PART

One

Market Microstructure

CHAPTER 1

Financial Markets: Traders, Orders, and Systems

This chapter describes a big picture of financial markets: who the traders are, what types of orders can be submitted, how these orders are processed, how prices are formed, and how markets are organized.

TRADERS

Let us start with the people who trade. They are called (well, you guessed it) *traders*. Those who trade for their own money (or their employer's money) are *proprietary traders*. Their ultimate goal is to make profits by buying low and selling high, whether it is long-term investment or day trading. Other traders execute orders for their clients. They are called *brokers* or *agency traders*. To denote the institutional character of a broker, the term *brokerage (firm)* is also used. For brokers, profits from trading may not be important since they receive commissions for trading and other services from their clients. Typical brokerage services include matching the clients' buy and sell orders, connecting to markets, *clearing* and *settlement*, providing market data and research, and offering credit. Most of the listed services are self-explanatory, but clearing and settlement may need some elucidation. Settlement is delivery of traded assets to the trading counterparts (buyers and sellers). The trading process (sometimes called the *transaction*) does not occur instantly. For example, settlement in the spot foreign exchange for most currencies takes two business days. Clearing denotes all brokerage actions that ensure settlement according to the market rules. These include reporting, credit management, tax handling, and so on.

The institutions that trade for investing and asset management (pension funds, mutual funds, money managers, etc.) are called the *buy-side*. The *sell-side* provides trading services to the buy-side. Besides brokers, the sell-side

includes *dealers* who buy and sell securities upon their clients' requests. In contrast to brokers, dealers trade for their own accounts. Hence, they have a business model of proprietary traders. Namely, dealers make profits by selling an asset at a price higher than the price at which they simultaneously buy the same asset.[1] Providing an option to buy and sell an asset simultaneously, dealers are *market makers* who supply *liquidity* to the market (see more about liquidity below). Traders who trade with market makers are sometimes called *takers*. Many sell-side firms have brokerage services and are called *broker-dealers*.

Harris (2002) provides a detailed taxonomy of various trader types. Here I offer a somewhat simplified classification. There are two major groups: (1) *profit-motivated traders* and (2) *utilitarian traders*. Profit-motivated traders trade only when they rationally expect to profit from trading. Utilitarian traders trade if they expect some additional benefits besides (and sometimes even instead of) profits. Investors who trade for managing their cash flows are the typical example of utilitarian traders. Indeed, when an investor sells (part of) his equity portfolio to get cash for buying a house, or invests part of his income on a periodic schedule, his trades may be not optimal in the eyes of pure profit-motivated traders. *Hedgers* are another type of utilitarian traders. The goal of hedging is to reduce the risk of owning a risky asset. A typical example is buying put options for hedging equities. Put options allow the investor to sell stocks at a fixed price.[2] The immediate expenses of buying options may be perceived as a loss. Yet these expenses can protect the investor from much higher losses in case of falling stock price. In the economic literature, utilitarian traders are often called *liquidity traders* to emphasize that they consume the liquidity that is provided by market makers.

Profit-motivated traders can be partitioned into *informed traders*, *technical traders*, and dealers.[3] Informed traders base their trading decisions on information on the asset fundamental value. They buy an asset if they believe it is underpriced in respect to the fundamental value and sell if the asset is overpriced. Since buying/selling pressure causes prices to increase/decrease, informed traders move the asset price toward its fundamental value. Traders who conduct thorough fundamental analysis of the asset values, such as the company's profits, cash flow, and so on, are called *value investors* (Graham & Dodd 2006). Note that fundamental values do not always tell the entire story. New information that comes in the form of unexpected news (e.g., discovery of a new technology, introducing a new product by a competitor, CEO resignation, or a serious accident, etc.) can abruptly challenge the asset price expectations. Also, estimates of the fundamental value of an asset may vary across different markets. Traders who explore these differences are called *arbitrageurs* (see Chapter 11).

