Securities Trading: Procedures and Principles
Draft Teaching Notes
Joel Hasbrouck

This draft: May 19, 2014

Teaching notes for Stern MBA Course FINC GB.3149: The Structure and Dynamics of Financial Markets.

Joel Hasbrouck is the Kenneth G. Langone Professor 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. Disclaimer: I regularly teach (for compensation) in the training program of a firm that engages in high frequency trading. I have recently served (without compensation) on a CFTC advisory committee. I give presentations at financial institutions for which I sometimes receive honoraria.
Preface

This manuscript is a compilation of teaching notes for a half-semester MBA course entitled The Structure and Dynamics of Financial Markets. The target audience is finance students planning careers in trading or investment management and information technology students who will be supporting trading and investment systems. The exposition draws on general economic principles, but the focus of the presentation is US equity markets.

Much of the content is institutional in nature. This 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 Given the speed with which markets evolve, however, it is likely (maybe even certain) that some of the details are out of date. Markets change their websites and documentation much more frequently than they change their trading procedures, however. So it is quite possible that even if the link documenting a particular practice is no longer current, the practice itself persists.

In teaching the course, I've come to appreciate the value of trading games and other interactive material. My web site (http://pages.stern.nyu.edu/~jhasbrou) contains links to BVIEW (a Java program to depict the dynamics of an actual limit order book) and 7EX (an Excel-based market simulator). Both are freely distributed. I've used a number of VECONLAB games (http://veconlab.econ.virginia.edu/). I've recent started using the Rotman Interactive Trader simulator (http://rit.rotman.utoronto.ca/).

The text is organized in chapters and sections.

Material supplemental to the text is indicated as a bordered inset.

DISCLAIMER

Preparation of these notes was not supported by any external funding. I should note, though, that I receive compensation for teaching short courses in the financial services industry, and I accept speaking honoraria. In addition, I have served in the role of uncompensated consultant to industry and government.
**Table of Contents**

Chapter 1. Introduction .......................................................................................................................... 1

Chapter 2. The US equity market .......................................................................................................... 3

Chapter 3. Limit order markets .............................................................................................................. 14

Chapter 4. Multiple Markets .................................................................................................................. 20

Chapter 5. Auctions .................................................................................................................................. 33

Chapter 6. Dealers and dealer markets ................................................................................................. 45

Chapter 7. Dark liquidity and dark mechanisms .................................................................................. 56

Chapter 8. Public information and trading halts .................................................................................. 61

Chapter 9. Private information ............................................................................................................. 73

Chapter 10. Economic models of private information ......................................................................... 81

Chapter 11. Trading costs ....................................................................................................................... 88

Chapter 12. Conditional orders ............................................................................................................. 103

Chapter 13. A first look at order splitting strategies INCOMPLETE: See powerpoint slides ............. 106

Chapter 14. Statistical models of order-price dynamics ....................................................................... 107

Chapter 15. Order splitting with price impact ....................................................................................... 112

Chapter 16. Fees, rebates and other inducements ................................................................................. 116

Chapter 17. Reg NMS ............................................................................................................................... 121

Chapter 18. The arithmetic of probabilities ........................................................................................... 125

References.................................................................................................................................................. 131
Chapter 1. Introduction

We demand a lot of our securities markets. When we plan our investment or hedging strategies, we rely on market prices to define the strategies are feasible (affordable). We then enter the markets to trade and implement these strategies. 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 liquidate the securities or “unwind” our hedge.

The traditional view of a perfect frictionless market is summarized in the conventional “Econ 101” supply and demand framework. Each buyer and seller is assumed to be “atomistic”, that is, small relative to the market. Because each trader is small relative to the market, and knows that they’re small, they believe that nothing they do will affect the market price. They willingly express their true preferences: when they are asked “how much would you buy if the price were \(x\)?,” for example, they answer honestly. (It does not occur to them 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 is equal to the quantity supplied defines the equilibrium price and quantity.

The process of arriving at the 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?” and “who wants to sell?” The auctioneer than 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, it’s a reasonable conjecture. Stocks are held by thousands of investors, and thousands more might be standing by as potential buyers.

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: ourselves and our counterparty. All of a sudden, the large-number perfect-competition abstraction seems less relevant. We’re behaving strategically, taking into account the fact that our actions can change the price. Most of the time there is no one acting as an “auctioneer”. Most trading occurs during continuous trading sessions. In these interactions, the market procedure and rules matter very much.

These notes are about these rules – the trading protocols – and the economic principles that shape them. Although we can’t avoid talking about the securities, the stocks, bonds and options that constitute the flow of trade, these notes are not primarily about them, their characteristics, or their uses. This is instead an explanation of how they are traded.

Although the material applies to most security markets, the presentation is heavily focused on US equity markets. It is useful to have one actual functioning market as a central example, and in this respect, the US equity market is a good choice. The US equity market is large and active, and also 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 are likely to play out elsewhere.
This is not to say that US markets are invariably at the forefront of technology. For example, into the 21st century, when the rest of the world had long since adopted decimal prices, US markets were still trading in “eighths”. 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.

Nonetheless, when the US finally did make the transition to electronic trading, it did so in a flexible and open fashion. US regulators mostly took the stance that a stock exchange was not a “natural monopoly”, and that there was much to be gained from a competitive race 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 (again, for better or worse) are spreading to the rest of the world.

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.

No single reference can coherently present the diverse lines of thought that comprise market microstructure. But 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. The approach in (Foucault, Pagano, and Roell, 2013) comes closest to the present material, but with more emphasis on economic models and principles. (Aldridge, 2013) discusses financial markets from a high-frequency trading perspective.

The citations in these notes will point you 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), i.e., the structure and characteristics of specific securities. In this regard, (Bodie, Kane, and Marcus, 2011) is a useful source to have at hand.
Chapter 2. The US equity market

These notes use the US equity (stock) market to illustrate trading mechanisms and principles. This chapter lays out the key features of the market and its operations.

2.1. The global context.

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 S&P (Standard and Poors) Global Broad Market Index covers about 10,000 companies and accounts for most of the value of equities world-wide. The 2011 end-of-year total market capitalization of these equities is approximately $31 Trillion. The S&P Total [US] Market Index roughly approximates the US component of the global index. The 2011 year-end market capitalization of this index is about $14 Trillion. So the US component is about 45%.

Besides capitalization, we’re often concerned with trading activity. Whereas market capitalization is a “point in time” measure, trading volume is a flow, shares (or dollars) traded per unit time. Usually, trading volume simply denotes the number of shares bought or, equivalently, the number of shares sold by all participants in the market. The most useful estimates of these figures are provided by the stock exchanges – the places where trade occurs. Table 1-1 summarizes market capitalization and trading volume for some of the world’s largest exchanges.

<table>
<thead>
<tr>
<th>Domestic equity market capitalization (end of year, $ Trillion)</th>
<th>Trading Volume (annual, $ Trillion)</th>
<th>Annual Turnover</th>
</tr>
</thead>
<tbody>
<tr>
<td>NYSE Euronext (US)</td>
<td>11.796</td>
<td>18.027</td>
</tr>
<tr>
<td>NASDAQ OMX</td>
<td>3.845</td>
<td>12.724</td>
</tr>
<tr>
<td>Tokyo SE Group</td>
<td>3.325</td>
<td>3.972</td>
</tr>
<tr>
<td>London SE Group</td>
<td>3.266</td>
<td>2.837</td>
</tr>
<tr>
<td>NYSE Euronext (Europe)</td>
<td>2.447</td>
<td>2.134</td>
</tr>
<tr>
<td>Shanghai SE</td>
<td>2.357</td>
<td>3.658</td>
</tr>
</tbody>
</table>

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

The correspondence between the S&P country estimates and the exchange estimates is close, but not exact. Exchanges trade products besides equities (such as exchange-traded funds and closed-end mutual funds), and there are other technical differences between the two sets of calculations.

There’s a general positive relation between the market capitalization figures in Table 1-1 and the dollar trading volume figures in Table 1-2. This is to be expected because shares in large firms would be widely held and actively traded. An intuitive measure of trading activity that looks beyond this effect is turnover.
Turnover is the ratio of dollar trading volume to market capitalization. For a single firm, this is equivalent to the ratio of share volume to the number of shares outstanding. Because volume is a rate, turnover has a time dimension. An annual turnover of 200% for a firm, for example, suggests that trading volume over the year could have resulted from the firm being bought and sold twice. This interpretation is only suggestive because it is an average estimate, and it is unlikely in practice that each share was actually bought and sold twice. It is more probable that most of the shares did not change hands during the year, while some shares were traded ten or more times. The turnover figures in the table are also averaged across firms, and here too we’d expect large variation, with some firms being very actively traded, and others not.

There are roughly 5,000 actively traded stocks in the US. The largest 1,500 have a total market capitalization of about $13T, and account for about 90% of the total. Most have a capitalization of around $100 Million but a small number are much larger. Exxon Mobile and Apple are both around $400 Billion.

2.2. Exchanges

An exchange usually 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.

Modern exchanges offer a complicated, diverse, and sometimes bewildering range of financial products and services. But most of their activity and value arises in 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, and the exchange 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 and regulation of the trading process.
- The trading 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 primary. The most important US listing venues are the NYSE, NYSE Arca, NYSE Amex, 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.

NYSE (sometimes called “NYSE classic” to differentiate the former New York Stock Exchange from other exchanges owned by NYSE Euronext) 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 the Amex when it grew a little larger, and ultimately step up to the NYSE. From a listing perspective, NYSE Arca represents an NYSE Euronext initiative to list companies whose profile comes closer to NASDAQ. From the NYSE Euronext website: “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-classic and Amex listings have ticker symbols of three letters or less, like IBM, GE, or C (Citigroup); most NASDAQ- and ARCA-listed securities have four-letter symbols, like MSFT (Microsoft), INTC (Intel), and QCOM (Qualcom).

**Exchange Ownership**

Until the end of the twentieth century, exchanges were historically 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 in 1990, exchanges began to reorganize themselves as for-profit-corporations, with publicly-traded shares. In this new 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.

Nowadays, although the exchanges continue to operate mostly independently, many are owned by large holding companies, and there have been many changes of name and ownership. NYSE Euronext owns the New York Stock Exchange, MKT (the former American Stock Exchange), Arca, the Paris, Amsterdam, Brussels and Lisbon Exchanges, among others. It also owns various derivatives exchanges, notably LIFFE (the London International Financial Futures Exchange). As of June, 2013, NYSE Euronext is on track to be acquired by the Intercontinental Exchange (“ICE”).

NASDAQ OMX owns the NASDAQ stock market, BX (the former Boston Stock Exchange), PHLX (the former Philadelphia Exchange), most of the Scandinavian stock exchanges (except for the Norwegian Stock Exchange), and eSpeed (a trading platform for US Treasuries).

Some of this consolidation has encountered resistance from regulators concerned about competitiveness. In 2011, a contemplated merger of the NASDAQ OMX and NYSE Euronext was effectively blocked by the US Department of Justice. A planned combination of the NYSE Euronext and DeutscheBourse (the principal German exchange) was similarly derailed.

### 2.3. Brokers

We can’t trade simply by visiting an exchange’s web site and giving a credit-card number. For a number of 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. Whatever the level of complexity, though, the broker is still representing the customer, acting as the agent.

*Prime brokers* provide transaction-related services for large and institutional customers. There are also *retail brokers* that focus on transaction-related services. They are sometimes described as “discount” or “online” brokers, as opposed to “full-service” brokers who provide more comprehensive investment management and advice.

### 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.

#### 2.4.A. Investment horizon

Investors are sometimes categorized on the basis of investment horizon. *Long-term investors* include institutions like endowment funds and individuals saving for retirement or a child’s education. Their horizons are long relative to real business cycles. Most long-term investors start from an indexed portfolio: stocks are held in proportion to their market capitalization. This is most easily accomplished with index mutual funds or index exchange-traded-funds (ETFs). Investors might deviate from this portfolio due to individual-specific risk exposures or an opinion on whether a company is correctly valued. Changes in risk exposure or valuation beliefs will induce these investors to revisit their portfolios and possibly trade. Most trading needs can be spread out over days, weeks or months. These investors are the most numerous and most important in terms of the value of their holdings. Their turnover, however, is low, and they don’t generate very much trade.

*Medium-term investors* have holding periods are on the order of a business cycle (3-5 years). These investors often seek to profit from changes in relative valuations of securities. Portfolio weights usually change over the business cycle, and the strategies are sometimes described as tactical asset allocation. Tactical asset allocation strategies can be based on beliefs about country, industry, company fundamentals or market-driven technical indicators.1

*Short-term traders* have holding periods ranging from minutes to a few months. They are a heterogeneous lot. News traders play short-term momentum or volatility – often in the wake of a public news announcement.

---

1 For a given company, fundamental investment information comprises considerations like sales, earnings, and so on. These data are associated with the business activities of the firm. For a country, fundamentals would include consumption, investment, and other measures related to real economic activity. Technical information is based on financial market data. One indicator, for example, might be the average of recent stock prices. Technical indicators are usually summary statistics of market prices and volumes.
2.4.B. Motive

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). Information effects have profound implications for the trading process, and are discussed more fully in a later chapter.

Non-informational motives include hedging, liquidity, and gambling.

**Hedging** trades aim at risk reduction. For example, high-level executives, as a result of stock and option grants, often have personal wealth that is more concentrated in the stock of a single firm (their own) than prudent diversification would suggest. A partial sale of these holdings will reduce their overall risk.

Alternatively, the typical stock-ETF arbitrage involves buying stock and selling an equivalent amount of the ETF. The arbitrage is not riskless until trades have established both positions. This can cause trading needs that must be satisfied immediately.

**Liquidity** motives stem from unexpected cash outflows and inflows. A mutual fund’s assets under management, for example, changes as customers invest or divest shares in the fund. On any given day, these are unlikely to be exactly offsetting, so the fund must trade.

Financial markets are sometimes compared to casinos. There are good arguments against this view. Financial markets are not, in the long run, zero-sum games. (In a zero-sum game, there’s a loser for every winner.) Financial markets perform important functions in allocating capital across time and different uses, and in risk management. These functions can create wealth, not simply redistribute it.

Nevertheless, financial markets can exert a powerful appeal to our gaming instincts. As in most gambling situations, there is the possibility of large gain, but the expected return is negative. (Even if financial markets are fair and efficient, trading is costly.)

Gambling is not completely antithetical to “rational” investing and hedging. An investor might buy a stock on the basis of exhaustive analysis, and yet still experience a thrill at the moment of purchase akin to that associated with the purchase of a lottery ticket, making a wager on a horse race, or betting on the outcome of a sporting event. In the words of John Maynard Keynes, “The game of professional investment is intolerably boring and over-exacting to anyone who is entirely exempt from the gambling instinct; whilst he who has it must pay to this propensity the appropriate toll” (Chapter 12: The State of Long-Term Expectation (The General Theory of Employment, Interest and Money).
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 a 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 a trade; they really bought and sold. On the other hand, since the trade occurred before it was reported, 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 process of bidding and offering is very market-specific. Sometimes bids and offers are made orally, “bidding 20.10 for 100 shares” or “offering 200 shares at 20.15,” for example. This might happen during face-to-face trading on a floor market, or we might hear it as a disembodied voice over the squawk-box next to our computer. Nowadays, though, the bid or offer are more commonly communicated made machine-to-machine.

If a bid or offer is firm, it is available for trading with no further questions asked (like, “Are you sure?”). If the bid is firm, a seller can “hit” the bid and trade with no further discussion. Similarly, if the ask is firm, a buyer can “lift” or “take” the ask. In US equity markets, bids and asks are presumed to be firm and available for immediate execution, unless qualified in some manner.

In some situations, the bid and offer might be indicative. With an indicative bid, a prospective seller has to contact the bidder directly and ask, essentially, “At what price are you really willing to buy.” A retail broker, for example, might post indicative bids and offers to their website, indicating that they are in the business of actively buying and selling, and encouraging customers to pick up the phone and get the real prices.

How long can we presume a bid or offer is firm? By custom, oral quotes are assumed to be “firm only as long as the breath is warm.” In electronic systems, though, bids and offers are assumed to be valid until they are explicitly withdrawn (or the market is closed).

The difference between the ask and bid quotes is the spread. Assuming that the bid and ask prices didn’t change, the spread is the cost incurred by someone who bought and immediately reversed the trade. Alternatively, if the bid and ask quote were posted by a single agent, the spread could represent the realized profit.

The average of the bid and ask quotes, also known as the bid-ask midpoint (BAM) is sometimes used as a convenient price indication. Particularly for infrequently-traded stocks, for which the last-sale price might be quite stale and the bid-ask spread might be very wide, the BAM is especially useful.
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 at the moment, or should we try to make our own price and await the arrival of a counterparty who finds our price agreeable?

We’ll start by assuming that we have a stock and a direction. Direction is shorthand for “buy” or “sell”, and to make things concrete, let’s assume that we’re buying. A buyer entering the market has 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 is $100 (per share) bid, offered at $101. She can buy immediately by paying $101, i.e., taking someone else’s price. Or she might make a bid of, say, $100.25. If an agreeable seller arrives, she’ll buy at 100.25.

The make or take decision is the choice: 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 and buying at a price higher than the original $101 offer.

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 actually 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 sometimes lead to nasty surprises, some markets 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 agent who is posting the bid or ask/offer and stands willing and available for trade. The passive side is also called the standing or resting side.

In any given trade, the active side might be the buyer or the seller. We refer to these situations differently. An active buyer hits the bid. An active seller 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. Indeed there are many instances of the expression online and in print. To the traditionalist, though, “hit the ask” sounds wrong, and may even be taken as a mark of the speaker’s ignorance.
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 seem more than a few. The hit/lift construction adds a little more information that helps clarify intent.

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 cognitive closure (resolution, removal of uncertainty).

2.7. The market in motion

The market is a dynamic system. The make-or-take decisions of traders determine the prevailing bid and ask prices, and generate trades. To illustrate how this plays out, Figure 2.1 shows activity for ticker AACC (Asset Acceptance Capital Corporation, a debt collection company) for few hours on April 25, 2011.

In the figure the bid is a solid line and the ask is a dashed line. (If you have a color copy, the bid and ask are blue and red.) Once a bid or ask is set, it persists until it is withdrawn or revised. The bid and ask therefore extend in time, and appear as step functions. A trade, which occurs at a single point in time, is indicated by a black dot. Some events may appear to take place simultaneously. This is largely an artifact of the data collection and dissemination process.

Between 10:15 and 12:00 there are many changes in the bid and offer. Mechanically, this is occurring as traders modify their limit orders. A limit order can be canceled at any time (unless it has already been hit), and the trader can replace it with another order at a difference price. This cancel and resubmit effectively reprice the order. Some markets accept order modification instructions that accomplish these steps with a single message.

At any given time, a limit order’s price reflects the trader’s belief in the “true” or “fundamental” value of the security, and also the trader’s desire to attract a counterparty. Both of these change as new information becomes available or the trader’s circumstances change. A buyer, for example, might raise his bid because he’s heard a rumor of an earnings increase or because he’s nearing some deadline for accomplishing the purchase.
Figure 2.1. Trading activity in AACC

The bid (National Best Bid) is a blue solid line; the offer (ask) price (National Best Offer) is a dashed red line; black dots represent trades. Source: NYSE daily TAQ.

Between 11:00 and 12:00, the bid progressively rises, and then suddenly drops. We can’t be certain that this is one bidder, but if it were, we might conjecture that the increases are intended to sweeten the deal for a prospective seller. That is, the bidder might be thinking, “If I raise the bid a bit more, one of the traders on the ask side of the market might decide to cancel his sell order and hit my bid.” We can also speculate about the reason for the sudden drop. Put yourself in the position of a potential seller. Seeing an increase in the bid, the seller might simply wait in the expectation of further increases. The bidder in turn might suddenly drop his price to signal, “This is as far as I go.” On the next round of increases, the seller might be less inclined to wait.

This is a rather involved story to explain a particular price path. It is a conjecture, and it might be wrong. But it is a sensible story, and considerations of this sort play out in all kinds of markets. More to the point, though, it illustrates a key aspect of the trading process. The buyer and seller must each make some conjecture about the others’ behavior.
Shortly after 12:00, the bid is hit and a trade occurs. To continue our narrative, "A seller decided not to wait any longer." The trade removes the bid, leaving exposed the next lower order as the market’s new best bid. The process begins again. A bidder (the same bidder?) increases the price. A seller (the same seller?) waits until the first trade price is reached, and then hits the bid.

Around 13:45, there is a different sort of behavior. One or more buyers lifts the ask quote. The ask quote does not bounce back, however, to its previous level. Other buyers appear, and lift progressively higher ask prices. More precisely, lower priced sell-limit orders are executed, leaving exposed the higher-priced orders. There are a few trades at the bid, but trades at ask prices dominate. ("The order flow is net buy", in market shorthand.)

Over the course of the next few minutes, both bid and ask prices increase. These increases are, at least relative to a short-run trading horizon, persistent. Persistent changes security prices are usually due to some new information. Sometimes the information consists of a public news release relates in a direct fashion to the company's earnings or financial prospects (see Chapter 8 on public information). Sometimes, though, the trading process itself generates the information. A net buy order flow causes a suspicion that the buyers know something, that some as yet unpublicized positive development has occurred.

At this point, we’ll move on from the study of AACC. We have by no means exhausted the possibilities, and other questions might well be occurring to the reader. Some of these will be addressed in the material that follows, but ultimately, short of dredging up the recollections of all the players, any narrative of this sort will inevitably remain incomplete.

2.8. 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.
- **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.
- **Resiliency**, in the sense of "bounce back," suggests that any price changes that might accompany large trades are transient and quickly dissipate.

Tightness and depth can be measured fairly precisely at a given time. Immediacy, though, for all but the smallest quantities, is speculative. Large trades are usually accomplished gradually, and we don’t know in advance how long the whole process might take. Resiliency is a similar, dynamic, property.

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 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.

The term 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”.

Transparency refers to the amount of information available about the market and trading process. In US equity markets, one generally knows 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 widely 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* is sometimes used to refer 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.
Chapter 3. Limit order markets

Most trading in equities, options and futures contractions is organized around limit order books. To describe these markets, it’s useful to start with the bids and offers discussed in the last section. Specifically, where do they come from?

