in misappropriation insider trading cases:

> … (b) Enumerated "duties of trust or confidence." For purposes of this section, a "duty of trust or confidence" exists in the following circumstances, among others:
>
> (1) Whenever a person agrees to maintain information in confidence;
>
> (2) Whenever the person communicating the material nonpublic information and the person to whom it is communicated have a history, pattern, or practice of sharing confidences, such that the recipient of the information knows or reasonably should know that the person communicating the material nonpublic information expects that the recipient will maintain its confidentiality; or
>
> (3) Whenever a person receives or obtains material nonpublic information from his or her spouse, parent, child, or sibling; provided, however, that the person receiving or obtaining the information may demonstrate that no duty of trust or confidence existed with respect to the information, by establishing that he or she neither knew nor reasonably

> should have known that the person who was the source of the information expected that the person would keep the information confidential, because of the parties' history, pattern, or practice of sharing and maintaining confidences, and because there was no agreement or understanding to maintain the confidentiality of the information.

Tipper-tippee liability

Often, as in the Winans case, the source of the information does not trade directly, but passes the information along to someone else. This separation does not remove legal liability. There might still have been an initial breach of a duty of confidentiality. The tippee (recipient of the information) may still be liable if they knew (or should have known) about the breach.

Sometimes the information can pass through many people before someone actually trades on it. In an SEC case involving IBM's takeover of Lotus, at least one person who traded only learned of the information after it had been passed on through seven other people (U.S. Securities and Exchange Commission, 2001).

In 1973, Raymond Dirks was a securities analyst who learned about fraud in the Equity Funding of America (an insurance company). Dirks alerted his clients, and they sold before the fraud became widely known, and Equity collapsed. Dirks' source of information was Secrist, a former Equity insider. The SEC initially censured (penalized) Dirks, but this was ultimately reversed by the US Supreme Court. In this case, the Court found that there had not been a breach of a fiducial duty because Secrist derived no benefit from his disclosures.

The question of benefit

It is a commonplace truth that we derive benefits from sharing information in our relationships. Recommendations and endorsements about everything from movies to medical providers get passed back and forth among family and acquaintances in the normal course of maintaining and strengthening the bonds. When we say that someone "owes us one" or that a good deed "went into the favor bank," we're using terms of commerce to describe a general system of social barter and exchange. Where, then, does valuable inside information fit within this scheme?

In 2012, Anthony Chiasson and Todd Newman were convicted of insider trading in a jury trial. The instructions to the jury limited their determination to whether Chiasson and Newman knew that the information on which they relied had been improperly disclosed. The question of whether the ultimate source of the information received a personal benefit was not, according to the instructions, relevant. In December 2014, a panel of justices from the 2nd US Circuit Court of Appeals reversed the conviction, however, arguing that the question of benefit was an important one, and that "the mere fact of friendship" does not suffice to establish a presumption of such a benefit.

Bassam Salman was accused of trading on advance knowledge of takeover information supplied by a friend who heard the news from his brother. In 2011 he was convicted. But following the Chiasson and Newman reversals, Salman appealed on the grounds that there was no personal benefit provided in exchange for the information. In 2016 the US Supreme Court upheld the conviction, noting

- "… A tipper breaches a fiduciary duty by making a gift of confidential information to 'a trading relative.'"
- "[Giving] a gift of trading information is the same thing as trading by the tipper followed by a gift of the proceeds."

The case established the illegality of what was called "friends and family" tipping.

In 2014 Matthew Martoma was convicted of insider trading in the stock of two companies developing pharmaceuticals for treating Alzheimer's disease. He obtained the information from

information that he obtained from Sidney Gilman, a Professor at the University of Michigan who was monitoring a clinical trial. Martoma appealed, arguing an absence of personal benefit: he and Gilman weren't friends, and Gilman received no monetary or similar benefit. The course upheld the conviction: "… a corporate insider personally benefits whenever he discloses inside information as a gift with the expectation that the recipient would trade on the basis of such information or otherwise exploit it for his pecuniary gain."

14.8. Practical advice

Many of us are or will be knowledge workers in a knowledge-based economy. Many careers involve the synthesis and use of information that is economically valuable, information that at times is non-public and material. While some situations may be fraught with ambiguity, there are several clear guideposts. Many of the situations discussed in this section fall under the broad umbrella of employee conduct. Apple's Business Conduct Policy, for example, states (in part):

> Never buy or sell stock when aware of information that has not been publicly announced and could have a material effect on the value of the stock. This applies to decisions to buy or sell Apple stock and to third party stock, such as the stock of an Apple supplier or vendor. It is also against Apple policy and may be illegal to give others, such as friends and family, tips on when to buy or sell stock when aware of material, nonpublic information concerning that stock.

Some of the FAQs address the kind situations that any of us might face.

> I have stock in companies that do business with Apple. Is this a problem?
>
> Probably not. However, it could be a concern if (1) you're influencing a transaction between Apple and the company, or (2) the transaction is significant enough to potentially affect the value of your investment.
>
> Does Apple's policy apply to buying or selling stock in other companies?
>
> Yes. For example, say you learn about a customer's nonpublic expansion plans through discussions about hardware purchases. If you purchase stock in the customer's company or advise others to do so, it could be viewed as insider trading.

Most regulated firms in the financial sector (such banks, brokerages, investment advisers) are required to have a designated compliance officer. The compliance officer should be consulted in any case of doubt.

Summary of terms and concepts

Disclose or abstain; fiduciary duty; tipper-tippee liability; misappropriation of information; short-swing profits; 10b5-1 plans and the criticisms; economic arguments and counterarguments in favor of insider trading.
Cases: Cady, Roberts (disclose or abstain); Texas Gulf Sulfur (disclose or abstain); Chiarella (the requirement of a fiduciary duty); O'Hagan, also Carpenter [Winans] (misappropriation of information); Chiasson and Newman (the requirement of some element of payment).

References

Bainbridge, Stephen M., 2000, Insider trading: an overview, UCLA School of Law, Available at: http://ssrn.com/abstract=132529.

Bainbridge, Stephen M., 2001, The law and economics of insider trading: a comprehensive primer, Available at: http://ssrn.com/abstract=261277.

Bainbridge, Stephen M., 2012, An overview of insider trading law and policy: An introduction to the insider trading research handbook, in Stephen M. Bainbridge, ed.: *Research Handbook on Insider Trading* (Edward Elgar).

Larcker, David F., Bradford Lynch, Phillip Quinn, Brian Tayan, and Daniel J. Taylor, 2021, Gaming the system: three "red flags" of potential 10b5-1 abuse, Available at: https://ssrn.com/abstract=3769567.

Morrison-Foerster, LLP, 2015, 2014 Insider Trading Annual Review, Morrison-Foerster, LLP, Available at: http://www.mofo.com/~/media/Files/ClientAlert/2015/02/150211InsiderTradingAnnualReview.pdf.

Shifflett, Shane, 2021, Executive Stock Sales Are Under Scrutiny. Here's What Regulators Are Interested In., Wall Street Journal (Dow Jones & Company).

U.S. Securities and Exchange Commission, 2001, Litigation Release No. 16848 (Lotus Securities).

Part IV. The Basics of Algorithmic Trading

A trading algorithm is a recipe, a structured procedure for trading, that is generally implemented on a computer directly connected to the market. Algorithms ("algos") lie on a spectrum. At the simple end are qualified orders (like immediate-or-cancel) that are a bit more reactive and adaptive to market conditions. These can be classed as conditional orders. More complex algos, though, generally rely on assumptions about how market prices and orders evolve, and particularly about how our own orders will affect the prices. These assumptions take the form of dynamic statistical models. Once we have a statistical model, we can formulate a trading problem and determine the optimal strategy. We apply this approach to the classic order splitting problem.

Chapter 15. Complex Orders

Earlier we looked at qualifiers such as immediate or cancel (IOC), all or none (AON), fill or kill (FOK), that modified the handling of standard market and limit orders. Conditional orders are those for which activation or execution depends on some market event, like the stock price hitting a pre-specified level.

The dividing line between qualified and conditional orders is not a precise one. The complexity of an order or order strategy lies on a continuum. Labeling the orders discussed in here as "conditional" is simply a way to indicate that they are more complicated than the qualified orders discussed earlier, but less complicated than the multi-stage algorithms that will be discussed later.

The rules defining conditional orders are straightforward, and it is usually easy to identify one dominant appealing feature. On the other hand, it can be very difficult to decide which type is optimal in any given situation. Moreover, in the right circumstances, with the right knowledge, they can often be "gamed" (or "tricked," to use a less sporting term). For each type, you might well ask, "How might I respond if my competitors (on the same side of the market) were using one of these algos?" Or "… if my potential counterparties (on the opposite side of the market) were using one?"

15.1. Stop orders

A stop order is *elected* (that is, becomes active) when there is a trade at or through the stop price. The stop price is different from the limit price. A stop sell order is also called a stop loss order.

> Example: "Sell 100 MSFT stopped at 24, limit 23" would typically be entered when MSFT is trading well above 24. When there is a trade in MSFT at 24 or below, the stop order is elected (becomes effective). It becomes "Sell 100 MSFT limit 23".

This is a stop limit order because it carries a limit price. Some markets accept stop market orders. Upon election, the order becomes "sell 100 MSFT at the market."

A stop loss order might be used when an investor has an accumulated profit on a long position. The stock was purchased; the price went up, and the investor wants to keep at least some of the profit in the event that the price drops. (With a trailing stop loss order, the stop price is reset relative to the highest price recently realized.) Alternatively, losses in a long margin position can mount rapidly, due to leverage. A stop loss order can provide an automatic exit.

A stop loss order is not really an order until the triggering event occurs, and the order is elected. It is sometimes thought that a stop loss order ensures a sale at the stop price. In fact, there is no such certainty. A market stop loss order becomes a market sell order when there is a trade at or below the stop price. If the bid is falling rapidly, a market sell order might execute well below the stop price. If the bid has fallen through the limit price of the stop loss order, the order won't be executed. Before it is elected, a stop order is not placed in the book, made available for execution, or accorded any time priority (except, perhaps, relative to other stop orders).

A stop buy order works in the opposite direction. "Buy 100 MSFT stopped at 28, limit 29" might be entered when MSFT is trading below 28. When there is a trade at 28 or higher, the order is elected, and it becomes "Buy MSFT limit 28." It may be an attractive option for an investor holding a short margined position. This investor faces the risk of a sudden price rise.

Stop orders are not displayed until they are elected. If a trader suspects that there are many stop-loss orders waiting at a particular price, he may aggressively short the security, driving the market t down to the stop price. The election of these orders triggers a wave of selling that can quickly drive the price down further. The short seller then covers at the lower price, realizing a profit. This practice, known as "gunning the stops," is usually considered manipulative, exposing the trader to legal and regulatory sanctions.

In this connection, the U.K. Financial Conduct Authority recently determined that five banks (Citibank, JPMorgan Chase, RBS, UBS, and HSBC) engaged in this practice in the FX market: "Traders [from these banks] shared the information … to help them work out their trading strategies. They then attempted to manipulate fix rates and trigger client "stop loss" orders." Retail traders, incidentally, are sometimes believed to herd in the placing of their stops, grouping at "natural" price points (five and ten dollar multiples, for example).

15.2. Pegged orders

With a pegged order, the price is set relative to some other price (typically the NBO, the NBB or the midpoint). If the reference price changes, the order is repriced.

> Example: If the market is $25 bid for 1,000; 2,000 offered at $25.10, we might see "Buy 500, pegged at the NBB less $0.05." This order will initially be priced at 24.95, but if the NBB changes, the order will be repriced. It will execute only when there is sell order that walks through the book.

A repriced order is added to the book at the new price behind all orders previously entered at that price. Effectively, the time stamp on a pegged order is the time of the most recent repricing, not when the pegged order was initially submitted.

Once identified, a visible pegged order is an inviting target. Suppose that market *A* has a buy order pegged to the NBB. A seller might enter an aggressive bid on market *B* (to raise the NBB), send a marketable sell order to market *A* (which would execute against the pegged order), and lastly cancel the bid on market *B*. This practice ("spoofing") is forbidden in most markets, but it may be difficult to detect. At least one exchange (BATS) requires pegged orders to be hidden.

15.3. Discretionary orders

This is a basically a limit order, but if the opposing quote gets within a specified range, the order is repriced to become marketable.

> Example: If the market is $25 bid for 1,000; 2,000 offered at $25.10, we might see "Buy 500 limit pegged at the NBB, with discretion price $0.04 above the NBB." This will initially go in as a limit order pegged at the NBB. If the NBB changes, the order will be repriced. But if there is ever an offer at or closer to NBB+$0.04, the buy order is repriced to be marketable.

If we suspect that potential counterparties (on the opposite side of the market) are using discretionary orders, we will manage our own orders less aggressively. For example, suppose that we're buying and the market is $50.00 bid for 100 shares (ours), 1,000 shares offered at $50.10. If we are up against a deadline to accomplish our purchase, we might be inclined to reprice our order at $50.10, lifting the market offer. If we suspect that some of the offers are discretionary, though, we would instead raise our bid incrementally, in the hope of triggering one.

15.4. Reserve ("iceberg") orders

A reserve order is partially hidden. The "tip" of the order is visible, the larger portion is hidden When the visible part of the order is executed, it is "refreshed" from the undisplayed portion. This refresh can be mechanical: "10,000 shares total to buy at $25. Start showing 500. When that executes, immediately show another 500, until the entire amount is filled."

This regularity, though, can render the reserve order easily detectable. 500 shares are shown; 500 shares trade; 500 more shares immediately appear, and so on. This is so obvious that it defeats the original purpose of disguising the larger amount. To control the predictability, it makes sense to randomize the refresh quantity. The NYSE Arca market offers such a feature, called a random reserve order. Some implementations of reserve orders also introduce a random delay in the refresh process.

15.5. Post-only

"Post-only" is a qualification that, when added to an ordinary limit order, ensures that the order will not execute when it is first submitted. This might strike you as mystifying. Aren't executions always desirable? We'll return to this question, but first let's look at an example.

The basic functionality of Nasdaq's post-only order can be illustrated as follows (https://nasdaqtrader.com/content/ProductsServices/Trading/postonly_factsheet.pdf). Suppose that the market in WXYZ is 10.12 bid, offered at 10.15, and that there are no hidden orders. An order "Buy limit 10.15" would normally execute at the offer price. "Buy limit 10.15, post-only", however, won't be executed. It can't go onto the book, though, because it would lock the market. It might be cancelled, or it might be repriced one tick lower (to 10.14) and then placed in the book. The choice, according to Nasdaq, "[depends] on the customer port setting," which in turn presumably depends on level of service and/or other attributes of the submitter.

Now since the order submitter is presumably aware of the best bid and ask, can't they decide on pricing before the order is sent? This is not always possible. With delay ("latency") an order intended to rest on the book might be executable when it arrives at the market, because market prices are quickly changing. But if the order is executable (at its limit price or better), what could be gained from forcing it onto the book? The answer turns on how the exchange charges for its services. To encourage traders to display bids and offers, many exchanges offer rebates for resting orders and charge for executable orders. This "maker/taker" pricing is described more fully in section 19.1. But for the present discussion, it suffices to note that in

maker/taker exchanges, it is better to be the resting limit order ("maker") than the executing order ("taker"). Post-only orders seek to capture the maker rebate.

In a sense, a post-only order is the opposite of an immediate-or-cancel (IOC, see 4.3). An IOC order seeks execution. If it can't be executed, it is cancelled in its entirety. Nothing ever goes to the book. A post-only order, on the other hand, avoids immediate execution and is handled by the exchange in a way that favors placement in the book.

15.6. Implementation

Is an algorithm implemented by customer, the broker, or the market center? Who actively manages it, and ensures that the instructions are correctly followed?

These questions actually arise with the primitive order types. In a floor market with bilateral trading (like a futures pit), even simple market and limit must be handled attentively. With a contract trading at a price around 80 (the units don't matter), a broker holding a customer limit order to buy limit 75 might think that it is safe to go out for lunch. But if in his absence, the price has dropped to 70 and then gone up to 85, the broker has "missed the market". The customer has a valid grievance. In fact, one of the early functions of the NYSE specialist was to act as agent for the book of limit orders left in his care.

Initially, pegged, discretionary and reserve orders were implemented on customers' or brokers' systems, and responded to data received from the market center. Over time, this functionality has moved to market centers' computers. The BATS market, for example, offers all these types. To keep the IT architecture clean (and fast), the systems that handle the special orders are often kept separate from the system that runs the basic limit order market.

More complex algorithms may involve multiple orders and rely on broader information feeds (such as prices for other securities or markets). These algos are often implemented by brokers, who make them available to customers. This reflects a broad diffusion of expertise. Algorithms that were formerly used only by sophisticated proprietary trading shops are now available at the retail level.

Summary of terms and concepts

Stop orders; trailing stops; pegged orders; discretionary orders; reserve/iceberg orders; post-only orders.

Chapter 16. Transaction Cost Analysis (TCA)

In the typical investment firm, the portfolio manager will make decisions about which securities to buy, sell, or hold. These decisions will be communicated to the firm's trading desk, which will make the trading decisions (how fast? which exchanges? which brokers? what mix of orders?).

At the end of the quarter the investor will see one number, hopefully a profit. He may neither know nor care what decisions and processes led to the final number. The firm's managers, though, will be aware of all the steps. Portfolio allocation decisions are usually based on judgements of long-term value and risk of the candidate stocks; implementation decisions involve short-run assessments of markets and trading opportunities. Investment and trading typically require different perspectives, models, data, and perhaps most importantly, different people. Was an investment profitable due to an astute portfolio manager, or to the expertise of the head trader? How should we divide the bonus?

Like most managerial accounting, measurement of trading costs can indicate to management the size and sources of costs. The process might also suggest ways in which they can be minimized. For the customers or beneficiaries of an investment fund, reporting of trading costs can help evaluate management. How do the fund's costs compare with its peers? There are also legal considerations. Asset managers usually have a fiduciary duty to act on behalf of their investors or beneficiaries. When managers make trading decisions, the costs are ultimately borne by these investors. US market centers are responsible for computing and reporting trading costs. These are posted to their web sites.

16.1. The Implementation shortfall approach to trading cost

In the portfolio implementation shortfall approach, we assume a separation between investment and trading decisions. Long term investment strategies are made by portfolio managers. They make clear decisions about what to buy, sell and hold. These decisions are implemented by a trading desk.

We compare:

- The performance of an actual portfolio (gain, loss, or return) and
- The performance of an imaginary paper portfolio in which all trades are made at benchmark prices.

A benchmark price is simply a reference price that is supposed to represent the security's true value at the time the buy or sell decision is made. The performance of the paper portfolio is obviously going to depend very strongly on our choice of benchmark. We will discuss the alternatives later in this chapter, but a common choice is the average of the bid and ask prices at the time of decision, and so for the moment let's fix on this value. This average falls halfway between the bid and ask, and so it is usually called the bid-ask midpoint (BAM).

The portfolio implementation shortfall is the difference. For example, if the return on the paper portfolio is 10% and the return on the actual portfolio is 9%, the implementation shortfall is 1%. The idea here is that with perfect markets (and a perfect trading desk), our trades would have been executed at BAMs. Any divergence between the actual and paper returns must be attributed to trading (implementation) costs.