Technical traders believe that the information necessary for trading decisions is incorporated into price dynamics. Namely, technical traders use multiple patterns described for historical market data for forecasting future price direction (see Chapter 10).

As it was indicated above, dealers (market makers) supply liquidity to other traders. In some markets, traders who are registered as dealers receive various privileges, such as exclusive handling of particular securities, lower market access fees, and so on. In return, dealers are required to always provide at least a minimum number of securities (in which they make the market) for buying and selling. Dealers make profits from the difference between the selling price and buying price that they establish. This implies that there are takers in the market who are willing to buy a security at a price higher than the price at which they can immediately sell this security. It seems like easy money providing that the price does not change and there are equal flows of buying and selling orders. Obviously, there is always a risk that dealers have to replenish their inventory by buying security at a price higher than the price they sold this security in the near past. This may be caused by a sudden spike in demand caused by either informed or liquidity traders. Similarly, dealers' loss may ensue when takers exert selling pressure. We shall return to the dealers' costs in Chapters 3 and 4.

ORDERS

When traders decide to make a trade, they submit *orders* to their brokers. Order specifies the trading instrument, its quantity (*size*), market side (buy or sell), price, and some other conditions (if any) that must be met for conducting a trade. When orders find their counterparts in the markets, a transaction occurs and it is said that orders are *matched* (or *filled*). Orders are submitted upon some market conditions. If these conditions change before an order is matched, the trader can cancel and possibly resubmit an order with other properties. Here the *latency* problem may become important. Traders do not receive confirmations of their trades instantly. If a trader attempts to cancel the order while it is being transacted, the cancellation fails.

There are two major order types: *market orders* and *limit orders*. Price is not specified for market orders and these orders are executed at the best price available at the order arrival time (i.e., a bid/offer order is filled at the current best offer/bid price). Limit buy and sell orders are *quoted* in the market with their *bid* and *ask* (or *offer*) prices, respectively. Prices of limit orders are sometimes called *reservation prices*. The highest bid and lowest ask (offer) currently present in the market are called *best bid* and *best ask* (*offer*), respectively. The difference between the best ask and the best bid is

the *bid/ask spread* (see the next section). The *bid/ask bounce* of transaction prices is caused by trades randomly initiated by buy and sell orders. As a result, sequential transaction prices fluctuate between the best ask and best bid prices. It is said that any price within the spread is *inside the market*. The half-sum of the best bid and best ask is called *mid-point price* (or just *mid-price*).

Limit orders specify the worst price (highest offer or lowest bid) at which traders agree to trade. If a better price is available for matching the limit order at the time of its arrival, the transaction is done at the better price. Limit orders are not guaranteed to be executed. For example, a limit buy order is placed below the best bid but the price does not fall that low. Limit orders that are not immediately filled are stored in the *limit order book* (LOB) until they are matched or cancelled (see more about the LOB below). It is important to remember that the aggregated order size at any price in the LOB is finite. Hence, a large market order may wipe out the entire LOB inventory at best price and get filled not at a single price but within some price range. As a result, the best price worsens—at least temporarily. It is said that large orders have *market impact*. Some markets do not permit market orders and limit orders are the only option.

In some markets, all limit orders are automatically cancelled at the end of trading day. To prevent such a cancellation, an option known as *good-till-cancelled* may be available. Usually, such an option has a limited duration (e.g., one month).

For a trader, the choice between a limit order and market order (when the latter order is permitted) may be non-trivial. For example, a trader assuming a *long position* can submit a market buy order and fill an order at current best offer price. In other words, the trader becomes a taker. Another option is to submit a limit order at a current best bid (or even at a lower) price—that is, become a maker. It is said that takers pay the spread, which is the price for immediacy. Indeed, there is a risk that the price will move in an adverse direction and the maker order will not be executed within the acceptable time horizon. We shall return to this problem in Chapter 13. As I have indicated above, taker order is always filled at the best available price. Namely, a bid/ask submitted with a price higher/lower than the best ask/bid is still filled at the current best ask/bid. In general, limit orders submitted across the market are called *marketable limit orders*. Why would anyone submit such an order? This may happen if a trader wants to make a sure shot with at least partial filling but not to pay beyond some limit price.