3.1. Sources of bids and offers

Who can make a bid or offer in a market? In a sense, the answer is “anyone”. But it doesn’t do much good to make a bid or offer if no one is aware of it and no one can execute against it. For a bid or offer to be meaningful, it must be publicized and it must be accessible for execution. In some markets, all bids and offers are made by professional intermediaries – dealers. Bids and offers made by customers are neither visible nor accessible.

In US equity markets it is easy for anyone to make a bid or offer. If the customer wants his bid/offer to be displayed, it will be. A visible bid or offer also has certain protections that enhance the likelihood that it will be executed. At any given instant, there may be many bids and offers pending in the market, originating from a wide variety of retail and institutional customers. Formally, these bids and offers correspond to customer limit orders. A collection of limit orders is called a limit order book (or simply, a book).

Sometimes the book of customer orders is the only source of bids and offers, and all trades have a book bid or offer on one side. In this scenario, the book is the central market mechanism. A market organized around a book is called a limit order market, or an order-driven market (in that the bids and offers primary reflect customer orders).

Worldwide at the present time, limit order markets dominate trading in equities, listed options and listed futures contracts. Another mechanism, the dealer market (also known as a quote-driven market) dominates trading in swaps, foreign exchange and bonds. We’ll discuss limit order markets in greater detail, and return to dealer markets later.

3.2. Execution procedures

Recall that a limit order specifies a direction (buy or sell), quantity (number of shares), and a price. “Buy 200 MSFT limit 25,” means “Buy 200 shares of MSFT, but don’t pay more than $25 per share.”

When this order reaches a market center, 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”) order.

Suppose that the book has one standing order: “Sell 200 MSFT limit 25.”

- An arriving order “buy 200 MSFT limit 25” is marketable. There is a trade at 25.
  
  The price is determined by the standing 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” \( \rightarrow \) trade of 200 shares (the offer amount).

  “Buy 50 MSFT limit 25” \( \rightarrow \) trade of 50 shares (the buy amount).
3.3. Priority within the book

For a given incoming marketable order there might be multiple standing 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 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 offerer for stepping in quickly with their best price.

Some books accept orders marked as hidden (nondisplayed). 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 arrived later. A displayed limit order is like an advertisement for the market, and so priority for these orders therefore makes sense from the exchange's viewpoint.

Some markets accord 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 share 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.

3.4. More complex interactions and order qualifications

Suppose the current state of the book is:

<table>
<thead>
<tr>
<th>Price</th>
<th>Quantity</th>
<th>Entered at</th>
<th>Trader</th>
</tr>
</thead>
<tbody>
<tr>
<td>50.12</td>
<td>1,000</td>
<td>9:30</td>
<td>Cathy</td>
</tr>
<tr>
<td>50.11</td>
<td>500</td>
<td>9:32</td>
<td>Bill</td>
</tr>
<tr>
<td>SELL</td>
<td>50.10</td>
<td>9:31</td>
<td>Amy</td>
</tr>
<tr>
<td>BUY</td>
<td>50.05</td>
<td>1,000</td>
<td>David</td>
</tr>
<tr>
<td></td>
<td>50.04</td>
<td>500</td>
<td>Ellen</td>
</tr>
<tr>
<td></td>
<td>50.03</td>
<td>400</td>
<td>Fred</td>
</tr>
</tbody>
</table>
Traders’ names are shown here to simplify reference to the order. In US equity markets, customer names are not visible. If the bid or offer comes from a professional intermediary, the firm might be identified.

If Gina sends an order to “buy 500 shares limit 50.12,” her order executes at multiple prices. (It “walks through the book.”) She buys 400 shares @50.10, 500 shares at $50.11, and 100 shares at $50.12.

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.

Sometimes qualifications are added to orders. 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.

3.5. Floor markets

Many of today's limit order markets were originally floor markets. In a floor market, all trading occurs in a real centralized space. Floor traders either buy and sell on their own account, or represent orders given to them by people away from the floor. Trading is mostly face-to-face and bilateral.

In a floor market, traders make oral bids and offers, and accept or reject those of other traders. The bids and offers are good “only as long as the breath is warm,” so trading requires an ongoing physical presence. A limit order is also a bid or offer, but most limit orders are intended to be more persistent. While one can imagine that trader representing a customer limit order might wait around patiently until the market price comes near, this would not be an efficient use of his time, at least certainly not for one order.

One solution is to collect a number of limit orders in one place (a book) and designate one person as agent for the book, charged with representing the orders. On the New York Stock Exchange, this role fell to a broker designated as a “specialist”. The limit order book was originally paper-and-ink, then electronic. Once order conveyance and maintenance of the book were electronic, the only thing required to achieve a modern market architecture was automatic execution. This would allow an incoming marketable to transact quickly and automatically against the book. It was also evident, however, that automatic execution would largely render unnecessary the specialist, floor traders
and the very existence of the floor itself. Most floor markets sensibly resisted automatic execution. 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.”

Not all floor markets had one single centralized book. In the Chicago futures pits, numerous brokers had their own books, consisting of limit orders entrusted to them by their customers and other brokers. For these markets, the transition to automation was a difficult one. It was arguably made possible by the ownership structure of the exchange. As the owners of the exchange, members had some incentive to maintain the economic value of the exchange, even at the expense of a diminished role for the floor. Furthermore, at least one exchange facilitated the transition with side-by-side electronic and pit trading.

Although floor-based exchanges are fading from memory, there are good reasons for studying them. They are among the oldest formal trading institutions, and their procedures have therefore stood the test of time. Their rules often facilitated efficient trading. Furthermore, the members designed many rules and imposed these rules upon themselves well before the advent of external government regulation. Many of the questions presently encountered by our dispersed electronic markets were first faced by the floor markets. In later chapters, we will explore floor exchanges’ approaches to market making and dark trades.

On the NYSE floor, established principles included last-sale reporting and price priority. The workings of time priority were a little more complicated. In the floor crowd, time priority was extended only to the first bid (or offer) at a new price. This may have reflected the difficulty of monitoring arrival times in a changing crowd of people: it’s usually clear who’s first, but it’s less clear who is tenth. Within the limit order book, however, time priority was strictly observed. Floor procedures also accorded an advantage to large orders. The principle of precedence according to size ranked resting orders that were large enough to accommodate an incoming order ahead of smaller orders.
Problems

Problem 3.1 Constructing the book
Under price-time priority, build the bid side of the book from the following limit buy orders.

<table>
<thead>
<tr>
<th>Trader</th>
<th>Arrival Sequence</th>
<th>Quantity</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amy</td>
<td>1</td>
<td>100</td>
<td>20.02</td>
</tr>
<tr>
<td>Brian</td>
<td>2</td>
<td>400</td>
<td>20.05</td>
</tr>
<tr>
<td>Chad</td>
<td>3</td>
<td>200</td>
<td>20.07</td>
</tr>
<tr>
<td>Dana</td>
<td>4</td>
<td>100</td>
<td>20.06</td>
</tr>
<tr>
<td>Emily</td>
<td>5</td>
<td>300</td>
<td>20.07</td>
</tr>
<tr>
<td>Frank</td>
<td>6</td>
<td>300</td>
<td>20.05</td>
</tr>
</tbody>
</table>

Problem 3.2 Constructing the book
Under price-visibility-time priority, build the offer side of the book from the following sell orders.

<table>
<thead>
<tr>
<th>Trader</th>
<th>Arrival Sequence</th>
<th>Quantity</th>
<th>Price</th>
<th>Display</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amy</td>
<td>1</td>
<td>100</td>
<td>10.10</td>
<td>Y</td>
</tr>
<tr>
<td>Brian</td>
<td>2</td>
<td>400</td>
<td>10.02</td>
<td>Y</td>
</tr>
<tr>
<td>Chad</td>
<td>3</td>
<td>200</td>
<td>10.01</td>
<td>N</td>
</tr>
<tr>
<td>Dana</td>
<td>4</td>
<td>100</td>
<td>10.02</td>
<td>Y</td>
</tr>
<tr>
<td>Emily</td>
<td>5</td>
<td>300</td>
<td>10.01</td>
<td>Y</td>
</tr>
<tr>
<td>Frank</td>
<td>6</td>
<td>300</td>
<td>10.10</td>
<td>N</td>
</tr>
<tr>
<td>Gina</td>
<td>7</td>
<td>200</td>
<td>10.00</td>
<td>N</td>
</tr>
</tbody>
</table>

Problem 3.3 Order outcomes
The state of the book is:

<table>
<thead>
<tr>
<th>Trader</th>
<th>Sequence</th>
<th>Price</th>
<th>Quantity</th>
<th>Display</th>
</tr>
</thead>
<tbody>
<tr>
<td>Irina</td>
<td>5</td>
<td>16.00</td>
<td>1,000</td>
<td>Y</td>
</tr>
<tr>
<td>Gina</td>
<td>2</td>
<td>15.07</td>
<td>500</td>
<td>Y</td>
</tr>
<tr>
<td>Esteban</td>
<td>1</td>
<td>15.07</td>
<td>100</td>
<td>Y</td>
</tr>
<tr>
<td>Dmitri</td>
<td>4</td>
<td>15.05</td>
<td>300</td>
<td>N</td>
</tr>
<tr>
<td>Sell</td>
<td>Alice</td>
<td>3</td>
<td>15.05</td>
<td>200</td>
</tr>
<tr>
<td>Buy</td>
<td>Bruce</td>
<td>2</td>
<td>15.03</td>
<td>400</td>
</tr>
<tr>
<td>Hans</td>
<td>1</td>
<td>15.01</td>
<td>300</td>
<td>Y</td>
</tr>
<tr>
<td>Cho</td>
<td>3</td>
<td>15.00</td>
<td>200</td>
<td>Y</td>
</tr>
<tr>
<td>Frank</td>
<td>4</td>
<td>15.00</td>
<td>300</td>
<td>N</td>
</tr>
<tr>
<td>Jing</td>
<td>5</td>
<td>14.50</td>
<td>2,000</td>
<td>Y</td>
</tr>
</tbody>
</table>

For each of the following incoming orders, describe the outcome (executions if any, and any changes in the prices, quantities and trader names at the BBO). The incoming orders represent different scenarios (not sequential arrivals).

a. Kathy: “Buy 300, limit 15.05.”
b. Lane: “Buy 600, limit 15.10.”
c. Maureen: “Buy 600, limit 15.05.”
d. Ollie: “Sell 500, limit 15.00.”
e. Petra: “Sell 500, limit 15.02.”
f. Rama: “Sell 500, limit 15.02, IOC.”
g. Sydney: “Sell 800, limit 15.01, FOK.”
Answers to chapter problems

Answer to Problem 3.1

<table>
<thead>
<tr>
<th>Priority</th>
<th>Trader</th>
<th>Arrival Sequence</th>
<th>Quantity</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Chad</td>
<td>3</td>
<td>200</td>
<td>20.07</td>
</tr>
<tr>
<td>2</td>
<td>Emily</td>
<td>5</td>
<td>300</td>
<td>20.07</td>
</tr>
<tr>
<td>3</td>
<td>Dana</td>
<td>4</td>
<td>100</td>
<td>20.06</td>
</tr>
<tr>
<td>4</td>
<td>Brian</td>
<td>2</td>
<td>400</td>
<td>20.05</td>
</tr>
<tr>
<td>5</td>
<td>Frank</td>
<td>6</td>
<td>300</td>
<td>20.05</td>
</tr>
<tr>
<td>6</td>
<td>Amy</td>
<td>1</td>
<td>100</td>
<td>20.02</td>
</tr>
</tbody>
</table>

Answer to Problem 3.2

<table>
<thead>
<tr>
<th>Priority</th>
<th>Trader</th>
<th>Sequence</th>
<th>Quantity</th>
<th>Price</th>
<th>Display</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Gina</td>
<td>7</td>
<td>200</td>
<td>10.00</td>
<td>N</td>
</tr>
<tr>
<td>2</td>
<td>Emily</td>
<td>5</td>
<td>300</td>
<td>10.01</td>
<td>Y</td>
</tr>
<tr>
<td>3</td>
<td>Chad</td>
<td>3</td>
<td>200</td>
<td>10.01</td>
<td>N</td>
</tr>
<tr>
<td>4</td>
<td>Brian</td>
<td>2</td>
<td>400</td>
<td>10.02</td>
<td>Y</td>
</tr>
<tr>
<td>5</td>
<td>Dana</td>
<td>4</td>
<td>100</td>
<td>10.02</td>
<td>Y</td>
</tr>
<tr>
<td>6</td>
<td>Amy</td>
<td>1</td>
<td>100</td>
<td>10.10</td>
<td>Y</td>
</tr>
<tr>
<td>7</td>
<td>Frank</td>
<td>6</td>
<td>300</td>
<td>10.10</td>
<td>N</td>
</tr>
</tbody>
</table>

Note: The offer side of the book is often displayed on a screen in reverse order: the “top” of the book is displayed on the bottom.

Answers to Problem 3.3

a. Kathy buys 200 from Alice @ 15.05; Kathy buys 100 from Dmitri @ 15.05; the market is 15.03 bid for 400 (Bruce), 100 offered at 15.07 (Esteban). Esteban’s order sets the best offer because Dmitri’s offer is hidden.

b. Lane buys 200 from Alice @ 15.05; Lane buys 300 from Dmitri @ 15.05; Lane buys 100 from Esteban @ 15.07; the market is 15.03 bid for 400 (Bruce), 500 offered at 15.07 (Gina).

c. Maureen buys 200 from Alice @ 15.05; Maureen buys 300 from Dmitri @ 15.05; Maureen has 100 shares left to buy, so the market is 15.05 bid for 100 (Maureen), 100 offered at 15.07 (Esteban).

d. Ollie sells 400 to Bruce @ 15.03; Ollie sells 100 to Hans @ 15.01; the market is 15.01 bid for 200 (Hans), 200 offered at 15.05 (Alice).

e. Petra sells 400 to Bruce @ 15.03; the market is 15.01 bid for 300 (Hans), 100 offered at 15.02 (Petra).

f. Rama sells 400 to Bruce @ 15.03. Rama’s remaining 100 shares can’t be executed, and (because of the IOC) they are canceled; the market is 15.01 bid for 300 (Hans), 200 offered at 15.05 (Alice).

g. Only 700 shares are available at 15.01 or better, so Sydney’s entire order is cancelled. The market is (still) 15.03 bid for 400 (Bruce), 200 offered at 15.05 (Alice).
Chapter 4. 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 that is sometimes called a “CLOB”.

Most present-day regulators, though, are reluctant to give one exchange a monopoly on trading. The trend, therefore, favors multiple exchanges, resulting in a market structure that is described as fragmented. 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.

Multiple markets can best be illustrated by a quick look at the US terrain.

4.1. US trading venues

Table 4.1 lists the largest venues and their trading volume (in million shares) on a typical recent day (Tuesday, January 17, 2012). Volume is reported separately for each listing venue.

<table>
<thead>
<tr>
<th>Listing venue</th>
<th>NYSE</th>
<th>NASDAQ</th>
<th>Amex</th>
<th>Arca</th>
</tr>
</thead>
<tbody>
<tr>
<td>New York (NYSE)</td>
<td>811</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chicago</td>
<td>10</td>
<td>4</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>CBOE</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>NYSE Arca</td>
<td>383</td>
<td>205</td>
<td>9</td>
<td>212</td>
</tr>
<tr>
<td>NASDAQ</td>
<td>530</td>
<td>466</td>
<td>11</td>
<td>166</td>
</tr>
<tr>
<td>NASD ADF</td>
<td>1,230</td>
<td>636</td>
<td>37</td>
<td>284</td>
</tr>
<tr>
<td>PSX</td>
<td>28</td>
<td>24</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>NYSE Amex</td>
<td></td>
<td></td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td>Boston</td>
<td>94</td>
<td>47</td>
<td>1</td>
<td>21</td>
</tr>
<tr>
<td>National Exch</td>
<td>17</td>
<td>10</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>BATS (and BATS Y)</td>
<td>386</td>
<td>222</td>
<td>7</td>
<td>153</td>
</tr>
<tr>
<td>Edge (A and X)</td>
<td>374</td>
<td>187</td>
<td>8</td>
<td>75</td>
</tr>
<tr>
<td>Composite</td>
<td>3,865</td>
<td>1,810</td>
<td>88</td>
<td>945</td>
</tr>
</tbody>
</table>

Source: Wall St. Journal Online (Market Data Center, U.S. Stocks, Markets Diary)

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, the most important facts are: 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.

For all listings, the largest entry is “NASD ADF”. “ADF” stands for “Alternative Display Facility”. This system provides a trade reporting channel for venues that aren’t exchanges. This is a catchall category that covers trades occurring by many diverse means.

By most accounts, the largest contributors to ADF-reported volume are “dark pools”. We will discuss the operations of dark pools in detail below. They are regulated, but in comparison with most markets they have low transparency. (It is difficult to observe the details of the trading process and the condition of the market. There are about fifty of them.)

“NASD” stands for the “National Association of Securities Dealers.” Formerly, this was the regulatory partner of NASDAQ. In 2007 it merged with its counterpart at the NYSE. The combined entity is the Financial Industry Regulatory Authority (FINRA): “...the largest independent regulator for all securities firms doing business in the United States ...

4.2. Trading in fragmented markets.

Traders face great challenges in navigating a fragmented market. Different markets might have different prices. Who, at the 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”. We will discuss this in some detail later, but for now we’ll just cover some key concepts and terminology.

In Reg NMS parlance, any place (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). 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. (http://www.essexradez.com/faq_data_feeds.html is a good online source of information about the various market information systems.)

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.

4.3. The NBBO

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, 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 whenever its quote changes (or is cancelled). This new record is an update that replaces all information on the previous record. 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 4.2 Example computation of the NBBO

<table>
<thead>
<tr>
<th>Quote record</th>
<th>Computations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>Bid</td>
</tr>
<tr>
<td>9:31</td>
<td>70.00</td>
</tr>
<tr>
<td>9:32</td>
<td>70.05</td>
</tr>
<tr>
<td>9:33</td>
<td>69.90</td>
</tr>
<tr>
<td>9:34</td>
<td>70.00</td>
</tr>
<tr>
<td>9:35</td>
<td>70.00</td>
</tr>
<tr>
<td>9:50</td>
<td>70.10</td>
</tr>
</tbody>
</table>

Table 4.2 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 in this determination that we are taking 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.

4.4. 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 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.
- 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.

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 buyer A is orally bidding 100,
buyer B is bidding 101, and trader C sells to A at 100. B can observe and protest. The remedy is also
usually clear: C’s sale to A stands, but C also “owes B a fill,” that is C must find shares to sell to B at
101.

4.5. Order protection Reg NMS

The policing of trade-throughs that occurs in a floor market has not proven feasible in dispersed
electronic markets. Briefly, network latencies coupled with the rapid pace of trading render it
difficult to determine the bids and offers extant at the precise instant that a trade occurred, much
less what the parties to the trade could have reasonably known.

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. We discuss below the order
protection rule of Reg NMS, one such attempt. It is a complex rule, but one worth understanding not
just for sake of compliance, but also as a window into the forces shaping today’s markets. It was the
most controversial part of Reg NMS; it occasioned many thoughtful comment letters; and was
ultimately adopted in a narrowly split vote of the SEC Commissioners. The Reg NMS trade-through
rule defines what bids and offers are protected, who has the responsibility for avoiding them, and
how this responsibility is discharged. We begin with the what and how.

The class of bids and offers considered to be protected under Reg NMS is an important but
narrowly defined set. To be protected, an order must be at the market center’s BBO, the top of the
market center’s visible book. Orders priced away from the market’s BBO and hidden orders (even if
they are superior to the BBO) are not protected. Reg NMS also stops short of forbidding trades
through these protected orders. It applies instead to the precursors of these trades, namely the
order routing decisions that are likely to cause them.

The distinction between routing and execution may not be immediately apparent. Table 4.3
describes, at a given point in time, the bid books on exchanges A and B (sh=shares):

<table>
<thead>
<tr>
<th>Bid</th>
<th>Exchange A's Book</th>
<th>Exchange B's Book</th>
</tr>
</thead>
<tbody>
<tr>
<td>$103</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$102</td>
<td>100 sh</td>
<td>100 sh</td>
</tr>
<tr>
<td>$101</td>
<td></td>
<td>300 sh</td>
</tr>
<tr>
<td>$100</td>
<td>300 sh</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.3 Example bids on two exchanges.
The bold-faced quantities mark the tops of A’s and B’s books. Based on what we can see, an order to sell 400 shares, limit $100 that is routed to A would generate executions of 100 sh @ $102 (trading through B’s bid at $103) and 300 sh at $100 (trading through B’s bid at $101). If the same order were routed to B, 100 shares would execute at $103, 100 sh @ $101, and 200 sh @ $101 (trading through A’s bid at $102).

If the seller wanted to avoid trade-throughs, she could send 100 sh to B (filling at $103), 100 sh to A (filling at $102), 100 sh to B @ $102, and then 100 more to B (filling at 101). By proceeding in this fashion, she is essentially acting as if the two books were consolidated, and her orders are walking through this consolidated book.

This strategy is sensible and perhaps even desirable, but it ignores an important market reality. Limit order books are dynamic and volatile. The latency (delay) spanning her sequential order transmissions (and the acknowledgements of the executions) is sufficiently long that her trading intentions might be completely frustrated. If she is reacting to some market development, others will be competing with her to execute against the bids in both markets. The bidders themselves, for that matter, may be racing to cancel their orders. In either case, in reaching for the visible bids, the seller's hands may close on thin air.

Latency drives a wedge between intentions and outcomes. The Reg NMS order protection rule does not apply to execution outcomes. These are not ultimately under the control of the trader sending the marketable orders or the exchange that executes them. Reg NMS applies instead to the routing decision, mandating minimum quantities that must be sent to each market.