The portfolio implementation shortfall, as originally proposed, includes both explicit and implicit costs. The key explicit costs are:

- Commissions, net of any rebates
- Taker fees and liquidity rebates not included in the commission (discussed in Chapter 19).
- Transactions taxes

In the actual portfolio, the explicit costs are paid out of a cash account, and they are usually clearly identified on the trade confirmations that the broker sends to the investor. The implicit costs are less visible. They include:

- Costs of interacting with the market (e.g., bid-ask spread or price impact costs), relative to the benchmark prices.
- Opportunity costs (the penalty associated with *not* completing intended trades). Examples of this include
    - The failure of a limit order to execute because the market has moved away from the limit price.
    - Failure to complete a hedging trade, which may leave the portfolio exposed to additional risk.
- Delay (failure to fill the order immediately).

The implicit costs can be very difficult to assess.

The implementation shortfall framework is generally attributed to Andre Perold (Perold, 1988). Perold originally proposed the portfolio implementation shortfall at the level of the portfolio: one number reflecting all the trades and activity. Nowadays, however, the implementation shortfall is usually computed for individual orders that are executed in full. It is defined as:

$$Implementation\ Shortfall = \begin{cases} Trade\ Price - Benchmark\ Price, & for\ a\ buy\ order \\ Benchmark\ Price - Trade\ price, & for\ a\ sell\ order \end{cases} \quad (1)$$

That is, for a buy order, the implementation is how much we overpay relative to the benchmark; for a sell order, how much less we receive. In the case where a large *parent order* is executed over time in a sequence of smaller *child orders*, an average trade price is used. This is more in line with "the implicit cost of interacting with the market" in Perold's original formulation. The Investment Technology Group (ITG) transaction-cost reports, for example, use the breakdown:

$$Total\ Cost = Implementation\ Shortfall + Commissions$$

Henceforth, we'll use this narrower definition.

16.2. Benchmark prices used in IS calculations

The implementation shortfall calculation depends crucially on the choice of the benchmark price. The possibilities are often grouped according to where they are (before/after/during) relative to the trade. *Pre-trade benchmarks* include:

- The NBBO midpoint at the time the trading or order submission decision was made.
- The previous day's closing price.

When a pre-trade benchmark is used, implementation shortfall is sometimes referred to as *slippage*. Examples of *post-trade benchmarks* are:

- The NBBO midpoint prior five minutes after the trade.
- The next day's opening price

Interval benchmarks are also sometimes used:

- Time-weighted average price (TWAP, "Tee Wap") over the day or duration of the order.
- Volume-weighted average price (VWAP, "Vee Wap") over the day or duration of the order.

For individual trades, the prior NBBO midpoint and the NBBO midpoint at trade time + five minutes are popular choices.

When the prior NBBO midpoint is used as a benchmark for a parent order, the whole sequence of child orders is judged relative to the initial midpoint, typically taken when the parent order is sent (by the portfolio manager) to the portfolio manager's trading desk, or when the parent order is sent to the broker.

An institution will typically lack the ability to directly monitor the NBBO midpoints. For these, they are dependent on their brokers' reports. The NBBO can change significantly over the course of a few milliseconds. An institution may suspect that the broker is "gaming" the measure, choosing from a set of nearly simultaneous NBBO records the one that gives the most favorable IS. TWAP and VWAP are easier to compute. VWAP, in particular, is very widely used as a benchmark.

16.3. Effective and realized costs

Effective and realized costs are implementation shortfall calculations for marketable orders, that is, orders that are executed on arrival.

The effective cost uses a benchmark price equal to the midpoint of the prevailing visible bid and offer. Letting m denote the bid-offer midpoint, letting p denote the execution price, the effective cost is defined as:

$$Effective\ Cost = \begin{cases} p - m, \text{for a marketable buy order} \\ m - p, \text{for a marketable sell order} \end{cases}$$

The effective spread is simply twice the effective cost. If all buys executed at the NBO, and all sells at the NBB, the effective and posted spread would be the same. Accurate timing can be difficult. An institution sending the order would probably take the midpoint at the time the order was sent. If the market center is doing the calculation, the midpoint is taken as of when the order was received. Usually, these times differ only by a few milliseconds, but in a volatile market even this small difference can meaningfully affect the outcome of the calculation

As a result of a dark trade, execution against a hidden limit order, or a concession by a dealer, the execution price may lie within the posted NBBO. The amount by which the quote is bettered is called price improvement:

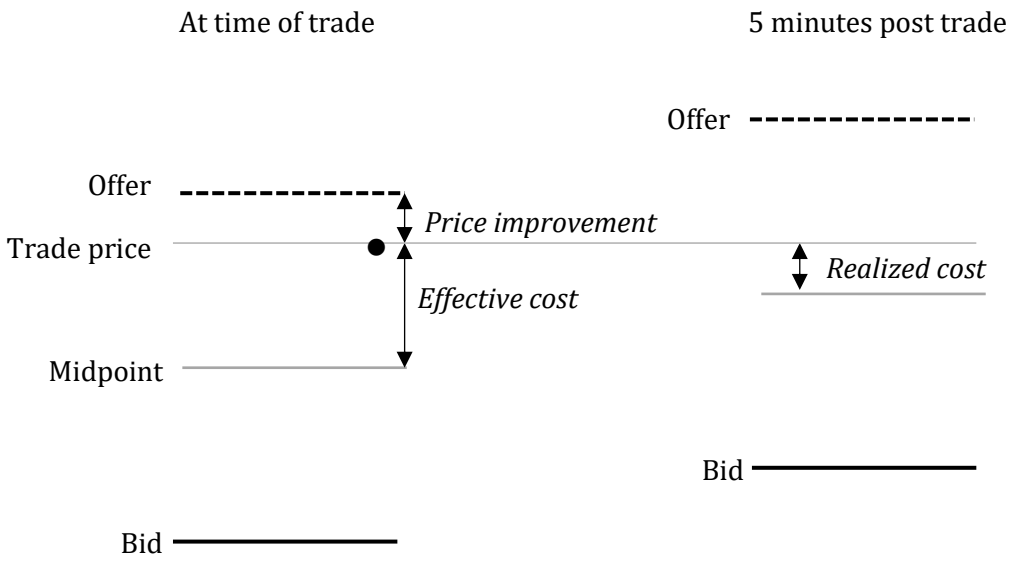
$$Price\ improvement = \begin{cases} NBO - p, & \text{for a marketable buy order} \\ p - NBB, & \text{for a marketable sell order} \end{cases}$$

The realized cost uses a post-trade benchmark. Any post-trade benchmark might be used, but for regulatory purposes, the SEC mandates the NBBO midpoint prevailing five minutes after the market receives the order (or the closing NBBO midpoint). Denoting this midpoint as m_5:

$$Realized\ Cost = \begin{cases} p - m_5, \text{for a marketable buy order} \\ m_5 - p, \text{for a marketable sell order} \end{cases}$$

Figure 16-1 depicts these quantities for a marketable buy order. The realized cost is on average less than the effective cost under the presumption that the price impact of the order is positive. The realized cost can be interpreted as the trading profit made by the dealer (or other trader) who acted as counterparty to the marketable order net of the impact of the trade, and assuming that they could reverse the trade at the then-prevailing quote midpoint.

Figure 16-1 Effective cost, realized cost and price improvement for a marketable buy order.



The quantity $Effective\ Cost - Realized\ Cost$ measures the movement of the quote midpoint from the trade to the five-minute mark. It is therefore an approximate measure of the price impact of the order. The correspondence is only approximate because the quote midpoint change over the interval is driven by *all* of the trades prior to the five-minute mark – not just the single trade in question.

Table 16.1 describes some sample calculations. The left portion of the table gives the NBBOs over the interval; the right portion contains information on three orders. The table is arranged so that the NBBO to the left of an order describes the NBBO prevailing at the order time. Note that the realized cost is in one instance negative. We would not expect this to be the case on average, because that would imply a trader generally buying before the price goes up and generally selling before a decline. For an individual trade, however, the five-minute delay used to set the benchmark price can contain substantial price variation, and almost anything is possible.

Table 16.1 Calculation of effective cost, realized cost, price improvement and VWAP for marketable orders.

Quotes					Trades								
Time	NBB	NBO	BAM	Spread	Order Time	Dir	Vol	Exec Price	Effective Cost	Price Improv	BAM_5	Realized Cost	Vol x Pr
10:14:00	19.81	19.87	19.840	0.06									
10:15:00	19.80	19.85	19.825	0.05	10:15:09	B	300	19.85	0.025	0.00	19.855	-0.005	5,955
10:16:00	19.78	19.83	19.805	0.05									
10:17:00	19.85	19.90	19.875	0.05	10:17:22	B	300	19.88	0.005	0.02	19.860	0.020	5,964
10:18:00	19.92	19.97	19.945	0.05									
10:19:00	19.91	19.97	19.940	0.06									
10:20:00	19.84	19.87	19.855	0.03	10:20:09	S	400	19.84	0.015	0.00	19.860	0.020	7,936
10:21:00	19.92	19.96	19.940	0.04									
10:22:00	19.84	19.88	19.860	0.04									
10:52:00	19.93	19.97	19.950	0.04									
					Total Volume:		1,000				Total (Vol x Pr):		19,855
												VWAP:	19.855

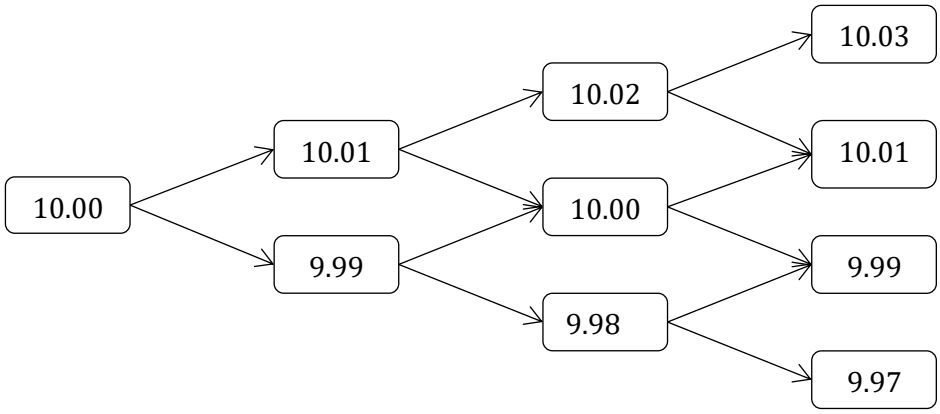
Notes: BAM is the bid-ask midpoint as of the indicated quote time; BAM_5 is the bid-ask midpoint five minutes after the indicated order time.

16.4. Implementation shortfall and limit orders

A general principle of trading is that urgency has a cost. If you want to trade cheaply, trade patiently and passively. Why go to someone else's price? Why not wait and let them come to yours? This principle argues in favor of limit orders and generally passive strategies. The principle is a sound one, and implementation shortfall analysis can usually measure the reward to patience. In practice, though, the computations might well be biased due to neglect of failed executions.

Figure 16-2 illustrates a typical situation. The setup uses a binomial model to describe the dynamics of the stock price. The offer price starts at 10.00. At every time step (say, every minute), the offer can go up by \$0.01 or down by \$0.01, with equal probabilities. Over the first three minutes, then, there are $2^3 = 8$ possible paths, all equally likely.

Figure 16-2



Suppose that we want to buy. One approach is to trade immediately by lifting the offer price of \$10.00 with a market order. Alternatively, we could try to trade passively. Suppose that we put in a buy order, limit \$9.99. If this order executes, we will save \$0.01 over a market order. For any pre-trade benchmark, \$0.01 will be difference in implementation shortfalls. It is therefore convenient to take \$10.00 as the pre-trade benchmark.

If we submit a hundred limit orders in this situation and analyze our costs only for the orders that fill, we'll conclude that the implementation shortfall of the limit order strategy is -0.01, that is, that we're trading at negative cost.

Now what is the probability of a limit order fill? This order will execute whenever a path hits 9.99. If we are willing to wait for three minutes, there are five paths on which this will occur. (Using *d* and *u* to denote down and up: *duu, dud, ddu, ddd,* and *udd*.) So, the probability of execution is $5/8 = 0.625$ – better than even chances.

But what happens when the limit order doesn't execute? If we don't care, then we're okay ignoring the execution failures. But if we don't care, why were we contemplating using a market order? Why were we in the market at all?

Suppose that we're required to purchase the stock within three minutes. This means that if the limit order hasn't executed by the third step, we'll have to cancel it, and lift the offer with a market order. The probability of the limit order failure is $1 - 0.625 = 0.375$.

If the limit order fails, what will we have to pay to acquire the stock? In one of the failure paths (*uuu*) the offer is 10.03; in two of the failure paths (*udu* and *uud*) the offer is 10.01. So, the expected offer conditional on a limit order failure is

$$E[offer|limit\ order\ fails] = \frac{1}{3} \times 10.03 + \frac{2}{3} \times 10.01 = 10.0167$$

Our overall expected cost of buying with the limit order strategy is

$$E[cost] = 9.99 \times P(LimEx) + 10.0167 \times P(LimFail)$$

where *LimEx* and *LimFail* denote the success and failure of the limit order. So:

$$E[cost] = 9.99 \times 0.625 + 10.0167 \times 0.375 = 10.00$$

This is exactly the cost of using a market order. A limit order priced at \$9.99 does not on average outperform a market order, and it has outcome uncertainty.

Perhaps the limit order was too aggressive. Suppose we submit an order to buy limit \$9.98. If we restrict our analysis to the outcomes in which the order executes, we'll measure an implementation shortfall of −\$0.02. There are two paths on which this order will execute (*ddd* and *ddu*). So the probability of execution is $2/8 = 0.25$. If

$$E[offer|limit\ order\ fails] = \frac{1}{6} \times 10.03 + \frac{3}{6} \times 10.01 + \frac{2}{6} \times 9.99 = 10.0067$$

and

$$E[cost] = 9.98 \times 0.25 + 10.0067 \times 0.75 = 10.00.$$

The practice of using a market order after a failed attempt with a limit order is called "chasing the market". The limit order fails to execute because the price has moved in the wrong direction (up, if we are trying to buy), and we have to "chase" a price that's running away from us.

The strategy of first trying limit orders to achieve a passive execution, and then switching to a market order at the deadline is a fairly representative limit order strategy. In fact, the Tokyo Stock Exchange has a specific order type, the funari order, that works in exactly this fashion.

The binomial model is used in many financial settings, particularly in option pricing, where great reliance is placed on its validity. In the present case, though, if this model's depiction of the world were accurate, no one would ever use a limit order. It is quite possible that execution would occur before the ask price moved to the limit price, as an urgent liquidity trader or other latent seller might well hit an aggressive bid.

It is moreover likely that the approach of imputing a market-order fill after a limit order failure overstates the opportunity cost. It is always feasible to submit such a market order, and the fact that this step is often not taken suggests that it is perceived as an unnecessary expense. There may be alternative securities or alternative hedges available that can substitute at a better price.

The main point of the analysis stands, however, which is that limit orders evaluated with no penalty for failure will always appear superior to market orders. Limit orders that are priced less aggressively will appear better yet.

16.5. Estimating trading costs

Most institutions compute implementation shortfalls for a sample of their trades, estimate averages, and compare these averages across brokers, algorithms, and routing destinations that the firm employs. These results then guide future order submission choices. This sort of analysis is sensible and useful. As often applied, though, the process neglects opportunity costs and the interaction with the firm's investment and trading strategies.

The opportunity costs of failed limit orders were discussed in the last section. The point generalizes, however, to most passive strategies, including those that follow dynamic limit order strategies. As an example, in the situation of Figure 16-2, the strategy of initially submitting an order to buy limit \$9.99 but repricing it if the offer moves away. So, if the offer moves to \$10.01, we'd reprice our buy at \$10.00. This is a pegged limit order (discussed further in section 15.2). It will have a higher execution probability than a limit order submitted once and for all at \$9.99, but it will have some nonzero failure probability, which must be considered in the implementation shortfall analysis.

Interactions with the firm's investment and trading decisions also affect the validity of the analysis. Ideally, following the protocol of drug clinical trials, we'd perform experiments on random orders in random stocks, with random quantities in random direction (buy or sell), and randomized execution strategies. In this way, we'd achieve unbiased estimates of "treatment" effects.

Some firms and investors *do* perform experiments, but they are expensive, as they involve buying stocks that one does not really wish to own, and trading in ways that are "obviously" inefficient. The more common practice is to analyze the orders actually generated by the portfolio managers, and the executions actually achieved by the trading desk.

An example shows why this can lead to problems. Suppose that the portfolio desk generates two kinds of orders: rebalancing orders that simply seek to keep the overall portfolio close to some desired allocation weights, and momentum orders that try to profit on short-term price movements. The rebalancing orders are sent to broker *R* with instructions to execute gradually over time; the moment orders are sent to broker *M* and flagged as urgent. In a cross-broker comparison of implementation shortfall, we'd expect broker *R* to have the lower costs. This is not due to any special ability, but rather to the sort of orders we send.

A more subtle interaction occurs when our strategies are also being used by others and our orders are correlated with others'. The price change associated with our order, then, reflects not only our order, but also the orders originating from all other trading strategies.

Institutions do not report their trading costs directly to investors or in SEC filings. They do, however, often share data with other institutions. Some firms (e.g., ITG, Abel-Noser, TAG) produce aggregate reports of trading costs. A mutual fund might report its trade data to one of these firms. The firm then compiles summary statistics (disguising the identity of any individual fund).

16.6. Implementation shortfall decompositions

Once we settle on a benchmark price, the implementation shortfall is easily computed. The resulting number, though, doesn't give us much information about the source of the costs. Nor does it give us guidance about how we might adjust our behavior to reduce our trading costs.

To illuminate the source of the costs, therefore, the implementation shortfall is often decomposed further. As a first step, the implementation shortfall is sometimes decomposed as:

$$Implementation\ shortfall = price\ impact + cost\ of\ delay \tag{2}$$

Price impact reflects the price movements induced by our executions. A portion of price impact may be temporary, as in the case where other market participants conclude that our orders are uninformed, and the price reverts. The impact may also be permanent if our orders are viewed as potentially informed. Price impact cannot be observed directly. (We traded, and the price moved, but did we *cause* the movement?) It is usually inferred from a statistical model of order-price dynamics.

The cost of delay is the remainder or residual. It captures the tendency of the price to move against us, apart from the impact of our orders. This tendency (also known as "slippage") is commonly observed. Wayne Wagner, perhaps the main proponent of this measure, estimates the cost of delay for institutional trades in US stocks around 2003 as 77 basis points (Wagner, 2003). By comparison, a $0.05 bid-ask spread on a $50 stock is 10 basis points. The source of delay costs, though, is unclear. We might be incurring them if others become aware of our intentions and trade ahead of us, something that might happen if our orders are being leaked.

A plausible alternative explanation, though, is that others are using strategies similar to ours or responding to the same signals. Under this mechanism, the aggregated flood of orders arriving at the market in a short interval will cause a large apparent price impact that the individual traders may attribute to delay. Across all these traders the total delay cost might

therefore appear to be very large. From another perspective, though, these are the total costs of losing the race for order execution, a race that can only have a few top-ranked finishers.

16.7. SEC Rule 605

SEC Rule 605 requires market centers (exchanges, dealers, dark pools, etc.) to provide detailed reports on the orders they receive and the outcomes. The reports cover order counts, share counts, execution rates, cancellation rates, effective spreads and realized spreads. (Effective and realized spreads are simple two times the effective and realized costs.) These statistics are broken down by stock, trade size, and pricing relative to the same-side quote. Monthly figures must be published on their websites. The SEC mandates a standardized format (U.S. Securities and Exchange Commission, 2001).