Some markets permit *hidden limit orders*. These orders have a lower priority in respect to visible orders at the same price but higher priority than the limit orders with worse price. Sometimes orders can be partially hidden. In the latter case, when the visible order part is filled, it is

replenished with the hidden amount and the order position in the LOB is preserved.

Cancel and replace limit orders allow traders to change the order size without losing the order position in the LOB.

Limit orders can be *pegged* in some markets. There are three ways to define pegged order. The first definition involves primary (market)—peg to the best price on the same (opposite) side of the market. Also, orders can be pegged to the bid/ask mid-price. The price of unfilled (e.g., due to latency) pegged orders moves along with their peg.

Some markets have an option to submit *market-on-open* and *market-on-close* orders. These orders are submitted in advance for executing at a new market opening and closing, respectively.

Stop orders can be treated as limit orders since they, too, specify the execution price. However, price has a different role in stop orders. It constrains possible loss rather than yields the realized profit. Indeed, a trader sells an instrument using a limit order at a higher price for locking in the profit after buying an instrument at a lower price. On the contrary, the sell stop order is filled when price falls to (or below) the order price. Hence, traders submit stop orders for mitigating the risk of possible adverse price moves.

Some other instructions may be provided with orders. *Fill-or-kill orders* are filled at their arrival in the market. Any portion of such an order that cannot be immediately filled is canceled. Another constraint is used in the *all-or-none orders*: These orders can be filled completely, or not at all.

So far, selling implies that the trader owns the selling asset (i.e., has a long position in it). *Short selling*, or acquiring a *short position*, means that the trader borrows an asset from his broker and sells it. This makes sense if there is expectation that the asset price will fall. Then the trader buys the same asset (presumably) at a lower price for returning it to the broker and pockets the difference. Two special order types are used to implement this strategy: *sell short* and *buy to cover*. Note that the so-called *uptick rule* forbids short selling unless a short order is submitted either at a price above the last traded price, or at the last traded price if that price was higher than the price in the previous trade. In the United States, the uptick rule was in effect for many years until it was canceled in 2007. However, discussions on the necessity of this rule resumed in 2009 in the context of introducing a stricter regulation of financial markets.

THE BID/ASK SPREAD

The size of the bid/ask spread is an important object of the microstructure theory, which shall be addressed in future chapters. Here is a list of common definitions and components of the spread (de Jong & Rindi 2009).

The *quoted spread* between ask A_t and bid B_t that is averaged over T periods equals

$$S^Q = \frac{1}{T}\sum_{t=1}^{T}(A_t - B_t) \tag{1.1}$$

In terms of the asset fundamental price, P_t^*, the averaged spread is

$$S = \frac{1}{T}\sum_{t=1}^{T}2q_t(P_t - P_t^*) \tag{1.2}$$

where q_t is 1 for buy orders and -1 for sell orders. Since the value of P_t^* is not observable, the *effective spread* in terms of mid-price $M_t = 0.5(A_t + B_t)$ is usually used:

$$S^E = \frac{1}{T}\sum_{t=1}^{T}2q_t(P_t - M_t) \tag{1.3}$$

Sometimes the *realized spread* is applied in post trade analysis:

$$S^R = \frac{1}{T}\sum_{t=1}^{T}2q_t(P_t - M_{t+1}) \tag{1.4}$$

It was already indicated that the bid/ask spread from the point of the view of a taker is the price for immediacy of trading. Now, let's examine the main components of the bid/ask spread, which are determined by dealers (makers).

First, the spread incorporates the dealers' operational costs, such as trading system development and maintenance, clearing and settlement, and so on. Indeed, if dealers are not compensated for their expenses, there is no rationale for them to stay in this business.