Under Reg NMS, only the top of a market center's visible book is protected. The top of A’s bid book is $102 bid for 100 shares; the top of B’s is $103 bid for 100 shares. These dictate the minimum quantities that must be routed to the two exchanges. If the seller’s intended order is “sell 400, limit 100”, then at least 100 shares, limit $100 must be routed to A and at least $100 shares limit 100 must be routed to B. The remaining 200 shares could be routed to either destination, even if some of the resulting executions might be (in the broadest vision of the consolidated book) trade-throughs.

For example, sell 300 shares, limit $100 to A, and sell 100 shares limit $100 to B is consistent with the rule. On A, 100 shares will execute at $102, and 200 more @ $100 “trading through” B’s bids at 102 and 101. These are execution outcomes, however, after the fact. Strictly speaking, the rule does not prohibit trade-throughs. Instead, it specifies the routing precautions that must be taken to minimize the likelihood of their occurrence.

We recall that a trade-through hurts not just the resting orders that were violated, but also the person who caused the trade-through. In the present case, under the rules, the trader had discretion over where to send the remaining 200 shares. Why might she have sent them to A, when better prices were apparently available on B? The most probable answer is that her experience with the two exchanges suggests that the bids on A are more likely to actually be accessible when her order arrives, that the bids on B’s book are (in current slang) phantom liquidity.

The preceding discussion focused on defining the protected quotes and the importance of the routing decision over the execution outcomes. The discussion now turns to the question of where the responsibility lies.

In the example above, the routing decision was in the hands of the seller. Many sophisticated traders will fact exercise this control. Their orders are said to be directed. But many traders (and
smaller brokers) won’t have data systems that can identify the tops of the various market centers’ books. They simply lack the information they need to comply with the rule.

Accordingly, Reg NMS imposes the primary responsibility for making the correct routing decisions on market centers. Before a market center can execute a marketable order it must check all other market centers’ books beforehand, to ensure that the execution would not trade through a protected quote. If this would apparently occur, the market center receiving the order is supposed to re-route the order to the center posting the superior quote.

In the above example, if exchange A received an order to sell 400 sh limit $100, it would have to route 100 sh to B (with the intention of executing against B’s protected bid at $103) before executing the remaining 300 shares against its own bids. If exchange B received the order to sell 400 sh limit $100, it would have to send 100 sh to A before executing the remaining 300 sh against its own book.

It might seem to be a matter of indifference as to who actually makes the routing decision, but in a rapidly moving market the requirement that exchanges always check the tops of others’ books leads to delays that might result in worse prices for the routed orders. Fortunately, the rule allows some flexibility on this point. A sophisticated trader can flag her orders with an Intermarket Sweep Order (ISO) qualifier. This tells the receiving exchange that the sender of the order is assuming the responsibility of checking and routing to satisfy the Reg NMS requirement, and that the receiving exchange should simply execute the orders without performing additional checks.

In the above example, the seller could satisfy Reg NMS by routing 100 sh to A limit $100 and 300 sh to B limit $100. If each order were marked ISO, they would be executed on arrival. Without the ISO designation, A and B would have to check each other’s bids before executing.

Throughout this discussion, we have consistently attached a limit price to the routed orders. The significance of this becomes apparent when we explore the rule’s details. According to the definition given in SEC Rule 300 (30):

> Intermarket sweep order means a limit order for an NMS [National Market System, that is, “actively traded”] stock that meets the following requirements:

(i) When routed to a trading center, the limit order is identified as an intermarket sweep order; and

(ii) Simultaneously with the routing of the limit order identified as an intermarket sweep order, 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. These additional routed orders also must be marked as intermarket sweep orders.

Most importantly, the routing requirement in paragraph (ii) applies to protected bids or offers that are priced superior to the limit price of the sweep order.

It is therefore okay in this example to send to B an order to sell 200 sh limit $102 ISO. We expect that 100 sh will execute at $103 and 100 more at $102. Now B’s bid at $102 is not at the top of B’s book, and so is not protected. A’s bid at $102 is protected because it is at the top of A’s book. It is
protected, however, against an execution elsewhere at a price lower than $102. The execution of 100 sh on exchange \( B \) at $102 does not constitute a trade-through. Only when the limit price of the order is below $102 is there any requirement to route anything to \( A \).

It is also permissible to send to \( B \) and order to sell 10,000 sh limit $102 ISO. The seller might send a large quantity if she suspects that there are many hidden bids at or better than $102. It would probably be prudent in this case to also designate the order immediate or cancel (IOC) because otherwise the unexecuted quantity will be posted as a large visible offer, but this is not a Reg NMS requirement.

Without an IOC, the remaining 9,800 shares would go on \( B \)’s offer book at $102. The situation of one market bidding a price (remember \( A \)’s bid of $102) while another market is offering the same price is called a locked market. You’d think that the logical outcome would involve either \( A \)’s buyer or \( B \)’s seller realizing that they could achieve an execution at the other exchange and re-routing their order accordingly. This might happen, but subtleties of exchange pricing may make both parties reluctant to move. This is discussed in Chapter 16.

The Reg NMS trade-through rule is a complicated one. Here are some more examples.

Table 4.4 describes a market center’s bid book. The market center’s visible bid would be 20.04 for 500 shares. Carmen’s and Dan’s orders are protected. Billy’s and Ava’s are not visible and are not protected. Note, too, that protection against trade-through is not equivalent to a guarantee of execution. A seller using an intermarket sweep order at a limit price of 20.03 would have to send an order to this market that would execute the visible bid. “Sell 500 limit 20.04” would comply. Upon arrival, however, this order would execute against Billy (100 shares), Carmen (200 shares) and Dan (200 shares). Dan would still have 100 shares to purchase at 20.04.

<table>
<thead>
<tr>
<th>Price</th>
<th>Shares</th>
<th>Time</th>
<th>Displayed?</th>
<th>Trader</th>
</tr>
</thead>
<tbody>
<tr>
<td>20.05</td>
<td>100</td>
<td>9:46</td>
<td>No</td>
<td>Billy</td>
</tr>
<tr>
<td>20.04</td>
<td>200</td>
<td>10:01</td>
<td>Yes</td>
<td>Carmen</td>
</tr>
<tr>
<td>20.04</td>
<td>300</td>
<td>10:02</td>
<td>Yes</td>
<td>Dan</td>
</tr>
<tr>
<td>20.04</td>
<td>800</td>
<td>9:45</td>
<td>No</td>
<td>Ava</td>
</tr>
</tbody>
</table>

The sender of an ISO need only worry about protected quotes better than the limit price. Changing the limit price might therefore change the ISO requirements. Consider the example in Table 4.5. A trader can send “sell 1,000 shares limit 20.05” to any single one of the exchanges (including \( C \)) without using an ISO. The ISO obligation does not arise because there are no protected quotes better than the limit price. (The trader might send the entire order to \( C \), for example, if she believes that \( C \) has substantial undisplayed size at or above 20.05.)
### Table 4.5 The bid books at exchanges A, B, and C

<table>
<thead>
<tr>
<th>Price</th>
<th>Exchange A Shares</th>
<th>Exchange B Shares</th>
<th>Exchange C Shares</th>
</tr>
</thead>
<tbody>
<tr>
<td>20.05</td>
<td>100</td>
<td>200</td>
<td></td>
</tr>
<tr>
<td>20.04</td>
<td>200</td>
<td>500</td>
<td>200</td>
</tr>
<tr>
<td>20.03</td>
<td>300</td>
<td>600</td>
<td>300</td>
</tr>
<tr>
<td>20.03</td>
<td>800</td>
<td>700</td>
<td>200</td>
</tr>
</tbody>
</table>

If the seller’s order is “sell 1,000 shares limit 20.04,” then both A and B have protected bids better than the limit price. At least two ISOs must be sent: one (for at least 100 shares) to A; the other (for at least 200 shares) to B.

If the order is “sell 1,000 shares limit 20.03” then an ISO order for at least 200 shares must be sent to exchange C. This might seem unnecessary. Between them, exchanges A and B are bidding for 1,000 shares at prices of 20.04 or better. For example, a 300 share order sent to A and a 700 share order sent to B (both limit 20.03) would execute against the displayed books without trading through C’s best displayed bid of 20.04.

The problem is that while the ISOs are being sent to A and B, the books might change. If the 500 share bid at B is cancelled (or executed by someone else), then “sell 700 limit 20.03” will at least partially execute at 20.03, trading through C’s protected bid. This possibility lies behind the requirement, if the limit price is 20.03 or lower, to send at least 200 shares to C.

#### 4.6. 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 U.S. 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 life 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. Section 17.1 provides further discussion.
Problems

Problem 4.1 Fragmented and consolidated markets

The bid sides of the books at exchanges A and B are as follows:

<table>
<thead>
<tr>
<th>Trader</th>
<th>Shares</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conrad</td>
<td>100</td>
<td>30.05</td>
</tr>
<tr>
<td>Brenda</td>
<td>100</td>
<td>30.02</td>
</tr>
<tr>
<td>Anna</td>
<td>100</td>
<td>30.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trader</th>
<th>Shares</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doug</td>
<td>100</td>
<td>30.04</td>
</tr>
<tr>
<td>Ellen</td>
<td>100</td>
<td>30.03</td>
</tr>
<tr>
<td>Franklin</td>
<td>100</td>
<td>30.03</td>
</tr>
<tr>
<td>Gina</td>
<td>200</td>
<td>30.00</td>
</tr>
</tbody>
</table>

Now suppose that the following sequence of sell orders occurs.

i. Haley sends, “sell 200, limit 30.00” to B.
ii. Inez sends, “sell 200, limit 30.00” to A.
iii. James sends, “sell 200, limit 30.00” to B.
iv. Kathy sends, “sell 200, limit 30.00” to B.

a. If each exchange only executes the orders that it receives, and never routes the orders to the other exchange, what executions occur? What trade-throughs occur? Which orders are left unexecuted?

b. Construct the consolidated (combined) bid book. If the orders in part a had been sent to the consolidated book, what executions would have occurred?

Problem 4.2 Protected quotes

The bid sides of the books at three market centers are as follows. There are no hidden orders.

<table>
<thead>
<tr>
<th>Price</th>
<th>Exchange A shares</th>
<th>Exchange B shares</th>
<th>Exchange C shares</th>
</tr>
</thead>
<tbody>
<tr>
<td>50.49</td>
<td>300</td>
<td>200</td>
<td></td>
</tr>
<tr>
<td>50.48</td>
<td>100</td>
<td>800</td>
<td></td>
</tr>
<tr>
<td>50.47</td>
<td>100</td>
<td>300</td>
<td>400</td>
</tr>
<tr>
<td>50.46</td>
<td></td>
<td>900</td>
<td>600</td>
</tr>
<tr>
<td>50.45</td>
<td>2,000</td>
<td>2,000</td>
<td>2,000</td>
</tr>
</tbody>
</table>

a. For each exchange, what are the protected bids (prices and quantities)?

b. A trader wants to sell 1,000 shares limit 50.45 using ISOs. What is the minimum quantity that must be submitted to A? to B? to C?

c. A trader wants to sell 6,000 shares limit 10.48 using ISOs. What is the minimum quantity that must be submitted to A? to B? to C?

d. If the trader submits sell 2,000 shares limit 10.45 to each exchange using ISOs, are there any trade-throughs of displayed bids that aren’t protected.
Problem 4.3 Protected quotes with undisplayed orders
The bid sides of the books at three market centers are:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>10.49</td>
<td>100</td>
<td>900</td>
<td>800</td>
<td>2,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10.48</td>
<td>300</td>
<td>100</td>
<td>200</td>
<td>200</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10.47</td>
<td>400</td>
<td>100</td>
<td>500</td>
<td>200</td>
<td>400</td>
<td></td>
</tr>
<tr>
<td>10.46</td>
<td></td>
<td>700</td>
<td></td>
<td>1,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10.45</td>
<td>800</td>
<td></td>
<td>200</td>
<td>2,000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. For each exchange, what are the protected bids (prices and quantities)?
b. A trader simultaneously sends the following sell orders, all with limit prices of 10.46, all marked ISO: sell 800 to A, sell 2,000 to B, sell 500 to C. What executions occur?
Answers to problems

Answer to Problem 4.1

a. In the fragmented market:
   i. On Exchange B, Haley sells 100 to Doug @ 30.04 and 100 to Ellen @ 30.03. Haley’s sale to Doug trades through Conrad’s bid on (exchange) A. After these trades the books in the two markets are:

<table>
<thead>
<tr>
<th>Exchange A</th>
<th>Exchange B</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Trader</strong></td>
<td><strong>Shares</strong></td>
</tr>
<tr>
<td>Conrad</td>
<td>100</td>
</tr>
<tr>
<td>Brenda</td>
<td>100</td>
</tr>
<tr>
<td>Anna</td>
<td>100</td>
</tr>
</tbody>
</table>

   ii. On Exchange A, Inez sells 100 to Conrad @ 30.05 and 100 to Brenda @ 30.02. The sale to Brenda trades through Franklin’s bid of 30.03 on B. After these trades, the books in the two markets are:

<table>
<thead>
<tr>
<th>Exchange A</th>
<th>Exchange B</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Trader</strong></td>
<td><strong>Shares</strong></td>
</tr>
<tr>
<td>Anna</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

   iii. On Exchange B, James sells 100 to Franklin @ 30.03 and 100 to Gina @ 30.00. The sale to Gina trades through Anna’s bid of 30.01 on A. After these trades, the books in the two markets are:

<table>
<thead>
<tr>
<th>Exchange A</th>
<th>Exchange B</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Trader</strong></td>
<td><strong>Shares</strong></td>
</tr>
<tr>
<td>Anna</td>
<td>100</td>
</tr>
</tbody>
</table>

   iv. On Exchange B, Kathy sells 200 to Gina @ 30.00; she has 100 left unexecuted. The sale to Gina trades through Ann’s bid of 30.01 on A.

b. The consolidated book is

<table>
<thead>
<tr>
<th>Price</th>
<th>Shares</th>
<th>Trader (originating exchange)</th>
</tr>
</thead>
<tbody>
<tr>
<td>30.05</td>
<td>100</td>
<td>Conrad (A)</td>
</tr>
<tr>
<td>30.04</td>
<td>100</td>
<td>Doug (B)</td>
</tr>
<tr>
<td>30.03</td>
<td>100</td>
<td>Ellen (B)</td>
</tr>
<tr>
<td>30.03</td>
<td>100</td>
<td>Franklin (B)</td>
</tr>
<tr>
<td>30.02</td>
<td>100</td>
<td>Brenda (A)</td>
</tr>
<tr>
<td>30.01</td>
<td>100</td>
<td>Anna (A)</td>
</tr>
<tr>
<td>30.00</td>
<td>200</td>
<td>Gina (B)</td>
</tr>
</tbody>
</table>

Against this consolidated book:

   i. Haley sells 100 to Conrad @ 30.05, and 100 to Doug at 30.04.
   ii. Inez sells 100 to Ellen @ 30.03, and 100 to Franklin @ 30.03
   iii. James sells 100 to Brenda @ 30.02 and 100 to Anna @ 30.01
   iv. Kathy sells 200 to Gina @ 30.00

Relative to the fragmented market, everyone trades and there are no trade-throughs.
Answer to Problem 4.2

a. A’s protected bid is 300 sh @ 50.49; B’s is 200 sh @ 50.49; C’s is 400 @ 50.47. It is not necessary for a protected quote to match the market-wide best bid or offer.

b. The limit price of 50.45 is below all of the protected bids, so the trader must send 300 to A, 200 to B, and 400 to C.

c. Only A’s and B’s protected bids are better than the limit price, so the trader must send at least 300 to A and 200 to B. There is no need to send anything to C.

d. On exchanges A and C, the final executions will occur at 50.45. But B will still have 200 shares left at 50.46 (a better price).

Answer to Problem 4.3

a. Only the top of the visible book is protected. A’s protected quote is 10.49 bid for 100; B’s is 10.49 bid for 800; C’s is 10.47 bid for 200.

b. On exchange A, 800 shares @ 10.49; on B, 2,000 shares @ 10.49; on C, 500 shares @ 10.47. Based on what was displayed, the trader would have expected that the orders would take out all of the displayed interest on A and B above 10.46. In fact, due to the hidden executions, everything executed at each markets best bid. In retrospect, the executions on exchange C could have been filled at better prices on the other exchanges.

FAQ

Why do all 500 shares get sold on market C at 10.47? Wouldn’t the only protected orders be the 200 visible orders? And if that is the case then the ISO wouldn’t require the remaining 300 shares to be sold from the hidden quantity at 10.47?

And isn’t it the case that all orders that are better than the ISO order must be executed in a market before they can move to the next market? – I was under the impression that just the protected order at the top of the book needed to be filled first. Is that incorrect?

The answer to both questions arises in the distinction between routing and execution. The trade-through rule only dictates what quantities must be routed (sent) to a market. The executions those orders actually receive depends on the state of the order books at those markets. Sweep orders are not sent sequentially. They are sent simultaneously. Each receiving market handles the ISO independently of what other markets are doing.

Relative to what is on the books at A and B, the entire execution of 500 shares at 10.47 at C could be judged a trade-through. That is, both A and B are holding buy orders at prices better than 10.47. But these orders aren’t protected unless they’re visible and at the top of the book. C will execute the incoming 500 share sell order instantly: once C’s front end sees that the order is marked “ISO”, C is relieved of the responsibility of checking other markets protect bids.
Chapter 5. Auctions

To this point, most of our analysis has focused on the limit order market. Certainly this is a very widely used mechanism. There are nevertheless many securities and settings where it is augmented, or even displaced, by other mechanisms. In this and the following two chapters, we describe some of these alternatives.

This chapter focuses on auctions. Auctions generally serve to concentrate buyers and/or sellers in time. The competition that results from this broad participation in the trading process can establish prices that are viewed as fair and reliable. They are widely used in the exchange of many goods and services. (Klemperer, 2004) discusses the common formats and their economic principles.

This chapter will survey auctions in financial markets. In this context, their most important role lies in the opening and closing of continuous limit order markets.

5.1. 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 few orders and a wide spread, with no strong interest on the buy or sell side of the market.

In many stocks, though, there is strong trading interest at 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 in the case of a stock where 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. This is most easily accomplished with an auction procedure. There are many variations, but most follow the basic model known as the double-sided single-price auction.

5.2. The basic setup

The double-sided single-price auction 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).

<table>
<thead>
<tr>
<th>Trader</th>
<th>Bid</th>
<th>Quantity</th>
<th>Cumulative demand</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alan</td>
<td>Market</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Beth</td>
<td>$8</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>Cam</td>
<td>$6</td>
<td>4</td>
<td>14</td>
</tr>
<tr>
<td>Dana</td>
<td>$4</td>
<td>4</td>
<td>18</td>
</tr>
</tbody>
</table>

By way of illustration, Table 5.1 contains a sample of ranked buy orders. Figure 5.1 shows their arrangement on the price-quantity graph. The height of each bidder's rectangle is their bid price; the width is their desired quantity. The thick line connecting the tops of the rectangles is the demand “curve”. The dotted line atop Alan’s market order simply indicates that his bid price is not the $10 shown, but infinity.
Sell orders are ranked from lowest price to highest, with a market sell order implicitly priced at zero. They are arranged on the price-quantity graph in this manner. A sample set of sell orders is given in Table 5.1, and graphed in Figure 5.2.

Table 5.2. Sample sell orders

<table>
<thead>
<tr>
<th>Trader</th>
<th>Ask</th>
<th>Quantity</th>
<th>Cumulative supply</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gina</td>
<td>Market</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Hari</td>
<td>4</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>Ilse</td>
<td>6</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>Jon</td>
<td>10</td>
<td>4</td>
<td>16</td>
</tr>
</tbody>
</table>

Figure 5.2 Construction of the supply curve
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. In the basic introductory economics version the demand and supply are smooth upward and downward sloping curves and the intersection consists of a single point. With discrete prices and quantities, though, these curves are actually staircase (step) functions.

Figure 5.3 depicts the most commonly encountered situation. Supply and demand cross at a single price ($6), but overlap in quantities. The two functions overlap at a price of 6. Quantity is maximized at 12. At the clearing price, all bidders at higher prices (Alan and Beth) get completely filled. All sellers at lower prices (Gina and Hari) get completely filled. Once this happens, Ilse has two shares left to sell. These go to Cam, who gets a partial fill (he originally wanted four shares).

Figure 5.3

The best unexecuted orders remaining after the clearing are Cam’s (bidding 6 for 2 shares) and Jon’s (asking 10 for 4 shares). If the auction is being used to open the market, these orders are added to the book, and establish the initial best bid and offer.

Normally the limit order book is integrated with the auction process. Orders in the book prior to the auction are incorporated in the auction, and orders unexecuted in the clearing are added to the book. Orders can be qualified, however, with indications that they should not be considered for the auction, or alternatively that they should be considered only for the auction (in which case unexecuted portions are canceled).

The discreteness of prices and quantities can cause complications. If Ilse’s asking price had been $7 (instead of $6), we’d have the situation indicated in Figure 5.4. The market-clearing quantity is 10 shares, but now there is some wiggle-room in prices. Beth is willing to buy at $8 or lower; Ilse is willing to sell at $7 or higher. Any price in this range will clear the market. There is a good case for simply taking the midpoint, $7.50 (“splitting the difference”). But there is also something to be said for picking the price closest to some previous price (like yesterday’s close).
In the idealized setting, when participants submit their true demand and supply functions, the double-sided auction yields an outcome that possesses a number of desirable properties. For one thing, the “right” people trade. After the result is announced, no remaining buyer will come forth and say, “I would have been willing to pay a higher price,” and no excluded seller will claim that she would have accepted a lower price.

The auction outcome also yields the highest overall welfare in the following sense. A person who bids $8 for a unit, and pays $6 in the auction has a **surplus** of $2. A seller who offers at $4 and receives $6 also enjoys a surplus of $2. The auction outcome maximizes the total surplus of all the buyers and sellers. No well-intentioned social planner or agency could do any better.

Auctions are familiar. They are used in many real-world non-financial settings. They appear to be robust and easy to run. For these reasons, we'd expect that they would enjoy widespread use in financial settings. They are indeed used to open and close continuous trading sessions, and Section 5.7 describes some other applications.

