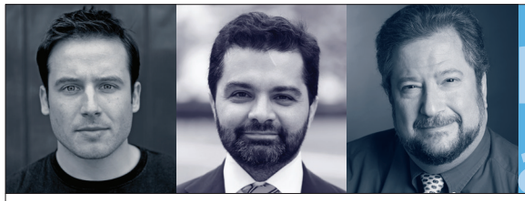

Richard Dewey, Ciamac Moallemi and Aaron Brown

Free Exchange is Not Free

The rise of crypto markets and smart contracts has fueled innovation in exchange mechanisms. This article explores core market design principles and their tradeoffs.

shaddim, CC BY-SA 2.5 <https://creativecommons.org/licenses/by-sa/2.5>, via Wikimedia Commons

The technical approach of setting up an exchange for traditional assets like stocks and bonds, or alternatives like cryptoassets, can seem straightforward. Customers send in orders and when an order arrives that can be matched with a previous order — say A wants to buy 100 shares for up to US$50/share and B wants to sell 100 shares for at least US$48/share, B can sell A 100 shares at US$49/share.

Traditional exchanges have historically used a variety of mechanisms for buyers and sellers to express their preferences and to be matched. The advent of crypto markets and, in particular, programmable smart contracts has brought an explosion in experimentation and innovation in the design of exchange mechanisms. At the same time, regulatory interest in what is called "market microstructure" has grown as well. We're going to focus on important core principles of market design and the tradeoffs inherent in each approach.

Three exchange types

The most popular exchange mechanism is a limit order book (LOB). Customers submit "limit" orders that specify quantity and a maximum price for a buy order or a minimum price for a sell order. Orders are executed as soon as they can be matched but may need to wait or "rest" in the order book. LOBs can be "lit" exchanges that publish resting orders, or "dark pools" that do not.

A fast-growing alternative market structure popular for decentralized trading in crypto is that of an automated market maker (AMM) like Uniswap. AMMs execute each order immediately as it comes in, without waiting for a match. The customer specifies only the quantity, the AMM determines the price for that quantity. A simple and popular model for AMMs is a constant product pool, where the product of the two assets remains constant before and after the trade.

For example, a constant product pool for exchanging USD for Bitcoin might have a 25 billion constant product, US$25 million time 1,000 BTC initially. If someone wants to sell one BTC, so BTC increases from 1,000 to 1,001, USD has to be 25 billion divided by 1,001 or US$24,975,024.98. The BTC seller gets the difference, US$24,975.02. A BTC buyer would get the same amount, as the pool would be restored to US$25 million and 1,000 BTC. But if a second seller came after the first, the dollars in the pool would shrink to 25 billion divided by 1,002, or US$24,950,099.80, so the seller would get only US$24,925.18. The pool price automatically adjusts based on supply and demand.

Another idea emerging in crypto and sometimes used in traditional markets is the frequent batch auction (FBA). This works like an LOB, except matching orders are not executed immediately. Instead, the exchange waits for either fixed time intervals, or until some liquidity threshold is met (such as at least US$1 million of executable orders). FBAs were previously used in equity markets such as Taiwan and have been advocated by academics as a means of mitigating the so-called "HFT tax".

It's natural to think that all three types of exchanges could co-exist, competing for trades. However, this fragments liquidity and may prevent each exchange from getting the necessary diversity of trader types. So, there are both economic forces and social benefits to concentrating transactions on one exchange type for each asset class, leaving the other exchange types to pick up niche business.

Four types of traders

It's convenient to analyze the three exchange mechanisms by considering four types of traders. Actual traders can mix characteristics of different types, but it's still a useful distinction. The simplest group trades to own assets — to acquire stocks, BTC, or cash. The trade is only a way to change assets owned. This is the group whose preferences determine fundamental long-term asset value. We'll call these long-term investors (LTIs) although in crypto they will include crypto users as well as pure investors.

The next group is liquidity providers (LPs). LPs trade to make profits by smoothing out short-term supply/demand imbalances. LPs bring capital and need sophisticated knowledge of short-term price dynamics. LPs generally want near-zero net market exposure.

Information traders (ITs) are the first to trade on short-term information such as news and/or price movements in correlated markets and generally reverse their trades once information is embedded in prices. ITs bring relative price information to market, but generally do not know much about long-term fundamental value.