Table 16.2 contains a portion of the Rule 605 statistics for BATS (specifically, the BATS "Z" exchange) for November 2009 for ticker EIHI (Eastern Insurance Holdings, Inc., a diversified insurance concern). In this period, BATS received only two market orders, but 363 marketable limit orders of size 100-499 shares. (This reflects the reluctance of traders and market centers to use unpriced orders.) There were 37,002 shares in these orders, of which 34,400 were cancelled, 1,122 were executed at BATS, and 1,480 were routed elsewhere. All the executions were accomplished in under nine seconds. For these shares the average effective spread was \$0.0930, and the average realized spread was \$0.0685. This implies effective and realized *costs* of \$0.0465 and \$0.0343. The difference, \$0.0122 is the implied price impact (see the figure above). In other words, an incoming 100 share buy order should raise the market share price by about \$0.0122.

Table 16.2. BATS Rule 605 statistics for EIHI, November 2009.

Order type	Order size	No. orders	Shrs in orders	Shrs canc	Shrs exec	Shares exec elsewhr	Eff spread	Rlzd spread	Shrs rcvng price improvmnt	Avg price imprvmnt
Market	1-499	2	200	0	100	100	0.050	0.045	200	0.0475
Mkt'l limit	1-499	363	37,002	34,400	1,122	1,480	0.093	0.069	1228	0.0227
Mkt'l limit	500-1,999	10	5000	0	0	5,000	0.198	0.190	0	

16.8. Further reading

Because high liquidity is sometimes viewed as functionally equivalent to "low trading cost," most liquidity measures are negatively related to most estimates of trading costs. (Holden and Jacobsen, 2014) provide an excellent recent survey of alternative measures and their relations.

Summary of terms and concepts

Implementation shortfall; explicit and implicit costs; effective cost; realized cost; price improvement; market impact; opportunity costs for unexecuted limit orders; delay costs; SEC Rule 605.

References

Holden, Craig W., and Stacey Jacobsen, 2014, Liquidity Measurement Problems in Fast, Competitive Markets: Expensive and Cheap Solutions, *The Journal of Finance* 69, 1747-1785.

Perold, Andre, 1988, The implementation shortfall: paper vs. reality, *Journal of Portfolio Management* 14, 4-9.

U.S. Securities and Exchange Commission, 2001, Staff Legal Bulletin No. 12R (revised): Frequently Asked Questions About Rule 11Ac1-5 [Rule 605], in Division of Market Regulation, ed.

Wagner, Wayne H., 2003, Testimony, House Committee on Financial Services, Subcommittee on Capital Markets, Insurance, and Government Sponsored Enterprises, March 12, 2003.

Chapter 17. Order Splitting

Updated August 2023

A $1 Billion mutual fund might hold 1% of its assets ($10 Million) in a single stock. At a representative share price of $50, this holding consists of 200,000 shares. The median trade size in US equity markets is about 200 shares. It is therefore almost certain that should the fund seek to sell or reallocate the holding, the shift will have to be accomplished in multiple trades. A larger order, like "sell 200,000 XYZ" will typically be divided among many smaller "child" (or "daughter") orders that are fed to the market over time. This is called order splitting.

For each child order, we face complex problems of how to peg, how much to show, what discretionary conditions should be used, whether to make our orders visible or dark, and so on. The larger problem, though, involving the number, size and timing of the child orders is more tractable.

An order splitting strategy is usually developed by minimizing the cost of buying (or maximizing the receipts from selling) relative to some benchmark price. Equivalently, we can view the problem as minimizing the implementation shortfall relative to some benchmark, following Chapter 16. The present chapter examines two situations. In the first, the benchmark is an average price computed over the working period or interval of the parent order. In the second instance, we're using a pre-trade benchmark like the bid-ask midpoint at the time the parent order is given to the trading desk. The second approach is also distinctive in that it explicitly focuses on the impact of our orders on the market price. We turn to each case in turn.

17.1. Order splitting with interval benchmarks

Among the candidates for benchmarks, we encountered the time-weighted average price (TWAP) and the volume-weighted average price (VWAP). Now if one is trying to buy as low as possible, or sell as high as possible, why should either of these, (or indeed, any benchmark) be considered as a *goal*? The answer is that sometimes investors are more concerned with monitoring and measuring the performance of their broker.

For example, suppose that at 9:30 on Monday, the NBBO midpoint stands at $10, the manager of the Essex fund sends a large buy order to a broker to be worked over three days, and that Essex ends up paying an average price of $10.20 per share. The implementation shortfall cost is $0.20 per share. Essex confronts the broker with this large trading cost, but broker simply replies, "Sorry, but the market was generally tending up over those three days." But if the volume-weighted average price over the period was $10.15, Essex can come back with, "Alright, the market trend can explain $0.15 of the implementation shortfall, but that still leaves $0.05. Other brokers working for other customers over this same period somehow managed to buy *their* shares $0.05 cheaper than ours."

Now perhaps Essex should be evaluating the original instruction. Why was the horizon three days? If it had been one day, the overall cost would probably have been lower. What was the manager thinking? There will be answers to these questions, but they will be qualitative, imprecise, and probably constructed after the fact. The implementation shortfall relative to VWAP, on the other hand, is easily measured, interpreted, and assigned to a particular agent (the broker). The measurement issue also arises at higher levels of delegation. It is likely that the Essex manager is herself bearing a fiduciary responsibility to manage and report trading costs. TWAP and VWAP are relatively easy to justify and to achieve.

TWAP is the simpler of the two. Because the passage of time is perfectly predictable, one simply trades at a constant rate over the working horizon. If 80,000 shares are to be purchased over four hours, then 20,000 must be purchased in each hour, 5,000 over each fifteen minutes, or 1,000 every three minutes. This does not, of course, guarantee that TWAP will be achieved, because it depends on the trades of others as well as our own. There also remains the question of how to manage the child orders. But the strategy of using a constant rate of trade is certainly the place to begin.

VWAP is more complicated. In aiming at TWAP we know the total duration of the trading day. If we knew the total volume on the day would be, say, V, we could aim at a fixed fraction of this volume. If we seek to buy v shares (assumed to be included in V, for simplicity), we'd want to buy a fraction (v/V) of each trade. For example, if we want to buy $v = 5{,}000$ shares on total volume $V = 100{,}000$ shares, we'd want to be buying 5 of every 100 shares traded. This would require participating in each trade. Although market makers are sometimes allowed rights of participation, this ability is not usually extended to other traders.

If the trading rate is constant over the day, then VWAP and TWAP are identical. They differ because the trading rate (volume per unit time) changes throughout the day. Some of this variation is regular. Volume tends to be "U"-shaped: elevated at and immediately after the open, declining and leveling off during mid-day, and rising again toward the end of regular closing hours. (Madhavan, 2002) provides an example (Microsoft).

We can estimate the average trading rate by examination of behavior on previous days. Table 17.1 illustrates the approach for a planned VWAP trade of 5,000 shares. We divide the day into thirteen half-hour intervals starting at 9:30 AM (columns A and B). In column (C) we enter the average volume in each interval, estimated over a sample of previous trading days. Column (D) reports the proportion of volume for each interval. That proportion is applied to the planned order size (5,000 shares) to get the planned trading volume in each interval (column E). This calculation distributes the trades over time in a way matches the average volume profile of the stocks throughout the day.

TWAP and VWAP strategies simply distribute orders over time. They do not explicitly take into account the total size of the trade relative to overall market activity. If a stock has an average daily volume of 1,000 shares, then a sell order for 50,000 shares being worked over the day will roil the market if the strategy is TWAP, VWAP or just about anything else. To avoid extreme market impacts, one can impose participation constraints. For example, "sell 50,000 VWAP, but our own trades should not exceed ten percent of the total volume."

Table 17.1 Planning a VWAP trade for 5,000 shares

(A) Interval	(B) Start	(C) Avg Vol	(D) Proportion	(E) Planned Trade Volume
1	9:30 AM	11,900	9.8%	492
2	10:00 AM	10,180	8.4%	421
3	10:30 AM	9,440	7.8%	390
4	11:00 AM	8,500	7.0%	352
5	11:30 AM	7,820	6.5%	323
6	12:00 PM	6,380	5.3%	264
7	12:30 PM	6,540	5.4%	270
8	1:00 PM	7,480	6.2%	309
9	1:30 PM	8,740	7.2%	361
10	2:00 PM	8,720	7.2%	361
11	2:30 PM	10,320	8.5%	427
12	3:00 PM	12,000	9.9%	496
13	3:30 PM	12,880	10.7%	533
	Total	120,900	100.0%	5,000

Participation constraints, though, can have unexpected consequences. In the May 6, 2010 "flash crash", the S&P 500 index futures contract fell about 6% over the course of a few minutes. The precipitating event was a large trade. According to the joint CFTC-SEC report (U.S. Commodity Futures Trading Commission and Commission, 2010):

> A large fundamental trader (a mutual fund complex) initiated a sell program to sell a total of 75,000 E-Mini contracts (valued at approximately $4.1 billion) as a hedge to an existing equity position. … This large fundamental trader chose to execute this sell program via an automated execution algorithm ("Sell Algorithm") that was programmed to feed orders into the June 2010 E-Mini market to target an execution rate set to 9% of the trading volume calculated over the previous minute, but without regard to price or time.

An algorithm that was more aware of the impact of its orders might not have behaved so wildly.

(Madhavan, 2002) discusses alternative ways of pursuing VWAP. Instinet runs a crossing session in which buyers and sellers are matched in advance, for a given quantity, before the start of the trading session. At the end of the day, VWAP is determined and the price is set. This option is only available, of course, if a matching counterparty can be found.

17.2. Order splitting with price impact

When working a large parent order, a pre-trade benchmark is known before the first child order is submitted. The benchmark might be the midpoint of the bid and ask immediately prior to the first trade. It could be taken from a much earlier point, though, perhaps yesterday's closing price.

In this analysis, there are some similarities to the situation discussed in the last section. As above, we divide the overall trading horizon into smaller intervals. For example, if the current time is 9:30am and we must trade the parent quantity within the next three hours, we might

divide the trading process into $t = 1, \ldots, T = 36$ five-minute subperiods (9:30 to 9:35, 9.35 to 9.40, …, 12:25 to 12:30). We depart from the last section, though, in that we try to model the behavior of the security price over the trading horizon, emphasizing the price impact of our trades. The framework in the following discussion is due to (Bertsimas and Lo, 1998). (Almgren and Chriss, 1997; Engle, Ferstenberg and Russell, 2012; Kissell and Glantz, 2003) offer alternative treatments.

The two-period case

The problem with $T = 2$ is useful for illustrating the approach and developing intuition. We'll begin with an initial price, p_0, which you might think of as the price from the 9:30 opening auction. The interval designated by $t = 1$ runs from 9:30 to 9:35. At the end of this interval, we assume that the price will be:

$$p_1 = p_0 + \lambda\, q_1 \tag{17.1}$$

where q_1 is the size of our trade in the interval, signed to indicate direction. For example, if $q_1 = 100$, then we've bought 100 shares and if $q_1 = -150$ then we've sold 150 shares. λ is the price impact parameter. The product $\lambda\, q_1$ measures how much our trade has moved the price. For example, if $\lambda = 0.0001$ then a 1,000-share purchase would move the price by $0.0001 \times 1{,}000 = \$0.10$ per share. Equation (17.1) is *recursive*. That is, we can apply it again, to find the price at the end of interval $t = 2$, that is, from 9:35 to 9:40:

$$p_2 = p_1 + \lambda\, q_2$$

Using substitution (for p_1), the two prices are

$$\begin{aligned} p_1 &= p_0 + \lambda\, q_1 \\ p_2 &= p_1 + \lambda\, q_2 = p_0 + \lambda(q_1 + q_2) \end{aligned} \tag{17.2}$$

The last equality expresses p_2 as the sum of the initial price, p_0, and the cumulative price impacts of our orders.

The model is obviously minimal: it is missing many realistic and desirable features. There's only one price in each period (no bid or ask). Our orders move prices, but those of other traders do not. In fact, the price responds only to our orders. There are no surprises from public information. We'll later discuss modifications to incorporate these features. But for the moment, view the model as a device to highlight and focus on one thing: the price impact of our orders.

To illustrate the approach, let's assume that we know the value of λ and the starting point p_0, and that our assignment is to buy a total of Q shares over $T = 2$ periods. We must plan our trades over these two periods, q_1 and q_2, so that $q_1 + q_2 = Q$, the quantity constraint. Our objective is to minimize the total purchase cost $C = p_1 q_1 + p_2 q_2$. This is equivalent to minimizing the average price per share $(p_1 q_1 + p_2 q_2)/Q$. It is also equivalent to minimizing the implementation shortfall $IS = (p_1 q_1 + p_2 q_2)/Q - M$ where M is any pre-trade benchmark. Formally, then, the problem is:

$$\underset{q_1, q_2}{Min}\ C = p_1 q_1 + p_2 q_2 \text{ subject to } q_1 + q_2 = Q \tag{17.3}$$

If p_1 and p_2 were known and constant, this would be easy. If $p_1 = p_2$ then our choice doesn't matter. We could set $q_1 = Q$ and $q_2 = 0$; we could set $q_1 = 0$ and $q_2 = Q$. For that matter, we could set q_1 and q_2 to anything as long as they sum to Q. If $p_1 \neq p_2$ then we simply buy all Q shares in the period with the lowest price.

Generally, though, the prices are not constant: p_1 depends on q_1; p_2 depends on q_1 and q_2. Using the prices given in equations (17.2) the cost is

$$C = p_1 q_1 + p_2 q_2 = (p_0 + \lambda\, q_1) q_1 + \big(p_0 + \lambda(q_1 + q_2)\big) q_2 \tag{17.4}$$

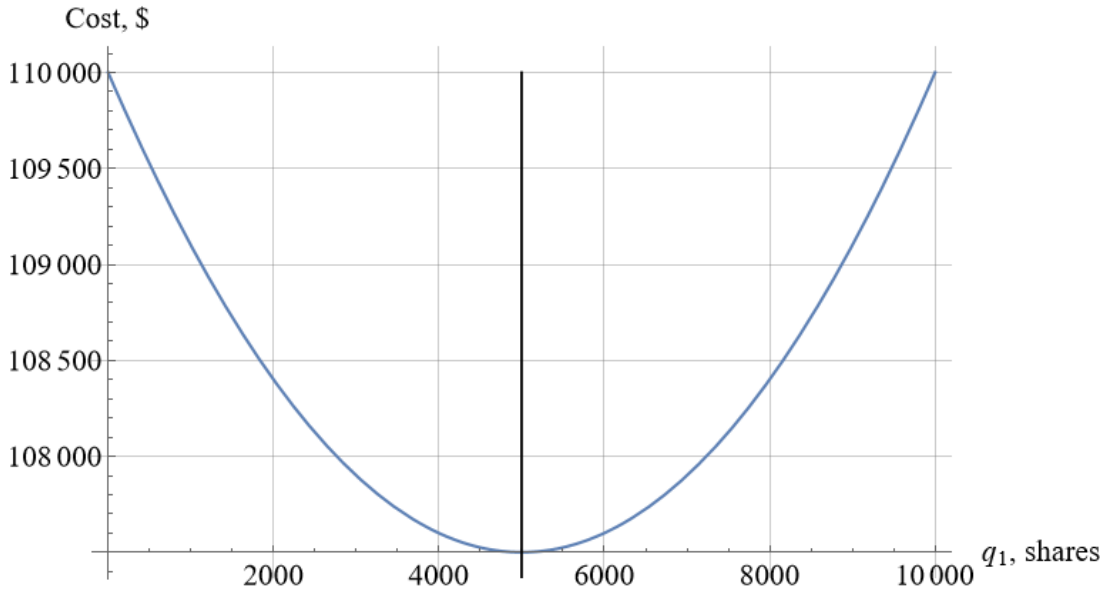
Taken with the constraint $q_1 + q_2 = Q$ this may be formulated as a classic constrained optimization problem. For present purposes, though, it is simpler to use the constraint to eliminate one of the variables. That is, setting $q_2 = Q - q_1$ in (17.4) gives:

$$\begin{aligned} C &= (p_0 + \lambda\, q_1) q_1 + (p_0 + \lambda[Q - q_1])[Q - q_1] \\ &= Q\, p_0 + \lambda(Q^2 - Q\, q_1 + q_1^2) \end{aligned} \tag{17.5}$$

This last expression expresses the cost as the combination of an outlay Qp_0 (assuming that we could buy all Q shares at the initial price) and an additional outlay that depends (through λ) on the impact of our trades.

For illustration, consider the numerical values, $\lambda = 0.0001, p_0 = \$10$ and $Q =$ 10,000 shares. (The current price is \$10 per share; we must buy 10,000 shares in the next two periods.) The cost is graphed in *Figure 17-1*. Note that the worst outcomes (highest costs) correspond to concentrating the trading in one period ($q_1 = 0$ or $q_1 = 10{,}000$). Visually, we can see that the best outcome (lowest expenditure) is obtained by trading that is evenly balanced ($q_1 = 5{,}000$ *and* $q_2 = Q - q_1 = 10{,}000 - 5{,}000 = 5{,}000$).

Figure 17-1 Trading cost and trade timing



Mathematically, we can find the minimum cost by setting the derivative of the cost with respect to q_1 equal to zero:

$$\frac{dC}{dq_1} = \frac{d}{dq_1}[Q\ p_0 + \lambda(Q^2 - Q\ q_1 + q_1^2)] = \lambda(-Q + 2q_1) = 0$$

This implies that the optimal trade size is $q_1^* = Q/2$. At the optimum the cost is $C^* = 3Q^2\lambda/4 + Q\ p_0$, or (with the numerical values suggested above), \$107,500.

Before we began to trade, the market price was \$10 per share, so 10,000 shares would have been valued at \$100,000. The extra \$7,500 that we incurred in executions is sometimes called *slippage*. Note that the cost of the optimal execution strategy goes up as the square of Q.

This approach can be generalized to consider richer and more realistic problems. We next consider some of the common modifications.