But the apparent simplicity of auctions is deceptive. Auction outcomes often depend on easy-to-overlook details of design and implementation. Klemperer gives some examples from cellular telephone spectrum auctions that embarrassed participants and designers ((Klemperer, 2002)).

So when we turn to how securities auctions are actually implemented, we encounter complexities. Supplementary rules and procedures often seen necessary to ensure that auction outcomes are sensible and stable.

**5.3. Clearing time**

When should we clear the market? The most straightforward approach is to set a fixed time, announced in advance. In many trading situations, though, a hard deadline simply induces traders to hold off entering their orders until shortly before the deadline.
To counter the deadline pile-up, one online US stock auction system experimented with a timed commission schedule. Orders that were entered early paid lower commissions (Wunsch Auction Systems, operated in the early 1990's).

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, Alan entered a market buy order for four shares. When the market cleared, he (and everyone else who bought or sold) traded at $6.

Now suppose instead that when the order entry period opens, Alan submits two orders: his intended four-share market order, and a second order for ten shares priced at $9. At this point, the bid side of the market would appear to be very strong (Figure 5.5). With original set of sell orders, the market would clear at $9. Given this, it is quite likely that the other bidders (Beth, Cam, and Dana) might believe that their (lower) contemplated bids are unlikely to be filled. They might simply decide that bidding is not worth the trouble, leaving Alan’s two orders standing as the only bids in the market.

Suppose, with the sell orders unchanged from the original example, Alan cancels his second order just before the market clears. This leaves the market shown in Figure 5.6. Alan buys his shares at $4 (instead of $6), to the dismay of the other potential buyers (who would have bid higher).

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 this sort of behavior, a market may randomize the clearing time. Even a small amount of uncertainty may be enough to discourage manipulations based on last-instant moves. (Would a bidder let the $10,000 tea set bid stand if there were a chance that it might actually win the auction?)

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); Euronext Amsterdam (Instinet, 2010a); Euronext Lisbon (Instinet, 2010b); and Xetra (Deutsche Börse Group, 2011).
In the U.S., however, neither the NYSE nor NASDAQ use randomized clearing times. Instead, they attempt to stabilize the price determination process by restricting the information available to auction participants and also by restricting order entry and modification immediately prior to the clearing.

The procedures for the NYSE and NASDAQ are documented on their websites and by (Bacidore, Berkow, and Wong, 2012). We consider the NYSE first.

5.4. The NYSE opening and closing auctions

The clearing occurs shortly after 9:30am. The NYSE designated market maker (DMM, described in the next chapter) oversees the process, and has the right (along with certain obligations) to participate in the auction.

An order entry phase starts at 7:30am. Prior to the open, the system publishes imbalance information. This information consists of matched volume (the largest feasible quantity) and any excess supply or demand. The price at which these quantities are computed, however, is not the crossing price in the sense of the last section. It is instead, a reference price, which is usually set to the previous day’s closing price. That is, if we were to take yesterday’s closing price as a starting point, would there be an excess of buyers or sellers?

The frequency of imbalance disclosure increases as the open draws closer: every five minutes between 8:30 and 9:00; every one minute between 9:00 and 9:20; every fifteen seconds between 9:20 and 9:35 (or the opening, if the opening occurs prior to 9:35).

Beginning at 9:28, the system also disseminates an indicative equilibrium (crossing) price. Shortly after 9:30, the DMM opens the stock, usually taking the other side of any imbalance at the opening price.

The closing auction has additional requirements. Market and limit on-close orders are only accepted until 3:45pm. After that, only orders in the direction that would offset an imbalance may be entered. On-close orders cannot be canceled after 3:45pm. Indicative clearing prices, and imbalances are disseminated every five seconds. At 4:00pm, the auction clears, under the control of, and with the participation of, the DMM. The crossing also includes eligible orders from the regular continuous limit order book.

5.5. The NASDAQ opening and closing auctions (“crosses”)

The NASDAQ opening cross is distinctive in that it operates concurrently with NASDAQ’s regular continuous limit order book and also that it disseminates minimal imbalance information.

The order entry phase begins at 7:00am and the cross occurs at 9:30. An order sent to the continuous book can be marked as eligible (for inclusion in the opening cross) or not. An order can be designated for opening cross only. In addition to the usual market or limit orders, the opening cross also uses imbalance only (IO) orders. These are only executed to redress imbalances. For example, in a situation like that depicted in Figure 5.3, an IO sell order would trade against Cam’s excess demand at the clearing price, while an IO buy order would not be executed.

Up until 9:28, orders may be submitted, modified or canceled. After 9:28 no orders may be modified or cancelled, and the only orders that may be submitted are IO orders. The system publishes no imbalance information until 9:28 (at which point extensive imbalance information is available. At
9:28 the system begins to publish extensive imbalance information. At 9:30, eligible orders in the continuous book are “swept” into the opening auction and an equilibrium price is determined.

The closing cross proceeds in much the same way. “On close” market and limit orders are accepted until 15:50. After 15:50, imbalances are published and the only orders that are accepted are IO orders.

During the continuous trading sessions, the U.S. equity market is very fragmented, with trades occurring across many different market centers (see Chapter 2). Auctions are different. They can be, and are, run independently in different centers (arriving at different prices). NASDAQ, for example, runs opening auctions in NYSE-listed stocks. The primary listing exchange, however, generally holds the dominant position. Most of the opening volume in NYSE-listed stocks occurs in the NYSE opening auction; most of the opening NASDAQ volume occurs in the NASDAQ opening cross. Auctions are also used to re-open continuous trading after a halt. Here, too, the primary listing exchange is dominant.

5.6. Use of auctions for illiquid securities

During regular trading hours, virtually all stocks in the US trade in a continuous market. Bids and offers may be placed, and matches may occur at any time.

There are strong 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.

But if trading interest in the stock is low, traders might not bother to participate in a continuous market. Why go to the trouble of entering, monitoring, and modifying a bid (as market conditions change), and exposing it to being picked off by low-latency traders, if it is unlikely that it will be hit by someone we might actually want to trade with?

To have any volume, a market needs to attract marketable orders. To do this, it needs reasonably priced limit orders. But traders won’t bother to post limit orders unless there’s a good flow of marketable orders. Below some critical mass, the book will simply be empty. There’s no trade and the market doesn’t get any trading revenue.

Auctions concentrate buying and selling interest: everyone comes together at the same time. The increased participation can support more trade and better prices. For these reasons, many markets use periodic call auctions for illiquid stocks, and don’t allow the stocks to trade via the continuous mechanisms.

The Euronext markets (Paris, Brussels, Amsterdam and others) run call auctions twice a day in illiquid securities (usually at the regular market open and close). These securities do not trade in the continuous market.
5.7. Other uses of auctions

This chapter has focused on the use of auctions in secondary markets, but auctions are often the first choice for other kinds of trades.

5.7.A. Primary market auctions of US Treasury securities

The Federal Reserve Bank of New York conducts auctions to sell newly-issued US government securities. The current format is a sealed-bid single-price auction. Bidders specify quantities and yields. Yields are equivalent to prices, but are easier to interpret in this market. Buyers of small quantities (including retail investors) can submit unpriced bids. These bids are like market orders in that the yield the buyer receives is set by the competitive bids.

5.7.B. Initial public offerings of corporate equity

Equity IPOs in the US are traditionally conducted via a book-building mechanism. The investment bankers handling the deal survey their customers, assess the demand for the stock, and fix an issuing price. The process is opaque: only the bankers see the customer bids. When the issue becomes available for trading in the regular market, there are often large price changes relative to issuing price. The most common pattern on the opening day is an increase (the opening day "pop"), but there are instances, including the recent Facebook IPO, where stock closed the opening day lower than the issue price.

For these and other reasons, some observers have argued for wider use of single-price auctions. There has been a steady but modest flow of auction IPOs, mostly for smaller firms. But even after the Google auction established the viability of the mechanism, traditional book building continues to dominate.

5.7.C. High-frequency auctions

The continuous limit order book rewards speed. Traders compete to be the first to act on new information. The rewards to the winners here are sufficiently large that substantial resources are spent gaining advantages of mere milliseconds. Some view trading at this time scale as providing few benefits to overall social welfare, while encouraging an expensive arms race among technologically adept participants.

An auction with no time priority concentrates buying and selling interest at a single point in time, and nullifies the reward for being first. It has been suggested that one possible way of keeping this property and also maintaining the essential advantages of continuous trading is to organize the market as a sequence of auctions with minute (or even sub-minute) periodicity ((Budish, Cramton, and Shim, 2013; Schwartz and Wu, 2013))

The Taiwan Stock Exchange formerly used high-frequency call auctions, but is now a continuous price-time priority limit order market.

5.7.D. CDS settlement auctions

Credit default swaps are effectively insurance policies that pay off when the reference firm defaults. The amount of the payoff is determined by the market price of the reference firm’s bonds: the lower the market price of the bond, the higher the CDS payoff. Typically, however, the notional amount of the CDS contracts outstanding is much larger than the total value of the outstanding bonds. Conceivably, a small purchase in the bond market could boost the price enough to dramatically
lower the CDS payoff. In response to this concern, the market has adopted a two-stage auction procedure that clears the market for CDSs and the bonds simultaneously (Helwege, Maurer, Sarkar, and Wang, 2009; Gupta and Sundaram, 2011; Chernov, Gorbenko, and Makarov, 2012).

Problems

**Problem 5.1 Clearing**

In an NYSE opening auction, the system receives the following on-open orders:

<table>
<thead>
<tr>
<th>Trader</th>
<th>Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dave</td>
<td>Buy 7 shares, market</td>
</tr>
<tr>
<td>Amy</td>
<td>Buy 10 shares, limit 20.00</td>
</tr>
<tr>
<td>Cathy</td>
<td>Buy 30 shares, limit 19.95</td>
</tr>
<tr>
<td>Brian</td>
<td>Buy 5 shares, limit 19.90</td>
</tr>
<tr>
<td>Ira</td>
<td>Sell 8 shares, market</td>
</tr>
<tr>
<td>Gabe</td>
<td>Sell 7 shares, limit 19.90</td>
</tr>
<tr>
<td>Haley</td>
<td>Sell 10 shares, limit 19.95</td>
</tr>
<tr>
<td>Fiona</td>
<td>Sell 30 shares, limit 20.10</td>
</tr>
</tbody>
</table>

a. Construct the supply and demand schedules. At a reference price of 20.08, what is the matched volume? What is the imbalance?

b. What is the indicative clearing price? What are the matched volume and imbalance at the clearing price?
Answer to Problem 5.1.

The cumulative demand and supply schedules constructed from these orders are:

<table>
<thead>
<tr>
<th>Demand:</th>
<th></th>
<th></th>
<th>Supply:</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Buyer</td>
<td>Price</td>
<td>Quantity</td>
<td>Cum</td>
<td>Buyer</td>
<td>Price</td>
</tr>
<tr>
<td>Dave</td>
<td>Mkt</td>
<td>7</td>
<td>7</td>
<td>Ira</td>
<td>Mkt</td>
</tr>
<tr>
<td>Amy</td>
<td>$20.00</td>
<td>10</td>
<td>17</td>
<td>Gabe</td>
<td>$19.90</td>
</tr>
<tr>
<td>Cathy</td>
<td>$19.95</td>
<td>30</td>
<td>47</td>
<td>Haley</td>
<td>$19.95</td>
</tr>
<tr>
<td>Brian</td>
<td>$19.90</td>
<td>5</td>
<td>52</td>
<td>Fiona</td>
<td>$20.10</td>
</tr>
</tbody>
</table>

They are graphed as:

\[\text{Graph showing demand and supply schedules.}\]

\(a\). At a reference price of 20.08, the matched quantity is 7. This is the minimum of 7 (the demand) and 8 (the supply). There is a net sell imbalance of \(25 - 7 = 18\).  

\(b\). The indicative clearing price is 19.95, with a matched volume of 25. This is the minimum of 47 (the demand) and 25 (the supply). There is a net buy imbalance of \(47 - 25 = 22\).
Chapter 6. Dealers and dealer 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.

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 U.S. 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?

Once we move beyond equities, we encounter markets where dealers are again central. The markets for foreign exchange (FX), so-called over-the-counter derivatives (swaps and off-exchange options), corporate bonds, and so forth, are dominated by their dealer networks. These markets have not for the most part experienced the IT revolution that has reshaped the exchange markets. Are they likely to be transformed, or will they be more resistant?

6.1. 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 selling when the customer wants to buy. 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 sometimes 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 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, 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 as high a price 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?

The process of simultaneously representing customer orders and trading on one’s own behalf is
called dual trading. In all markets it is highly regulated. But we’ll shortly see that there are a lot of
gray areas.

6.2. 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 is that the beneficiaries of the DMM are the shareholders of the
listed companies, and they in turn (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). By reason of this moral hazard problem, U.S.
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.

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 U.S. exchange specialists and NASDAQ market makers.
6.3. Designated market makers past and present

6.3.A. Exchange Specialists

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. Around 2005, the specialists were relabeled designated market makers. They still play a role, but their importance and centrality to the overall market is greatly diminished. Our discussion will concentrate on their role as defined over the last half of the 20th century, arguably their era of greatest importance.

The specialist was an independent trader, buying and selling on his own account (not an Exchange employee). During this period, 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 could truly be characterized as 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. Certain of these appear to be particularly relevant for current markets:

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, on all market makers.

Avoiding “destabilizing trades.” The specialist was not supposed to trade actively, that is, by hitting/taking public limit orders.

Public priority. A specialist was supposed to yield to customer orders 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.

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 completion against his own. As a result, the specialists’ trades did not generally constitute the preponderance of trading activity.

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 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, $49.78. (Until roughly the end of the 20th century, the tick size in U.S. equity markets was $\frac{1}{8} = \$0.125.)

6.3.B. 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 NASDAQ market-makers agreed to yield to public orders, display them, and make them available for execution.

6.3.C. 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). Reg NMS mandates trade-through protection of bids and offers. This is subject to certain restrictions. As noted above, it only applies to the top of the 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.

Beginning in 2007, the US entered a financial crisis in which one major investment firm (Lehmann) failed, others were merged into stronger partners (sometimes under regulatory pressure), and government guarantees were deemed necessary to restore confidence in the financial system as a whole. 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:

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? (U.S. Securities and Exchange Commission, 2010).

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. Commodities 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 a traditional exchange specialist would have been expected to maintain a bid and an offer, their successors did not. While a specialist would have been forbidden 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.

6.4. Dealer markets

The last section discussed dealers in the context of a limit order market. In this capacity, they are supplemental. They post bids and offers, and may even possess technological advantages over the customers in doing so. The structure of the limit order market, though, still provides opportunities for direct customer trade.

The dealer can also be the defining feature of a market. In a dealer market (also called a quote driven market), the dealer network is the sole mechanism. The important dealer markets include: foreign exchange (FX), over-the-counter derivatives, swaps, government bonds, and corporate bonds. Measured by notional trading volume, these markets are substantially larger than the equity markets.

As an example of a trade in a dealer market, a hedge fund wishing to buy Euros may contact the FX desks (of large banks) and ask for a two-sided market: the dealer's bid and ask quotes. The quotes are often oral, and good only at the time they are made. The quotes are “take it or leave it”. The dealer/customer arrangement is sustained by reputation: The dealer will always make a market. The customer must (sometimes) trade.

Dealers are linked by computers and telecommunication, but there is little centralized coordination. Customers have access only to dealers with whom they've previously established a relationship. A dealer is always the counterparty to a customer trade.
Dealers will usually 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 pre-existing 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 reports trades in foreign exchange or US government bonds.

A dealer making a market (even against random, uninformed customer orders) will over time build up large long or short positions. Dealers lay off positions (against other dealers) in an interdealer market. Only dealers can directly participate in the interdealer market. Prices are generally not visible to non-dealer customers. The arrangement is sometimes described as an “inner” and “outer” market. Figure 6.1 shows the connections that might exist in such an arrangement.

Several mechanisms facilitate trade in the interdealer market. 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 Bank Y will be making the call. The 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.

Customers can often obtain access to web-based systems that have the look and feel of an order book market. The screens will show bids and offers, and it’s possible to click and obtain an automatic execution. These systems do not, however, constitute a window on the interdealer market or any other comprehensive activity. They are simply platforms where the customers of a particular dealer can execute against the quotes posted by that dealer. These quotes typically track those in the interdealer market, but are usually outside of (and therefore inferior to) them.

We encountered above the dealer’s conflict of interest: acting on the customer’s behalf while simultaneously trading against the customer. When the dealer is supplemental to a limit order market, this conflict of interest is managed by exchange rules (like public priority). In the typical dealer market, however, regulation is more remote and less binding. Instead, the conflict is managed (at least partially) by assigning the broker and dealer functions to different individuals. One person, typically called a sales trader, is responsible for maintaining a relationship with the customer and serving as the point of entry for the customer’s orders. Another person, the position trader, functions as the dealer.
6.5. Dealers' quoting strategies

How does a dealer decide how to set her bid and offer prices? The answer turns out to be important, not just for dealers, but for anyone facing a make-or-take decision. If we've decided to place a limit order, for example, we need to determine our limit price, that is, our bid or offer. In starting to think about this, it is natural to consider the strategies of those traders who are most experienced in placing bids and offers because it is a crucial decision in their business operations, the dealers.

To sustain his reputation, the dealer must always post a bid and offer, offering liquidity to incoming customers. Furthermore, ignoring the interdealer market for the moment, the dealer does not have the option of taking (hitting a customer's bid or lifting a customer's offer). The opportunity simply does not arise. Customers contact the dealer to solicit a bid and offer. The dealer never calls the customer for this purpose. This means that the dealer can only trade passively, at his bid or offer.

We will also simplify matters in one other respect. We'll assume that the dealer doesn't face much competition. The dealer's customers are captive. They have established relationships with the dealer and cannot easily and quickly set up accounts at other dealers. A dealer in this position is said to possess market power and sometimes described as monopolistic. (Remember, though, that a traditional monopoly involves one seller and many buyers who have no alternative. A monopolistic dealer is also a monopsonist, a sole buyer.)

The dealer’s available actions involve setting her bid price at a given time $t$, $B_t$, and her ask price $A_t$. These two prices are the only things under her direct control. How should the dealer set $B_t$ and $A_t$ to maximize her trading profits?

Ideally, she would like buyers and sellers who don’t care about their trade prices, so she could set her bid very low and her ask very high. She’d also hope that sellers and buyers would alternate.
First a seller would arrive, hitting her (low) bid; then a buyer would arrive, lifting her (high) ask. She’d pocket the difference between the bid and ask (the spread, her “turn”). Then the next pair, seller and buyer, would arrive, and so on, generating an endless stream of profits.

The first problem here is that in fact buyers and sellers do care about the price. Even a monopolist has to accept the fact that a potential buyer might find the asking price unacceptable. When the dealer lowers her bid price, sellers get discouraged and fewer of them will actually trade. Similarly, a high asking price discourages buyers. The trade-off she faces can be summarized as follows:

- A wide bid-ask spread ensures that the dealer’s turn, the profit on each arriving pair of buyers and sellers, is large. But there will be fewer such pairs. The wide spread discourages them and they will arrive less frequently.
- A narrow bid-ask spread leads to a smaller profit on each pair, but there will be more pairs.

The dealer sets the bid-ask spread to maximize the average rate of trading revenue. For example, let’s assume that all customers want trade a quantity of one unit (share or contract). If a setting a spread of $0.50 “per share” leads to an average customer arrival rate of two per hour (that is, two buyers and two sellers per hour), the average hourly revenue is $0.50 \times 2 = $1. If setting a spread of $0.10 per share causes this arrival rate to rise to twenty per hour, the average hourly revenue is $0.10 \times 20 = $2. Between these two possibilities, the narrow $0.10 spread is the better choice, but it might not be the best choice. The dealer can calibrate the parameters by experimenting with various spreads.

<table>
<thead>
<tr>
<th>Spread</th>
<th>Arrival rate of buyers (per hour, also the arrival rate of sellers)</th>
<th>Hourly revenue</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0.50</td>
<td>2</td>
<td>$1.00</td>
</tr>
<tr>
<td>$0.10</td>
<td>20</td>
<td>$2.00</td>
</tr>
</tbody>
</table>

A second problem is that even if the average rate of buyers and sellers is equal, any given trading day might well exhibit long runs dominated by buyers or sellers, leading to large changes in the dealer’s position. After a morning dominated by sellers, the dealer will have accumulated large holdings (a long position) in the security. These large positions are costly because they must be financed (usually by borrowing on margin). They also represent significant risk. If there is a large market swing driven by some fundamental information, there will be a corresponding change in the value of the position (possibly a large profit, but possibly a large loss). For this reason, dealers manage their positions closely, keeping their net holdings close to zero.

This is particularly important when the trading day is drawing to a close and any remaining positions must be held overnight (or over a weekend). While overnights and weekends are less information-intensive that regular trading days, they can still experience large events. The classic trader’s expression is, “Go home flat.”

Now remember that the final position can’t be closed out by simply hitting someone else’s bid or offer. We’re ignoring the interdealer market, and the dealer can’t solicit bids and offers from the customers. Instead, the position must be reduced indirectly and passively, that is, by setting the bid and offer asymmetrically to encourage an order flow imbalance in the desired direction.

We can do this because the bid only affects the arrival rate of sellers (buyers don’t look at the bid), while the offer only affects the arrival rate of buyers. Suppose that the fundamental value of the
security is $10. The bid and offer are placed relative to this value, and the dependence of the arrival rates is given in the following table.

<table>
<thead>
<tr>
<th>Bid</th>
<th>Arrival rate of sellers</th>
<th>Offer</th>
<th>Arrival rate of buyers</th>
</tr>
</thead>
<tbody>
<tr>
<td>$9.75</td>
<td>2</td>
<td>$10.25</td>
<td>2</td>
</tr>
<tr>
<td>$9.95</td>
<td>20</td>
<td>$10.05</td>
<td>20</td>
</tr>
</tbody>
</table>

These values are consistent with those used earlier. That is, if we set the bid and offer at $9.75 and $10.25 respectively, the spread is $0.50 and the rates of arrival (for both buyers and sellers) are 2 per hour; if the bid and offer are $9.95 and $10.05, the spread is $0.10 and the arrival rates are 20. In either of these cases we expect that order flow will be on average balanced.