We'll refer to the last group of market participants as noise traders (NTs), a term coined by Fischer Black. This group is often ignored by economists and the name is misleading as the group

can run the gamut from naive, uninformed participants to highly sophisticated quantitative hedge funds. NTs are not concerned with long-term fundamental value, short-term news, or supply/demand imbalances. They hold positions longer than LPs and ITs, shorter than LTIs. The main idea Fischer Black sought to identify with NTs is the idea of trading on historical price patterns. The simplest noise trading is trend-following, but it could also include complex systems that incorporate hundreds of signals and sophisticated portfolio optimization.[1]

Evaluating exchange mechanisms

LTIs are the essential end-users of most exchanges, while LPs, ITs, and NTs can be viewed as intermediaries that provide services for LTIs and seek to extract profits. A good general-purpose exchange will attract lots of LTIs and execute their trades at fair prices, easily, quickly, reliably, and cheaply. LPs have no social benefit other than the liquidity they provide, so the exchange should restrict LP profits to the minimum required to attract enough liquidity.

ITs are more controversial. Many LTIs would prefer to exclude all information from trading — this is usually phrased as making trading "a level playing field." A big asset manager might prefer to do all trading with everyone using information that had been available at the beginning of the day, if that were somehow possible, so ITs extracted no profits from the markets. The gains and losses from this relative to normal trading would net out close to zero in the long run. On the other hand, there is a social benefit in having market prices incorporate all information.

NTs are also more complex than LPs. They help intermediate between participants with different time horizons and holding periods, and different information sets. This is why economists often overlook them by assuming uniform LTIs. Anyway, it's not clear whether an exchange mechanism with high NT activity and profits is better or worse than one with low NT activity and profits.

The most important characteristic of an exchange mechanism is to facilitate the maximum number of mutually beneficial exchanges. This is a complex outcome of secondary characteristics, which are also beneficial in themselves. So, we'll first run through how the three types of exchanges rank on these simpler secondary characteristics before trying to compare them in terms of gross beneficial exchange volume.

Composable

The driving force for AMMs in crypto is they are the most easily composable system, that is, they can be easily snapped into place as modules in smart contracts or larger systems. A program reads the pool parameters, chooses a quantity, and orders a trade, which in most cases is executed quickly. Settlement is instant.

LOBs offer an interesting case study in composability. A program must read complex order book information, which changes in sub-millisecond intervals. The program then must choose orders — not just quantities but price as well — after which it must wait for a result. Simple algorithms are exploited by high-frequency LPs. Settlement is a separate process and sometimes fails. Moreover, current blockchains are too slow and expensive to implement a full LOB.

The driving force for AMMs in crypto is they are the most easily composable system

FBAs are intermediate in this respect, the main distinction being between synchronous and asynchronous composability. There is plenty of time to review the results of the last auction, no expensive services needed, no worries about high-frequency exploitation. But that time to wait between auction settlements does not allow synchronous composability. In FBAs, prices must be specified as well as quantity, but it's much simpler with FBAs than LOBs. Execution may be slower than AMMs — depending on the specifics of the two exchanges compared — but is more predictable than LOBs.

Credit and capital

A fundamental trade-off in all finance is between credit risk — potential losses from entities not honoring promises — and capital — stores of economic value tied up to guarantee performance. LOBs, as implemented in centralized exchanges, typically extend full credit, traders do not have to post cash or assets for orders. This is important, since only a small minority of orders are eventually executed, so posting full capital for all orders would be prohibitive. Credit is often provided by upstream entities like clearinghouses or brokers rather than the exchange itself. Thus, exchange trades sometimes "fail" when a party defaults at settlement time.

AMMs seldom offer credit since trades settle immediately upon execution. This eliminates the problem of trade failures but does require capital to be reserved throughout the trade process. This is a relatively small amount of capital for liquidity takers, since most AMM orders result in execution. However, those providing liquidity may have much more onerous capital requirements.

FBAs for traditional assets usually offer credit like LOBs, but in crypto the few examples of FBAs demand full collateralization. Capital is less of an issue for FBAs compared to LOBs because a much larger fraction of orders result in execution, and the time from trade input to decision (at which time capital is released) is small on average. Crypto FBAs are still nascent, but most designs generally settle immediately upon execution, so credit is not offered.

User-friendliness and power

A classic technology trade-off is between making something easy and natural for the majority of users, versus allowing advanced features for a minority of sophisticated users. LOBs typically offer market orders for the least-demanding users, limit orders for more sophisticated participants, and many complex order types mainly created for high-frequency traders.[2] Pricing is simple for market and limit orders — a fixed exchange fee — complex for other orders.