Dealer's inventory costs, too, contribute into the bid/ask spread. Since dealers must satisfy order flows on both sides of the market, they maintain inventories of risky (and sometimes undesirable) instruments. The inventory microstructure models will be discussed in Chapter 3. Glosten & Harris (1988) combine the operational and inventory costs into a single *transitory* component, since their effect on security price dynamics is unrelated to the security value. Another spread component reflects the dealer's risk of trading with counterparts who have superior information about true security value. Informed traders trade at one side of the market and may profit from trading with dealers. Hence, dealers must recover their potential losses by widening the spread. Not surprisingly, dealers pass these losses to uninformed traders.[4] This component of the bid/ask spread is called the *adverse-selection* component since dealers confront one-sided selection of their order flow. The adverse selection will be discussed in Chapter 4.

LIQUIDITY

Liquidity is a notion that is widely used in finance, yet it has no strict definition and in fact may have different meanings. Generally, the term *liquid asset* implies that it can be quickly and cheaply sold for cash. Hence, cash itself (i.e., money) is the ideally liquid asset. Real estate and antique on the other hand are not very liquid.

In the context of trading, liquidity characterizes the ability to trade an instrument without notable change of its price. A popular saying defines liquidity as the market's *breadth, depth, and resiliency*. First, this implies that the buying price and selling price of a liquid instrument are close, that is, the bid/ask spread is small. In a deep market, there are many orders from multiple makers, so that order cancellations and transactions do not affect notably the total order inventory available for trading. Finally, market resiliency means that if some liquidity loss does occur, it is quickly replenished by market makers. In other words, market impact has a temporary character. As we shall see in Chapter 13, analysis of market impact dynamics is a rather complex problem.

Various liquidity measures are used in different markets. For example, Xetra's (the European electronic trading system) liquidity measure corresponds to the relative market impact costs for the so-called *round trip* (simultaneous buying and selling of a position) for a given order size (Gomber & Schweickert 2002). Barclays Capital derives the FX liquidity index using the notional amounts traded for a fixed set of FX spreads and aggregated using a weighting by currency pair (quoted by Bank of England 2009). Sometimes inverse liquidity (*illiquidity*), based on the price impact caused by trading volume, is used (Amihud 2002):

$$\text{ILLIQ} = \frac{1}{N}\sum_{k=1}^{N} |r_k|/V_k \tag{1.5}$$

In (1.5), r_k and V_k are the return and trading volume for time interval k. The notion of liquidity is rooted in the Kyle's model (1985), which will be discussed in Chapter 4.

MARKET STRUCTURES

Markets differ in their organization and trading rules. Some markets that are highly organized and regulated by government agencies are called *exchanges* (or *bourses*). In the United States, the trading of stocks, bonds, and several other securities is regulated by the *Securities and Exchanges Commission* (SEC). However, trading of commodities (including spot,

futures, and options) is regulated by another government agency, the *Commodity Futures Trading Commission* (CFTC).

Historically, exchanges were founded by their members (dealers, brokers) for trading among themselves. In our days, many exchanges have become incorporated. Still, in most cases only members can trade at exchanges. An alternative to exchanges is the *over-the-counter* (OTC) *markets* where dealers and brokers can trade directly.

Market structure is defined with specifics of execution systems and with the type of *trading sessions* (Harris 2002; de Jong & Rindi 2009). There are two major execution systems: *order-driven markets* and *quote-driven markets*. In terms of trading sessions, order-driven markets can be partitioned into *continuous markets* and *call markets*. Many order-driven markets are *auctions* in which trading rules ensure that trading occurs at the highest price a buyer is willing to pay and at the lowest price a seller is willing to sell at. The process of defining such a price is called *price discovery* (or *market clearing*).