But if we set the quotes asymmetrically about the fundamental value, we can encourage an imbalance. For example, if we set the bid at $9.75 and the offer at $10.05, we expect that more buyers will arrive (lifting our offer) than sellers. In fact, we expect an excess buyer arrival rate of $20 - 2 = 18$ sellers per hour. If we had accumulated a long position of 9 units, setting our bid and offer in this fashion would enable us to get flat (back to zero) in thirty minutes (on average).

To take an example in the opposite direction, if we were short 18 units, setting the bid to $9.95 and the offer to $10.25 would get us back to zero in one hour (on average).

In this discussion, we've treated expected revenue maximization and position management as separate strategies. This is the case if we set our quotes to maximize revenue, and then whenever we hit our position limit, we switch over to a position management strategy until we get back to zero.

It is usually better to consider both objectives simultaneously, focusing more on revenue maximization when our position is near zero and stressing position management when we're near our limits. Amihud and Mendelson explore these sorts of combined strategies (Amihud and Mendelson, 1980).

Models of trader behavior that feature the use of asymmetrically placed bids and offers to manage positions are called inventory control models. Originally developed for dealer markets, they are now also used to analyze trading strategies in limit order markets (Obizhaeva and Wang, 2005).

### 6.6. Market segments

**Government bonds**

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. U.S. Treasuries trading is concentrated on two systems eSpeed and BrokerTec. (eSpeed is presently a subsidiary of NASDAQ.)

**Corporate bonds**

In fundamental structure the US corporate bond market is a dealer market. Nevertheless, under SEC pressure, the industry has adopted last-sale reporting. All bond trades must be reported within...
fifteen minutes 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).

*Foreign Exchange (FX)*

The dealer network is dominated by the FX desks of major banks. In addition to spot currencies (for immediate delivery) they also make markets forward (for delivery at a mutually-agreed time in the future). As noted earlier, many of these dealers maintain web sites that offer customers immediate executions at the dealer’s bid and offer.

The largest interdealer FX markets are run by Reuters and EBS (for “Electronic Brokerage Services,” now owned by ICAP). Historically these venues were open only the major banks. Presently, customers may (for a fee) obtain sponsored access. (That is, their credit is guaranteed to a pre-set limit by a participating bank.)
Chapter 7. Dark liquidity and dark mechanisms

“Darkness” generally refers to pre-trade quote transparency. In US equity market data, we infer the occurrence of a dark trade when:

- An execution is reported (“prints”) at a price inside of the National Best Bid and Offer.
- An execution prints at the NBB or the NBO, but it occurs at a market center whose quote is inferior to the NBBO.
- An execution is reported on NASD’s ADF (Alternate Display Facility).

Note that a “dark trade” is not an unreported trade. The darkness reflects the fact that neither the buyer nor the seller is posting a visible bid or ask. Post-trade, the execution is reported like any other.

Dark executions typically arise from one of the following mechanisms:

- A hidden (undisplayed) limit order in a limit order market (like NASDAQ’s ISLAND) that also handles visible orders.
- A NASDAQ market maker trades against a customer order at the NBBO at a time when the MM’s own quotes are behind the NBBO.
- The trade occurs in a crossing network or dark pool that posts no quotes of its own, but matches buyers and sellers at prices determined by the NBBO.

The second and third of these mechanisms probably account for much greater trading volume than the first. They also differ from the first in one other important respect. Whereas the hidden limit order is executed at its own limit price, the MM and the crossing network/dark pool trades occur at prices determined by the NBBO. This reliance will turn out to be problematic.

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

7.1. The logic of dark mechanisms.

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 buyer is willing to pay, say, $10, why not advertise the bid to attract sellers?

It’s 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 of these 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.

7.2. Undisplayed 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 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.

7.3. 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 occur 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, and the print will probably occur on the ADF. 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. (The use of preference as a verb violates the rules of grammar, but it is nevertheless a well-established industry practice.)

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.

7.4. Crossing networks and dark pools

As with preferenced and internalized orders, this class of mechanisms uses prices set by the NBBO. Point-in-time crossing networks came first. Many of them are still around, but some of them evolved into continuous crosses, and ultimately 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).

Here are the particulars.

*Point-in-time crossing networks*

Orders are submitted anonymously. The orders communicate direction (buy or sell) and quantity (or sometimes a quantity range), but not prices. They are held in the system undisplayed. Buyers and sellers are paired up. Trade occurs at some externally determined reference price, typically the prevailing NBBO midpoint, the day’s closing price or the day’s value-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 11 daily point-in-time crosses in US equities, and is also active in Europe and Asia; ITG Posit runs intraday NBBO midpoint crosses; The NYSE MatchPoint system runs a closing cross.

Point-in-time matches are used by institutional investors who seek to trade in size. Reported trade sizes are large.

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.

*Dark pools and continuous crossing networks*

Orders communicate direction, quantity and limit price. Orders aren’t displayed. Crossings are not scheduled: a trade occurs when a marketable order arrives. In a regular limit order market, the execution would be priced off of the standing order. That is, if “bidding 50 for 200 shares” is resting in the book, an incoming order of “sell 200 shares, limit 45” executes at 50. In the dark pool, however, the orders are matched at 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).

Dark-pool order confidentiality is difficult to verify, and some users harbor doubts (Domowitz, Finkelsteyn, and Yegerman, 2009).

In one case, a dark pool turned out to have ties to a proprietary market-making operation. The SEC file (at http://www.sec.gov/litigation/admin/2011/33-9271.pdf) explains:

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”].

Pipeline advertised that it had “no prop[rietary trading] desk gaming [customer] orders.” Its advertising and other public statements repeatedly claimed that the trading opportunities on the ATS were “natural,” that the ATS would not reveal the side (i.e., whether an order was to buy or to sell) or price of a customer order before a trade was completed, that the ATS denied “arbitrageurs” and “high-frequency traders” information needed to “front run,” that it provided a refuge from “predators,” and that it prevented “pre-trade information leakage.”

These claims were false and misleading in that the Affiliate was on the other side of the vast majority of trades executed on the ATS. The Affiliate sought to predict the side and price of Pipeline customers’ orders and then trade on the same side as those orders in other trading venues before filling them on the ATS.

[Pipeline] assured its customers that any affiliate trading on the ATS would receive no preferential treatment and would “have no access to order or trade data of other subscribers.” In fact, Pipeline occasionally revealed to the Affiliate, after the trades were consummated, order and trade data of other customers. Pipeline also provided the Affiliate with certain access and information that improved the Affiliate’s ability to trade advantageously, including certain electronic connectivity to the ATS, input into designing ATS rules for deterring predatory conduct, and information about ATS operations and Pipeline’s methods of policing for predatory behavior that were not known to other customers.

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.

7.5. 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 bid 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).
After extensive debate Canadian regulators adopted restrictive measures on dark trading. Generally, large dark executions (above $100,000) are permitted at or within the NBBO. Smaller orders can be executed in the dark only if there is price improvement relative to the visible bid or offer. For example, if the Toronto Stock Exchange bid in Toronto Dominion Bank (TD) is 82.00, a retail sell order could not execute below 82.01. That is, unlike in the U.S., it is not permissible for a broker dealer to simply match the 82.00 bid.
Chapter 8. Public information and trading halts

This and the next two chapters deal with price formation: the fundamental factors bearing on security value, their interaction with the trading mechanism, and valuation effects that arise from the trading process. As a rough approximation, the important distinctions are those of scale in time and size. Fundamental factors are related to the returns on the real assets underlying the securities, like revenues and earnings. They are large and long-term, often playing out over years.

The role of these fundamentals is usually summarized by a valuation model that relates them to an idealized or hypothetical intrinsic value. These models are usually specific to the security and won't be discussed here in any detail. The trading process does not directly affect the intrinsic value, but as information about the fundamentals is generated and revised, the trading process often plays a major role in transmitting changes in the intrinsic value to the market price.

Other factors, though, seem more closely determined by the trading process. Who holds the security? Who actively trades it? How is trading conducted? These considerations can give rise to other price effects that are short term, on the order of minutes or days. They are reflected, for example, in bid-ask spreads.

The discussion starts in this chapter with public information and the classic statement of market efficiency. It will become apparent that most markets can handle information consisting of "small" data, even when the individual facts accumulate in a one-sided fashion to cause large price changes. Most mechanisms can also handle large information developments with scheduled announcement timings (like earnings results). Unexpected and sudden revelations, however, can render the trading process chaotic. The discussion therefore turns to the logic and implementation of trading halts, and the arguments for and against. Subsequent chapters consider the determinants of bid and ask prices, and the role of private and "inside" information.

8.1. Market efficiency

Investors form beliefs about securities' values largely on the basis of public or "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").

Can all security price changes be attributed to value-relevant information? Of course not. The burgeoning academic field of behavioral finance has cataloged more than a few lapses from pure rationality. For example, Hirshleifer and Shumway note: "Psychological evidence and casual intuition predict that sunny weather is associated with upbeat mood. ... Sunshine is strongly significantly correlated with stock returns ... These findings are difficult to reconcile with fully rational price setting" (Hirshleifer and Shumway, 2003). Warren Buffet, one of the most successful
investors of the modern era, has been quoted: "I would be a bum on the street with a tin cup if markets were efficient."

Nevertheless, while we allow for effects of exuberance, animal spirits and similar effects on security prices, information still stands as the lead influence.

The economic force driving market efficiency is competition among investors to accurately assess and interpret the information available to them. Even if we agree with Warren Buffet and reject the pure form of the hypothesis, we can certainly observe this competitive process at work.

To understand the dynamics (changes over time) in security prices, it is useful to consider several sorts of public information and their effects on the market. The significance of a news event will lie somewhere on a scale between very small and very large, so we’ll start by looking at these two extremes.

8.2. Small information arrivals

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 extremely liquid. The bid-ask spread is generally $0.01; sizes at the bid and ask are large. Trading volumes are large.

There are other ways to invest in the index. For individual investors, an index mutual fund is a good choice. The investments and divestitures can only be made, however, at daily close prices. (They can’t be traded within the day.) Index futures contracts are another alternative, but the minimum size is too large for many small investors, and it is somewhat more difficult to set up a brokerage account to trade futures contracts (relative to stock-only brokerage accounts).

"SPY" is one of a family of trusts called Standard and Poors Depositary Receipts (abbreviated SPDR, and pronounced “spider”). Related SPDR ETFs mirror extended market indices and specific components of the index (sector SPDRs).

Figure 8.1. 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).
As a broad index, SPY is driven by information that affects the macro economy, such as commodity prices, interest rates and exchange rates. Much of the time, this information consists of frequently-arriving news. Each individual item is of minor importance, but over time the accumulations can become large enough to notice, and to move prices.

Figure 2.1 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). This behavior is what very much in line with the accumulation-of-small-information principle.

The price plot also possesses several other distinctive features.

- There are periods when the price seems to bounce around some constant value, without moving too far away.
- Some periods seem to exhibit momentum. That is, we get a long run of price drops or price increases.
- Large price changes are often followed by a "bounce back," a "correction" in the opposite direction.

Copyright 2014, Joel Hasbrouck. All rights reserved. Draft of May 19, 2014.
In all of these instances, there is a local low price, at which we might hypothetically have bought, and a nearby high price at which we might have sold (or sold short). These trades would have been profitable. Unfortunately, we're discerning these patterns after the fact. It is well-known that sample paths of random walks often possess these and similar features, merely by chance. A more meaningful question is whether or not these patterns are predictable and consistent. Attempts to find such patterns usually involve automated search and back-testing, and invariably fall well short of establishing definitive results.

One additional feature of the plot, though, illustrates something that is commonly found in security price dynamics. Although we've been concentrating on small price changes, the SPY plot clearly displays some shifts that are large and almost immediate, gaps or jumps marked by few intervening trades. These changes, too, can result from public information, but of different sort. We turn to these events in the next subsection.

8.3. Major public information announcements and developments

Against a background of slow accumulation of minor news, we're sometimes subjected to major developments: a takeover bid, results of a therapeutic drug test, the filing or settling of a major lawsuit, etc.

In exploring how the market reacts to major announcements, it's useful to start by emphasizing 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 acknowledged condition.)

In contrast to gradual diffusion, significant announcements are associated with sudden and large price changes, or “jumps”. The details of the market response to public news differ according to whether or not the announcement event was anticipated. A press conference is usually a pre-scheduled event; a natural disaster is usually completely unexpected. We’ll discuss the expected and unexpected announcements separately.

8.3.A. Anticipated announcements

Looking back at Figure 8.1, at 8:30am there is a sudden jump in the price of the SPY. What happened?

Economic analysts in the US government and elsewhere measure economic quantities that reveal the current state of the economy and suggest its evolution in the near future. 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. The following table is a partial record from briefing.com.
Table 8.1. Public announcements on April 15, 2011.

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

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 8.2 provides more detail in the neighborhood of 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.

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. More often than not, 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 took established a position with the intent of betting on the impact of the announcement will unwind.
- Any kind of an 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.

Because major 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 publicly 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.

The Michigan Consumer Sentiment index is compiled by a unit of the University of Michigan. According to a Wall Street Journal article, “[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

Copyright 2014, Joel Hasbrouck. All rights reserved. Draft of May 19, 2014.
time. Those who pay for Thomson Reuters’s regular news services get the report two seconds later” (Mullins, Rothfeld, McGinty, and Strasburg, 2013).

8.3.B. Unanticipated announcements

When the information comes from an independent research firm, the announcement may be unexpected by most market participants. Acorda Therapeutics (ticker symbol: ACOR) is a NASDAQ-listed pharmaceutical firm. Figure 8.3 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 8.3 Trading in Acorda, April 14, 2011

Apparently 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 Empire announcement discussed above. 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. 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 newsflow 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.”

Clearly, not everyone gets the same “public information” at the same time. In the interest of fairness, some regulators and market operators believe that when a significant event is pending or when a significant surprise has just occurred, the market should be closed, until most market participants have received the news.

8.4. Trading halts, circuit breakers, and price limits

At its best the trading process involves many participants reacting thoughtfully and deliberately to the unfolding of public information. A “fair and orderly” market stands as the ideal. 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 closing the market.

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.

In the US, 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. The set of halt codes for NASDAQ-listed stocks is typical (Table 8.2).
<table>
<thead>
<tr>
<th>Trade Halt Code</th>
<th>Trade Halt Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>Halt - News Pending. Trading is halted pending the release of material news.</td>
</tr>
<tr>
<td>T2</td>
<td>Halt - News Released. The news has begun the dissemination process through a Regulation FD compliant method(s).</td>
</tr>
<tr>
<td>T3</td>
<td>Halt - Resumption Time ... Two times will be displayed: (1) the time when market participants can enter quotations, followed by (2) the time the security will be released for trading.</td>
</tr>
<tr>
<td>T5</td>
<td>Single Stock Trading Pause in Effect. Trading has been paused by NASDAQ due to a 10% or more price move in the security in a five-minute period.</td>
</tr>
<tr>
<td>T6</td>
<td>Halt – Extraordinary Market Activity. Trading is halted when extraordinary market activity in the security is occurring ...</td>
</tr>
<tr>
<td>T7</td>
<td>Single Stock Trading Pause/Quotation Only Period. Quotations have resumed for affected security, but trading remains paused.</td>
</tr>
<tr>
<td>T8</td>
<td>Halt – Exchange-Traded-Fund (ETF). Trading is halted in an ETF due to the consideration of, among other factors: 1) the extent to which trading has ceased in the underlying security(s); 2) whether trading has been halted or suspended in the primary market(s) for any combination of underlying securities accounting for 20% or more of the applicable current index group value; 3) the presence of other unusual conditions or circumstances deemed to be detrimental to the maintenance of a fair and orderly market.</td>
</tr>
<tr>
<td>T12</td>
<td>Halt - Additional Information Requested by NASDAQ Trading is halted pending receipt of additional information requested by NASDAQ.</td>
</tr>
</tbody>
</table>

Halts are common. Table 8.3 gives a sample of actual halts for a recent date.


<table>
<thead>
<tr>
<th>Halt Date</th>
<th>Halt Time</th>
<th>Issue Symbol</th>
<th>Issue Name</th>
<th>Reason Code</th>
<th>Resumption Date</th>
<th>Resumption Quote Time</th>
<th>Resumption Trade Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>01/25/2012</td>
<td>15:04:42</td>
<td>NUVA</td>
<td>NuVasive Inc</td>
<td>T1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>01/25/2012</td>
<td>09:19:08</td>
<td>GLRE</td>
<td>Greenlight Capital Re, Ltd.</td>
<td>T3</td>
<td>01/25/2012</td>
<td>12:45:00</td>
<td>12:50:00</td>
</tr>
<tr>
<td>01/25/2012</td>
<td>09:00:38</td>
<td>PNNW</td>
<td>Pennichuck Corporation</td>
<td>T12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>01/25/2012</td>
<td>08:10:27</td>
<td>INCB</td>
<td>Indiana Community Bancorp</td>
<td>T3</td>
<td>01/25/2012</td>
<td>08:40:00</td>
<td>08:45:00</td>
</tr>
<tr>
<td>12/19/2011</td>
<td>13:29:39</td>
<td>FEED</td>
<td>AgFeed Industries, Inc.</td>
<td>T12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11/29/2011</td>
<td>07:01:23</td>
<td>BQI</td>
<td>Oilsands Quest Inc Common Stock</td>
<td>T2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


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 run by the primary exchange (see Chapter 5).

U.S. market-wide circuit-breakers are triggered by a drop in the S&P 500 index of 10%, 20%, or 30% from the previous day’s close. With a 10% drop, trading is halted for one hour, or (if after 2:30pm) 30 minutes. The higher levels of decline trigger longer closures.

The market-wide circuit-breakers were instituted in response to the October 1987 “market break,” a sudden price decline across virtually all stocks. On May 6, 2010, a similar event occurred (the “flash crash”), but all stocks were not affected (U.S. Commodities Futures Trading Commission and Commission, 2010). In response, single-stock circuit breakers were instituted. These were subsequently replaced by current policies, usually described as a limit-up/limit-down (LULD) mechanism.

The limit-up/limit-down rule sets a price band which, once hit, triggers a five-minute “pause”. At the end of the pause, the primary listing exchange reopens the stock. For most stocks, the band is ±5% of a reference price. The reference price is the average trade price over the previous five minutes, or (if there are no trades) the previous reference price.

Trading halts and price limits are controversial.

The arguments against halts and limits are:

- 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 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 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.
Chapter 9. Private information

9.1. Overview

Although a vast amount of information is public, a much larger amount is private. A stock might have millions of holders, and millions more potential holders, each of whom brings to the market his or her own opinions about the stock’s value as well as knowledge about their own financial needs, aspirations and trading styles.

At their best, financial markets can aggregate diverse private information, reflecting in the price of a security a consensus estimate of what the security is worth. The estimate is all the more reliable because traders, in taking a long or short position, are putting money at risk.

In the worst cases, though, financial markets can be completely shut down by private information. If every potential trader believes that their counterparty possesses superior information, the trading process will seem too stacked against them to actually go through with a trade. The financial crisis was marked by markets (most notably asset-backed securities) where normally liquid mechanisms failed.

Private information does not simply affect overall market quality in a general way. It is a very real and immediate consideration for every trade we consider. Our counterparty is not some collective abstraction called “the market”. It is a real person or institution who might have information superior to ours, and against whom we will tend to lose. And they, of course, are similarly wondering about us.

This chapter takes public informational efficiency as the starting point. How might this be extended to embrace private information? What is the role of trading process? What are the implications for trading strategies?

9.2. Efficiency and private information

The public-information version of the efficient market hypothesis usually strikes people as a reasonable first approximation. 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 ones.

But why should the price of a security reflect private, non-public information? The broad answer is that private information creates motives for trade, and in the trading process the price moves and information is revealed.
The simplest case arises when one person possesses significant value-relevant information not known to the rest of the market. 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 actually 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 of the 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 actually purchase the shares, there would be no reason for the market prices to move. Secondly, for the trader to actually 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. Valuable private information is advance knowledge of public information.

When information is not common knowledge, it is said to be asymmetric (across market participants). Not all cases of asymmetric information, however, involve someone with superior information. Sometimes everyone holds partial information not commonly known, a small piece of the valuation puzzle, so to speak. Suppose that we’re trying to guess the total number of marbles in an urn, the marbles come in ten colors, there are ten of us guessing, and each of us knows the exact count for “our” color. As a group we know the total number, but if we can’t communicate, then the game is still a matter of chance. The information in this example is similar to that possessed by a group of security analysts, each of whom is an expert on one division of a multi-industry firm.

In many situations a financial market can effectively impound private information as well as public. The price of a security is determined by traders’ orders, and from the trading process a consensus view emerges. It is said in such situations that the market price aggregates diverse information.

The necessity of trade explains the framing of US 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:

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. (U.S. Securities and Exchange Commission, 2013)

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. Avoid 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.

### 9.3. The dealer’s perspective.

When market participants differ in the amount and quality of their information, the market is said to exhibit asymmetric information. “Asymmetry” refers to inequality across traders. 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.)

Almost all traders who place bids and offers are affected by asymmetric information, but we focus on a dealer because in some ways the dealers problem in simpler (as in section 6.5). The dealer has to post a bid and offer; he can’t use market orders; and if he doesn’t at least break even on average, his business fails.

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’ problem 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 eliciting offsetting orders (as in section 6.5) or using the interdealer market. 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 dealers 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 is more costly than if there were only one 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.

How should the dealer set his bid price? If his bid is actually hit, the seller might be uninformed (“liquidity”) seller, but the seller might also be an informed trader with bad news. The latter possibility will lead the dealer to revise his opinion about the stock: he thinks the stock is worth a little less than he did before his bid was hit. When he has the opportunity to rebid, he’ll move his bid down.

The dealer won’t know immediately whether or not he actually lost money on the trade. If later in the day the firm announces some bad news, the dealer will look back over his purchases and suspect that at least a few sellers were informed. If there is no news, the dealer will be more inclined to view the sellers as uninformed (and therefore profitable).