On one hand, this allows LOBs to cater to

diverse users. On the other hand, it creates complexity and suspicion. The suspicion has led to rules that multiply complexity, leading to even more suspicion. Concepts like flash crashes, spoofing, manipulation, phantom liquidity, shadow liquidity, and others are bandied about.

AMMs, at the opposite extreme, are simple and transparent for all users. That avoids complexity and suspicion but can seem inflexible to power users. FBAs are intermediate here. They typically allow only two order types (limit buys and sells) and settle all of them at the same price, reducing any games. However, LOB-type problems can arise not within one auction, but by sophisticated traders with strategies that work across auctions.

Public price and liquidity information

All three types of exchanges reveal transactions after they happen, giving price information to the public. FBAs provide the most reliable transaction prices because they are based on large auctions that aggregate many orders, not just single transactions. On the other hand, the price is updated only once per batch rather than with every transaction. FBAs typically also give information about unfilled orders near the transaction price, so the public can gauge liquidity.

On the good side, all of these orders are real, as they were subject to execution. LOB orders can represent phantom liquidity, orders are very likely to be canceled before anyone can trade against them. AMMs provide full transparency on the available liquidity, and rarely suffer from phantom liquidity. On the negative side for FBAs, the order information is from the recent past, while LOBs, in theory anyway, offer information about current liquidity.

Market impact

In an ideal exchange, an order that carried no information would not affect the price. In all real exchanges, orders affect price both because counterparties have to allow for the possibility the orders contain information, and because LPs require compensation to absorb short-term imbalances. Thus, market impact is inextricably related to the appeal an exchange has to LPs and ITs.

AMMs embed an observable schedule of market impact that depends on order size. The amount of market impact depends on the amount of capital allocated to the exchange and thus the overall available liquidity.

Thus, market impact is inextricably related to the appeal an exchange has to LPs and ITs

LOBs have a conceptually similar dynamic schedule shown in the order book, but liquidity at price levels far from the best bid or offer is typically latent and not visible. Hence, good execution depends on trader skill in spreading out trading activity over time and using information about supply and demand for external sources. The amount of market impact depends mainly on the sophistication and capital of LPs attracted to the exchange.

The liquidity schedule for FBAs can only be guessed by looking at previous auctions. Low market impact requires LTIs to submit orders away from the likely transaction price, or for LPs to devote significant capital to the market.

Putting it all together

All of the features discussed above contribute to maximizing beneficial transactions. But it is not a case where the whole is equal to the sum of the parts. A feature is only useful if the exchange itself succeeds.[3] The exchange mechanism that wins out will be the one that can attract enough liquidity to set off a virtuous spiral, as more liquidity attracts more users, who, in turn, bring more liquidity.

LOBs have an enormous entrenched advantage in the traditional financial system. SEC Chairman Gary Gensler has recently started a movement to switch to FBAs. Even if that succeeds, it will likely be focused on retail stock trading and be slow to implement. Many people think AMMs or FBAs would be beneficial, or at least offer some advantages.

The main challenge for AMMs is to attract sufficient LP capital. It seems plausible that AMMs will succeed for less liquid assets, which benefit from the fact that AMMs do not require the participation of sophisticated active market makers.

FBAs are easier to build because they do not require capital. FBAs are likely to compete based on features like portfolio trading, zero-knowledge orders, and more solid price and liquidity information.

Of course, one cannot discount the emergence of new exchange mechanisms as well. What emerges from all this is not likely to be one simple exchange model, but some sort of hybrid with some upstream and downstream applications to address problems and provide additional functionality, and some niche exchanges to bypass the main liquidity pool. But even at superspeed crypto innovation that is likely a decade away.

ENDNOTES

1. Economists are apt to ignore NTs because they think of some long-term fundamental price existing, which the exchange is supposed to identify to create a market equilibrium. Of course, that fundamental price is constantly changing, and the market identification has noise and lags (thought of as error), so the actual price process is complex. Finance students, in contrast, are taught that the path is critical. Securities like options depend not on any fundamental value, but on the volatility of the price path of the underlying. Capital asset prices in general depend on the Beta of the asset with a market portfolio, another path concept. Fundamental value is an abstraction of little meaning since no one knows it to a useful accuracy (and this is even more true in crypto than with traditional assets).
2. A market order executes the trade immediately as the best available price for the least-demanding users. A limit order executes the trade if it can be done at or below a maximum buy/at or above a minimum sell price.
3. Specific disadvantages of a mechanism can be addressed with special rules or off-exchange procedures. Some users will use alternative exchange mechanisms for niche requirements, but it seems likely that prices will be set, and most transactions executed at a single exchange type.