Designing Electronic Markets: On the Impossibility of Equitable Continuously-Clearing Mechanisms with Geographically Distributed Agents

Dhananjay K. Gode
Leonard N. Stern School of Business
New York University
427 Tisch Hall
40 West 4th Street
New York, NY 10012
Dgode@stern.nyu.edu

Shyam Sunder
Yale School of Management
135 Prospect St. New Haven CT 06520-8200
(203) 432 6160
(203) 432 6974 FAX
shyam.sunder@yale.edu

Abstract

The twin goals of (1) designing continuously-clearing market mechanisms in which geographically distributed agents trade, and (2) giving all agents equal access to the market and information generated in the market are mutually incompatible. The frequency with which orders are matched to determine trades and the equality of access to the market are two important attributes by which markets are judged. Insurmountable technological constraints imply a tradeoff between these attributes. In applications sensitive to equal access, continuously clearing market designs, and their information processing advantages, may have to be sacrificed in favor of call markets where orders are batched before they are matched to determine trades. Change in design of markets, in turn, call for different bargaining strategies which might be effective in each environment.

Revised February 2000
Copyright ©1999, All Rights Reserved. Please do not quote without permission.
Comments are welcome.

Designing Electronic Markets: On the Impossibility of Equitable Continuously-Clearing Mechanisms with Geographically Distributed Agents

Computer and communications technologies, with their order-of-magnitude leaps in performance, have brought about important changes in markets. Achieving the ideal of continuous markets, global in extent, and lightning fast in response, has appeared to be within reach. In this paper, we show that there is a theoretical upper bound on the frequency with which markets can be cleared. The upper bound depends on the communication delays. Recognizing these limitations suggests that, in applications sensitive to equal access, the new technology might be best employed in the form of call markets instead of continuous markets to which much of the current efforts are directed. Changes in market design have important implications for bargaining strategies of trading agents.

In the financial sector of the economy, any acceptable market design must provide all traders equal access to the market, and to the information generated in the market. Equal access can be defined in many different ways. At the minimum, equal access requires that the communication delay for every trader be equal. Several methods can be used to achieve this equality. In all methods, the lower bound of the cycle time depends on the maximum transmission time to any of the traders in the market. In agent-based financial markets, even the smallest of differences in access can have major implications for the profits generated by the agents

The ideal of continuous markets is unachievable because the speed of transmission is bounded from above by the speed of light. Even at these speeds, transmission of messages to geographically dispersed agents at global scale leads to delays that are sufficiently large to render the ideal of a continuous market at a global scale effectively unachievable. Electronic implementation of continuous markets must necessarily deviate in significant ways from floor trading, and potentially induce different trader behavior and strategies, and price patterns in the market. Consequently, we explore the possibility that an alternative market organization--a call market--may exploit the advantage of electronic technology better than the current attempts that are directed at imperfect recreation of the continuous environments of physical exchange floors.

Continuous Versus Call Markets

Trading activity on the trading floors of New York, American, and most other U.S. stock exchanges as well as the trading pits of commodity exchanges in Chicago and elsewhere, may occur in spurts. However, there is no lower limit on the amount of time that must pass between two consecutive market actions. In other words, a trade can, and does take place as soon as a bid and an ask match, without waiting for any further bids and asks. These markets, in which a transaction is finalized as soon as a match occurs, are called continuously-clearing markets. For brevity, we will call them continuous markets. Within the limits of the information processing capacity of the humans who operate them, most exchanges in U.S. operate as continuous markets.

Call markets, on the other hand, accumulate bids and offers for a period of time, the length of time being determined by a pre-specified rule. When the market is called, a single price is computed based on the accumulated bids and asks to maximize the number of units traded. All possible transactions are executed at this single price and the price and allocations are announced. Immediately following the call, accumulation of bids and asks for the following call resumes. This process determines the opening price and transactions in the New York Stock Exchange and many exchanges.

Continuous markets provide immediate execution of market orders. Thus, they allow prices to adjust quickly to changing market conditions. However, these prices are subject to volatility arising from the sequence in which orders arrive at the exchange. Since call markets accumulate orders before matching them they are less prone to the volatility arising from the sequence in which the orders arrive at the market. However, call markets do not provide instantaneous price discovery.