Notice though, that the bid moves well before the announcement. The fact that the bid has been hit (especially if it's been hit repeatedly) is enough to raise suspicions of impending bad news. The trade (at the bid) is in and of itself a public signal suggesting private knowledge.

9.4. Causation and detection

A customer lifts the ask (buys) and subsequent prices rise. Conversely, if the customer hits the bid (sells), subsequent prices will be lower. In this sense, incoming orders cause price movements.

The price reaction to a buy order might not seem unexpected. After all, 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?

One puzzle arises in connection with 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.
A second mystery arises from the obvious limitations of the trading process. Ultimately the value of the security is determined by fundamentals: the firm’s revenues, costs, profits, and payouts to investors. Unless the security holders at some point take an active and direct role in the firm’s management, these fundamentals won’t change. Such shareholder activism certainly occurs, but it typically plays out in proxy fights or directorial elections. These may be well-publicized, but there are infrequent. It is difficult to connect day-to-day order flows, and the slow turnover of security ownership with changes in the security’s fundamentals.

These two facts, prices that respond in the short-run to order flow, and values that don’t depend on ownership (except perhaps in the long run), seem contradictory. Why should a buy order drive up the price if the new owner doesn’t plan on changing anything that might affect value?

These two things can be reconciled, though, once we recognize the power of information. The simple story is this. The sources of value that justify the new higher price reflect information that is already known, at least to a small number of market participants. The possibility that the purchases might have originated from such informed participants explains the price impact. Orders reveal, through the trading process, what is already privately known. Order flow is a signal, a window onto things that aren’t yet public.

This also explains the apparent paradox of scale. 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.

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 Chapter 13.

9.5. Other considerations

Private information touches on many aspects of market design and regulation. Here are some of the more current and important ones.

9.5.A. Should trades be publicly reported?

All trades in US equities are reported to the “tape” and broadly disseminated. This practice was adopted by the New York Stock Exchange and other floor exchanges as a measure imposed by the membership on the membership, well before the contemplation of any government securities regulation. Publication of trades is an advertisement of a market’s transparency and suggests fairness. To the extent that it draws in more trading activity, it is also a good business practice for the exchange.
Yet a surprising number of markets do not practice trade publication, or do so only partially. Transactions in US corporate bonds are reported to TRACE system, and are promptly available on the web and elsewhere. But far from being a self-imposed industry practice, the TRACE system was forced on a reluctant dealer community by securities regulators. Trades in US Treasury debt (bills, notes and bonds) are not reported, although a subset of them may be visible to subscribers of (mostly institutional) trading systems. The same state of affairs also characterizes foreign exchange markets, all sorts of swaps, and other “over-the-counter” derivatives.

Trade reporting is a public good, but imposes costs on the counterparties to the trades. Consider the first trade in mining company example. Observing Clarence’s purchase induces other sellers to move their quotes, making Clarence’s subsequent purchases more expensive. Certainly he would prefer that the trade not be reported. Furthermore, if the trade is unreported, the seller has an informational edge in that she has learned something from Clarence’s purchase, and can perhaps trade against the asks of other (ignorant) traders.

So why do market practices differ? History suggests an explanation. Traders in physically-convened floor markets find it easier to observe others’ behavior. Lapses in trade reporting do not go unnoticed. Most stock and futures exchanges, although presently electronic, reflect their floor heritage in their maintenance of trade reporting. The foreign exchange market, in contrast, was never a centralized floor.

9.5.B. 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.

9.5.C. The social cost of informed trading

In our simple model, 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. Most of us have given audience in social situations to those who declaim in great depth and with strong confidence on the present state and future direction of the market. 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.

9.5.D. 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.

The creation of private information by delay can be subtle. The Globex limit order market for US futures is an electronic platform, on which trades are generally reported promptly and automatically. It recently came to widespread attention, however, that the public trade report slightly lagged the confirmation messages that were transmitted to the parties involved in the trade. A limit order seller, for example, would know that his ask had been hit slightly before the rest of the market. This was quickly recognized as a potential problem and corrected. (Patterson and Strasburg, 2013; Patterson, Strasburg, and Pleven, 2013)

9.5.E. 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 or not 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.
Chapter 10. Economic models of private information

This chapter describes a more formal discussion of markets with private information. The central construct is known as a sequential trade model. The name derives from one distinctive feature of these models, that the sequence of events involves a cycle in which the market maker posts quotes, a single customer arrives and buys, sells, or does nothing, the market maker revises her quotes, the next trader arrives, and so on. The class of these models is sufficiently broad and flexible to incorporate many features of actual securities markets.

These models aim to provide a formal framework for thinking about the trading process. They are usually designed to highlight a particular feature of the market, rather than serve as comprehensive models of reality. The advantage of this approach is that it clearly exposes the incentives and strategies used by agents. Within their own set of rules and assumptions, they are considered partial equilibrium models. That is, the result describes a balance among the players that reflects their feasible actions, the knowledge they start with, and information they glean from the trading process. The assumptions are usually confining and unrealistic, but once we understand the workings of the models, it is often easy to guess the direction things would take if the assumptions were relaxed.

This chapter relies on a deeper background in probability and statistics than most of the book. The results aren’t essential to later chapters, and it may be skipped on first reading. An appendix reviews the required probability material.

10.1. The elements

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: $V=\text{Low}$ or $\text{High}$ 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.

Giving ourselves a role as the passive trader, we 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 $V = \text{Low}$, the insider will hit any bid above Low, making a profit for himself, and a loss for us. If $V = \text{High}$, the insider will lift any offer below High, and we lose again.

The withdrawal of the passive trader is not only an modeling inconvenience, it is 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: $V = \text{Low}$ or $V = \text{High}$. 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 $V = \text{High}$, he'll buy at any ask price below $\text{High}$; if $V = \text{Low}$, he'll sell at any bid above $\text{Low}$.
   b. With 80% probability, the trader is uninformed. Uninformed traders buy or sell with equal probability.

Figure 10.1 describes the sequencing of the random events in the model. (The setting of bid and ask quotes is not indicated.)

---

**Figure 10.1 Sequencing of random events**

The transitional probabilities used in the text example are given in italics.

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. Our analysis is clearly going to involve probabilities, and we'll have to use some of the tools from Chapter 18.
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 or not 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 10.1. They are transitional probabilities, reflecting the likelihood of moving from one box (“node”) to another. From them, 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 an informed trader is:

$$Pr(V = High, Informed) = 0.5 \times 0.2 = 0.10$$

The probability that $V = High$, 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 10.2. Note that the total probability of a buy is 0.5, as is the total probability of a sell.

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:
\[
P(V = \text{Low}|\text{Sell}) = \frac{P(V = \text{Low, Sell})}{P(\text{Sell})} = \frac{0.1 + 0.2}{0.5} = 0.6
\]

You can check that \(P(V = \text{Low}|\text{Buy}) = 0.4, P(V = \text{High}|\text{Buy}) = 0.6, \) and \(P(V = \text{High}|\text{Sell}) = 0.4.\)

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

\[
E[V] = 0.5 \times \text{Low} + 0.5 \times \text{High}
\]
\[
= 0.5 \times 100 + 0.5 \times 150 = 125
\]

The expected value conditional on a sell is,

\[
E[V|\text{Sell}] = 0.6 \times \text{Low} + 0.4 \times \text{High}
\]
\[
= 0.6 \times 100 + 0.4 \times 150 = 120
\]

In this way, we (and other uninformed market participants) can "read" the order flow, to draw inferences about what the informed participants already know.

10.2. Setting the initial bid and offer

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 pretty close to \(E[V|\text{Sell}].\)

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 \(E[V|\text{Buy}].\)

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 \(\text{Low} = 100\) and \(\text{High} = 150,\) let \(\text{Low} = 90\) and \(\text{High} = 160.\) 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.

10.3. The second trade and order impact

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.

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 probability 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 10.3.

---

Figure 10.3. 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 

\[
P(V = \text{Low}|Sell, Sell) = \frac{P(Low, Sell, Sell)}{P(Sell, Sell)} = \frac{0.12 + 0.24}{0.52} = 0.692
\]

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,

\[
E[V|Sell, Sell] = 0.692 \times \text{Low} + 0.308 \times \text{High} = 0.692 \times 100 + 0.308 \times 150 = 115.38
\]

Repeating the argument 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, 
\( P[V = \text{Low}|Sell, Buy = 50\%] \), and 
\( E[V|Sell, Buy] = 0.5 \times \text{Low} + 0.5 \times \text{High} = 0.5 \times 100 + 0.5 \times 150 = 125 \). 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.

### 10.4. The dynamics of prices and orders

We can use the model to study any sequence of bids and offers over any horizon. SimpleSequentialTradeModel02.xlsx (on the website) is set up to perform these calculations. Figure 10.4 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.
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, it also reduces (over time) the uncertainty about the true value. After a period of one-sided order flow, the spread narrows.
Chapter 11. Trading costs

Are you a trader or an investor?

In one sense, the distinction is irrelevant. Total profits are completely determined by the size and timing of the cash flows (purchases and sales). Labeling one component of the cash flows as “trading”, and the other as “investment” won’t affect the total. Still, some measurable expenses such as commissions, bid-ask spreads, and price impacts seem more closely related to how we bought or sold, with everything else being attributed to longer-run investment returns.

Even if we could separate trading costs and investment characteristics, we’d ideally want to make our trading and investment decisions jointly, taking into account any interactions. The goal would be to plan a portfolio and execution strategy by simultaneously considering the long term security attributes and the costs of actually buying and selling the securities.

But trading and investment decisions require different data and different ways of thinking. The decisions usually fall to different groups of people. The practicalities of delegation, management and evaluation seem to force trading and investment processes to be considered separately and distinctly.

The distinction is important quantitatively because trading costs can be a very large drag on portfolio performance. As a rough approximation, the one-way trading cost on a stock is about 50 bp (0.5%). This includes explicit costs (like commissions) and implicit costs (like the bid/ask spread). If we buy or sell a $100 stock, it will cost us $0.50 per share.

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. U.S. market centers are responsible for computing and reporting trading costs. These are posted to their web sites.

11.1. (Portfolio) Implementation shortfall

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 a hypothetical paper portfolio in which all trades are made at notional ("benchmark") prices.

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 if we had a perfect trading desk, our trades could be executed at the notional prices. Any divergence must be attributed to trading (implementation) costs.

The portfolio implementation shortfall, as originally proposed, includes both explicit and implicit costs.
Explicit costs

- Commissions, net of any rebates
- “Take” fees and liquidity rebates (if not included in the commission).
- Transactions taxes

Implicit costs

- Costs of interacting with the market (e.g., bid-ask or price impact costs), relative to the benchmark prices. These costs can be estimated with statistical models of the sort discussed in Chapter 13.
- 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 may leave the portfolio exposed to additional risk.
- Delay (failure to fill the order immediately).

The explicit costs are usually easy to measure. They tend to be reported on the confirmation, the report given to the investor client by the broker. 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, “implementation shortfall” is usually defined as:

\[
Implementation\ Shortfall = \begin{cases} 
\text{Trade Price} - \text{Benchmark Price}, & \text{for a buy order} \\
\text{Benchmark Price} - \text{Trade price}, & \text{for a sell order}
\end{cases}
\] (5.1)

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:

\[
Total\ Cost = Implementation\ Shortfall + \text{Commissions}
\]

Henceforth, we'll use this narrower definition.
11.2. Benchmark prices used in IS calculations

The implementation shortfall calculation depends crucially on the choice of the benchmark price.

**Pre-trade benchmarks**

- 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*.

**Post-trade benchmarks**

- The NBBO midpoint prior five minutes after the trade.
- The next day’s opening price

**Pre-/post-trade benchmarks**

- Time-weighted average price (TWAP, “Tee Wap”) over the day or duration of the order.
- Value-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.

11.3. 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.

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.
Letting $m$ denote the bid-offer midpoint, letting $p$ denote the execution price, the effective cost is defined as:

$$\text{Effective Cost} = \begin{cases} p - m, & \text{for a marketable buy order} \\ m - p, & \text{for a marketable sell order} \end{cases}$$ \hspace{1cm} (5.2)

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.

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:

$$\text{Price improvement} = \begin{cases} \text{NBO} - p, & \text{for a marketable buy order} \\ p - \text{NBB}, & \text{for a marketable sell order} \end{cases}$$ \hspace{1cm} (5.3)

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 $m_5$:

$$\text{Realized Cost} = \begin{cases} p - m_5, & \text{for a marketable buy order} \\ m_5 - p, & \text{for a marketable sell order} \end{cases}$$ \hspace{1cm} (5.4)

Figure 11.1 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 11.1 Effective cost, realized cost and price improvement for a marketable buy order.**

---

The quantity $\text{Effective Cost} - \text{Realized Cost}$ 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 11.1 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 11.1 Calculation of effective cost, realized cost, price improvement and VWAP for marketable orders.

<table>
<thead>
<tr>
<th>Quotes</th>
<th>Trades</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>NBB</td>
</tr>
<tr>
<td></td>
<td>Vol</td>
</tr>
<tr>
<td></td>
<td>Dir</td>
</tr>
<tr>
<td>10:14:00</td>
<td>19.81</td>
</tr>
<tr>
<td>10:15:00</td>
<td>19.80</td>
</tr>
<tr>
<td>10:16:00</td>
<td>19.78</td>
</tr>
<tr>
<td>10:17:00</td>
<td>19.85</td>
</tr>
<tr>
<td>10:18:00</td>
<td>19.92</td>
</tr>
<tr>
<td>10:19:00</td>
<td>19.91</td>
</tr>
<tr>
<td>10:20:00</td>
<td>19.84</td>
</tr>
<tr>
<td>10:21:00</td>
<td>19.92</td>
</tr>
<tr>
<td>10:22:00</td>
<td>19.84</td>
</tr>
<tr>
<td>10:52:00</td>
<td>19.93</td>
</tr>
<tr>
<td>Order Time</td>
<td>Dir</td>
</tr>
<tr>
<td>10:15:09</td>
<td>B</td>
</tr>
<tr>
<td>10:17:22</td>
<td>B</td>
</tr>
<tr>
<td>10:20:09</td>
<td>S</td>
</tr>
</tbody>
</table>

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.

Total Volume: 1,000

Total (Vol x Pr): 19,855

VWAP: 19.855

Copyright 2014, Joel Hasbrouck. All rights reserved. Draft of May 19, 2014.
11.4. 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 11.2 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 11.2

Suppose that we want to buy. At the start of things we can 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’ll 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: $dud$, $udd$, $ddd$, 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?

Often limit orders are used early in strategies that have a time deadline. If there has been no execution of the limit order by the deadline, then we must submit a market order to achieve the execution without further delay.

The Tokyo Stock Exchange actually offers a single order type, a *funari* order, that implements this strategy. A *funari* order is a limit order that rests in the book during a continuous trading session. But if it hasn’t been executed by the end of the session, it is converted to a market-on-close order \{Tokyo Stock Exchange, 2011 #1508\}.

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 \(1 - 0.635 = 0.375\).

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

\[
E[\text{offer}|\text{limit order fails}] = \frac{1}{3} \times 10.03 + \frac{2}{3} \times 10.01 = 10.0167
\]

Our overall expected cost of buying with the limit order strategy is

\[
E[\text{cost}] = 9.99 \times P(\text{LimEx}) + 10.0167 \times P(\text{LimFail,})
\]

where \(\text{LimEx}\) and \(\text{LimFail}\) denote the success and failure of the limit order. So:

\[
E[\text{cost}] = 9.99 \times 0.625 + 10.0167 \times 0.375 = 10.00
\]

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

\[
E[\text{offer}|\text{limit order fails}] = \frac{1}{6} \times 10.03 + \frac{3}{6} \times 10.01 + \frac{2}{6} \times 9.99 = 10.0067
\]

and

\[
E[\text{cost}] = 9.98 \times 0.25 + 10.0067 \times 0.75 = 10.00.
\]

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’re trying to buy), and we have to “chase” a price that’s running away from us.
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. Other aspects of order-price dynamics will be discussed in Chapter 13.

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.

11.5. 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 11.2, 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 12.2). 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).

11.6. Implementation shortfall decompositions

Once we settle on a benchmark price, the implementation shortfall is easily computed using equ (5.1). 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:

\[
\text{Implementation shortfall} = \text{price impact} + \text{cost of delay}
\]

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 even if we did not in fact achieve any executions. This may occur because others are using strategies similar to ours, or responding to the same signals. It may also arise because others are detecting our intentions and trading ahead of us.

11.7. 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 11.2 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 of 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 11.2. BATS Rule 605 statistics for EIHI, November 2009.

<table>
<thead>
<tr>
<th>Order type</th>
<th>Order size</th>
<th>No. orders</th>
<th>Shares in orders</th>
<th>Shares cancelled</th>
<th>Shares executed</th>
<th>Shares executed elsewhere</th>
<th>Effective spread</th>
<th>Realized spread</th>
<th>Shares receiving price improvement</th>
<th>Average price improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Market</td>
<td>1-499</td>
<td>2</td>
<td>200</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>0.050</td>
<td>0.045</td>
<td>200</td>
<td>0.0475</td>
</tr>
<tr>
<td>Marketable limit</td>
<td>1-499</td>
<td>363</td>
<td>37,002</td>
<td>34,400</td>
<td>1,122</td>
<td>1,480</td>
<td>0.093</td>
<td>0.069</td>
<td>1228</td>
<td>0.0227</td>
</tr>
<tr>
<td>Marketable limit</td>
<td>500-1,999</td>
<td>10</td>
<td>5000</td>
<td>0</td>
<td>0</td>
<td>5,000</td>
<td>0.198</td>
<td>0.190</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
Problem 11.1 Cost computations

The NBBO / Bid-ask midpoint record for XYZ is:

<table>
<thead>
<tr>
<th>Time</th>
<th>NBB</th>
<th>NBO</th>
<th>BAM</th>
<th>Spread</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:01:00</td>
<td>35.05</td>
<td>35.10</td>
<td>35.075</td>
<td>0.05</td>
</tr>
<tr>
<td>11:02:00</td>
<td>34.94</td>
<td>34.99</td>
<td>34.965</td>
<td>0.05</td>
</tr>
<tr>
<td>11:03:00</td>
<td>34.83</td>
<td>34.86</td>
<td>34.845</td>
<td>0.03</td>
</tr>
<tr>
<td>11:04:00</td>
<td>34.77</td>
<td>34.81</td>
<td>34.790</td>
<td>0.04</td>
</tr>
<tr>
<td>11:05:00</td>
<td>34.77</td>
<td>34.85</td>
<td>34.810</td>
<td>0.08</td>
</tr>
<tr>
<td>11:06:00</td>
<td>34.74</td>
<td>34.78</td>
<td>34.760</td>
<td>0.04</td>
</tr>
<tr>
<td>11:07:00</td>
<td>34.75</td>
<td>34.78</td>
<td>34.765</td>
<td>0.03</td>
</tr>
<tr>
<td>11:08:00</td>
<td>34.80</td>
<td>34.86</td>
<td>34.830</td>
<td>0.06</td>
</tr>
<tr>
<td>11:09:00</td>
<td>34.72</td>
<td>34.75</td>
<td>34.735</td>
<td>0.03</td>
</tr>
<tr>
<td>11:10:00</td>
<td>34.75</td>
<td>34.83</td>
<td>34.790</td>
<td>0.08</td>
</tr>
<tr>
<td>11:11:00</td>
<td>34.66</td>
<td>34.69</td>
<td>34.675</td>
<td>0.03</td>
</tr>
<tr>
<td>11:12:00</td>
<td>34.74</td>
<td>34.78</td>
<td>34.760</td>
<td>0.04</td>
</tr>
<tr>
<td>11:13:00</td>
<td>34.82</td>
<td>34.88</td>
<td>34.850</td>
<td>0.06</td>
</tr>
</tbody>
</table>

What are the effective cost, realized cost, price improvement and price impact for the following trades?

a. An order arriving at 11:02:44, buy 100 shares limit 35.00, was executed at 34.98.
b. An order arriving at 11:07:02, sell 100 shares limit 34.70, was executed at 34.73.
Answer to problem

a. The quote prevailing when the order arrived was set at 11:02:00; five minutes subsequent to arrival is 11:07:44, and the prevailing quote at that point was set at 11:07:00. These quotes are highlighted below:

<table>
<thead>
<tr>
<th>Time</th>
<th>NBB</th>
<th>NBO</th>
<th>BAM</th>
<th>Spread</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:01:00</td>
<td>35.05</td>
<td>35.10</td>
<td>35.075</td>
<td>0.05</td>
</tr>
<tr>
<td>11:02:00</td>
<td>34.94</td>
<td>34.99</td>
<td>34.965</td>
<td>0.05</td>
</tr>
<tr>
<td>11:03:00</td>
<td>34.83</td>
<td>34.86</td>
<td>34.845</td>
<td>0.03</td>
</tr>
<tr>
<td>11:04:00</td>
<td>34.77</td>
<td>34.81</td>
<td>34.790</td>
<td>0.04</td>
</tr>
<tr>
<td>11:05:00</td>
<td>34.77</td>
<td>34.85</td>
<td>34.810</td>
<td>0.08</td>
</tr>
<tr>
<td>11:06:00</td>
<td>34.74</td>
<td>34.78</td>
<td>34.760</td>
<td>0.04</td>
</tr>
<tr>
<td>11:07:00</td>
<td>34.75</td>
<td>34.78</td>
<td>34.765</td>
<td>0.03</td>
</tr>
<tr>
<td>11:08:00</td>
<td>34.80</td>
<td>34.86</td>
<td>34.830</td>
<td>0.06</td>
</tr>
<tr>
<td>11:09:00</td>
<td>34.72</td>
<td>34.75</td>
<td>34.735</td>
<td>0.03</td>
</tr>
<tr>
<td>11:10:00</td>
<td>34.75</td>
<td>34.83</td>
<td>34.790</td>
<td>0.08</td>
</tr>
<tr>
<td>11:11:00</td>
<td>34.66</td>
<td>34.69</td>
<td>34.675</td>
<td>0.03</td>
</tr>
<tr>
<td>11:12:00</td>
<td>34.74</td>
<td>34.78</td>
<td>34.760</td>
<td>0.04</td>
</tr>
<tr>
<td>11:13:00</td>
<td>34.82</td>
<td>34.88</td>
<td>34.850</td>
<td>0.06</td>
</tr>
</tbody>
</table>

The effective cost is 34.98 – 34.965 = 0.015; the realized cost is 34.98 – 34.765 = 0.215. The price improvement is 34.99 – 34.98 = 0.01; the price impact is 34.765 – 34.965 = –0.200.
b. The quote prevailing when the order arrived was set at 11:07:00; five minutes subsequent to arrival is 11:12:02, and the prevailing quote at that point was set at 11:12:00. These quotes are highlighted below:

<table>
<thead>
<tr>
<th>Time</th>
<th>NBB</th>
<th>NBO</th>
<th>BAM</th>
<th>Spread</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:01:00</td>
<td>35.05</td>
<td>35.10</td>
<td>35.075</td>
<td>0.05</td>
</tr>
<tr>
<td>11:02:00</td>
<td>34.94</td>
<td>34.99</td>
<td>34.965</td>
<td>0.05</td>
</tr>
<tr>
<td>11:03:00</td>
<td>34.83</td>
<td>34.86</td>
<td>34.845</td>
<td>0.03</td>
</tr>
<tr>
<td>11:04:00</td>
<td>34.77</td>
<td>34.81</td>
<td>34.790</td>
<td>0.04</td>
</tr>
<tr>
<td>11:05:00</td>
<td>34.77</td>
<td>34.85</td>
<td>34.810</td>
<td>0.08</td>
</tr>
<tr>
<td>11:06:00</td>
<td>34.74</td>
<td>34.78</td>
<td>34.760</td>
<td>0.04</td>
</tr>
<tr>
<td>11:07:00</td>
<td>34.75</td>
<td>34.78</td>
<td>34.765</td>
<td>0.03</td>
</tr>
<tr>
<td>11:08:00</td>
<td>34.80</td>
<td>34.86</td>
<td>34.830</td>
<td>0.06</td>
</tr>
<tr>
<td>11:09:00</td>
<td>34.72</td>
<td>34.75</td>
<td>34.735</td>
<td>0.03</td>
</tr>
<tr>
<td>11:10:00</td>
<td>34.75</td>
<td>34.83</td>
<td>34.790</td>
<td>0.08</td>
</tr>
<tr>
<td>11:11:00</td>
<td>34.66</td>
<td>34.69</td>
<td>34.675</td>
<td>0.03</td>
</tr>
<tr>
<td>11:12:00</td>
<td>34.74</td>
<td>34.78</td>
<td>34.760</td>
<td>0.04</td>
</tr>
<tr>
<td>11:13:00</td>
<td>34.82</td>
<td>34.88</td>
<td>34.850</td>
<td>0.06</td>
</tr>
</tbody>
</table>

The effective cost is $34.765 - 34.73 = 0.035$; the realized cost is $34.760 - 34.73 = 0.030$; the price improvement is $34.75 - 34.73 = 0.02$; the price impact is $34.765 - 34.760 = 0.005$. (Note that the signs of these calculations are different for buy and sell orders.)
Chapter 12. Conditional 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 the key feature of appeal. 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?”