Drift

Sometimes we find ourselves attempting to trade during a time when security price is trending. This might happen when the price is responding to new information. The trend is captured by a drift parameter α, which might be positive or negative depending on the direction of the trend. The price equations are modified as:

$$p_1 = p_0 + \alpha + \lambda\, q_1$$
$$p_2 = p_1 + \alpha + \lambda\, q_2 = p_0 + 2\alpha + \lambda(q_1 + q_2)$$

In these expressions, the drift captures predictable price changes that are not affected by our trades. The execution cost then becomes:

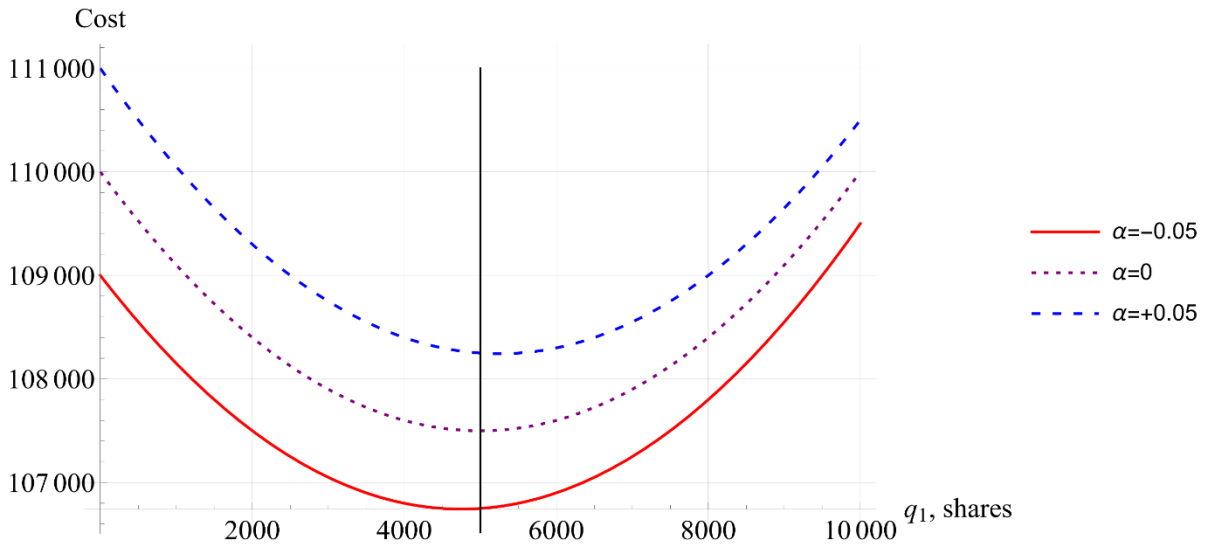
$$C = p_1 q_1 + p_2 q_2 = (p_0 + \alpha + \lambda\, q_1)q_1 + \big(p_0 + 2\alpha + \lambda(q_1 + q_2)\big)q_2$$

The constraint is unchanged: $q_1 + q_2 = Q$. Substituting with $q_2 = Q - q_1$ gives:

$$C = Qp_0 + Q(2\alpha + Q\lambda) - (\alpha + Q\lambda)q_1 + \lambda q_1^2$$

We continue with the numerical values used in the prior example: $\lambda = 0.0001, p_0 =$ \$10 and $Q = 10{,}000$ shares. For the drift, we consider three cases: $\alpha = -0.05$, $\alpha = 0$, and $\alpha = +0.05$. Figure depicts the three cost curves. The middle curve, with $\alpha = 0$ is the no-drift case, a parabolic curve centered at $q_1 = 5{,}000$ shares. The other curves are skewed left or right, depending on the sign of the drift. The top curve (blue dashes) corresponds to a positive drift. It is skewed right, with a minimum that lies slightly above 5,000 shares. Since we are buying in a rising market, we should speed up our purchases, setting $q_1 > 5{,}000$. The bottom curve (solid red) corresponds to negative drift. We are buying in a dropping market, and so will do better by buying fewer shares in the first period ($q_1 < 5{,}000$) and buying more in the second period. Even with non-zero drift, though, we are penalized by price impact for large trades.

Figure 17-2 Trading costs and drift



Mathematically, we can find the optimum by setting $dC/dq_1 = 0$ and solving for q_1^*:

$$\frac{dC}{dq_1} = \frac{d}{dq_1}[Qp_0 + Q(2\alpha + Q\lambda) - (\alpha + Q\lambda)q_1 + \lambda q_1^2] = -(\alpha + Q\lambda) + 2\lambda q_1 = 0$$

$$\Rightarrow q_1^* = \frac{Q}{2} + \frac{\alpha}{2\lambda}$$

At the optimum, the execution cost is

$$C^* = Qp_0 + \frac{1}{4}\left(6Q\alpha - \frac{\alpha^2}{\lambda} + 3Q^2\lambda\right)$$

Risk

In the basic model, prices are driven solely by our trades. This simplifies the analysis, but at the cost of ignoring important effects that are outside of our control. We can improve things by adding disturbances to our price equations. The basic model is modified as:

$$\begin{aligned} p_1 &= p_0 + \lambda\, q_1 + u_1 \\ p_2 &= p_1 + \lambda\, q_2 + u_2 \\ &= p_0 + \lambda\, q_1 + u_1 + \lambda\, q_2 + u_2 \end{aligned}$$

The *u*s are random terms that reflect a diverse set of unpredictable effects, such as new public information and the price impacts of other traders' orders. Like the price impact effects, they are cumulative and permanent. They are unpredictable, which suggests that they are (on average) zero. Formally, $Eu_t = 0$ for $t = 1{,}2$. They have constant variance, $Var(u_t) = \sigma_u^2$. They are

mutually independent, that is, u_1 doesn't depend on u_2 or vice versa. (Instead of independence, its often sufficient that they are uncorrelated.)

This treatment of risk leaves much of the original analysis intact. The basic cost function is now written as:

$$C = p_1 q_1 + p_2 q_2 = (p_0 + \lambda\, q_1 + u_1) q_1 + (p_0 + \lambda(q_1 + q_2) + u_1 + u_2) q_2 \qquad (17.6)$$

We can't directly minimize this because the us are unknown. Instead, we can take the expectation:

$$\begin{aligned} EC &= (p_0 + \lambda\, q_1 + Eu_1) q_1 + (p_0 + \lambda(q_1 + q_2) + Eu_1 + Eu_2) q_2 \\ &= (p_0 + \lambda\, q_1) q_1 + \big(p_0 + \lambda(q_1 + q_2)\big) q_2 \end{aligned}$$

In expectation the us are zero. The *expected* cost EC has the same form as the original cost function, so the analysis and optimization are identical. That is, if our objective is to minimize the expected cost, the optimal strategy sets $q_1^* = Q/2$.

There remains, of course, uncertainty about the actual cost. We can measure this by computing $Var(C)$. Recall that if x is a random variable, then a linear function of x, say $y = ax + b$ (for fixed a and b) has $Var(y) = a^2 Var(x)$. Using $q_2 = Q - q_1$, and rearranging (17.6) to isolate the terms involving u_1 and u_2,

$$\begin{aligned} C &= \cdots + (q_1 + q_2) u_1 + q_2 u_2 + \cdots \\ &= \cdots + Q u_1 + (Q - q_1) u_2 + \cdots \end{aligned}$$

$$\Rightarrow Var(C) = [Q^2 + (Q - q_1)^2] \sigma_u^2$$

In the above, the ellipses stand terms in C that are non-random (and therefore won't contribute to the variance). The second equality uses the relation $Q = q_1 + q_2$.

Figure 17-3 presents several perspectives. The expected cost in panel (a), as a function of the first-period trade is identical to the actual cost in the base case (without risk), minimized at $q_1 = 5{,}000$ shares. Panel (b) gives the standard deviation of the cost, $SD(C)$, a measure of risk. Note that we achieve the lowest risk by trading everything immediately, setting $q_1 = 10{,}000$.

This result sets up some competing objectives. No strategy minimizes both risk and expected cost. There is a trade-off between the two. As in general portfolio theory, mathematical analysis can only take us so far. Ultimately, the trader must be guided by their own risk-return preferences.

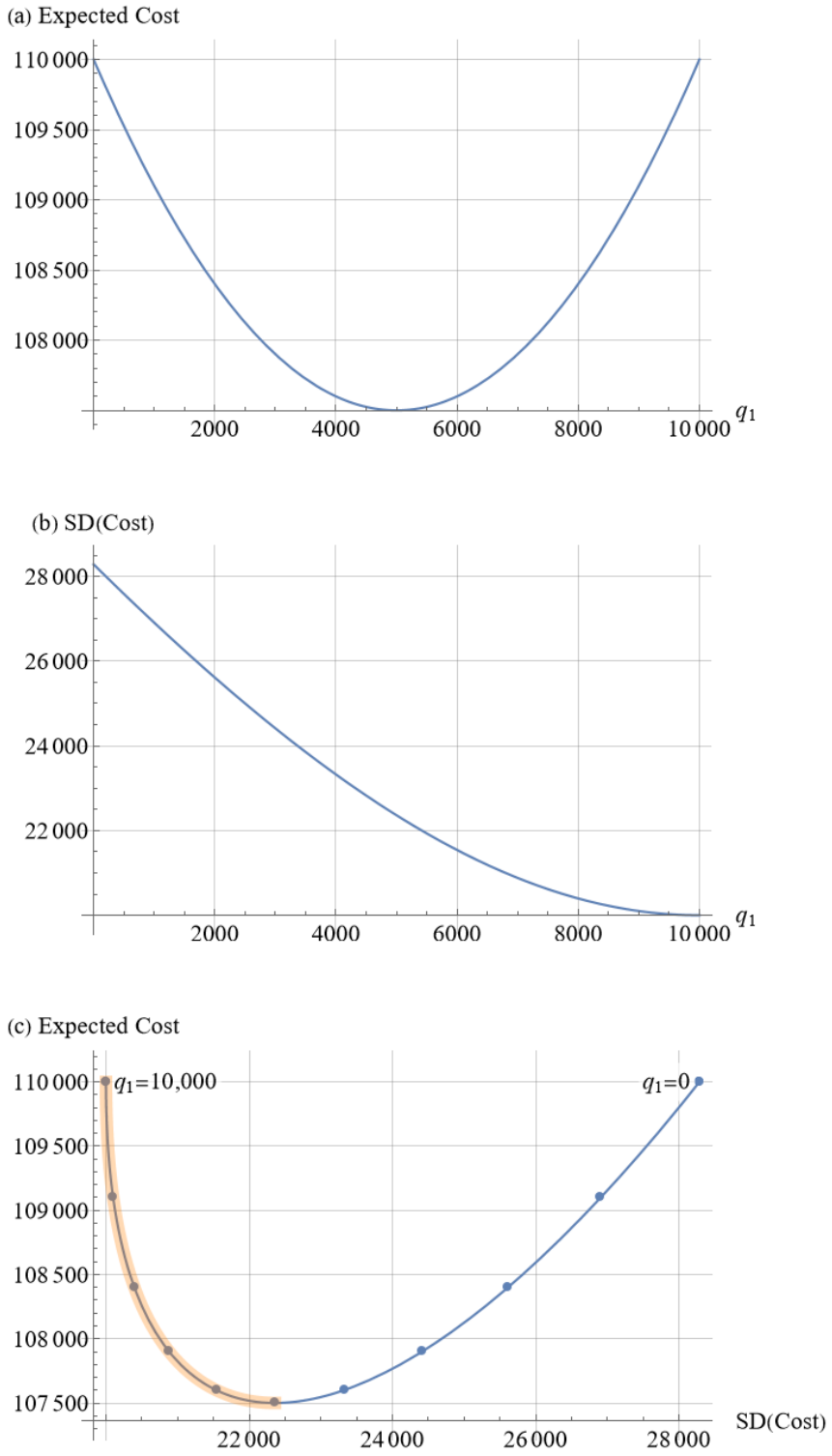
The connection to portfolio theory can be developed further. Classic investment portfolio situations are often depicted in graphs with expected return on the vertical and risk on the horizontal axes. Panel (c) depicts a similar curve for the two-period trading case. SD(C) is on the horizontal and EC is on the vertical axis. The curve is constructed as a parametric plot. That is, we vary q_1 between 0 and 10,000 shares. (The dots in the figure correspond to values of $q_1 = 0, 1000, 2000, \ldots, 10000.$) For each choice we compute the expected cost EC and $SD(C)$, and plot these values. As q_1 changes, we trace out a curve.[1]

[1] The comparison with portfolio analysis involves one key difference. In portfolio theory graphs (such as the capital market line or the security market line), the investor prefers to move in a "northwest" direction (higher expected return, lower risk). In the current situation, the trader usually prefers to move in a "southwest" direction (lower expected cost, lower risk).

All points on this curve are feasible, but if traders are risk-averse, some points are clearly better than others. The two choices, $q_1 = 0$ and $q_1 = 10{,}000$ both have an expected cost of \$110,000, but $q_1 = 10{,}000$ has the lower risk. For all levels of expected cost, the left-hand portion of the curve, indicated by shading, has less risk. In portfolio theory we denote a portion of the risk-return curve as the efficient portfolio frontier. In trading situations, the shaded portion of the curve in panel (c) is sometimes called the *efficient trading frontier* (Almgren and Chriss, 1997; Kissell and Glantz, 2003).

As a final caution, our choices of trading strategies help us balance risk and return in large samples. They don't guarantee the best possible outcome in any one situation. Buying all 10,000 shares immediately is not generally a good idea due to high impact costs. But if a large positive price shock in the second period drives the price higher ($u_2 \gg 0$), an immediate purchase will look (in retrospect) like sheer brilliance.

Figure 17-3 Expected Cost and Risk



The bid-ask spread

In the basic model, $p_1 = p_0 + \lambda q_1$, a purchase of shares corresponds to $q_1 > 0$. The purchase price is higher by λq_1. This increase continues into the second trade through the p_1 term on the right-hand side of $p_2 = p_1 + \lambda q_2$. That is, the λ impact terms arise from new information (the private information revealed by the trade. These informational effects are persistent and permanent. Some price impact effects, however, are temporary. Even in a market with no private/asymmetric information, the ask lies above the bid. From an economic viewpoint this spread allows the market maker (or limit order trader) to cover the non-informational costs of providing liquidity.

For simplicity, the bid-ask spread is a constant *s*. We now view the price as consisting of two components: a bid-ask quote midpoint, denoted m_t, and (as before) the actual transaction price, p_t. Starting from an initial midpoint m_0, the first period values are:

$$\begin{aligned} p_1 &= m_0 + \lambda q_1 + Sign(q_1)\frac{s}{2} \\ m_1 &= m_0 + \lambda q_1 \end{aligned}$$

As in the basic case, $p_1$contains a price impact term λq_1. The spread appears in the last term. It depends on the direction of the incoming order. The $Sign$ function (sometimes called the signum function) is defined for a general variable x as:

$$Sign(x) = \begin{cases} -1, \text{if } x > 0 \\ 0, \text{if } x = 0 \\ +1, \text{if } x > 0 \end{cases}$$

For a purchase, $q_1 > 0$, so $Sign(q_1) = +1$: we pay the half-spread. (If we were selling, then q_1 would be negative, $Sign(q_1) = -1$, and we be receiving a lower price.) The reason for modeling m_1 and p_1 separately becomes clearer when we move to the second period:

$$\begin{aligned} p_2 &= m_1 + \lambda q_2 + Sign(q_2)\frac{s}{2} \\ m_2 &= m_1 + \lambda q_2 \\ &= m_0 + \lambda q_1 + \lambda q_2 \end{aligned}$$

p_1 and p_2 both contain λ terms that are driven by cumulative trades: p_1 depends on λq_1; p_2 depends on λq_1 *and* (via the m_1 term) λq_2. These λ terms are also present in m_1 and m_2. The bid-ask term in p_2, however, depends only on $Sign(q_2)$, and not on $Sign(q_1)$. Bid-ask terms in the price are temporary. They affect only the current price, and do not carry over into future valuations.

The $Sign$ function is inconvenient for formal optimization because it is not differentiable everywhere. As long as our trades are in the same direction, this is not a concern. If we know that $q_1, q_2 > 0$, we can replace $Sign(q_t)$ by +1 throughout. (If we're always selling, we set $Sign(q_t)$ to -1.) If our strategy admits a mix of buying and selling, though, it is likely that we'll have to investigate all possible configurations of trade direction.

The distinction between permanent and temporary/transient effects arises frequently in economic studies at all horizons. (Do business cycles have effects on economic growth that last, say, for a hundred years? Are we still affected by the 2020 COVID shutdowns? By the 2007-

2008 Global Financial Crisis? By the US Panic of 1893?) Permanent and transitory effects might be easy to model conceptually, but the distinctions are difficult to draw in retrospect (and nearly impossible to discern in real time).

17.3. Extending the trading horizon

"I'd like to sell 20,000 shares," says the portfolio manager. "How quickly?" asks the trading desk. That is, are we dealing with $T = 2$ five-minute periods (ten minutes), $T = 78$ (one trading session), or $T = 780$ (ten trading sessions)? Two-period problems can be simplified because they can be reduced to one decision variable. Multiperiod problems are slightly more involved.

We are still, nevertheless, working in a manageable situation. Our price models are recursive. It's always the same model, applied and reapplied. Taking the basic model with risk as our point of departure, the price equation

$$p_t = p_{t-1} + \lambda\, q_t + u_t$$

is applied at all times $t = 1, \dots, T$. It is true that the recursions and derivations might involve, by the time we've made all the substitutions, many distinct terms, but these can often be reduced symbolic summations.

The cost to acquire Q shares over trading horizon T is

$$C = \sum_{t=1}^{T} p_t q_t \text{ where } Q = \sum_{t=1}^{T} q_t$$

The solution to the minimization of the expected cost in the two-period case is $q_1^* = q_2^* = Q/2$. This generalizes to the T-period case: $q_1^* = q_2^* = \cdots = q_T^* = Q/T$. The cost when following this rule is therefore:

$$C^* = \frac{Q}{T} \sum_{t=1}^{T} p_t$$

By recursive substitution,

$$\begin{aligned}
p_1 &= p_0 + \frac{\lambda Q}{T} + u_1 \\
p_2 &= p_0 + \frac{2\lambda Q}{T} + u_1 + u_2 \\
&\vdots \\
p_t &= p_0 + \frac{t\lambda Q}{T} + u_1 + \cdots + u_t \\
&\vdots \\
p_T &= p_0 + Q\lambda + u_1 + \cdots + u_T
\end{aligned}$$

Since the us have zero expectation, the expected price $Ep_t = p_0 + t\lambda Q/T$. The expectation of the sum is

$$\sum_{t=1}^{T} Ep_t = Tp_0 + \frac{\lambda Q}{T} \times \frac{T(T+1)}{2} = Tp_0 + \frac{\lambda Q(T+1)}{2}$$

where we have used the summation formula $1 + 2 + \cdots + n = n(n+1)/2$. The expected cost is

$$EC^* = \frac{Q}{T}\sum_{t=1}^{T} Ep_t = \frac{Q}{T}\left(Tp_0 + \frac{\lambda Q(T+1)}{2}\right) = Qp_0 + \frac{\lambda Q^2(T+1)}{2T}$$

The first term, Qp_0, is the hypothetical cost of buying the Q shares at the pre-trade price. The second term is the cost of slippage. It increases as the square of Q, but decreases in T. (In the limit, as $T \to \infty$, the second term approaches $\lambda Q^2/2$.

Turning now to the risk, the only terms in $\sum_t p_t$ that will contribute to the variance of the sum are the us. So

$$\begin{aligned} Var\left(\sum_t p_t\right) &= Var(Tu_1 + (T-1)u_2 + \cdots + u_T) \\ &= \sigma_u^2(T^2 + (T-1)^2 + \cdots + 1) \\ &= \sigma_u^2 \frac{1}{6} T(T+1)(2T+1) \end{aligned}$$

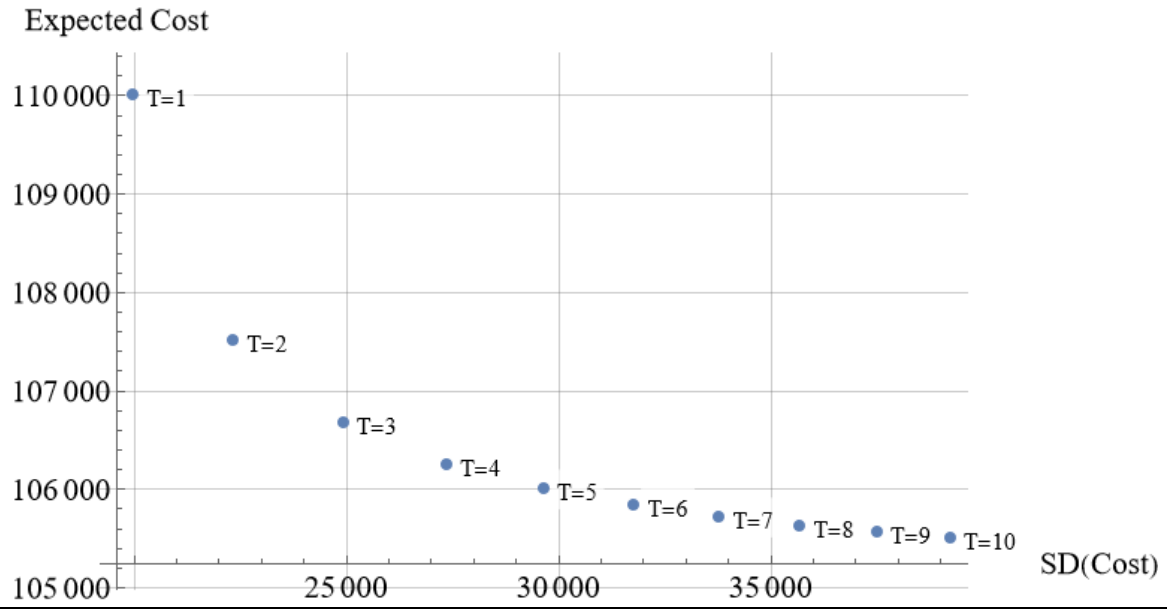
where we have used the summation rule $1 + 4 + 3 + \cdots + n^2 = (1/6)n(n+1)(2n+1)$. Then

$$Var(C^*) = \left(\frac{Q}{T}\right)^2 Var\left(\sum_t p_t\right) = \sigma_u^2 \frac{Q^2}{T} \times \frac{1}{6}(T+1)(2T+1)$$

This increases in the square of Q, but also in the square of T.