Another form of order-driven markets is *crossing networks*. Price discovery is not implemented in crossing networks. Instead, prices used in matching are derived from other (primary) markets. Hence, the term *derivative pricing rules* is used.[5] Orders submitted to crossing networks have no price and are prioritized according to their arrival time. The first advantage of crossing networks is that trading in these systems is (or at least is supposed to be) completely confidential. Another advantage is that trading there does not have an impact on price in the primary markets. Hence, crossing networks are attractive to those traders who trade orders of large size (*blocks*). On the other hand, trading in the dark may encounter significant *order imbalance*. It is usually calculated as a difference between aggregated demand and aggregated supply and is also called *excess demand*. As a result, the portion of filled orders (*fill ratio*) may be rather small. Another drawback of crossing networks is the imperfection of the derivative pricing that may be subject of manipulations. More specifics on different market structures will be detailed in the following sections.

Continuous Order-Driven Markets

In continuous markets, traders can submit their orders at any time while the market is open. Trading hours vary in different markets. For example, the New York Stock Exchange (NYSE) and NASDAQ are open on Monday through Friday; they start at 9:30 A.M. EST and close at 4:00 P.M. EST. On the other hand, the global FX spot market is open around the clock during the workdays and is closed only for a few hours on weekends (see more on FX markets in Chapter 2).

In order-driven markets, traders trade among themselves without intermediary market makers (dealers). In other words, every trader can become a market maker by placing a limit order. Those limit orders that are not immediately matched upon arrival are entered into the LOB according to the price-time priorities. Price priority has the primary precedence, which means that an order with a better (or more aggressive) price is placed before orders with worse prices.

Time priority means that a new order is placed behind the orders that have the same price and entered the market earlier. Matching of a new taker order with maker orders present in the LOB occurs upon the *First In, First Out* (FIFO) principle—that is, older maker orders are filled first. It is said that the first order with the best price is on *top of the order book*. Hence, the higher/lower the bid/offer order price is, the closer this order is to the top of the LOB.

In some markets, size precedence rule is used. Sometimes the largest order is executed in case of a parity of several orders, but sometimes the priority is given to the smallest order. Another option is *pro rata* allocation. Say the aggregated bid size exceeds the aggregated offer size. Then all bid orders are partially filled proportionally to their size.

Consider a few examples of matching in an order-driven market. Let the LOB have the following bids:[6] B1 –100@10.25 (best bid), B2 – 200@10.23, and the following offers: O1 – 200@10.30 (best offer), O2 – 200@10.35. A market buy order of a size less or equal to 200 will be filled at the price $P = 10.30$. If the size of the market buy order equals 200, it completely matches O1 and the bid/ask spread increases from $s = 10.30 - 10.25 = 0.05$ to $s = 10.35 - 10.25 = 0.1$.

A market buy order of size 300 will be matched as 200@10.30 + 100@10.35. What if you want to buy 500 units, which is higher than the entire offer inventory? Then you can submit a limit order that will be stored in the LOB until a new seller decides to match it. For example, you may want to submit a bid of 500@10.35. This would result in the immediate matching of 300 units and the remaining 200 units becoming the new best bid.

If a bid is submitted inside the market, that is, $10.25 < P < 10.30$, it is placed before B1 and becomes the new best bid. The bid/ask spread then decreases from $s = 0.05$ to $s = 10.30 - P$. If a bid is submitted with a price in the range $10.23 < P \leq 10.25$, it is placed in the LOB between the bids B1 and B2. Finally, a bid with a price $P \leq 10.23$ is placed behind the bid B2.

Oral Auctions

In *oral auctions* (or *open-outcry auctions*), traders (brokers and dealers) gather in the same place (floor market). Traders are required to

communicate (using shouting and hand signals) their trading intentions and the results of trading to all market participants. This ensures great transparency of the trading process.

In oral auctions, order precedence rules and price discovery are similar to those in continuous order-driven markets. However, there may be some additional secondary precedence rules. In particular, public order precedence gives priority to public traders in respect to floor traders.