12.1. 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”.

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 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 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. Retail traders, incidentally, are widely believed to herd in the placing of their stops, grouping at “natural” price points (five and ten dollar multiples, for example).

12.2. 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.

12.3. Discretionary orders

This is 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.

12.4. 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 obviousness is contrary to the original intent 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. It might also make sense to randomize the refresh time.

12.5. 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.
Chapter 13. A first look at order splitting strategies
INCOMPLETE: See powerpoint slides

For most stocks the depth of the book (shares bid for or offered) at any given instant is too low to permit execution of a large market order at an acceptable price. As a result, many institutional orders are broken up into smaller orders that are submitted to the market over time. These are sometimes described as parent/child orders.

Splitting strategies are influenced by all sorts of trade-offs. Since small orders generally receive better prices than larger orders, the best strategy would seem to involve splitting the parent order into many small child orders which are submitted over a long time horizon. The drawback of doing this is that the longer we defer execution, the more risk of an adverse price movement. Another important factor is that due to order price impact our earlier trades will make later trades more expensive. These trade-offs will be analyzed in Chapters 14 and 15.

The present chapter examines situations where the benchmark price used in the implementation shortfall is TWAP or VWAP (time- or value-weighted average price), and the execution strategy is aimed at achieving (or ideally bettering) these benchmarks.
Chapter 14. Statistical models of order-price dynamics

This chapter develops a small number of statistical models that describe how security prices evolve and how order flow affects these dynamics. From one perspective these models are short-term predictive tools. Entire books have been written about stock price forecasting, and it is probably a safe bet that the diversity of techniques and paucity of reliable results are unmatched by any other field of scientific endeavor.

The question guiding the present analysis, though, is not, “Which way are stock prices headed?” The concern is instead, “How will my orders and trades affect prices?” The reason this second question is interesting is that many of the trading strategies that we’ll discuss in the next chapter involve sequences of trades. In a sequence, early trades may affect the prices of later ones. Clarifying this interaction and accounting for it will be an important guiding principle.

The present analysis doesn’t take a position on any prediction tool – including your current favorite. The results can peacefully coexist with most prediction models. Along the way I’ll point out how the predictions can be incorporated.

The models here tend to be descriptive and statistical, rather than theoretical and economic. The distinction is driven by practicalities. We earlier explored a simple model of informed trading that exhibited order impacts and information-dependent bid-ask spreads. It is difficult, however, to map the model directly to a real stock. (It is 9am. What will be today’s High and Low values for Microsoft?) Instead, we’ll take the intuition that orders move prices and add it to a statistical model. The statistical model, in turn, will be estimated from recent data that presumably reflect how the stock actually behaves.

The development proceeds in stages, starting with simple models, and moving to more complex ones.

14.1. The random walk.

The simplest model of security prices is the random-walk:

\[ p_t = p_{t-1} + u_t \]

where \( t = 1, 2, \ldots \) might index minutes, seconds, or some similarly brief time interval; \( p_t \) is the stock price at the close of interval \( t \), and \( u_t \) is a random perturbation, disturbance, or error. \( u_t \) reflects new information that arrives over the interval. The news might be good or bad. More formally, \( E_t u_t = 0 \): as of the end of last interval, we expected this period’s news to be neutral.

We’re often concerned with the volatility of the price dynamics, characterized by the standard deviation of \( u_t \), \( \sigma_u \). If \( p_t \) is redefined, for the moment, as the logarithm of the stock price, then \( p_t - p_{t-1} \) is approximately the proportional (“percentage”) change in the stock price over the interval, and \( u_t \) is (approximately) a proportional disturbance.

In log price form, an annual volatility of, say, \( \sigma_u \approx 40\% \) corresponds to a monthly volatility of \( \sqrt{0.4^2/12} \approx 11.5\% \), or a daily volatility of of \( \sqrt{0.4^2/250} \approx 2.5\% \).

Note: there are about 250 trading days in a year. Most of the cumulative volatility in financial markets arises on trading days. (Some people read this as “trading causes volatility.”)
We now consider estimation. The model has one parameter, the volatility $\sigma_u$. Suppose that we have a set of $T$ prices $\{p_1, p_2, \ldots, p_T\}$. Then $u_t = p_t - p_{t-1}$ for $t = 2, \ldots, T$ (we lose $u_1$ because we don’t have $p_0$). The usual estimate for the variance would be $\hat{\sigma}^2_u = \frac{1}{T-1} \sum_{i=2}^{T} u_i^2$.

This is the mean sum-of-squares.

There are many ways of estimating volatility, however. In option pricing the above estimate is called the historical volatility. From the market prices of options, one can compute an implied volatility, which may be a better forecast of the volatility that will arise in the future.

---

Stock prices lie on a discrete grid. Bids and offers in U.S. equity markets cannot normally specify an increment finer than $0.01$, that is, there can’t be more than two places to the right of the decimal. Other markets adopt similar conventions. While accepting this discreteness as a feature of the data, however, we typically don’t incorporate it as a feature of the statistical model. This simplifies estimation. It doesn’t usually cause problems, but we must remember some of our statistical intuitions implicitly invoke continuous models, and well-behaved ones at that. With a normally-distributed (Gaussian) random variable, for example, the intervals defined by the mean $\pm \sigma$, $\pm 2\sigma$, $\pm 3\sigma$ contain on average $68.2\%$, $95.4\%$, and $99.7\%$ of the outcomes. This generally won’t be true for most of the discrete distributions encountered in security pricing.

### 14.2. Drift and alpha

An obvious shortcoming of the model is that it doesn’t allow for any expected capital gains. So we introduce a drift term:

$$p_t = \mu + p_{t-1} + u_t$$

If $p_t$ is redefined as the logarithm of the stock price, then $p_t - p_{t-1}$ is approximately the proportional (“percentage”) change in the stock price over the month, and $\mu$ is the expected percentage price appreciation. An annual return of $20\%$ per year works out to about $\mu = 0.017$ ($1.7\%$) per month. The implied daily expected return is about $\mu \approx 0.0008$ ($0.08\%$) per day (using $250$ trading days per year).

The importance of the random-walk model is due to:

- The concept of informational efficiency. This is an economic argument that a rational market incorporates all known information into the security price, so all future movements must arise from information not previously known. This implies that price movements should be unpredictable. The random-walk model is a simple construct with unpredictable movements.

- Statistical tests generally show that the model is “difficult to reject.” There aren’t solid grounds for preferring any simple alternative model.
In monthly, weekly or daily data, the model is sufficiently close to reality that it is useful for many purposes, notably option pricing. For trading (intraday) horizons, the model needs enhancement to allow for drift (alpha) and order impact. We turn now to the specifics of these modification.

When we’re using the random-walk model applied to, say, monthly log prices, the drift term $\mu$ is the expected price change over the month. Ignoring dividends this is the expected return over the month. Expected returns over monthly and longer horizons are determined by economic fundamentals, risk, and peoples beliefs about these things. At intraday horizons, we take a broader view of the drift.

The drift corresponds to predictable short-term price changes. The extent of the predictability depends on the information, the model, and the faith placed in the model. It may be different for different market participants.

In finance, predictable over- or under-performance (net of compensation for risk) is often called “alpha” ($\alpha$). In a trading context, to distinguish this predictability from effects associated with long-run fundamentals, it is sometimes called “short-term alpha”. This parameter, then, incorporates the output of any supplemental prediction model. The present analysis assumes that it is constant over the horizon of the trading problem, but with a little additional effort it can be viewed as time-varying ($\alpha_t$).

The random walk with drift has two parameters, $\mu$ and $\sigma_u$. The first difference of the price is $\Delta p_t = p_t - p_{t-1} = \mu + u_t$. So we can estimate $\mu$ and $\sigma_u^2$ as the mean and variance of the $\Delta p_t$, something that might easily be accomplished with a calculator or spreadsheet.

In practice things are bit more complicated. If the drift is attributed to the expected long-run return, its mean is likely to be quite close to zero. As a result, while useful estimates of $\sigma_u$ can be obtained from a sample of 30 days, important changes in $\mu$ require a few years of data to discern.

### 14.3. Permanent order impacts

Next, due to private information, trades can have cumulative and permanent effects on prices. Active sells can drive the price down and active buys tend to drive the price up. How permanent is “permanent”? This is debatable. For present purposes, “permanent” means “doesn’t go away over reasonable trading horizons”, an attribute sometimes described as “quasi-permanent”.

We can modify the random-walk model to incorporate short-term alpha and a permanent price impact. Let $S_t = \text{net buy volume} = (\# \text{shares actively purchased} - \# \text{shares actively sold})$. Then:

$$p_t = p_{t-1} + \alpha + \lambda S_t + u_t \quad (7.1)$$

where $\lambda$ is the price impact parameter. We’d expect $\lambda > 0$. Later, when we use this model to make predictions, $S_t$ for future times will include our own trades. But for the moment, think of the model as describing what’s happened in the past.

By rearranging, we obtain

$$\Delta p_t = p_t - p_{t-1} = \alpha + \lambda S_t + u_t \quad (7.2)$$

This looks like a simple linear regression of $\Delta p_t$ on $S_t$. The model is typically estimated over short time intervals (like one minute). By aligning the trade prices with the record of the bids and asks,...
we can compute $S_t$ as the number of executed shares priced at the ask less the number of shares priced at the bid. The model parameters $(\alpha, \lambda, \text{ and } \sigma_u^2)$ can then be estimated using a spreadsheet (or calculator) regression command. Not only is the model convenient to estimate, but it will also turn out to be convenient for solving the order splitting problem.

Of course, models chosen for convenience aren’t necessarily the most realistic. The intuition of the private information model only suggested a generally positive relation between order flow and quote changes. Why should the order flow variable be net buy volume? Why should the relation be linear?

This basic model admits many variations. The extent to which a regression fits the data is usually summarized in its coefficient of determination, $R^2$. This quantity varies between zero (no fit at all) and one (perfect fit). We can investigate alternative specifications and see which changes seem to improve the fit.

The basic specification is linear in signed order flow. This linearity implies that the price change caused by a 10,000 share purchase should be 100 times the impact of a 100 share purchase. In practice it is usually much less.

One way of lessening the impact of large trades is to simply ignore size altogether. Net buys, defined as $S_t' = (\#\text{ of active buys} - \#\text{ of active sells})$, is based on counts. A buy counts as “+1” no matter how many shares are involved.

The use of count-based variables goes against the intuition that a large order should move the price at least somewhat more than a small order. The desired property is an incremental impact that diminishes with size: the additional price impact associated with adding one share to a 10-share order should be less than if one share were added to a 10,000 share order. The declining incremental impact implies a concave relation. At present, the preferred specification involves a signed square root:

$$S_t'' = \text{Sign}(S_t)\sqrt{|S_t|}.$$ 

The $\text{Sign}$ function is defined by $\text{Sign}(x) = +1$ if $x > 0$; $-1$ if $x < 0$; and $0$ if $x = 0$. For example, if $S_t = -10,000$ shares, then $S_t'' = (-1)\sqrt{10,000} = -100$. This form is suggested by supportive evidence from various studies in different markets. It is sometimes called the “square root law” due to its apparent universality (Bouchaud, Farmer, and Lillo, 2008).

Sometimes $S_t, S_t', \text{ and } S_t''$ are used together in the specification (which then becomes a multiple regression). If the goal is fitting a set of data, after all, why not use all reasonable variations?

The problem is that the ultimate goal is not explaining the past, but forecasting the future. Market models are notoriously prone to a phenomenon known as “over-fitting”. (The analyst carefully crafts a statistical model that achieves an “in-sample” $R^2$ of 0.99, but then falls apart when applied to a different sample or used for out-of-sample prediction.)

14.4. Temporary order impacts

All of the action in the model (or the variants discussed) takes place within interval $t$, suggesting that the effect of an order on price is instantaneous. There are a number of reasons why we might want to allow for more flexibility in the timing.
Different market participants receive information in differing timeliness, process information at
different speeds, and trade over different horizons. The business-cycle investor lies at one extreme,
the high-frequency trader at the other. The existence of slow participants opens the possibility that
some of the reaction to new information might involve a gradual price adjustment over time.

The market might also be subject to temporary price movements due to liquidity needs of large
investors. In contrast to the lagged information adjustment effect, though, these effects might well
involve “overshooting” and reversion. A large mutual fund, for example, that experiences a large
unexpected investor outflow might need to trade in size quickly. As the fund’s large sell orders hit
the market, the price is pushed down. Bidders initially suspect negative private information and
react accordingly. But once the selling order flow abates, and it appears that there are no pending
news announcements that could explain the drop, bidders will return and the price will rise.

A simple specification that allows for a temporary impact is:

\[ p_t = p_{t-1} + \lambda_0 S_t + \lambda_1 S_{t-1} + u_t \]  

(7.3)

The total impact coefficient for an order is \( \lambda_0 + \lambda_1 > 0 \). If the process is dominated by lagged
adjustment, then typically \( \lambda_0 > \lambda_1 > 0 \). If there is overshooting, then \( \lambda_1 < 0 \).

### 14.5. Estimating the trade price impact using daily data

The statistical models here are properly estimated with intraday, high-frequency data. These
datasets, however, are large and complex. With some simplifying assumptions, we can form an
estimate from the more common daily price and volume data.

Consider the model given in equ. (7.2). On average, \( \alpha \approx 0 \) and \( u_t \approx 0 \). If we assume that these
approximations hold period-by-period, we can write

\[ \Delta p_t = \lambda S_t \]

and, taking the absolute value of each side,

\[ |\Delta p_t| = \lambda |S_t| \]

Now \( S_t \) is the signed order flow over the interval. If the interval is short, we can approximate its
absolute value by the trading volume, \( V_t \approx |S_t| \). Rearranging, the implied value of \( \lambda \) is:

\[ \lambda = \frac{|\Delta p_t|}{V_t} \]

This suggests estimating \( \lambda \) by

\[ \hat{\lambda} = \left( \frac{|\Delta p_t|}{V_t} \right) \]

where the overbar indicates an average taken over a sample of intervals. Since price and volume
data are widely reported at a daily frequency, it is convenient to use a sample of daily data. This
estimate is a variant of the Amihud illiquidity measure, often denoted \( I \) (Amihud, 2002).
Chapter 15. Order splitting with price impact

A one billion dollar 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.

The TWAP and VWAP strategies (discussed in Chapter 13) 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.”

Participation targets, 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. Commodities 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 have not have behaved so wildly. We now turn to a discussion of algorithms that incorporate these effects.

15.1. Order splitting with pre-trade benchmarks

A typical pre-trade benchmark is the bid-ask midpoint at the time the parent order was released to the trading desk, denoted $BAM_{pre}$. The implementation shortfall for, say, a buy order is then $IS = (average ~ execution ~ price) - BAM_{pre}$. Since the $BAM_{pre}$ doesn’t change over the execution process, however, we can ignore it in the following sense. Any strategy that minimizes the average execution price will also minimize the implementation shortfall.

A representative trading problem in this case involves buying a given number of shares before some deadline in a way that minimizes the overall price. The problem is dynamic because trades and market conditions evolve over time. It is stochastic (that is, driven by random behavior) because we have to plan a strategy before we know the future path of prices. The solutions are achieved by applying principles of stochastic dynamic programming.

Copyright 2014, Joel Hasbrouck. All rights reserved. Draft of May 19, 2014.
The basic approach is due to (Bertsimas and Lo, 1998). To see it at work, suppose that the price dynamics are given by equ. (7.1), and consider the problem of minimizing the cost of buying $S_{Total}$ shares over two periods, (period 0 is “right now”). This cost is simply

$$ C = p_1 S_1 + p_2 S_2 $$

We start by forecasting the prices forward using equation (7.1):

$$ p_1 = \alpha + p_0 + \lambda S_1 + u_1 $$
$$ p_2 = \alpha + p_1 + \lambda S_2 + u_2 $$
$$ = 2\alpha + p_0 + \lambda S_1 + \lambda S_2 + u_1 + u_2 $$

We substitute these forecasts into the cost expression

$$ C = p_1 S_1 + p_2 S_2 = S_1(\alpha + p_0 + \lambda S_1 + u_1) + S_2(2\alpha + p_0 + \lambda S_1 + \lambda S_2 + u_1 + u_2) \quad (8.1) $$

The expected cost is obtained by setting $u_1$ and $u_2$ to zero (as, in expectation, they are):

$$ EC = S_1(\alpha + p_0 + \lambda S_1) + S_2(2\alpha + p_0 + \lambda S_1 + \lambda S_2) \quad (8.2) $$

We want to minimize $EC$ subject to the constraint $S_1 + S_2 = S_{Total}$. We can eliminate this constraint by letting $S_2 = S_{Total} - S_1$. Substituting this into (8.2) gives

$$ EC = \lambda S_1^2 - S_1(\alpha + \lambda S_{Total}) + S_{Total}(2\alpha + p_0 + \lambda S_{Total}) $$

Now the expression involves just $S_1$. The optimal value, $S_1^*$, is obtained by computing the derivative, setting it to zero, and solving:

$$ \frac{dEC}{dS_1} = -\alpha + 2\lambda S_1^* - \lambda S_{Total} = 0 $$

or

$$ S_1^* = \frac{\alpha + \lambda S_{Total}}{2\lambda} \quad \text{and} \quad S_2^* = -\frac{\alpha - \lambda S_{Total}}{2\lambda} $$

In the special case of $\alpha = 0$, $S_1^* = S_2^* = S_{Total}/2$. If $\alpha$ is slightly positive, $S_1^*$ will be a bit larger and $S_2^*$ will be a bit smaller. Essentially, we’re accelerating our purchases, to partially avoid the anticipated price increase. We don’t, however, try to accomplish everything in the first period. Order impact considerations still come into play. If $\alpha$ is very positive, we might initially buy more than we need ($S_1^* > S_{Total}$), and we’ll sell off what we don’t need in period 2.

Using the optimal values in (8.2), the expected cost becomes

$$ EC^* = -\left(\frac{\alpha + \lambda S_{Total}}{4\lambda}\right)^2 + S_{Total}(2\alpha + p_0 + \lambda S_{Total}) $$

Note that this is quadratic. Expected trading cost rises as the square of the total amount purchase. In the special case of $\alpha = 0$
\[ EC^* = p_0S_{Total} + \frac{3\lambda S_{Total}^2}{4} \]

Since we don't know (at time 0) the prices at which the trades will be made, our trading plan involves risk. The total cost from equation (8.1) evaluated at the optimal quantities is:

\[ C^* = \frac{1}{4} S_{Total} (4p_0 + 3\lambda S_{Total} + 4u_1 + 2u_2) \]

All the uncertainty here is due to the \( u_t \). Isolating them gives

\[ C^* = S_{Total} (\cdots + u_1 + \frac{1}{2} u_2) \]

From this, the variance of the trading cost is

\[ Var(C^*) = S_{Total}^2 \left[ Var(u_1) + \frac{1}{4} Var(u_2) \right] = \frac{5}{4} S_{Total}^2 \sigma_u^2 \]

Risk also rises as the square of the total amount traded.