An example would illustrate the point. Suppose there are two buyers (B1 and B2) and two sellers (S1 and S2) in a market. B1 and B2 are willing to buy a unit of the object being traded for $16 and $10 respectively. S1 and S2 are willing to sell a unit for $4 and $8 respectively. For simplicity assume that these traders do not act strategically and the buyers submit their values as bids and the sellers submit their costs as asks. Suppose the midpoint is chosen as the clearing price if there is a feasible range of clearing prices. In a call market the market clearing price after all these four orders have been received will be $9. The clearing price in the call market will not depend on the sequence in which these

orders are submitted to the market. The clearing prices in the continuous market will, however, depend on the sequence in which the orders are submitted. If orders are submitted in the sequence: B1, S2, S1, and B2, then the clearing prices will be $12 (= (16+8)/2) and $7 (= (4+10)/2) respectively. If the price of the bid or ask submitted first is taken to be the clearing price, then the clearing prices will be $16 and $4 respectively. Thus, the sequence in which orders arrive at the market can cause prices in the continuous market to be volatile. This effect is likely to persist even if traders act strategically.

The above example makes it appear that the call market is superior to the continuous market. This may indeed be true if the demand and supply are static. However, if the demand and supply are changing constantly, then a call market has the disadvantage of slower price discovery. The market prices in the call market are unable to reflect the effect of changes in demand and supply until a new clearing price is set at the end of the call.

The Model

Let us assume the simplest of network configurations, a star network, shown in Figure 1. Each of the N trading stations on the periphery is connected to the exchange located at the center by a direct dedicated communication link. The star configuration and the assumption of direct dedicated lines abstract away from the complicating problems of traffic loads, and consequent variability and dependencies across channels. It also abstracts away from any correlation between market activity and the communication delay on any given channel because of congestion.

Associated with each link i is a deterministic communication delay d_i. In any global network, some trading stations will be closer to the hub than the others, introducing dispersion in d_i across trading stations. The generality of the results presented here does not depend on the assumption of this dispersion. The delay is assumed to be deterministic for simplicity, and the results would generalize to the stochastic case.

Implementation of a continuous market requires that each bid or ask be processed in the exact order in which it is received at the exchange. Any discretization of time into slices for the purpose of network operation can create the possibility that more than one

bids/asks will arrive at the exchange in a single time slice, making it impossible for the exchange to assign an appropriate time priority to them. Thus, the first requirement for implementation of a continuous market on a computer network is that the time slicing be smaller than the smallest possible interval between two consecutive bids or asks. In other words, the time grid over which bids and asks originate must not be finer than the time grid over which the exchange operates. If no constraints on the time grid on which bids and asks originate is feasible or desirable, the strict implementation of a continuous exchange on computers is impossible. Something must give.

Things that can give fall into two categories. First, we give strict time priority. If two or more bids or asks arrive at the exchange within the same time slice, the exchange will process them in some order to be determined by, for example, the highest-bid/lowest-ask first rule or by randomization.

Second, the concept of a continuous market may be abandoned in favor of a call market. As described above, all bids and asks will be accumulated for a number of time slices, either specified in advance, or determined by some other rule (e.g., when a pre-specified number of bids and asks have been accumulated). Call markets usually operate on price and time priority, and discretization of time will place the same limits on the implementation of time priority in call markets as it does in the continuous markets discussed in the preceding paragraphs. However, in call markets, time is the second sort key after price, while in the continuous markets, time is the first sort key. Call markets do not have to sort by time, they may, instead, ration or randomize among all buyers/sellers who stand at the margin.

Frequency of calls could be fixed in advance, or determined endogenously by a rule. As the frequency is increased, the market organization converges closer to the continuous form. We therefore parameterize the market organization by c, the amount of time between two consecutive calls.

Trader i who operates in a call market that runs at intervals of time c, with communication delay d_i, has a total of $(c - 2d_i)$ available for making decisions between two consecutive calls after subtracting the time needed to receive information about the results of the earlier call, and the time it takes for the orders for the next call to get to the exchange in time.