Figure 17-4 depicts the trade-off. Initially, there are large gains from increasing the trading horizon. In moving from $T = 1$ to $T = 2, 3,$ or 4, as risk increases, the expected cost falls sharply. At longer horizons, though, the decreases in expected cost are small. Like the efficient trading frontier shown in *Figure 17-3*, this curve reflects a risk-return tradeoff, but the two differ in the driving decision parameter: q_1 in *Figure 17-3*, T in *Figure 17-4*.

Figure 17-4 Risk, cost, and the trading horizon



Summary of terms and concepts

Parent and child orders; VWAP; alpha; order impact; efficient trading frontier.

References

Almgren, Robert, and Neil A Chriss, 1997, Optimal liquidation, Available at.

Bertsimas, Dimitris, and Andrew W. Lo, 1998, Optimal control of execution costs, *Journal of Financial Markets* 1, 1-50.

Engle, Robert F., Robert Ferstenberg, and Jeffrey Russell, 2012, Measuring and modeling execution cost and risk, *Journal of Portfolio Management* 38, 14-28.

Kissell, Robert, and Morton Glantz, 2003. *Optimal Trading Strategies* (American Management Association, New York).

Madhavan, Ananth, 2002, VWAP Strategies, *Institutional Investor Guides: Trading* 2002, 32-39.

U.S. Commodity Futures Trading Commission, and U.S. Securities and Exchange Commission, 2010, Findings Regarding the Market Events of May 6, 2010, (Washington D.C.).

Part V. Special Topics

Chapter 18. Market Infrastructure: custody, clearing, and settlement [Incomplete]

To this point in these notes a trade has been viewed as a quickly arranged event transferring a security (and usually, in the other direction, a payment) at a well-defined point in time. In fact, whether arranged on an exchange floor or a modern limit order book, the thing that we've been calling a "trade" is usually only a preliminary agreement. The actual transfers often occur substantially later (maybe days) after the original "trade." That is, it takes time for ownership and payments to catch up with the trading process.

Any failure along the way can be extremely disruptive. If the stock shares have been bought and sold many times, a failure at the beginning can invalidate presumptions about ownership and bank balances throughout the chain, leading to systemic financial problems. The systems we build to avoid these failures play an important role in maintaining widespread confidence in the financial system.

Our overview here begins with custody. Before even considering the possibility of trade we need to verify that a potential buyer has sufficient cash and that a potential seller in fact owns the shares. A retail customer will have cash on deposit with a broker. When that customer submits a buy order the broker can verify that the account balance is sufficient to fund the purchase. Also, in a retail account, any shares owned by the customer are in fact held (on the customer's behalf) by the broker. Before processing a sell order, the broker would make sure that the offered shares were in fact owned. In this role, the broker is acting as a custodian of the buyer's funds and the seller's shares. In managed investments (such as pension or mutual funds) custody often falls to a third-party custody bank.

If the first step of the start-to-finish trading process is custody, the last step is settlement. At settlement, transfer of ownership and payment occur and become legally irrevocable ("no backsies," in playground-speak). In a retail account settlement, the customer's broker is the destination for the transferred ownership and the source of the payment. In an institutional trade, settlement would occur between the buyer's and seller's custody banks.

Clearing broadly occupies the space between custody and settlement. It comprises the verification and confirmation of the terms of trade (buyer and seller identities, quantity, and price),

and also the settlement plan (identities of the custodians, timing of the settlement, and so forth).

In filling in the details of custody, clearing, and settlement, we will encounter many players that assist and facilitate these activities (clearing houses, sub-custodians, and so forth). Each may enter the transaction at an appropriate time, perform a service, and (of course) collect a fee. Isn't this inefficient? In the information age, couldn't we simply bundle everything in one silo, eliminate the redundant checks, the duplicate verifications, and the overlapping regulators? Possibly. Combining functions, though, can cover up performance failures until they become catastrophic events. Separating activities, isolating them behind their own firewalls, and having the entities report to different regulators can help contain small failures and keep them from growing into larger problems.

18.1. Ownership and custody

18.2. Settlement

18.3. Clearing

18.4. Crises

Custody, clearing, and settlement are sometimes considered to be "boring". When all the wheels of trade are turning smoothly and silently, perhaps they are. But it bears emphasis that most of today's custody, clearing and settlement arrangements have arisen and evolved in response to disasters that have left the affected markets in pieces. Examples abound.

The U.S. back-office crisis

Historically, stock shares were paper certificates bearing the name of the owner. Settlement would involve physically moving the certificates (from the seller's broker to the buyer's broker or custodian) followed by registration of the new ownership (by the corporation that issued the shares).

For simplicity and consistency markets generally standardize their settlement conventions. In the early 1940s the US equity market followed "T+2" settlement: a trade occurring at any time on day T would be settled as of the close of business on day T+2. In 1946 this changed to T+4 and, in 1968 to T+5.

As trading volume grew, the settlement system could not keep pace. Failures to deliver purchased shares (called simply "fails") were common. The SEC reported that "One out of every 8.4 transactions was a fail," (Seligman, 1995). Sometimes this occurred by inattention or error. (Brooks, 1973) reports "… stock certificates were turning up 'stuffed behind pipes in ladies' rooms, at the bottom of trash baskets, in the backs of filing cabinets with old letters.'" In other cases, the certificates had been stolen.

In response, the NYSE began closing early on most days, and closing entirely on Wednesdays. Attempts were made to computerize and automate the processes. The paperwork crisis did not really abate until trading volume declined for unrelated reasons. But even though there was no quick fix, the groundwork was laid for improved systems that we'll examine in the next section.

The Bombay stock exchange

Prior to 1992, the Bombay stock exchange was India's dominant equity market. But in 1992 the market was hit by instances where forged or nonexistent shares were posted as collateral for

loans. A new stock exchange was chartered (the National Stock Exchange) which supplanted the BSE. In 1996, the exchange started a new clearing system.

Credit Herstatt

Currency trading (foreign exchange) may involve transfers of currencies where one side is in a different time zone and different regulatory regime that the other side. On June 26, 1974, the Herstatt Bank (a German institution) was trading deutschemarks (DM) against the US dollar. On balance, Herstatt was receiving DM (in Frankfurt) and paying dollars in New York. Most of Herstatt's counterparties had released the DM and expected to subsequently receive dollars. Before the dollar transfers occurred, however, German regulators closed the bank, leaving the US counterparties unpaid.
In response, banks and regulators began work on the CLS Bank.

Cryptocurrencies

2022 saw a wave of institutional failures that saddled many traders with large losses and left many unable to establish ownership.

18.5. Custody

Brooks, John, 1973. *The Go-Go Years* (Ballantine Books, New York).
Seligman, Joel, 1995. *The transformation of Wall Street (revised)* (Northeastern University Press, Boston).

Chapter 19. Pricing, Fees, and Rebates

Our markets and other institutions of trade are not funded by kind donations. Exchanges and brokers incur costs in providing their services, and to survive they must recover these costs through customer fees. In this, of course, they are no different from most other firms in the economy.

The structure of these fees (and rebates), however, is complex. Some are direct; some are indirect. Sometimes they are levied on customers in a fairly transparent fashion; sometimes they are buried in other charges. A customer cannot simply focus on the "bottom line" commission because as her order passes from broker to market center, fees and rebates change hands that affect how her order is handled. Many industries, of course, have complicated pricing arrangements. In this arena, though, the opaqueness of the pricing is especially remarkable as it stands in sharp contrast with the transparency, availability and uniformity of the *security's* bids, offers, and reported trade prices.

This chapter focuses on two arrangements: exchange maker/taker pricing and payment for retail order flow. Both are topics of ongoing regulatory interest.

19.1. Exchange pricing

It is not surprising that exchanges charge for executions. What might seem perplexing, though, is that this cost is coupled with a subsidy.

Maker/taker pricing

Under the maker/taker model, market and marketable limit orders pay a small "taker" fee, and limit orders that are added to the book and subsequently executed receive a small consideration called a "maker" payment or "liquidity rebate". The taker fee is usually larger than the maker rebate, with the market center capturing the difference.

The terms on NASDAQ's limit order system were recently as follows:

- An executed limit order ("maker") generally received \$0.0029 per share (that is, \$0.29 per one-hundred share "standard" round lot) if the limit order was visible, and \$0.0015 per share (\$0.15 per hundred shares) if the limit order was hidden.
- A taker pays \$0.0030 per share "liquidity removal fee".

The practice of providing a rebate for supplying liquidity and charging a bit extra for removing it is consistent with a view that liquidity is worth rewarding. NASDAQ keeps \$0.0001 per share (\$0.01 on a hundred shares).

Maker/taker fees can distort posted prices. Suppose that trader A is offering \$10.00 per share (visible), and trader B lifts that offer. Net of fees and rebates, A receives \$10.0029, and B pays \$10.0030. Effectively, except for the \$0.0001 that goes to NASDAQ, it's as if A had priced his/her offer at \$10.0030.

Different market centers have different price schedules. We may be looking at posted offers from multiple exchanges that appear to be the same. But without consulting the fee/rebate schedule, we don't know which offer is really the best.

To reward the more consistent liquidity suppliers ("de facto market makers"), maker fee schedules can have quantity premia. The BATS-X liquidity rebate starts at \$0.0025 per share but rises to \$0.0029 for a member whose average daily volume is at least 1% of the consolidated average daily volume (June 1, 2012 fee schedule). Maker fees for executions against non-displayed orders are lower (\$0.0017): displayed size is a better advertisement for the exchange. The taker fee is a uniform \$0.0029.

Inverted pricing (taker/maker pricing)

Some exchanges offer inverted maker/taker fees. The BATS-Y exchange *charges* \$0.0030 for providing liquidity and *pays* \$0.0020 for removing it. At first glance this defies competitive logic. But suppose that BATS-X and BATS-Y are both offering at \$10. If a trader intends to lift the offer, she will certainly send her order, or at least the first part of it, to BATS-Y. (It's better to receive \$0.0020 then pay \$0.029.)

Now consider another situation. BATS-X has 10,000 shares offered at \$10, and the BATS-Y book is empty. We're considering where to send a \$10 limit sell order. If we send it to BATS-X, we'll receive \$0.0025. But we'll only get this if our order is executed, and there will be 10,000 shares ahead of us. If we send our order to BATS-Y, we'll pay \$0.0030, but our order will be at the front of the book. We also know that BATS-Y is a more attractive place to send a market order.

It is not a feasible alternative to send to BATS-X a sell order priced at $\$10 - \$0.0030 = \$9.9997$ because quotes on an increment finer that \$0.01 aren't permitted by the sub-penny part of Reg NMS (discussed in Section 20.4).

Routing charges

The SEC has not directed industry to set up a single consolidated access system. When US equity markets began to go electronic, most of the newer markets (like Inet) built systems so their subscribers could send in orders directly. The SEC envisioned a network of point-to-point connections, rather than a centralized hub-and-spoke system: "private linkages approach". This is in fact largely what has happened.

Not everyone has direct connections. If there are n nodes a point-to-point network needs $n(n-1)/2$ connections. There are over 200 market centers, so there would need to be about 20,000 links. A smaller number of market centers and brokers have developed their routing capabilities (speed, intelligence, number of connections) as a means of differentiation. Rather than access all market centers directly, a trader might set up direct links to a few centers but go

through a broker's routing system (or NASDAQ's) to access the others. Market centers charge for routing orders out to other centers (typically around $0.0030).

Current events

Spatt (2020) discusses the current exchange fee structure. He notes that for the operators of multiple exchanges, the pricing rules vary in a way that tends to segment the market, dividing traders (the exchanges' customers) into clienteles. The pricing rules are complex, and the breakpoints in the rules might be tailored to particular dealers. The US SEC proposed a pilot experiment (the Access Fee Pilot) in which stocks would be randomly assigned to different regulatory regimes (including various levels of fee caps and prohibitions on rebates). A group of exchanges brought a lawsuit against the SEC. The court sided with the exchanges. As of August 2021, the study is on hold.

19.2. Payment for (retail) order flow

Note: the handling of retail orders is the focus of current SEC regulatory interest. The Commission has proposed a rule that would require dealers to expose retail orders to auctions before internalizing them. See https://www.sec.gov/rules/2022/12/order-competition-rule.

Some retail investors believe that the brokers send their orders to exchanges where they interact in some central fashion with all other buyers and sellers, large and small. This may have once been true, but current practice is very different. An SEC Concept Release states, "A review of the order routing disclosures required by Rule 606 of Regulation NMS of eight broker-dealers with significant retail customer accounts reveals that nearly 100% of their customer market orders are routed to OTC market makers," (U.S. Securities and Exchange Commission, 2010).

This arrangement is described in Section 8.3 in connection with dark trading. Typically, the broker-dealer receiving a marketable customer buy order will sell directly to the customer at the NBO; a marketable sell order will receive the NBB.

This arrangement is driven by considerations of private information. As described in Section 13.2, a dealer loses to incoming informed traders, but profits from incoming uninformed traders. Retail traders are, as a group, less informed, and are therefore more desirable counterparties. Why, though, should a broker send a retail order to a dealer? Or, if the order will be sent, which dealer should receive it?

One factor bearing on these decisions is a payment from the dealer to the broker in exchange for the order. For various reasons discussed below, this compensation, akin to a referral fee, is controversial. The SEC permits the practice, but it requires disclosure.

Whereas Rule 605 is aimed at accountability by market centers (see Section 16.7), Rule 606 applies to brokers. They must report what orders they received, what they did with those orders, and other aspects of their relationships with market centers.

Charles Schwab is a full-service broker with large retail customer base. The 606 report for 2009 Q 3 documents these relationships (From Schwab website, January, 2010). Table 19.1 summarizes the Charles Schwab's customer orders ("Securities Listed on the NYSE/Network A Eligible Securities"). Table 19.2 describes where the orders were sent. The 606 reports also disclose the broker's arrangements with market venues:

> UBS is a market maker in certain NASDAQ, OTC and listed equity securities. Part of the consideration Schwab received for the sale of its capital markets business to UBS in 2004 related to execution services agreements with UBS and Schwab's commitments to route most types of equity and listed options orders through UBS for eight years.

> However, Schwab does not earn rebates or other consideration from UBS or other firms or exchanges for equity and options orders routed through UBS or routed by Schwab directly.

Table 19.1 Charles Schwab's Rule 606 report (order counts)

Non-directed orders as percentage (%) of total customer orders:	95.9%
1. Market Orders as % of total non-directed orders	35.2%
2. Limit Orders as % of total non-directed orders	57.3%
3. Other Orders as % of total non-directed orders	7.5%

Table 19.2 Charles Schwab's Rule 606 report (order routing destinations)

Venue	% of Non-Directed Order Flow Rec'vd
UBS Securities LLC	95.1%
INET/NASDAQ	3.9%
Citadel Derivatives Group LLC	0.7%

Other OTC Market Makers *do* receive payment. From E*TRADE's 2009 Q3 606 report:

> E*TRADE receives payment from its affiliate, E*TRADE Capital Markets, LLC ("ETCM"), a wholly owned subsidiary of E*TRADE Financial Corp. ("ETFC"), for directing listed equity order flow.
>
> Payments received from ETCM averaged approximately $.0006 per share [six cents on a 100-share order].
>
> ETCM executes on a principal basis and may have profited or lost in connection with such transactions.

Brokers are supposed to act in their clients' interests. Do side payments provide incentives to disregard those interests? From a comment letter on an earlier concept release (AGS Specialist Partners):

> When order flow is earned through better execution, customers are rewarded with tighter spreads and the efficiency of the marketplace is improved. When order flow is guaranteed through cash payments or through other means of bribery, customers suffer as spreads widen and the marketplace becomes less efficient. Prepaid order flow, if anything, gives MMs an incentive to widen their quotes as order flow is guaranteed regardless of performance.

From a Reg NMS comment letter (George A. Carroon):

> As a Member of the New York Stock Exchange, and a Specialist for more than forty years, … I also encourage the Commission to give careful consideration to the issue of payment for order flow, which, in the opinion of many, can only be considered commercial bribery.

The dispute over payment for order flow spilled into public view during Senate hearings in 2014 (Alden, 2014):

> TD Ameritrade, a brokerage firm that handles vast numbers of stock trades for average investors, promises to execute those orders on the best possible terms. But in practice, TD Ameritrade routes a large number of the customer orders to the exchanges that pay it the most, Steven Quirk, an executive at the firm, said at a Senate hearing on Tuesday.
>
> …
>
> The issue has divided the stock exchange sector. The president of the New York Stock Exchange, Thomas W. Farley, asserted on Tuesday that a system called maker-taker payments, in which exchanges pay rebates to brokerage firms for orders, created "inherent" conflicts. But the chief executive of the BATS Global Markets exchange company, Joseph P. Ratterman, played down that concern, saying the conflicts could be managed.
>
> But when Mr. Quirk was questioned, the talk was less hypothetical. Senator Carl Levin, Democrat of Michigan, who leads the panel, pointed to data from the fourth quarter of 2012 that showed that TD Ameritrade directed all nonmarketable customer orders -- meaning, orders that could not immediately be consummated based on the market price -- to one trading venue, Direct Edge. It so happened that Direct Edge paid the highest rebate.
>
> "Your subjective judgment as to which market provided best execution for tens of millions of customer orders a year allowed you to route all of the orders to the market that paid you the most," Mr. Levin said. "I find that to be a frankly pretty incredible coincidence."

Current events

The 2014 hearings did not reflect well on the brokerage industry, but payment for order flow was not then subjected to further investigation or regulation.

Since then, zero-commission trading has become the industry norm, at least for retail customers. With no commission revenue, brokers are even more dependent on payment for order flow. This stress has refocused regulatory attention. In December 2020, the SEC charged Robinhood (the original zero-commission broker) with inadequate disclosure of its payment for order flow practices and delivering sub-par executions for its clients (U.S. Securities and Exchange Commission, 2020). In testimony before the House Financial Services Committee, SEC Chair Gary Gensler indicated that the SEC's concerns on these matters extended more broadly than one firm (Gensler, 2021).

19.3. Further reading

For maker-taker pricing: (Foucault, Kadan and Kandel, 2013; Malinova and Park, 2015; Yao and Ye, 2014). For payment for order flow: (Battalio, 2003; Battalio and Holden, 2001; Chordia and Subrahmanyam, 1995; Parlour and Rajan, 2003).

Summary of terms and concepts

Taker fees (also called liquidity removal, liquidity access); rebates for adding liquidity (also called maker rebates); inverted ("taker/maker") pricing and the logic behind it; payment for order flow; [SEC] Rule 606 information.