The analysis extends to more than two periods. If we're spreading the purchase over three periods:

\[ S_1^* = \frac{\alpha}{\lambda} + \frac{S_{Total}}{3}, S_2^* = \frac{S_{Total}}{3}, S_3^* = -\frac{\alpha}{\lambda} + \frac{S_{Total}}{3} \]

If \( \alpha = 0 \), the optimal strategy over \( T \) periods is

\[ S_t^* = \frac{S_{Total}}{T} \text{ for } t = 1, ..., T \]

Table 15.1 summarizes the results for expected cost and risk. As \( T \) goes up, expected cost goes down, but risk rises. In portfolio theory there is a fundamental trade-off between risk and return. There is a similar trade-off here. In portfolio theory the trade-off line is called the "efficient portfolio frontier"; here it is sometimes called the efficient trading frontier.
Table 15.1 Trading horizon, expected cost and risk

<table>
<thead>
<tr>
<th>T</th>
<th>( EC^* )</th>
<th>( Var(C^*) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( p_0 S_{Total} + \lambda S_{Total}^2 )</td>
<td>( S_{Total}^2 \sigma_u^2 )</td>
</tr>
<tr>
<td>2</td>
<td>( p_0 S_{Total} + \frac{3\lambda S_{Total}^2}{4} )</td>
<td>( \frac{5}{4} S_{Total}^2 \sigma_u^2 )</td>
</tr>
<tr>
<td>3</td>
<td>( p_0 S_{Total} + \frac{2\lambda S_{Total}^2}{3} )</td>
<td>( \frac{14}{9} S_{Total}^2 \sigma_u^2 )</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Chapter 16. Fees, rebates and other inducements

The institutions of trade are not funded by kindly 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 fairly 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. Perhaps all industries have similarly Byzantine 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.

16.1. 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.

16.1.A. Maker/taker pricing

This occurs with an arrangement known as maker/taker pricing. Under this 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 (Inet) are (or recently were) as follows:

- An executed limit order (“maker”) generally received $0.0029 per share (2.9 “mils” per shares) if the limit order was visible, and $0.0015 (1.5 mils) per share if the limit order was hidden.

- A taker pays $0.0030 (3.0 mil) per share “liquidity removal fee”.

NASDAQ keeps $0.0001 (0.1 mil) per share. 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 “good”.

Maker/taker fees can distort posted prices. Suppose that on Inet 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 2.5 mils, but rises to 2.9 mils 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 (1.7 mils): displayed size is a better advertisement for the exchange. The taker fee is a uniform 2.9 mils.

16.1.B. Inverted pricing

Some exchanges offer inverted maker/taker fees. The BATS-Y exchange charges 0.3 mils for providing liquidity, and pays 0.2 mils 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 0.2 mils then pay 2.9 mils.)

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 2.5 mils. 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 0.3 mils, 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 − 0.3 mils = $9.9997 because quotes on an increment finer that $0.01 aren’t permitted by the sub-penny part of Reg NMS (discussed below).

16.1.C. Routing charges

The SEC has not directed industry to set up a single consolidated access system. When U.S. 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 3 mils).

16.2. Payment for (retail) order flow

Some retail investors believe that the brokers send their orders to exchanges where they interact in some central fashion with all the other buyers and sellers, large and small. Certainly at one point in living memory, this was for many stocks the reality.

Current practice, however, is very different. The 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 7.3 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 information. As in Chapter 9, 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 requires disclosure.

Whereas Rule 605 is aimed at accountability by market centers (see Section 11.6), Rule 606 applies to brokers. They must report what orders they received, what they did with those orders, and other aspects of their disclose 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 (Schwab website, Jan, 2010).

Table 16.1 summarizes the Charles Schwab’s customer orders (“Securities Listed on the NYSE/Network A Eligible Securities”)

<table>
<thead>
<tr>
<th>Non-directed orders as percentage (%) of total customer orders:</th>
<th>95.9%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Market Orders as % of total non-directed orders</td>
<td>35.2%</td>
</tr>
<tr>
<td>2. Limit Orders as % of total non-directed orders</td>
<td>57.3%</td>
</tr>
<tr>
<td>3. Other Orders as % of total non-directed orders</td>
<td>7.5%</td>
</tr>
</tbody>
</table>

Table 16.2 describes where the orders were sent.

<table>
<thead>
<tr>
<th>Venue</th>
<th>% of Non-Directed Order Flow Rec'vd</th>
</tr>
</thead>
<tbody>
<tr>
<td>UBS Securities LLC</td>
<td>95.1%</td>
</tr>
<tr>
<td>INET/NASDAQ</td>
<td>3.9%</td>
</tr>
<tr>
<td>Citadel Derivatives Group LLC</td>
<td>0.7%</td>
</tr>
</tbody>
</table>

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.

Copyright 2014, Joel Hasbrouck. All rights reserved. Draft of May 19, 2014.
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.

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.

Problems

Problem 16.1 Maker/taker pricing for marketable buy orders

The PAX Exchange charges a $0.003 per share access ("taker") fee and pays a ("maker") liquidity rebate of $0.002 per share. How do these fees/rebates apply to an execution in which Tae’s limit order (Sell 100 shares, limit $25.00) is hit by Sam’s order (Buy 100 shares, limit $25.00)?

Problem 16.2 Taker/maker pricing for marketable buy orders

(Continuation) For each execution, the ZAP Exchange charges a per share maker fee of $0.0029, and rebates $0.0018 to the active (marketable) order. On PAX, the offer is $25.00 for 10,000 shares, and the book at ZAP is empty. Vanessa enters an order on ZAP to sell 100 shares, limit 25.00. How do ZAP’s fees/rebates apply if Sam’s buy order is directed to ZAP instead of PAX?
Answers

Answer to Problem 16.1

Tae receives the rebate, $100 \times 0.002 = $0.20$, which brings the amount she receives from the sale to $2,500.20$. Sam pays the access fee, bringing his net payment to $2,500.30$. PAX keeps the $0.10$.

Answer to Problem 16.2

Since Vanessa’s order is the first in an empty book, she’s at the front of ZAP’s offer queue. Sam’s order executes against Vanessa’s. Vanessa pays maker fee, so the net amount she receives from the sale is $2,500 - 0.29 = 2,499.71$; Sam pays $2,500 - 0.18 = 2,499.82$. ZAP keeps $0.11$.
Chapter 17. Reg NMS

The SEC’s Regulation NMS established the framework governing 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 the shape of the market to come.

“NMS” stands for “National Market System”. In 1975, Congress passed a major securities act that directed the SEC to facilitate development of a national market system (one comprehensive market for 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. Reg NMS was positioned and promoted as a culmination of the 1975 Act.

17.1. Guiding principles and background

17.1.A. Market competition vs. order competition.

As a general principle, a market is most efficient (produces the most reliable prices, resists manipulation) when participation is deep. Liquidity is a network externality, i.e., something valuable that’s generated by connectedness.

Is a stock market a natural monopoly? If the consolidation of buyers and sellers (achieved in an opening auction, for example) is a good thing, perhaps we should adopt one single system. The point has merit, but over the years the SEC often wrestled with the NYSE and NASDAQ over anti-competitive practices. Some examples:

- NYSE Rule 390 (repealed in 2000) prevented member firms from trading except on the NYSE.
- NYSE Rule 500 made it extremely difficult for a company to voluntarily delist from the NYSE.
- In 1990’s, NASDAQ market makers were found to be engaging in widespread collusion to keep the spreads at least $0.25.

From about 1990 onwards, the SEC had allowed many new electronic markets to start up, eventually bringing them under the regulatory umbrella of “Automated Trading Systems”. Many of these markets were cheaper than the NYSE and NASDAQ, and they worked well. Clearly, competition among markets was bringing many benefits.

A consistent thread running through all SEC market regulation (including Reg NMS) has been balancing order and market competition. This is also known as the fragmentation vs. consolidation debate.

17.1.B. Long-term investors vs. short-term traders

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.” (Reg NMS, final rule, p. 37500).

We now turn to the individual rules. There are four main parts: the order protection rule; the access rule; the subpenny pricing rule; and market data rules. (The regulation also mandated a consolidation and renumbering of preexisting rules, but this was mainly a legal formality.)

At first glance, the rules seem technical. In fact, they reflect and promote some fundamental principles.

17.2. The order protection rule

A trade-through is an execution outside of the market 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.

The workings of this rule are discussed at length in Chapter 4. But for present purposes, several other aspects of the rule are worth noting.

The order protection rule was the most contested part of Reg NMS. It received the most 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: The rule isn’t needed. Brokers, in fulfilling their fiduciary duty of best execution, will ensure that their customers will always get the best price, and so trade-throughs won’t occur. If brokers don’t do this, the customer can always get a new broker. In any event, the problem is a matter between the broker and the customer, not something that needs to be addressed by federal statute.

The counter: Investors need protection. They lack the means to monitor what their brokers are doing. Further, the brokers themselves may not know at any given instant which particular market has the best quote.

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.

17.3. 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 3 mils ($0.003) per share (see Section 11.7).
17.4. 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 (cf. “Common Sense”) 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.

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.

As noted above, however, the effectiveness of the sub-penny rule has been partially undone by sub-penny maker/taker fees.

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 executed at $25.005. This is permissible.

17.5. 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 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.
17.6. Summary

Reg NMS provided the legal framework for competition among market centers. 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.)
Chapter 18. The arithmetic of probabilities

Financial markets continually generate information that can help us make trading decisions. The information, though, hardly ever implies direct and obvious recommendations that we can follow with complete certainty. Instead we find ourselves weighing probabilities of outcomes. Each new pieces of information shifts our beliefs a bit. Our inferences are of the form, “Given that x just happened, what is stock z worth?” In real world markets, x might be press release from company z, but often it is simply a purely market-related event, such as a trade or the narrowing of the bid-ask spread. We draw conclusions based on the manipulation of probabilities, expectations, and their conditional counterparts.

This chapter is a precursor for the material that follows. If you’ve studied statistics, you can skim or even skip it. You can always come back to it if subsequent developments turn out to be confusing.

18.1. Probabilities and conditional probabilities

An example illustrates the principles. Suppose that we’re offered the opportunity to buy a painting attributed to a famous artist “Roth Jackson”. A collector advises us that works attributed to Jackson are sometimes found to be forgeries. The only way, he says, to be completely sure is to perform a detailed chemical analysis.

The chemical analysis is extremely expensive. There is, however, an expert who knows Jackson’s style, and who can tell at a glance if the work is real. Is the expert accurate? It turns out that in a court case where his opinion on a large number of pieces taken from the gallery was subsequently verified by chemical analysis, he was correct most of the time. To be more precise, when the expert said that the painting was real, 90% of the time he was right. While this falls short of perfect accuracy, we reason, it is surely much better than we’d do guessing on our own. We are somewhat reassured.

The probability of something is often written $P(\text{something})$. We’ll use real and forged (in italics) to label the actual condition of the painting, so $P(\text{real})$ denotes the probability that the painting is in fact real. The expert’s opinion is labeled as a quote, either “real” or “forged”.

The expert’s track record is given to us as a conditional probability: given that expert said “real,” 90% of the time the painting was real. A conditional probability is written with a vertical bar: $P(\text{real} | \text{real}) = 90\%$. The order matters: the condition (the “given”) appears after the bar, and the “outcome” appears before.

Now a 90% accuracy might seem very satisfactory, particularly if the opinion will allow us to avoid the expensive verification. So we resolve to consult the expert. But as an afterthought we ask the collector, “Well, just how serious is the forgery problem?” The collector replies, “Very serious. In fact, one out ten Jackson’s offered for sale is a forgery.”

Okay, if the expert say “real”, there’s a 90% chance that it is in fact real. But if we don’t consult the expert and just blindly purchase, the chance we’ll get a real painting nine out ten times (that is, 90%) — the same as without the expert’s opinion.

How can the expert’s opinion seem accurate and convincing from one perspective, and useless from another? The answer turns on conditional probabilities. As evidence of the expert’s accuracy, we’re given $P(\text{real} | \text{real}) = 0.90$. It will turn out, though, that a more meaningful measure would be $P(\text{real} | \text{real})$. 

Copyright 2014, Joel Hasbrouck. All rights reserved. Draft of May 19, 2014.
To see this, suppose that whenever he’s presented with a painting the expert just flips a coin, and if it comes up heads he says that the painting is “real”. That is, \( P(\text{"real"}) = 1/2\). Since the actual state of the painting doesn’t affect the expert’s judgment, conditioning on the actual state doesn’t affect the probability: \( P(\text{"real"}|\text{real}) = P(\text{"real"}|\text{forged}) = 1/2\). We can depict the sequence of events as a tree.

![Figure 18.1](image)

The numbers in parentheses are probabilities. We start with a painting. There is a 0.90 probability that the painting is real, and 0.10 probability that it forged. (If you like, think of starting with 100 paintings, of which we’d expect 90 to be real.) The probability of the painting being real and being judged “real” is

\[
P(\text{real},\text{"real"}) = P(\text{real}) \times P(\text{"real"}|\text{real}) = 0.90 \times 0.50 = 0.45
\]

The left-hand side is a joint probability, the probability that the painting is real AND the expert says “real”. In a joint probability the order doesn’t matter: \( P(\text{real},\text{"real"}) = P(\text{"real"},\text{real})\). Joint and conditional probabilities are linked by Bayes Rule. For two events \( A \) and \( B \):

\[
P(A|B) = \frac{P(A,B)}{P(B)}
\]

“The conditional probability is the joint probability divided by the total probability (of the conditioning event).”

Our conjecture that the expert is flipping a coin is completely consistent with the accuracy claim:

\[
P(\text{real}|\text{"real"}) = \frac{P(\text{real},\text{"real"})}{P(\text{"real"})} = \frac{0.45}{0.45 + 0.05} = 0.90
\]

Note that there are two paths that lead to a painting being judged “real”. 45% of the time, a real painting is judged “real” and 5% of the time, a forgery is judged “real”.

Now suppose that the expert thinks that flipping a coin is too much work. He decides to simply say that any painting he’s presented with is “real”. It is still true that \( P(\text{real}|\text{"real"}) = 0.90\). (Verify this by revising probabilities in the figure.)
The expert probably realizes, of course, that he needs to vary his conclusions to give the appearance of some judgment at work. So he pretends to be deep in thought and flips a modified coin (or uses some other random outcome) that comes up heads ("real") 90% of the time. That is, 
\[ P(\text{"real"}) = P(\text{"real"}|\text{real}) = P(\text{"real"}|\text{forged}) = 0.90. \]

This might be more convincing in that he’ll call 90% of the paintings “real”, but it doesn’t help us. The problem is not that he’s using the wrong probabilities to generate an opinion. The problem is that the outcome of any game of this sort is independent of whether the painting is in fact real or forged.

If the expert does in fact possess some proficiency, this will show up in a conditional probability that changes when he renders an opinion, that is, 
\[ P(\text{real}|\text{real}) > P(\text{real}) = 0.90. \]
For example, suppose that the expert’s judgment is, if not perfect, at least partially meaningful. When the painting is real, he says "real" 80% of the time, and when the painting is forged, he says “forgery” 80% of the time. That is, 
\[ P(\text{"real"}|\text{real}) = P(\text{"forged"}|\text{forged}) = 0.80. \]
Now the event tree looks like:

![Figure 18.2](image)

Using these new numbers in Bayes’ Rule,
\[
P(\text{real}|\text{"real"}) = \frac{P(\text{real}, \text{"real"})}{P(\text{"real"})} = \frac{0.72}{0.72 + 0.02} = 0.973
\]
That is, if the expert says that the painting is "real", we’ve learned a bit; we’re a little more certain of the painting’s authenticity. Of course, if the painting isn’t real, it must be a forgery, so 
\[
P(\text{forged}|\text{"real"}) = 1 - P(\text{real}|\text{"real"}) = 0.027
\]
Similarly, if the expert says “forged”, the probability that the painting is actually forged is
\[
P(\text{forged}|\text{"forged"}) = \frac{P(\text{forged}, \text{"forged"})}{P(\text{"forged"})} = \frac{0.08}{0.18 + 0.08} = 0.308
\]
Does this last calculation strike you as a bit off? How could listening to someone who is right 80% of the time lead us to correctly identify a painting in the market as a forgery only 31% of the time? Isn’t this worse than we’d expect from coin flipping? The numbers are correct, but there is a powerful psychological bias at work, which we’ll consider in section 18.4.

18.2. Expectations and conditional expectations

In financial markets, different outcomes are often associated with differences in values or profits. Suppose that a real Jackson painting is worth $10,000, while a fake is worth $1,000. Without knowing anything beyond the fact that 10% of the paintings are forgeries, we expect

$$ E[\text{Value}] = 10,000 \times P(\text{real}) + 1,000 \times P(\text{forged}) = 9,100. $$

The $E[\text{...}]$ notation is read “expectation of”. It is the average of the values in each outcome, weighted by the likelihoods of the outcomes. The simple average of $10,000 and $1,000 is $5,500. The expected value is closer to $10,000 because it is more likely than not that the painting is real. $E[\text{Value}]$ is called an unconditional expectation because it looks at all the possibilities using the initial (“prior”) probabilities.

If an expert who’s right 80% of the time said “real”, what would the painting be worth? Not $10,000, because we’re still not completely sure that the painting is real. But it would certainly be worth a bit more than the unconditional expectation of $9,100. More precisely,

$$ E[\text{Value} | \text{"real"}] = 10,000 \times P(\text{real} | \text{"real"}) + 1,000 \times P(\text{forged} | \text{"real"}) $$

$$ = 10,000 \times 0.973 + 1,000 \times 0.027 = 9,756.76. $$

A similar calculation shows that $E[\text{Value} | \text{"forged"}] = 7,230.77$. This may seem high, but remember that even if the expert says “forged,” there is a good chance that the painting is real.

18.3. Conditioning on information and strategy

Suppose that the painting is being offered to us at $8,500.

If the seller knows the painting to be real, they’re selling it at a loss. If they know the painting is forged, then an offer price of $8,500 would impose a loss on the buyer. Even if the seller knows nothing more than we do about the authenticity, the price is still a bit low. A fair price would be $9,100.

In any trading situation, it is always good practice to think about your counterparty. Sometimes this consideration leads us to simply walk away from a deal, or make a counteroffer. Sometimes there are special considerations that can explain an otherwise dubious price. In this case, suppose that the painting was left in the attic of a house, the new owner is overwhelmed by the task of moving in, and simply prices the painting in the hope of a quick sale.

Now if we buy for $8,500 something that is really worth $9,100, we’ve made a $600 expected profit. The profit is notional or “paper” because it isn’t really a cash profit until we’ve managed to sell the painting at $9,100, but we’ll ignore this detail.

Should we call in the expert? Suppose that the examination fee is $100. We might reason that the expert just adds another cost to the deal, reducing our profit from $600 to $500.
This isn’t quite right, though, because the expert’s opinion might change our decision. If the expert says “real”, we’ll buy the painting and our profit will be $9,756.76 − $8,500 − $100 = $1,156.76. If the expert says “forged,” though, our expectation of the painting’s value is $E[\text{value} | f] = 7,230.77. Since this is below the offering price, we wouldn’t make the purchase. Our profit would be −$100 (a loss because we’re still out the $100 expert’s fee).

In summary, if we call in the expert, our expected profit is

$$E[\text{profit}] = 1,156.76 \times P(\text{"real"}) + (-100) \times P(\text{"forged"})$$

$$= 1,156.76 \times (0.72 + 0.02) + (-100) \times (0.18 + 0.08) = 830$$

If we don’t call in the expert, our expected profit is only $600, so it pays to call in the expert.

The point here is that conditional probabilities and expectations aren’t simply determined by some statistical assumptions working away in the background of a situation. They might also depend on the actions taken by you and other players.

18.4. The psychology of conditional probabilities

In this example, with an expert who is right 80% of the time, the probability that the painting is forged given that the expert says “forged” is $P(\text{forged} | \text{"forged"}) = 0.308$. Were you expecting something more like $P(\text{forged} | \text{"forged"}) = 0.80$?

If so, you are not alone. In a situation like this, we tend to focus on the expert’s accuracy, recalling that $P(\text{"forged"} | \text{forged}) = 0.80$. We forget that the probability of a forgery in the full sample is relatively low to being with: $P(\text{forged}) = 0.10$.

In the language of statistics, the $P(\text{forged}) = 0.10$ is termed a prior belief. $P(\text{forged} | \text{"forged"}) = 0.308$ is an updated or posterior belief. The posterior probability strikes us as too low because we aren’t giving enough weight to the prior probability. This is an illustration of a well-established psychological tendency to “underweight the prior” (also known as the base rate fallacy).

The taxicab problem is a classic illustration. As stated by (Kahneman, 2011):

A cab was involved in a hit-and-run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data:

- 85% of the cabs in the city are Green and 15% are Blue.
- A witness identified the cab as a Blue cab. The court tested his ability to identify cabs under the appropriate visibility conditions. When presented with a sample of cabs (half of which were Blue and half of which were Green) the witness made correct identifications in 80% of the cases and erred in 20% of the cases.

Question: What is the probability that the cab involved in the accident was Blue rather than Green?

Kahneman notes that the correct answer is 41%, but the most common survey response is 80%.

Psychologists have identified other cognitive biases.

- Over-weighting recent experience. A common question is a basic statistics course goes something like this: “Tosses of a fair coin have just come up heads 10 times in a row. What is
the chance that the next toss will come up heads?” The correct answer is, of course, 50%.
Most people will lean toward a higher number, generalizing from their recent experience.
• Representativeness heuristic. People often generalize from small samples.
• Overconfidence. People overestimate the accuracy of their judgments.

Kahneman discusses these and other tendencies.
References


Copyright 2014, Joel Hasbrouck. All rights reserved. Draft of May 19, 2014.

[https://www.instinet.com/includes/pdf/exchange/Euronext_Amsterdam_.pdf].

Instinet, 2010b. Euronext Lisbon.  
[https://www.instinet.com/includes/pdf/exchange/Euronext_Lisbon_.pdf].


Tel Aviv Stock Exchange, 2013. Trading Schedule.  