If the market design must permit a minimum of d decision time to every trader, $(c - 2d_i) > d$ for all i, or $c > (d + 2d_{max})$. Both components of this lower bound on call durations are dependent on technology. To the extent that the mechanism for making trading decisions is computerized, or supported by computerized decision aids, d may be reduced to a small amount of time. The minimum possible communication delay for the most far flung traders on the globe, d_{max}, is of the order of one second. There is not much room for further reduction in this delay, given the physical distance on the globe. This means that the minimum possible value of call duration of a global market, c_{min} would be about two seconds. If call markets are seen as an approximation of the continuous market, approximation closer than 2-second calls is not achievable. If traders on existing physical exchanges are used to faster response times (as they no doubt are in Chicago pits where hand signals and eye contact are much faster) they may find this alternative too slow for their what they already have. Whether the advantages of this alternative outweigh the disadvantages requires further analysis.

We have already discussed above some of the advantages and disadvantages of call markets. Calls of longer duration create deeper markets that yield more precise price discovery because such markets are able to accumulate a larger set of orders before they determine the price. Our analysis here suggests a new and different reason for increasing the duration of calls. Traders who are farthest (communication) distance away the exchange have less time available to them for making decisions between calls than the traders located closer to the market do. Suppose we use the following expression as a measure of the "fairness" or "equity" of the market:

$$(c - 2d_{max}) / (c - 2d_{min})$$

This measure would be 1 if and only if communication delays for all traders are equal and less than 1 otherwise. However, as the value of c is increased progressively, this fairness measure also converges to 1. In other words, in a call market with large inter-call durations, relatively small differences in communication delays can become inconsequential.

However, depending on how close one wishes to get to perfect fairness in market design, this method may require a large increase in call duration. Longer call duration has its own costs. First, while the accumulation of more bids and offers over the longer

durations yield a more precise discovery of price, this discovery takes place less often, and therefore, the more precise discovered price is also more stale. In fast moving markets, especially in the financial sector where the information role of price discovery is crucial, the trade off among consequences of duration is of critical importance.

An alternative to increasing the call duration when delays are not stochastic

Instead of increasing the call duration c, the fairness measure can be made to approach 1, by making d_{min} equal to d_{max}. This can be done by adding delays to the transmission of information and the receipts of orders from traders who are "close" to the exchange. This can be implemented as follows:

1. Synchronizing clocks of all trader machines. Of course, there are limits to the precision to which the clocks can be synchronized.
2. Embedding a time lock into the transaction information transmitted to all traders so that the information is readable by the traders only after time lock is deactivated at the time shown. The time lock should be set so that there is sufficient time for the information to reach all participants before the deactivation of the time lock.
3. Delay the transmission of orders from traders who are "close" to the exchange so that they are at par with traders who are "far" from the exchange.

It is difficult to make this scheme work if the delays are stochastic. In that case we can only ensure fair access on average. If the stochastic process underlying these delays is non-stationary, then this scheme of adding delays can get complicated.

Is providing equal access a desirable goal?

The exchanges of today do not provide equal access to all participants. Faster market access comes at a price. The price paid reflects how close, in a geographical sense, the participant or his agent is to exchange. Why should electronic markets be any different? Why should designers of electronic markets worry about providing equal access to participants?

In fact, making traders pay for faster or more convenient access to markets may be economically efficient. One of the important roles of markets is to facilitate trade

between buyers who value the objects the most and the sellers who can produce them at the lowest cost and exclude sellers who cannot produce at the lowest cost and buyers who do not value the object as much. The total profits earned by traders are maximized when markets achieve such an allocation. In fact, one of the important performance measures of markets is their allocative efficiency as measured by the ratio of actual total profits of all traders to the maximum possible total profits of all traders.

If access to the market is priced, then traders that are likely to benefit most from trading are likely to be willing to pay most for access. This is likely to increase the chances of trade among traders who stand to benefit most from trading while minimizing the intrusion from traders who do not benefit as much from trading. This, of course, depends on the assumption that traders can choose the price of market access as frequently as the changes the benefits they are likely to derive from trading. A low-cost seller is unlikely to remain a low-cost seller forever. It is optimal from an allocative efficiency point of view that traders who are likely to benefit from trading at any given point are able to purchase better market access.