References

Alden, William, 2014, At Hearing, Brokerage Firms Are Called Out for Conflicts: [Business/Financial Desk], June 18, 2014, New York Times.

Battalio, Robert H., 2003, All else equal?: A multidimensional analysis of retail, market order execution quality, *Journal of Financial Markets* 6, 143-162.

Battalio, Robert, and Craig W Holden, 2001, A simple model of payment for order flow, internalization, and total trading cost, *Journal of Financial Markets* 4, 33-71.

Chordia, Tarun, and Avanidhar Subrahmanyam, 1995, Market making, the tick size, and payment-for-order flow: theory and evidence, *Journal of Business* 543-575.

Foucault, Thierry, Ohad Kadan, and Eugene Kandel, 2013, Liquidity cycles and make/take fees in electronic markets, *Journal of Finance* 68, 299-341.

Gensler, Gary, 2021, Testimony Before the House Committee on Financial Services, May 6, 2021.

Malinova, Katya, and Andreas Park, 2015, Subsidizing liquidity: The impact of make/take fees on market quality, *Journal of Finance* 70, 509-536.

Parlour, Christine A, and Uday Rajan, 2003, Payment for order flow, *Journal of Financial Economics* 68, 379-411.

Spatt, Chester S., 2020, Is equity market exchange structure anti-competitive? , Available at.

U.S. Securities and Exchange Commission, 2010, Concept release on equity market structure.

U.S. Securities and Exchange Commission, 2020, Press Release: SEC Charges Robinhood Financial With Misleading Customers About Revenue Sources and Failing to Satisfy Duty of Best Execution, December 17, 2020.

Yao, Chen, and Mao Ye, 2014, Tick Size Constraints, Market Structure, and Liquidity, University of Illinois at Urbana, Available at.

Chapter 20. Reg NMS

The SEC's Regulation NMS heavily influences the form and operations of the US equity markets. The rules were initially proposed in 2004, adopted (after comment and revision) in 2005, and phased in over 2006. The process provided an opportunity for public debate and speculation on how the markets would and should evolve.

"NMS" stands for "National Market System". In 1975, Congress passed the Securities Acts Amendments (of 1975), which directed the SEC to facilitate development of a "national market system". The amendments did not define the system precisely, but they were widely interpreted as calling for one comprehensive market for US stocks. At a time when trading was dominated by floor-based exchanges, the 1975 act held out a vision of multiple trading centers that would be electronically linked, with orders flowing to the trading venues with the best prices. Three decades later, Reg NMS was positioned and promoted as a culmination and fulfillment of the 1975 Act.

The final version of Reg NMS (U.S. Securities and Exchange Commission, 2005), the online comment letters, and the dissenting Commissioner opinions offer a detailed discussion of the logic behind the rule.

20.1. Background

Market and order competition

In the 1980s and 1990s, the two principal US exchanges (NYSE and NASDAQ) found themselves beset by competition from new markets that took full advantage of modern technology. One of these, Instinet, had evolved into something close to a modern all-electronic limit order book, but participation was limited to large institutions (and NASDAQ dealers). Island, another limit order market, attracted more retail traders. Other systems included Optimark, Wunsch Auction Systems, and Delta Options. The SEC termed this process *market competition*, that is, competition among exchanges and similar venues. By the 1990s, it was clear that market competition was bringing many benefits to investors, in the form of lower trading costs and new ways of trading.

The pursuit of market competition, though, led to fragmentation. The new environment with many competitive markets was confusing, as investors who formerly faced simple choices in routing their orders (like, NYSE vs. NASDAQ) found themselves forced to consider a multitude of possible market destinations. The problem is that when orders are resting in many places, it's difficult to determine where the best prices are. The interplay of individual orders, described as *order competition*, works best when all orders go to a single destination. Enhancing order competition favors consolidation.

In designing Reg NMS, the SEC sought to balance market competition and order competition. The aim is sometimes (but not in the actual text of the rule) called *virtual consolidation*, that is, separate competitive markets that function as one cohesive market, and present to the user one unified view of the market.

Other trade-offs involved balancing the interests of the different kinds of traders. US equities markets, for example, have an unusually large number of retail traders. Should market regulation favor retail or institutional clienteles? Retail traders are more numerous, but aren't institutions like mutual funds and pension funds representing the interests of their beneficiaries, who tend to be individuals? This debate was affected by the SECs regulatory experience, in the 1990s, involving "two-tier" markets in which institutions received favorable bids and asks, while retail investors faced large spreads.

Yet another apparent trade-off involved balancing the interests of short-term traders (like dealers) vs. long-term investors (presumably the dealers' customers). The SEC's response in this case was more one-sided. The SEC views its mandate as acting on behalf of companies trying to raise capital for long-term projects and investors with long-term horizons. Market makers, intermediaries, and day traders are not considered to have a valid interest (for regulatory purposes) *except* insofar as they contribute to or facilitate the long-term investors' interests. From the final rule: "… it makes little sense to refer to someone as 'investing' in a company for a few seconds, minutes or hours."

Alternative trading systems (ATSs)

The new markets posed many regulatory challenges. To begin with, what should they be called? In the set of available regulatory categories, a "market" pretty much had to be an "exchange," that is, something that was owned and operated by its members. The new markets did not have members, at least in the traditional sense, and most could not withstand the expense of complying with all the legal and regulatory overhead that seemed to come with being labeled an exchange. The SEC initially granted them "no action" letters, provisional assurances that they could operate without fear of legal action. But as the numbers of these quasi-exchanges grew, the SEC needed a clearer and more consistent approach. In 1997, the Commission adopted Regulation ATS (Alternative Trading Systems).

Reg ATS provided several levels of regulatory requirements, depending on how much trading volume the system executed, as a proportion of the total volume across all market centers. Under the present (amended) rule, an ATS in the smallest size category (under 5% of the total volume): files an initial notice of operation and quarterly reports; keeps records (an audit trade) of its activities; and does not call itself an exchange. At or above 5%, the ATS must also link with a registered market (like an exchange), comply with the market's rules, and disseminate (that is, publish) its best bids and offers on the consolidated quote system. (Orders in a dark pool that aren't shown to anyone don't have to be published.) Also (at or above 5%), the

ATS must establish objective written standards, applied in a nondiscriminatory fashion, for admitting and denying access to its system ("fair access").[1, 2]

The SEC maintains a current list of ATS on its website. As of April 2015, there were slightly under one hundred registrants. Most dark pools and crossing networks are registered as ATSs, but a market does not need to be dark. Most markets trade US equities, but some trade bonds. MTS BondsPro, for example, is a lit market for US dollar-denominated bonds.

The structure of Reg NMS

Reg NMS has four main components:

- the order protection rule
- the access rule
- the subpenny pricing rule
- the market data rules

(The regulation also mandated a consolidation and renumbering of preexisting rules, but this was mainly a legal formality.) Each seeks to contribute to the balance of market and order competition.

20.2. The order protection rule

This is sometimes (but inaccurately) called the "no trade-through" rule. Recall that a trade-through is an execution outside of a market's best bid or offer. This part of Reg NMS has two main provisions:

- Market centers will put in place procedures to avoid trade-throughs.
- A bid/offer is not protected against trade-through unless it gives automatic execution.

In Chapter 5 we encountered the basics of this rule. Here we consider some of the finer points.

We noted earlier that although the order protection rule is sometimes characterized as prohibiting trade-throughs, this isn't quite correct. The rule simply says that markets need to avoid them. Nor does the rule require a broker to route a customer order to the exchange at the best visible quote (at the NBBO). It only says that an exchange shouldn't execute the order through the NBBO. Finally, it does not guarantee that the customer always gets the best price. No reasonable rule could accomplish this. The automatic execution condition for protections was far-reaching. It made a distinction between "fast" and "slow" markets. It forced floor markets (notably the NYSE) to quickly become electronic.

Intermarket sweep orders (ISOs)

Sometimes a trader wants to execute a large quantity, quickly, against all of the protected quotes in all of the market centers. If uncoordinated orders are simply routed to each destination, some markets might refuse to execute their components, believing that these executions will cause trade-throughs, even though the protected bids and offers have already been executed by the same trader. In this situation, Reg NMS permits the trader to use an intermarket sweep order (ISO). Functionally, this allows the trader to "sweep" all protected bids or offers,

[1] In the original rule, the fair access requirement did not apply until the ATS reached 20% of the total volume. The 20% threshold was dropped to 5% in 2005.

[2] The SEC has recently proposed amendments to require that the operator of the ATS provide more detail about its ownership and affiliates, and also that the ATS have written "safeguards and procedures to protect subscribers' confidential trading information," ((U.S. Securities and Exchange Commission, 2016)

notifying the individual markets of his intent, and relieving them of the responsibility to check for executions that would trade through protected bids or offers.

Formally, from the rule's text:

> An intermarket sweep order is … a limit order that meets the following requirements:
>
> (1) The limit order is identified as an intermarket sweep order when routed to a trading center; and
>
> (2) Simultaneously … one or more additional limit orders, as necessary, are routed to execute against the full displayed size of any protected bid, in the case of a limit order to sell, or the full displayed size of any protected offer, in the case of a limit order to buy, for the NMS stock with a price that is superior to the limit price of the limit order identified as an intermarket sweep order.

Here's how it works. The ISO provision specifies that the quantity to be routed to a market must accommodate all protected bids or offers. That is, the rule specifies a minimum. The eventual outcome may result in executions that are, in the usual sense, trade-throughs, as long as they aren't through protected quotes.

For example, consider a two-market case, with the bids given as in Table 20.1. It would satisfy the rule to submit:

- Sell 400 shares, limit $100, ISO to Exchange A, and
- Sell 100 shares, limit $100, ISO to Exchange B

The 100-share sell order sent to Exchange B suffices to execute B's protected bid at $103. On Exchange A, 100 shares would execute against A's protected bid of $102, and 300 shares would execute deeper in A's book, at $100. Now this last execution trades through orders in B's book that are priced at $102 and $101, but because these orders weren't at the top of B's book, they weren't protected.

Table 20.1 Two-market example. Bids on two exchanges.

Bid	Exchange *A*'s Book	Exchange *B*'s Book
$103		**100 sh**
$102	**100 sh**	100 sh
$101		300 sh
$100	300 sh	

The order protection rule was the most contested part of Reg NMS, and it received the strongest comment letters. In the final vote, two (of the five) SEC Commissioners voted against it. (Their dissent is available on the SEC website.) The objection was that the rule wasn't needed. It was asserted that brokers, in fulfilling their fiduciary duty of best execution, would ensure that their customers would always get the best price, and so trade-throughs wouldn't occur. If brokers didn't do this, the customer could always get a new broker. In any event, the problem should remain a matter between the broker and the customer, not something that needs to be addressed by federal statute. The counterargument was that investors needed protection. They lacked the means to monitor what their brokers were doing. Furthermore, the brokers themselves might not know at any given instant which particular market had the best quote.

20.3. Access rule

For a market to have its quotes protected under the order protection rule, anyone must be able to execute against those quotes quickly and *inexpensively*. The market must give everyone equal access. No discrimination in favor of subscribers. The rule also states that any "access fee" can't be higher than $0.0030 per share (see Section 19.1).

20.4. The sub-penny rule

Until the end of the 20th century, trading in US equity markets was conducted in eighths ($0.125). Congress's Common Cents Pricing Act of 1997 established the penny as the price increment in US stock market. As a transitional step, exchanges went first to 1/16ths, "steenths". (Quickly now: if you're a buyer, would you rather have five steenths or three eighths? If a seller, thirteen steenths or three quarters?) In 2001 the transition to penny pricing was completed.[3]

It is sometimes said that the tick size is the price of time priority. Suppose the bid is $10.00 for 20,000 shares. If I put in a bid at $10.00 for 100 shares, I'm at the end of a 20,000-share line (due to time priority). If I put in a bid at $10.01, I move to the front of the line. If the tick size were $0.000001, I could jump by the queue by paying an extra $0.000001 x 100 shares =$0.0001. The effectiveness of the sub-penny rule has been partially undone by sub-penny maker/taker fees (see Section 19.1).

The sub-penny rule applies to bid and ask quotes, not trades. If the NBBO is $25.00 bid, offered at $25.01, a dark pool trade that is priced at the midpoint will be executed at $25.005. This is permissible.

The US Congress usually delegates rulemaking to the SEC. The 1997 act was an exception, but it was not the last time Congress involved itself with tick-size. The Small Cap Liquidity Reform Act of 2014 directed the SEC to study the desirability of *raising* the pricing increment for some stocks to $0.05.

20.5. Market data rules

There are two main sources of exchange revenue: listing fees and data fees. Historically, they were about equal. Nowadays, data fees dominate.

The Consolidated Tape Association runs CTS and CQS and shares its revenues with the data providers (exchanges and non-exchange SIPs, securities information processors) under complex formulas. Prior to Reg NMS, exchanges were a fixed fee for reporting a trade (no matter how large). This led to "tape shredding" breaking up large trades into small 100-share trades to maximize the data revenue. Reg NMS removed this incentive. Also, prior to Reg NMS, an exchange was paid for posting a quote, no matter how aggressive it was. If the NBB was $10.00 for 10,000 shares, an exchange would get paid for posting a bid of $1 for 100 shares. Reg NMS revised the formula, so that revenue is based on size and time at NBBO.

What sort of data can the exchanges charge for? The basic principles were laid out in 1999 SEC report ("Seligman Commission"). The BBO and trade reports should be provided at low cost. All other data can be priced at what the market will bear, e.g., order book data, historical data, short-sale data, and so on.

When Reg NMS was being debated, the modern era of high-frequency trading was just beginning. It passed notice that an exchange might want to discriminate among its market data customers on the basis of speed. Is it presently permissible for an exchange to intentionally

[3] The title of the act is highly referential: Cents vs. "sense"; "common sense" used to mean practical knowledge; "Common Sense" was also a pamphlet published during the American Revolutionary War.

slow down market data transmission speed for one class of customers? Probably not. Can exchanges charge extra for high-speed transmission channels that outrun the consolidated trade and quote systems? This is certainly okay.

20.6. An appraisal

For the most part Reg NMS has functioned well. It has survived ten years with no substantial modifications. Market center competition has been robust. An innovative new exchange, IEX, was approved in 2016. Order competition is also vigorous. Trading volumes are large, and bid-ask spreads are low. (O'Hara and Ye, 2011) make case that market quality has not been impaired by fragmentation. Other commentators are more critical (see, for example, (Hatheway, Kwan and Zheng, 2013)). They claim that while the decline in bid-ask spreads has been good for small retail traders, institutional traders are finding large trades more costly. Some people believe that the market fragmentation that has occurred under the rule has encouraged high frequency trading (discussed in the next chapter). Access fees remain problematic and have proliferated in variety.

A fragmented market is a complicated place. The virtual consolidation vision of Reg NMS has certainly not resulted in virtual simplicity. For this reason, some commentators believe that while market competition may have been necessary during a period of technological innovation, that phase is now largely over, and it is time to encourage consolidation, perhaps even moving to one consolidated limit order market. That market, though, would be a monopoly, wielding enormous economic and political power. This might be counterbalanced by aggressive regulation, but experience is not encouraging. Throughout much of the last century, the SEC struggled to open up US exchanges to more competition (Seligman, 1985, 1995).

In providing a framework for market competition, Reg NMS has been influential. Shortly afterwards, the European Union adopted the Markets in Financial Instruments Directive ("MiFID"). As a result, the older exchanges (e.g., the Paris and Amsterdam Bourses) have been placed in competition with new entrants (e.g., BATS Europe, Chi-X, etc.).

Not all markets have followed the Reg NMS lead. US futures exchanges have remained consolidated. Their regulator (the CFTC) and key market constituents have not shown strong interest in reshaping the exchanges along Reg NMS lines. The differences in regulatory and market design philosophies that split US equity and futures markets are striking given the strong resemblances in what they trade. If we want to invest in the S&P 500, we can buy the SPY ETF on an equities exchange, or we can go long an S&P stock index contract on a futures exchange. The profit/loss outcomes will be very similar, but the trading environments are very different.

Summary of terms and concepts

Market competition; order competition; alternative trading system (ATS); electronic communications network (ECN); order handling rule; Reg NMS; order protection rule; access rule; subpenny pricing rule; protected quotes; intermarket sweep orders (ISOs); tick size; The Common Cents Pricing Act of 1997; Small Cap Liquidity Reform Act of 2014.

References

Hatheway, Frank M., Amy Kwan, and Hui Zheng, 2013, An empirical analysis of market segmentation on U.S. equities markets, SSRN, Available at: http://ssrn.com/abstract=2275101.

O'Hara, Maureen, and Mao Ye, 2011, Is market fragmentation harming market quality?, *Journal of Financial Economics* 100, 459-474.

Seligman, Joel, 1985. *The SEC and the Future of Finance* (Praeger Publishers, New York).

Seligman, Joel, 1995. *The transformation of Wall Street (revised)* (Northeastern University Press, Boston).

U.S. Securities and Exchange Commission, 2005, Regulation NMS (Final Rule Release No. 34-51808; June 9, 2005).

U.S. Securities and Exchange Commission, 2016, Regulation of NMS Stock Alterntive Trading Systems (Release No. 34-76474).

Chapter 21. High Frequency Trading (HFT)

The technology that exchanges presently use to manage continuous limit order books and run opening or closing auctions is far superior to what was available near the end of the last century. In fragmented markets, this is also true of the technology used to link market centers. Present day information, access and routing systems are more advanced, possessing superior speed, capacity, and reliability.

At the start of transition to electronic markets, it was believed that technology would weaken exchanges' floor constituencies, that electronic access would reduce the advantage of being physically present at the center of trading. Although we'll see that this prediction eventually proved false in some interesting respects, it was initially correct: the floor trading crowd was largely left behind, unable to keep up with automatic execution.

More broadly, though, electronic markets were initially viewed as fostering equality among participants. Morris Mendelson and Junius Peake, two early advocates, wrote that, ""Bids and offers [will] compete instantly, equally, and fairly regardless of their origin since, once entered into the system, they [will] be instantly reflected in the summary displays and capable of being executed. … All brokers, dealers and investors [will] have equal, simultaneous access to all data," (Mendelson and Peake, 1979). In this vision, everyone from retail customer to hedge fund trader sits at a computer terminal, equal but for their inborn quickness of mind.

As events unfolded, however, equality did not come to pass. The reach of technology did not stop at improvements within the market centers, it extended to the markets' users, that is, the brokers, investors, and traders. We might distinguish, therefore, between market technology (inside the market center) and trading technology (outside). Advances in trading technology altered the relative economic power and standing of key players. The changes have not been neutral. Technological shifts rarely are.

The effects of trading technology have been most profound with respect to speed. Simply put, in most of the trading mechanisms and strategies discussed throughout this book, speed confers an advantage. The first trader to hit a bid or lift an offer, the first to cancel a stale bid or offer, whether in a floor crowd or electronic limit order book, generally wins. Even in an auction, the bidder with the fastest technology can submit or cancel his order after everyone else.

Traders whose strategies depend on speed are generally lumped together as "high frequency traders." The term is widely used, but it carries some misleading connotations. "High frequency" implies that they trade often. They might, but it's probably more accurate to say that when they decide to trade, they trade quickly. For this reason, (Hasbrouck and Saar, 2013) suggest "low latency," that is, "with minimal delay."

Since technology plays a pervasive role in the subject, it also serves as a good starting point.

21.1. Technology

What does a trader need to win, place, or at least finish with the field in the speed race?

The first step in building a fast system is to eliminate (at least for trading purposes) the person entering orders on the screen and keyboard. The human blink reflex, in response to a loud noise (for example), requires about 200 milliseconds (that is, 0.2 seconds). Add in a bit more time to press the "send" key (assuming that the order screen is already correctly populated) and we might well be facing a response time of one second or more. Present markets are generally accessible through their application program interfaces (APIs) that allow a program written in a familiar language (like Java or C) to directly access a market's information and order entry ports.