Open-outcry auctions used to be the main trading venue in the past. In our days, most of floor markets have deployed electronic trading systems that have replaced or are used along with open outcry.

Call Auctions

In call auctions, trading occurs at predetermined moments in time. Call auctions can be conducted several times a day (so-called fixings) or at the openings and closings of continuous sessions. Orders submitted for a given call are batched and executed simultaneously at the same price. Prior to auction, all submitted orders are placed according to the price-time precedence rules. Aggregated demand and supply are calculated implying that a trader willing to buy/sell at price P will also buy/sell at a price lower/higher than P. The auction price is defined in such a way that yields a maximum aggregated size of matched orders. Consider an example of price discovery in a call market with the orders listed in Table 1.1.

The maximum trading volume here corresponds to the price of 10.80 (see Table 1.2). The aggregated supply at this price (with a size of 800) is matched completely. However, part of the aggregated demand ($900 - 800 = 100$), is not filled within the current call. Since order B5 is the last in the list of buyers involved in this fixing, it is filled only partially.

TABLE 1.1 An Example of the Pre-Auction Order Book

Buyers			Sellers	
Order	Size	Order Price	Order	Size
		9.95	S1	700
B1	100	9.90	S2	300
B2	200	9.85	S3	400
		9.85	S4	200
		9.85	S5	100
B3	600	9.80	S6	300
B4	500	9.75	S7	500
B5	600	9.70		

TABLE 1.2 Price Discovery in the Order Book Listed in Table 1.1

Price	Aggregate Demand	Aggregate Supply	Trading Volume	Excess Demand
9.95 and higher	0	2500	0	−2500
9.90	100	1800	100	−1700
9.85	300	1500	300	−1200
9.80	**900**	**800**	**800**	**100**
9.75	1400	500	500	900
9.70 and lower	2000	0	0	2000

If the rule of maximum aggregated size of matched orders does not yield a unique price, the auction price is chosen to satisfy the rule of minimum order imbalance (i.e., the minimum number of unmatched orders). If even the latter rule does not define a single price, the auction price is chosen to be the closest to the previous auction price.

The advantage of call auctions is that the entire interest in a particular instrument is concentrated at the same time, and is visible to all traders. On the other hand, continuous markets offer a flexibility of trading at traders' convenience.

Quote-Driven Markets and Hybrid Markets

In the quote-driven markets, only dealers submit maker orders. All other traders can submit only market orders. Price discovery in these markets means that market makers must choose such bid and ask prices that will cover their expenses (let alone generate profits) and balance buy and sell order flows. The theoretical models of dealers' strategies are discussed in Chapters 3 and 4.

Some markets combine quote-driven and order-driven systems in their structures. NYSE and NASDAQ are examples of such *hybrid markets*.

SUMMARY

- Two major trader types are profit-motivated traders and liquidity traders. Profit-motivated traders trade only if they expect to receive gains from trading. Liquidity traders may have other reasons, such as maintaining cash flow and hedging.
- Profit-motivated traders can be partitioned into informed traders, technical traders, and dealers. Informed traders base their trading

decisions on information on the asset fundamental value. Technical traders make their decisions using patterns in historical market data. Dealers provide liquidity on both sides of the market and profit from the bid/ask spread.

- Market liquidity is a measure of market breadth, depth, and resiliency.
- There are two major order types: market orders and limit orders. Price is not specified for market orders and these orders are executed at the best price available at the order arrival time. Limit orders are filled only at their (or a better) price.
- There are two major execution systems: order-driven markets and quote-driven markets. Order-driven markets can be partitioned into continuous markets and call markets (auctions).
- In the continuous order-driven markets, all traders can submit limit orders. Unfilled limit orders are placed in the limit order book according to the price-time precedence.
- In call auctions, orders submitted for a given call are batched and executed simultaneously at a price that yields maximum trading volume.
- In the quote-driven markets, dealers submit maker orders while other traders can submit only market orders.