The benefit of providing equal access is that "by leveling the playing field" the exchanges may be able to attract more dispersed set of market participants. Wider market participation increases the size and depth of the market, which improve allocative efficiency as well as liquidity.

Speed of trader decision making versus the speed of communications

It is generally assumed that faster communication lines and faster communication switches will automatically expand the scope of the market. Indeed, the introduction of telegraph, telephone, and now the world-wide web has expanded the scope of markets by allowing people to participate from far-flung places to trade in a timely and convenient way. This has been possible because the speed of trader decision making has been limited by human cognition while the communications have become faster with improvements in technology. If, however, the decision making becomes more automated, then small differences in communications delays will become significant.

Information Structures and Agent Strategies

Continuous and call markets have very different information structures and therefore call for very different agent bargaining or trading strategies that need to be explored. We shall briefly consider two aspects of these differences.

Public Information in Continuous Markets. In continuous markets, typically, the current (i.e., the highest) bid and the current ask (i.e., the lowest ask) are made public to all participants at all times. Beyond this inside spread information, the information about the entire book (bids below the current bid and asks above the current ask) may, but rarely are, made available to the participants. When a transaction takes place, the price and volume (but rarely the identity of the transacting parties) made public information.

Public Information in Call Markets. In such markets, the time of the next call, and clearing price and volume made public after each call. The number and prices of unaccepted bids and asks, and the identities of the transacting parties can be, but is rarely made public. During the interval when the bids and asks are being accumulated before the call, the prices and quantities associated with each bid and ask (i.e., the book) can be made public. Whether it is in fact made public is a crucial design variable for the call market.

The decision making by trading agents in the two kinds of markets is different in important respects. In a continuous market, it is always possible for an agent to guarantee a transaction by simply accepting the current bid or ask. In a call market, on the other hand, there is no way for the trading agent to ensure that he or she will be sure to transact in the next market clearing (though one can always increase the probability of this event by submitting a higher bid and a lower ask).

This advantage of continuous markets is counterbalanced by the market protection call markets provide to the trading agents against errors. In a call market, the agent is protected from even large errors in submitting bids or asks because all transactions at the time of clearing take place at a single price. The deeper the market, better is the protection call markets provide against such errors. In continuous markets, on the other hand, no such protection is available against bidding errors. Agents are punished immediately for any errors in bidding because other participants in the markets immediately snap up low asks and high bids relative to the prevailing market conditions.

Given these differences between the informational and strategic environments of continuous and call markets, very different trading strategies are called for on part of the trading agents. Detailed investigation of these differences, and their links to market rules is an interesting item for e-commerce research agenda.

Concluding Remarks

Ensuring equal access can not be left to communication engineers. The specifications of the market processes dictate a particular network design. Not all market processes can be implemented to guarantee equal access. It is easier to provide equal access in call markets.

It appears that as a result of computerization markets will have a wider geographical reach. This belief is predicated on the assumption that it will be easy to resolve the conflict between geographical dispersion and equal access for all the participants concerned. There are theoretical limits on the speed of communication. New technologies have reversed the traditional relationship between the speed of trader decision making and the communication and processing of messages; the latter now takes far less time than the former. With automated decision making (in markets populated by agents) the decision making speeds will increase dramatically, making it difficult to provide equal access to the market for geographically dispersed participants.

Technology changes both market institutions and market participants. Rapid advances in communications technology have resulted in markets with geographically dispersed participants. The New York Stock Exchange and Chicago Mercantile Exchange are typical examples of exchanges located in a specific geographical location. In such markets the traders on the "floor" of the exchange have an advantage over the traders away from the exchange because they receive the trading information before others and they can execute their orders faster than others.