Once we have optimized the speed of our programs, we might think about how to make the hardware faster. The off-the-shelf Intel and AMD chips that power our personal computers are too slow, even when multiple processors are used in parallel, so programmers turn to advanced customizable logic chips. The curious might start with "A Low-Latency Library in FPGA Hardware for High-Frequency Trading (HFT)," (Lockwood, Gupte, Mehta, Blott, English and Vissers, 2012).

By using APIs and bit of clever programming, we might be able to generate an order in a few ms. But now the order must be communicated to the market center. At this point we run up against a physical limit – the speed of light ($3 \times 10^8\ m/sec$). If our offices are in Chicago and the market is in New York (a separation of about 1,200 km), one-way transmission time is at least $\frac{1.2\times10^6 meters}{3\times10^8\ meters/second} = 0.004\ sec = 4\ ms$; Shanghai to New York is about 12,000 km (about 40 ms).

Now since we can't circumvent the speed of light, the next best alternative is to position our computer closer to the market. How close do we need to be? Is the same city good enough? The same street? The same *building*?

Modern market data centers are composed of blade servers: racks filled with boxes that are filled with slots. The blades that fit into these slots are circuit boards that hold processors and memory. One blade constitutes "the market". Ideally, we'll place our blade (the one holding our trading logic) in the same server in a slot adjacent to the market blade. Our computer is now said to be collocated (with the market). The practice is generally termed "collocation". Through collocation we can achieve fast two-way communication with the market where we're collocated. This facilitates single-market strategies. We can lift a new offer, for example, immediately after it is posted.

Our collocation does not protect us, however, from delays involving information produced in other places, such as other markets. Suppose that we're collocated with market *X* and we'd like to hit *X*'s bid quickly when market *Y*'s bid (in the same stock) is hit. In learning about market *Y,* we are again faced with transmission latencies. We could collocate a second computer with market *Y*, but this won't help us communicate between the two markets.

Faced with the intermarket communication problem and the light-speed barrier, we can at most achieve marginal improvements in speed. (Laughlin, Aguirre and Grundfest, 2013) note:

> Using relativistically correct millisecond-resolution tick data, we document a 3-millisecond decrease in one-way communication time between the Chicago and New York areas that has occurred from April 27th, 2010 to August 17th, 2012. We attribute the first segment of this decline to the introduction of a latency-optimized fiber optic connection in late 2010. A second phase of latency decrease can be attributed to line-of-sight microwave networks, operating primarily in the 6-11 GHz region of the spectrum, licensed during 2011 and 2012. Using publicly available information, we estimate these networks' latencies and bandwidths. We estimate the total infrastructure and 5-year operations costs associated with these latency improvements to exceed $500 million.

Or consider the following press release (Hibernia Networks, 2015):

> DUBLIN, IRELAND – September 24, 2015 - Hibernia Networks, a leading provider of high-speed global telecommunications services, announces that its new Hibernia Express transatlantic cable has an actual tested latency of better than 58.95ms (milliseconds) from New York to London, which is faster by more than half a millisecond off the original projected speed. The new 4,600km ultra-low latency submarine cable is the first transatlantic cable build in over a decade.

The cost of the cable was generally thought to be in excess of $300 Million.

Clearly our pursuit of speed carries a high price tag. In what senses, if any, are these expenditures justifiable?

21.2. The private and public value of speed

At long horizons, faster information is undoubtedly better information. We are learning about the future sooner. This can produce real benefits. The US government publishes Weekly Weather and Crop Bulletins that summarize recent weather conditions and their impact on key crops. It would be wonderful if some improvement in forecasting technology allowed us to predict in January the content of the coming August report. If we discovered and owned the forecasting technology, we could make large profits in the grain futures markets. We'd establish our positions, publicly release our forecasts, and reverse our trades. The private gains (our profits) would be substantial. But the public (social) gains arising from better planting and harvesting decisions would probably be even larger.

Someone making the case for allowing us to trade on our information might appeal to the principle of market efficiency. It would be argued that the possibility of trade gives us the incentive to produce the information, and that our profits capture only a small portion of the overall benefit. Such statements have frequently arisen in the public debate on high frequency trading, claiming that a faster market is a more efficient market. (We are invoking here the simplest view of private informational efficiency, ignoring the complications discussed in Chapter 13.)

Given equal accuracy, a July forecast of August conditions is less valuable than a January forecast. Farmers might still be able to make better harvesting decisions, but July is probably a little late to change plantings. Here too, though, if we could trade on the information, we'd realize sizable profits.

Now consider a "forecast" of the crop report that is accessible (privately, to us) ten seconds before the public release. Our trading profits would still be substantial, but the social gains would be virtually nonexistent. It is difficult to imagine a ten-second time difference affecting any farmer's decision.[1]

[1] The plot of the comedy "Trading Places" turns on illicit advance knowledge of an orange juice crop report.

The social value of information derives from its importance in affecting real production and consumption activities. Production and consumption decisions, though, are made at time horizons of days, weeks and months. The private value of information in trading decisions is much more time sensitive. Financial markets reward those who move "first", not necessarily those who are "fast" in any absolute sense. If it takes everyone else one minute to hit a bid, I can be first with a response time of fifty-nine seconds; if the normal response time is one ms, though, being first requires reaction within microseconds.[2]

21.3. The high frequency traders

So, who are the high-frequency traders? It depends on how broadly we define the category. Our mutual funds and pension funds routinely use computerized algorithms to implement their trading decisions. Many of these algorithms are available to retail investors. Retail investors, even if they are communicating with their brokers by telephone, would find that the best handling of their orders required them to be passed by sophisticated routing technology. (This is sometimes called agency algorithmic trading because the brokers are deploying it in service to and as an agent for their customers.) In a broad sense, by the standards of the late twentieth century, we are all high frequency traders.

Public discussions of high frequency trading, though, generally refer to a smaller set of players, proprietary trading firms that rely on the fastest technology. Although most of them were born in the current century, some have already become large market participants. In 2013, several of them (Global Trading Systems, Hudson River Trading, Quantlab Financial, and Tower Research Capital) formed a trade group (modernmarketsitiative.org). Their most important role is "de facto" market making, having displaced traditional market makers (like the NYSE specialists). While many observers feel that they have functioned well overall in this capacity, concerns persist (see Section 7.4).

One parallel is especially striking. On the old floor exchanges, members enjoyed better access and information that off-floor customers. This was viewed as an inherent limitation of a market that was convened in a small physical space. Although the new electronic markets held the promise of equal access, old ways hung on. The new "floor" is a blade server, and traders who are on the floor (that is, collocated) still have advantages relative to off-floor customers.

There are, of course, some important differences. The old floor had formidable barriers to entry. If you lacked the good fortune to have inherited a membership, you would have to purchase one. In the latter twentieth century memberships ("seats") on the NYSE sold for several million dollars. Current collocation prices appear to be much lower (see (NYSE, 2016)). When barriers to entry are lower, an economist would expect more competition, and lower profits for HFT firms.

---

[2] It has always been thus. "For some years prior to [the introduction of the telegraph in 1846], William C. Bridges, a stockbroker… maintained a unique private 'telegraph' system between Philadelphia and New York. By the ingenious device of establishing stations on high points across New Jersey on which signals were given by semaphore [flags] in the daytime and by light flashes at night, discerned with the aid of telescopes, information on lottery numbers, stock prices, etc., was conveyed in as short a time as *ten minutes* between the two cities. Some of the mysterious movements in the stock markets were ascribed to this pioneer financial news bureau," ((Barnes, 1911), italics mine). Such was the practice of high-frequency trading in the nineteenth century.
The reader may have elsewhere encountered the story that Nathan Rothschild profited by quickly trading on the news of Wellington's victory at Waterloo (1815), conveyed to him by carrier pigeon. Were it true this anecdote it would be very on point, but (Ferguson, 1998) refutes it.

Many commentators, however, remain critical of high-frequency trading. Some of the accusations are like those that have leveled at human brokers and market-makers, notably that they front-run (that is, trade ahead of) customers' orders, and destabilize markets (see recurrent blog postings at themistrading.com and nanex.net).

21.4. THOR and the start of IEX

As the speed advantages of the fastest traders become ever shorter, the claims to enhanced market efficiency become less persuasive. Even when the information is of a fundamental nature, like the Michigan Consumer Sentiment Index, the social gains from release a few seconds early to select traders is doubtful (see Section 10.1). We are even more skeptical when the information advantage involves market data.

The tension simmered in industry symposia and regulatory hearings for years, but burst into public awareness with the 2014 publication of Michael Lewis' *Flash Boys* (Lewis, 2014). The book was a sweeping indictment of many financial industry practices (including payment for order flow). The book pushed into prominence a firm (IEX), which went on to play a major role in the HFT regulatory debate.

IEX was started in 2012 by Brad Katsuyama, head of electronic trading at Royal Bank of Canada's (RBC's) Capital Markets Group. Initially IEX was a dark pool (an alternative trading system, or ATS). In 2014 it filed an application to become a full-fledged national securities exchange. The application was approved in June 2016.

In *Flash Boys*, Katsuyama recounts a demonstration from 2007, while he was at RBC:

> I'd say, 'Watch closely. I am about to buy one hundred thousand shares of AMD. I am willing to pay forty-eight dollars a share. There are currently one hundred thousand shares of AMD being offered at forty-eight dollars a share— ten thousand on BATS, thirty-five thousand on the New York Stock Exchange, thirty thousand on Nasdaq, and twenty-five thousand on Direct Edge.' You could see it all on the screens. We'd all sit there and stare at the screen and I'd have my finger over the Enter button. I'd count out loud to five . . .
> " 'One . . .
> " 'Two. . . . See, nothing's happened.
> " 'Three. . . . Offers are still there at forty-eight . . .
> " 'Four. . . . Still no movement.
> " 'Five.' Then I'd hit the Enter button and— boom!— all hell would break loose. The offerings would all disappear, and the stock would pop higher."
> At which point he [Katsuyama] turned to the guys standing behind him and said, "You see, I'm the event. I am the news."

The fact that a lifted offer conveys information is, of course, unsurprising (to us, and certainly to Katsuyama as well). But there is more to the story. Why would the offerings all disappear? It was not because Katsuyama's orders were getting filled. It turned out that the only market center where the orders were consistently executed was BATS. This suggested that the BATS execution was the key event in the causal chain that ended with offers on other exchanges vanishing.[3]

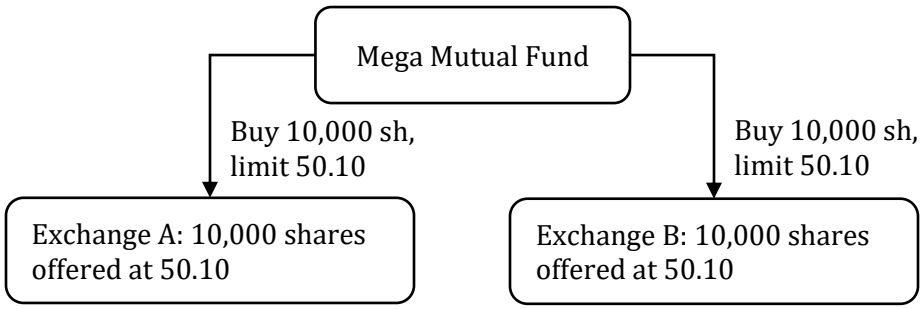
[3] To comply with Reg NMS, the order would have been entered as an intermarket sweep (ISO), with the components sent to all exchanges "simultaneously." With an ISO, of course, nothing is guaranteed, and prudence would dictate that the components be flagged IOC (immediate or cancel) to prevent unintended display of unexecuted portions.

If the offers on other exchanges were simply being cancelled, there might be little cause for concern. A market maker might simply place offers on all exchanges, intending to cancel the others after one of them executed. If I am selling a boat by posting "for sale" messages on five different bulletin boards in my building, nobody would presume that I have five boats for sale. In situations where this needs to be formally stated (as for auto dealerships), the ad will generally include a statement that the one item is being offered "subject to prior sale."

If the offers on other exchanges were being executed, however, we'd be more concerned. The purpose of Reg NMS was to promote virtual consolidation. In a single limit order book, if we submit a large order, it is impossible for another trader to jump in front of us as soon as we have executed a portion of our order. If this becomes possible in a fragmented market, it suggests that the linkage mechanisms are falling short.

To illustrate, suppose exchanges A and B are each posting offers of 10,000 shares (of anything) at $50.10, and that Mega Mutual Fund wants to sweep the top of two markets' ask books (Figure 21-1). But now suppose that transmission speed to A is fast, but transmission to B is slow. The difference in latencies (delays) opens a window of opportunity for a faster predatory trader "Clarence". As soon as there's an execution on A, Clarence buys the shares on exchange B. By the time that Mega's order arrives at B, the shares it wanted have been purchased by Clarence (Figure 21-2).

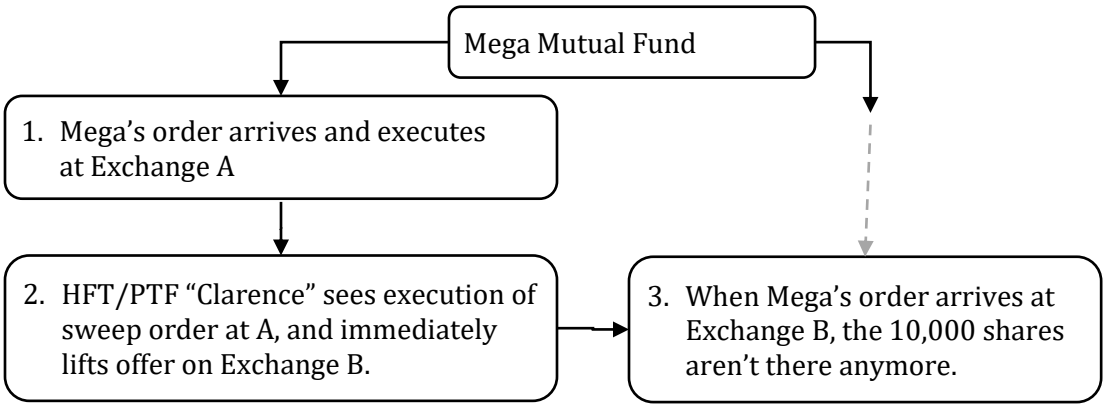
Figure 21-1



The market has seen the offer lifted twice (at each exchange), and this will tend to drive the ask higher. Perhaps Clarence will offer the shares a cent higher (at 20.11). If Mega buys them at this higher price, Clarence will make $100.

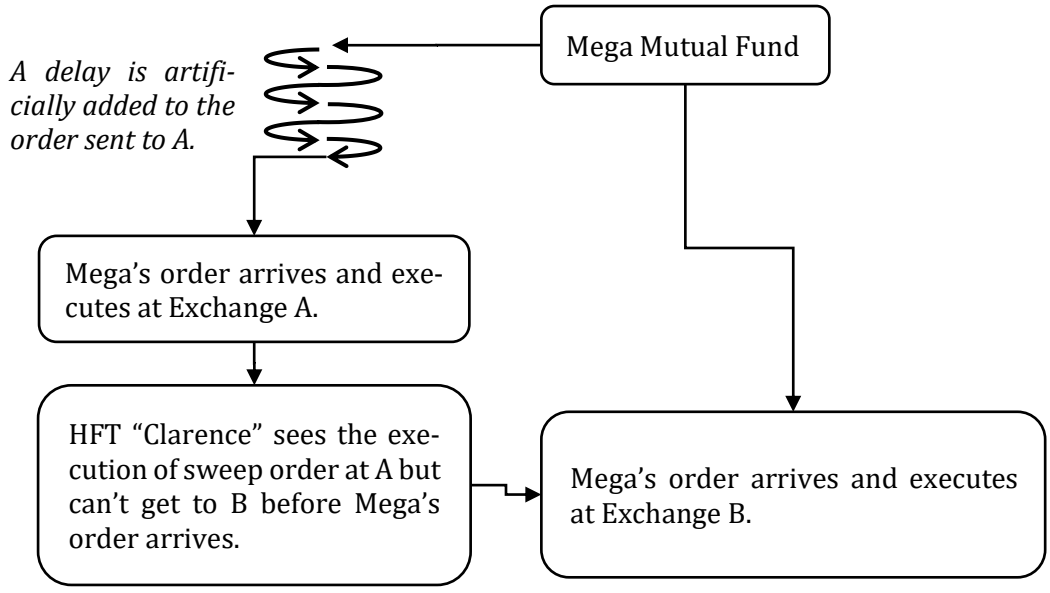
Clarence's strategy is one form of latency arbitrage. Strictly speaking it is not an arbitrage because it involves some risk: the price might drop while Clarence is holding the shares. Given two big trades at the ask, though, this isn't very likely, at least in the short run. Whether or not we call it an arbitrage, it is clear that the profit is deriving not from production of any fundamental information about the stock, but rather from Clarence's anticipation of A's order.

Figure 21-2



The RBC remedy was to delay transmission of the order to A so that it arrived at the same time as the order to B (Figure 21-3). The delay synchronizes the order arrivals so that Clarence cannot preemptively purchase the shares.

Figure 21-3



Although RBC could have introduced the delay by inserting a timer in the software, it decided instead to employ a length of fiber optic cable long enough to ensure that an order submitted at one end would emerge from the other end 350 microseconds later. The cable was compactly coiled in a way that resembled a spool of fishing line. Pictures of coil (once widely available) gave RBC's solution visibility and concreteness, attributes usually absent from the HFT debate. RBC called its system the Tactical Hybrid Order Router (THOR, after the figure from Norse mythology).

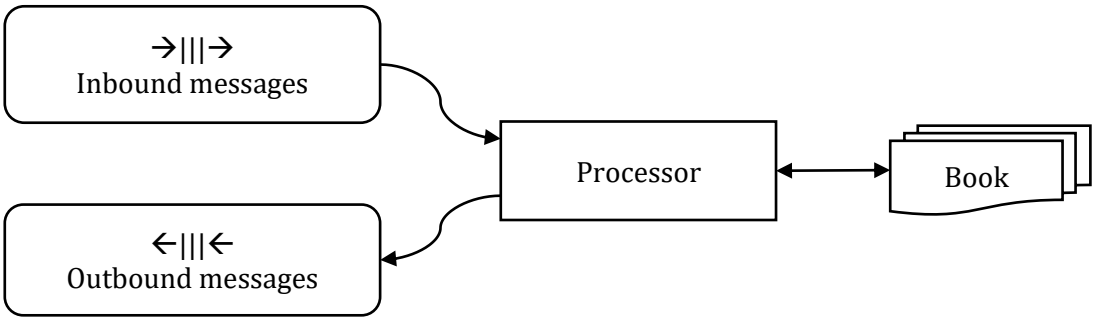
THOR was based on selective delay, the idea that by slowing down one part of the market, things could effectively be reordered in time to eliminate the possibility of latency arbitrage, that is, to recreate a level playing field. It was an effective response to the latency arbitrage directed at sweep orders, but it would not have been effective against similar strategies. In the situation described above, for example, if exchange A's executions generally occurred before exchange B's, Clarence could use A's trades as a signal, to pick off Exchange B's stale bids and offers.

IEX was formed with the intent of applying the delay principle more broadly, to construct a market center where latency arbitrage would not be possible, even when the market center was embedded in the complicated Reg NMS framework of competing exchanges and trade-through prevention. We'll return to the IEX story, but since delay plays such a prominent role, we need to step back and look at how other delays have been introduced and how they are viewed.

21.5. Delay

To explore the role of delay, let's start with a simple limit order market (Figure 21-4). The figure explicitly maps the inbound and outbound message queues, the book, and the processor that links them and makes the decisions. The inbound messages are orders and cancellations; the outbound messages convey the state of the book and report executions. The book itself is depicted as a separate data structure, distinct from the queues and processor. We can imagine the processor functioning as a loop, taking a message from the inbound queue, processing it (executing, cancelling from, or adding to the book). The queues are FIFO (first-in, first-out), which implies that the outcomes of this market are determined by the order of messages in the inbound queue.

Figure 21-4



In any given trade, the buyer and seller would generally prefer that publication of the execution be delayed as long as possible. A trade is often a piece of a broader plan, such as when a large order is split into smaller orders. Delaying publication of a trade makes it more difficult for others to see the big picture.

Disclosure is nevertheless usually viewed as an important aspect of transparency. In one floor market the rule states, "The buyer and seller in a pit transaction must report immediately to the pit observer any change in the last sales price or last quotation and it shall be their duty to make certain that such change in quotation is properly posted," (Chicago Mercantile Exchange, Rule 528). In US securities markets, FINRA rules also require prompt publication.

There are nevertheless some interesting exceptions. In the 1990's, NASDAQ reportedly delayed trade publication to the east coast of the US, to offset transmission delays to the west

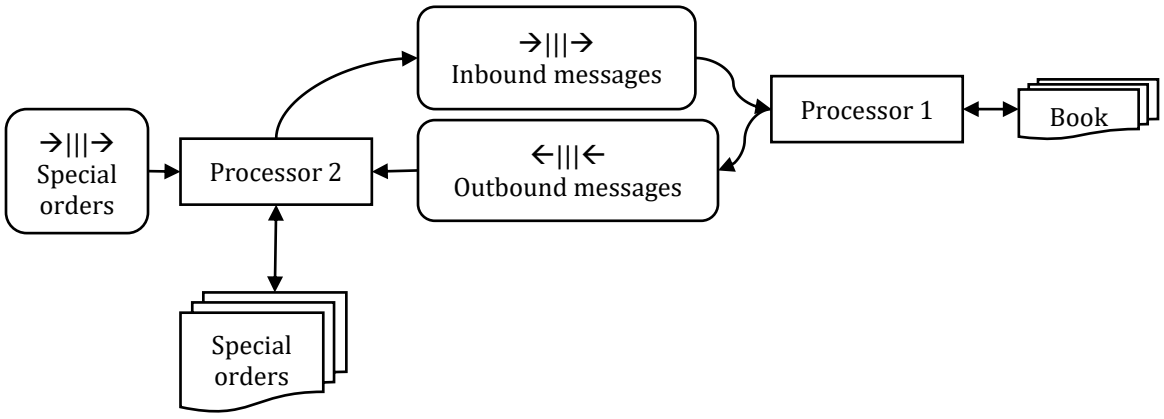
coast. The London Stock Exchange allowed large trades to be published the following day. The intent here was to give dealers an opportunity to work off large positions that they acquired when they accommodated a customer.

There have also been some accidental delays. When a trade occurs, it is usually disseminated to the outside world via broadcast on a market information system. But there are also reports (called confirmations) that are sent back to buyer and seller directly, informing them that the execution has occurred, and alerting them that clearing and settlement procedures will follow. The CME was embarrassed when it came to notice that confirmations were being sent out before public dissemination of the execution (Patterson, Strasburg and Pleven, 2013). The NYSE was fined $5 Million by the SEC because some subscribers received quote updates in advance of transmission to the consolidated feed (U.S. Securities and Exchange Commission, 2012). In both cases, the timing advantages would have been on the order of a few milliseconds, but they were nevertheless perceived as sufficient to support latency arbitrage. The differentials also arose in complex computer systems, and so are difficult to detect.

Intentional inbound delays are rarer. THOR used a selective delay on orders sent to BATS, but the intent was not to delay processing on BATS, but instead to delay other traders' awareness of the BATS execution.

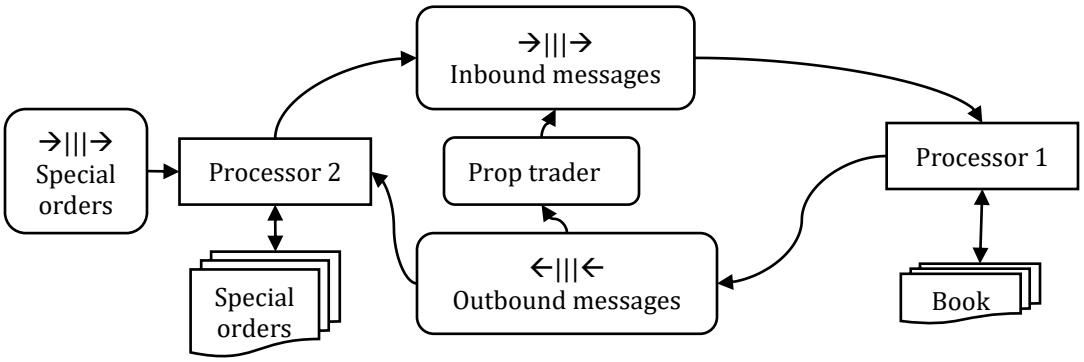
The processing of special orders or conditional orders, such as stopped, pegged, discretionary and so forth, introduces additional complications because that the processor has to monitor the output messages. A stop order is elected in response to an execution; a pegged order is repriced in response to a quote change; a discretionary order may be repriced to be marketable. If there are many of these orders, they may impose a substantial workload. Sometimes they are handled in a second system or processor, as shown in Figure 21-5. NASDAQ's system, RASH (for routing and special handling) was separate from the INET system that managed the book, for example.

Figure 21-5



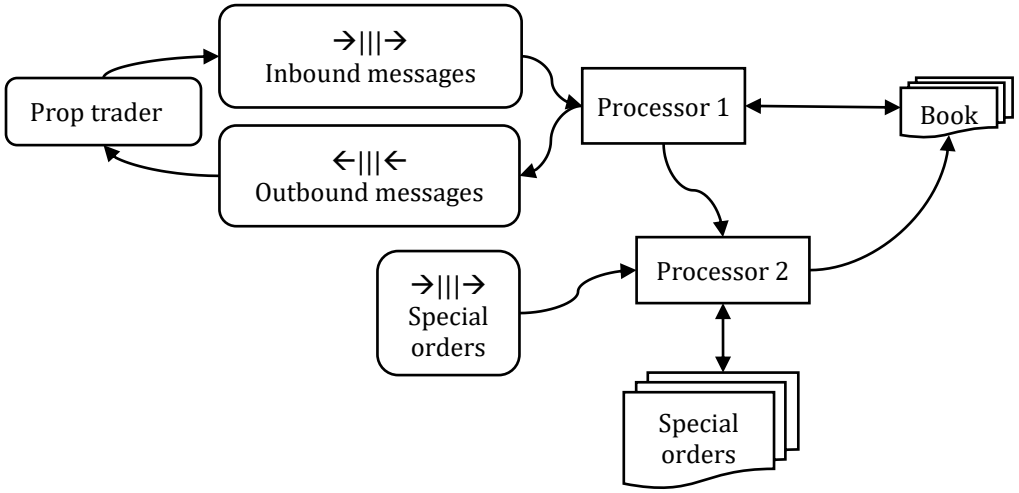
The processing of special orders may also introduce latencies that can be exploited. If a proprietary trader can insert himself advantageously, he might be able to act before all of the special and conditional orders have been updated. For example, if there are many offers pegged to the NBO, repegging all of them when the NBO rises might be so time-consuming that the prop trader might be able to lift some before they can be repriced (Figure 21-6).

Figure 21-6



One approach is to move the special order processing within the inbound/outbound message queues. This moves the prop trader farther away from the order updating processors (Figure 21-7).

Figure 21-7



21.6. The IEX Exchange Application

In its exchange application, IEX proposed the imposition of delays on inbound and outbound messages. Were this a consolidated market and IEX the only exchange, the proposal would have seemed straightforward.

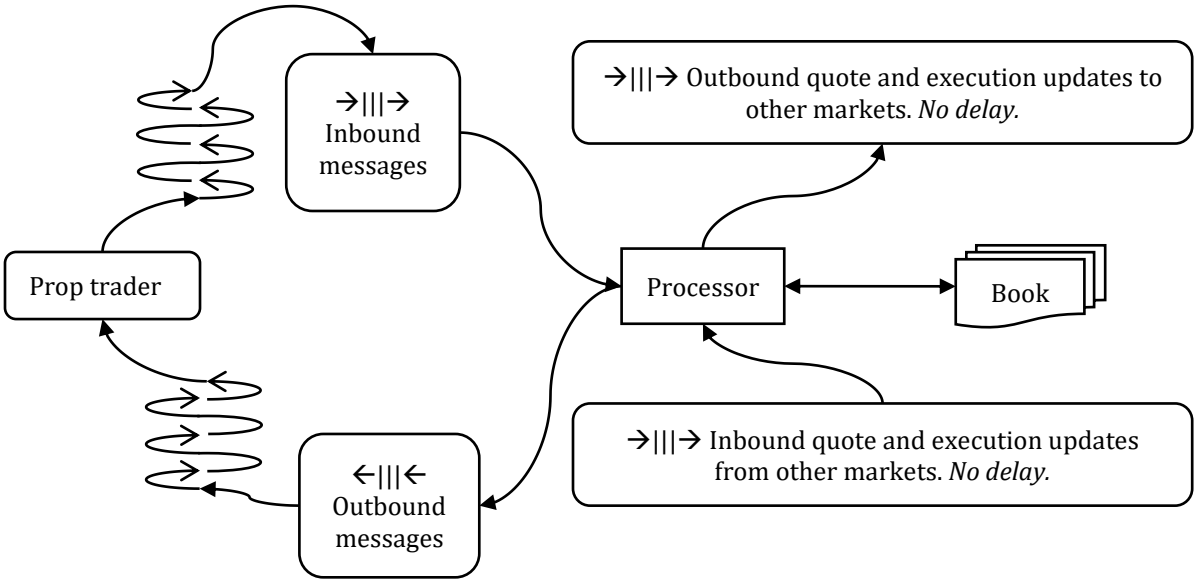
Compatibility with Reg NMS posed difficulties. IEX would certainly want its visible bids and offers to be protected, but in order to be protected, they had to be immediately and electronically accessible. Was this consistent with deliberate delay? The SEC decided that in this case, the delay was "de minimis", that is, so trivial or minor as to be consistent with the intent of Reg NMS, particularly given that the avowed purpose of the delay was to make markets fairer.

Furthermore, for purposes of determining and disseminating the NBBO, IEX's links with other markets would not be subject to the delay.

A summary of IEX's proposal, summaries of comment letters, and the SEC's analysis are given in order approving the exchange application (U.S. Securities and Exchange Commission, 2016). The following discussion draws extensively on this document.

What emerged is diagrammed in Figure 21-8. For clarity, the diagram has all orders (including special orders) going through one processor. IEX's delays on inbound and outbound message queues are actually 350 microseconds. The prop trader (in fact, all customers submitting orders) experiences the delay. Market information, though, at least that conveyed through other market centers, is not delayed in either direction. That is, IEX receives unimpeded updates from other exchanges, and it promptly updates other exchanges. The SEC decided that IEX's bid and offer quotes are protected on the same terms as those of other exchanges.

Figure 21-8



Pegged orders played a prominent role in the SEC's decision: "[The] advantage IEX provides to pegged orders is … designed to ensure that pegged orders on IEX operate as designed .... To accomplish this, IEX slows down incoming order messages by 350 microseconds to allow it to update resting pegged orders when the NBBO changes, so that the resting pegged orders are accurately pegged to current market prices. Without this protection, pegged orders resting on IEX have the potential to be subject to 'latency arbitrage' …" In the SEC's view, the repegging of orders should have priority over incoming attempts to execute against them. Protecting customer orders against latency arbitrage arguably just as valuable as protecting them against trade-throughs, but the view represents a new direction in regulation.

IEX's exchange application included one other novel feature, a new order type, called the discretionary pegged ("D-Peg") order. It is basically a nondisplayed (hidden) limit order, pegged to the NBBO midpoint, that (like other discretionary orders) can actively take the opposing quote. The discretionary orders discussed in section 15.3, however, followed a pre-set rule. For example, a buy limit order is pegged to the NBBO midpoint, but if the NBO ever comes within $0.02 of the midpoint, take it. With the IEX D-Peg order, the buy order will become a

"taker" only if IEX determines that the market is stable, that is, not in the middle of a downward movement, a condition described as a crumbling [bid] quote.

Of course, ascertaining whether a price trend will continue is usually a judgement subject to a high margin of error. ("IBM has closed higher in the last five trading sessions. What are the changes that it will also close higher today?") Why would we expect IEX to be in any position to make an accurate prediction?

IEX's advantage comes from the delays on inbound and, to a lesser extent, outbound messages. Suppose that I have a computer screen that is showing the current NBBO. If you are looking at a screen where the NBBO is delayed by a minute, then from your perspective, I will appear to be clairvoyant, capable of foreseeing bids and offers one minute "into the future". The delay on the outbound side hurts anyone trying to discern the current state of the market; the delay on the inbound side penalizes anyone trying to react to a perceived market condition.

In the SEC's analysis, the delays are supported by two lines of argument. The first is primarily operational, the allowance of a grace period to update pegged orders so that they can be repriced while removed from the threat of execution. The second is informational, allowing the exchange's order management software a look-ahead to protect a particular order type.

The SEC's approval was controversial. The order type being favored is nondisplayed, that is, dark. Some commentators therefore believed that the decision would tilt order strategies toward dark mechanisms. Others have suggested that in competitive response all exchanges will offer similar order types. This would lessen IEX's advantage but would within each exchange shift the balance of orders toward dark pegged orders, a step that would further stress lit markets and displayed limit orders.[4]

21.7. Further reading

The SEC order approving IEX's exchange application, taken together with the comment letters, provide a thorough discussion of current issues. Both are posted to the SEC's website.

Practitioner sentiments on HFT are strong and divided. The pro-HFT view is articulated by the trade organization Modern Markets Initiative (https://www.modernmarketsinitiative.org/). Critical views are more likely to be found on the blogs of Themis Trading (www.themistrading.com).

HFT has proven to be a fertile subject for research. (Jones, 2013) provides an excellent review article; also see the introduction to the Journal of Financial Market's special issue (Chordia, Goyal, Lehmann and Saar, 2013). Many of the earlier papers drew favorable conclusions about HFT, generally finding a positive association between HFT and market quality. It was more difficult to establish causality. (Were the HFT's simply better market makers, or were they becoming more active in markets that were already improving for unrelated reasons?) Later papers, pointing to practices like advance knowledge of public news announcements and latency arbitrage, were more skeptical of the benefits (Biais, Foucault and Moinas, 2015; Clark-Joseph, 2012; Foucault, Hombert and Rosu, 2016; Jarrow and Protter, 2012; Weller, 2016). Angel and McCabe (2013) discuss fairness issues related to HFT. Baruch and Glosten (2013) suggest that fast technology encourages traders to used mixed (randomized) strategies in placing their limit orders and to frequently revise their bids and offers.

Gode and Sunder (2000) point out that geographically dispersed traders can't access a market on an equal basis, and suggest moving to call markets. Budish, Cramton and Shim

[4] Nasdaq considered offering a limit order type that would be subject to a short no-cancel window. In exchange for this commitment to persistence, the order would have priority over a same-price order that could be cancelled at any time (Michaels, 2016).

(2015) suggest that periodic batch auctions without time priority would eliminate overinvestment in speed (also see Section 6.5).

References

Angel, James J., and Douglas McCabe, 2013, Fairness in Financial Markets: The Case of High Frequency Trading, *Journal of Business Ethics* 112, 585-595.

Barnes, Andrew Wallace, 1911. *History of the Philadelphia Stock Exchange* (Cornelius Baker, Philadelphia).

Baruch, Shmuel, and Lawrence R. Glosten, 2013, Flickering quotes, Columbia University, Available at.

Biais, Bruno, Thierry Foucault, and Sophie Moinas, 2015, Equilibrium fast trading, *Journal of Financial Economics* 116, 292-313.

Budish, Eric B., Peter Cramton, and John J. Shim, 2015, The high-frequency arms race: frequent batch auctions as a market design response, *Quarterly Journal of Economics* 130, 1547-1621.

Chordia, Tarun, Amit Goyal, Bruce N. Lehmann, and Gideon Saar, 2013, High-frequency trading, *Journal of Financial Markets* 16, 637-645.

Clark-Joseph, Adam, 2012, Exploratory trading, Department of Economics, Harvard University, Available at.

Ferguson, Niall, 1998. *The World's Banker: The History of the House of Rothschild* (Weidenfeld & Nicolson).

Foucault, Thierry, Johan Hombert, and Ioanid Rosu, 2016, News Trading and Speed, *The Journal of Finance* 71, 335-382.

Gode, Dhananjay K., and Shyam Sunder, 2000, On the impossibility of equitable continuously-clearing markets with geographically distributed traders, Yale School of Management, Available at: http://www.som.yale.edu/faculty/Sunder/Research/Experimental%20Economics%20and%20Finance/Presentations%20and%20Working%20Papers/Network/Design13march2000.pdf.

Hasbrouck, Joel, and Gideon Saar, 2013, Low-latency trading, *Journal of Financial Markets* 16, 646-679.

Hibernia Networks, 2015, Hibernia Express transatlantic cable route connects New York to London in under 58.95 milliseconds, (Dublin, Ireland).

Jarrow, Robert A., and Philip Protter, 2012, A dysfunctional role of high-frequency trading in electronic markets, *International Journal of Theoretical and Applied Finance* 15.

Jones, Charles M., 2013, What do we know about high-frequency trading? , Columbia Business School, Available at.

Laughlin, Gregory, Anthony Aguirre, and Joseph Grundfest, 2013, Information transmission between financial markets in Chicago and New York, Cornell University, Available at: http://arxiv.org/abs/1302.5966.

Lewis, Michael, 2014. *Flash Boys* (W. W. Norton & Company).

Lockwood, J. W., A. Gupte, N. Mehta, M. Blott, T. English, and K. Vissers, 2012, A Low-Latency Library in FPGA Hardware for High-Frequency Trading (HFT), 22-24 Aug. 2012, 2012 IEEE 20th Annual Symposium on High-Performance Interconnects.

Mendelson, Morris, and Junius W. Peake, 1979, The ABCs of trading on a national market system, *Financial Analysts Journal* 35, 31-34 and 37-42.

Michaels, Dave, 2016, Nasdaq tries to appeal to investors lured by rival IEX, August 14, 2016, (The Wall St. Journal).
NYSE, 2016, New York Stock Exchange, Price List 2016.
Patterson, Scott, Jenny Strasburg, and Liam Pleven, 2013, High-speed traders exploit loophole, May 1, 2013, Wall St. Journal (Dow Jones, New York).
U.S. Securities and Exchange Commission, 2012, SEC Charges New York Stock Exchange for Improper Distribution of Market Data.
U.S. Securities and Exchange Commission, 2016, In the Matter of the Application of Investors' Exchange, LLC for Registration as a National Securities Exchange.
Weller, Brian M., 2016, Efficient prices at any cost: does algorithmic trading deter information acquisition? , Department of Economics, Duke University, Available at.

Chapter 22. Cryptocurrency Markets [Incomplete]

Introduction

22.1. The landscape

22.2. Decentralized Exchanges: Uniswap v. 2

22.3. Decentralized Exchanges: Uniswap v. 3

Equation template table:

$$a = b + c \tag{22.1}$$