Guaranteeing fair access is not enough, the system has to meet the cost, efficiency and speed criteria. A system that is excessively slow but fair would not be used by the participants. In most situations however fair access can be attained only at the expense of speed and efficiency. Network designers may have to limit the scope of the networks to

alleviate the problem of fair access. Thus access problems may impose an upper limit on the globalization of electronic markets. How severe are these problems

With sufficient computing power and communication channels, it is possible to increase the frequency of call markets to such an extent that they appear to be close to continuous markets. However, if the decision-making speeds of computerized agents or computer-assisted traders also increase by similar proportions, then this approximation cannot approach continuous markets. Thus, aside from other reasons that have been advanced elsewhere for organizing securities exchanges as call markets, a new, very different theoretical reason in favor of call markets arises from this analysis. Electronic global exchanges should be call markets, not merely because they have more depth, but because there is no other practical alternative. The reason Globex (a project of Chicago Mercantile Exchange to trade futures contracts electronically) was delayed for so long is because its creators chased the ideal of continuous market ignoring the impossibility of that ideal.
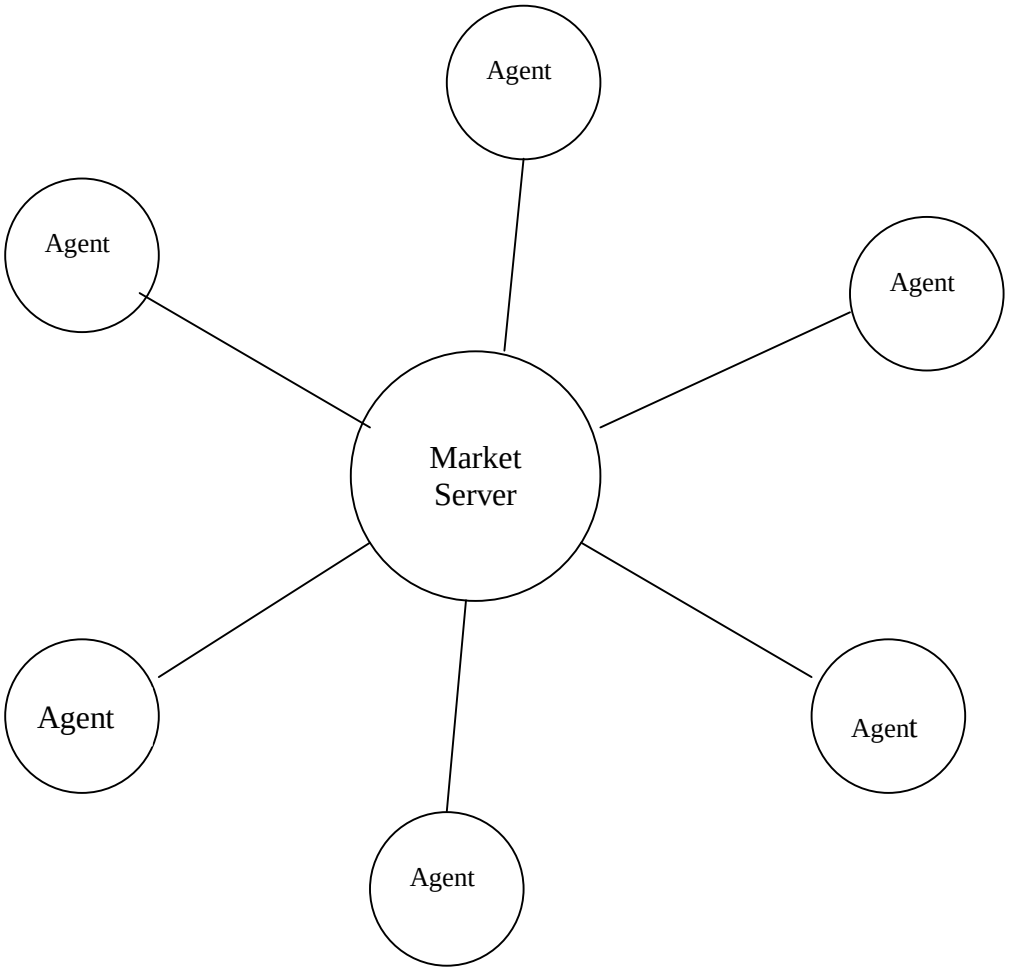
Figure 1: Star Network